

Mixed-Initiative Conversational Intelligence in the Era of Large Pre-Trained Models

Maximillian L. Chen

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy  
under the Executive Committee  
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

© 2025

Maximillian L. Chen

All Rights Reserved

## **Abstract**

Mixed-Initiative Conversational Intelligence in the Era of Large Pre-Trained Models

Maximillian Chen

With the rise of large pre-trained models, the idea of intelligent conversational agents has quickly gained attention in the public eye. Such conversational agents promise impressive capabilities in a multi-turn interaction setting, whether it be knowledge-grounded question answering, reasoning for code generation, or navigating real-world tasks like restaurant booking. Despite the greatly improved capabilities of agents to perform directed instruction-following tasks, agents built around large pre-trained models have not yet exhibited the ability to navigate mixed-initiative environments. Contrary to single-initiative settings, here an agent must recognize when it is appropriate to execute different actions to redirect the flow of the conversation (e.g. via clarifying questions or argumentative strategies) to maximize the chance of conversational success. This dissertation proposes novel methods to address core issues presenting bottlenecks to the mixed-initiative intelligence of such agents.

This dissertation is structured to address three core challenges. The first section introduces the challenges of multi-turn conversational modeling, as in-domain data can be expensive or infeasible to obtain. We propose a framework for synthesizing large-scale conversational data even for novel tasks by leveraging the instruction-following capabilities of existing large language models. In the second section, we discuss the challenges of action planning. Here, we propose a seminal line of work on inference-time strategy optimization by leveraging large language model prompting for search and simulation adapted for Monte-Carlo Tree Search. We then introduce a novel perspective towards

action optimization called implicit action recognition, and propose a novel model alignment algorithm called Action-based Contrastive Self-Training. The final section of this dissertation focuses on the challenges of multimodal user modeling, as modern conversational agents rapidly look to become more ubiquitous by expanding towards all interaction modalities. This work builds on the previous sections' progress towards improved data curation and implicit action recognition. Here, we introduce a novel task for mixed-initiative spoken conversation modeling, as well as a simple yet effective approach to adapt models to different users' speaking patterns.

The methods proposed in this dissertation address tractable real-world challenges and serve as the foundation for further exploration in mixed-initiative conversation modeling.

## Table of Contents

Acknowledgments . . . . .	xix
Dedication . . . . .	xxi
Chapter 1: Introduction and Background . . . . .	1
1.1 Overview . . . . .	2
1.1.1 Challenge: Data Scarcity . . . . .	2
1.1.2 Challenge: Pragmatic Action Planning . . . . .	3
1.1.3 Challenge: User Modeling in Multimodal Interaction . . . . .	5
Chapter 2: Resource Curation for Conversation Modeling . . . . .	8
2.1 Introduction . . . . .	8
2.2 Related Work . . . . .	11
2.3 Conversation Generation . . . . .	12
2.3.1 Writing Conversation Examples . . . . .	12
2.3.2 Creating Conversations via Prompting . . . . .	13
2.4 Human Evaluation Setup . . . . .	14
2.4.1 Static Conversation Evaluation . . . . .	14
2.4.2 Interactive Conversation Evaluation . . . . .	15
2.5 Human Evaluation Setup . . . . .	16

2.5.1	Interactive Evaluation . . . . .	17
2.6	Synthetic Conversation Evaluation Experiments . . . . .	17
2.6.1	Evaluation of Conversation Quality . . . . .	17
2.6.2	Fine-Tuning with Synthetic Conversations . . . . .	19
2.6.3	Triadic and Multi-Party Conversations . . . . .	21
2.7	Discussion . . . . .	23
2.7.1	Considerations for Dyadic Dialogue . . . . .	23
2.7.2	Considerations for Multi-Party Dialogue . . . . .	24
2.8	Error Analysis . . . . .	25
2.8.1	Dyadic Conversations . . . . .	25
2.8.2	Triadic Conversations . . . . .	28
2.9	Limitations . . . . .	31
2.9.1	Controllability . . . . .	31
2.9.2	Cost of Human Effort . . . . .	31
2.9.3	Computational Costs . . . . .	32
2.9.4	Prompt Design . . . . .	32
2.9.5	Ethical Considerations . . . . .	32
2.10	Summary . . . . .	33
Chapter 3: Explicit Mixed-initiative Action Planning via Prompt-based Monte-Carlo Tree Search in Multi-turn Conversations . . . . .		
3.1	Introduction . . . . .	35
3.2	Related Work . . . . .	37
3.3	Methodology . . . . .	38

3.3.1	Problem Definition . . . . .	39
3.3.2	Dialogue Planning as a Stochastic MDP . . . . .	39
3.3.3	GDP-ZERO . . . . .	40
3.4	Experiments . . . . .	44
3.4.1	Prompting Implementation Details for P4G . . . . .	44
3.4.2	Adapting GDP-ZERO Setup to P4G . . . . .	45
3.4.3	Static Evaluation . . . . .	46
3.4.4	Interactive Human Evaluation . . . . .	49
3.4.5	Ablation Studies . . . . .	51
3.5	Analysis of GDP-ZERO Dialogues . . . . .	52
3.6	Example Interactive Conversations . . . . .	54
3.7	Limitations . . . . .	54
3.8	Ethical Considerations . . . . .	56
3.9	Summary . . . . .	57
Chapter 4: Implicit Mixed-initiative Action Planning via Action-based Contrastive Self- Training in Multi-turn Conversations . . . . .		
4.1	Introduction . . . . .	59
4.2	Related Work . . . . .	61
4.2.1	Mixed-Initiative Conversational Agents . . . . .	61
4.2.2	Learning for LLM Alignment . . . . .	62
4.3	Methods . . . . .	64
4.3.1	Problem Setup . . . . .	64
4.3.2	<i>ACT</i> : Action-Based Contrastive Self-training . . . . .	65

4.4	<i>AmbigSQL: Modeling Ambiguity in Conversational Text-to-SQL</i> . . . . .	69
4.4.1	AmbigSQL Construction . . . . .	70
4.4.2	Examining Clarification Need in AmbigSQL . . . . .	71
4.4.3	Examples . . . . .	72
4.5	Experimental Setup . . . . .	74
4.5.1	Datasets . . . . .	74
4.5.2	Conditional Generation Model Details . . . . .	75
4.5.3	Action Classifier Details . . . . .	76
4.5.4	User Simulator Details . . . . .	78
4.5.5	Evaluation Setup . . . . .	80
4.5.6	Baselines . . . . .	84
4.6	Experimental Results . . . . .	88
4.6.1	Conversational QA with Tabular Grounding . . . . .	88
4.6.2	Conversational QA for Machine Reading Comprehension . . . . .	90
4.6.3	Conversational Text-to-SQL Generation . . . . .	90
4.6.4	<i>ACT</i> In-The-Wild: Learning Without Dialogue Action Supervision . . . . .	91
4.6.5	Ablation Studies . . . . .	92
4.7	Limitations, Ethical Considerations, and Broader Impacts . . . . .	93
4.7.1	Discussion of Limitations . . . . .	93
4.7.2	Ethical Considerations . . . . .	96
4.7.3	Broader Impacts . . . . .	97
4.8	Summary . . . . .	97



Chapter 5: Improving User Modeling in End-to-End Multi-Modal Spoken Conversation Modeling via Data-centric Multi-task Learning . . . . .	99
5.1 Introduction . . . . .	99
5.2 Related Work . . . . .	101
5.3 Data-Centric Multi-Task Learning for Cross-Modal Understanding . . . . .	103
5.4 Efficient Multimodal Adapters via Audio Representation Projection . . . . .	105
5.5 A Corpus for Ambiguous Spoken Conversational Question Answering (ASK-QA) .	106
5.5.1 Dataset Construction . . . . .	106
5.5.2 Evaluation . . . . .	109
5.5.3 Conversation Examples . . . . .	110
5.6 Experiments . . . . .	111
5.6.1 ASK-QA: Spoken Knowledge and Multi-Turn Spoken Dialogue . . . . .	112
5.6.2 Spoken-SQuAD: Spoken Knowledge and Textual Questions . . . . .	114
5.6.3 SD-QA: Textual Knowledge and Spoken Questions . . . . .	116
5.6.4 Ablation Studies . . . . .	117
5.7 Limitations . . . . .	118
5.8 Summary . . . . .	119
Conclusion or Epilogue . . . . .	121
References . . . . .	125
Appendix A: Supplementary Material on Conversation Synthesis . . . . .	149
Appendix B: Supplementary Material on Explicit Action Planning . . . . .	159

Appendix C: Supplementary Material on Implicit Action Planning . . . . .	169
C.1 Additional Qualitative Examples . . . . .	169
C.2 Additional Experimental Details . . . . .	171
C.2.1 Compute Resources . . . . .	171
C.2.2 Training Hyperparameters . . . . .	171
C.2.3 Assets Used . . . . .	172
Appendix D: Supplementary Material on Data-centric Multi-task Learning for Multimodal Understanding . . . . .	176
D.1 Additional Details on Data-Centric Multi-Task Learning . . . . .	176
D.2 Extended Experimental Results . . . . .	177
D.2.1 Additional Experiments on ASK-QA . . . . .	177
D.2.2 Additional Experiments on Spoken-SQuAD . . . . .	177
D.3 Training Details . . . . .	179
D.4 Risks and Ethical Considerations . . . . .	180
D.5 Assets Used . . . . .	180

## List of Figures

2.1	Pair of dyadic conversation excerpts about hometowns (upper) and pair of triadic conversation excerpts about Ithaca, NY (lower). In both pairings, one conversation is synthetically generated and the other is collected from humans. The top-left is taken from DailyDialog, whereas the top-right is generated synthetically. The bottom-left is generated synthetically and the bottom-right is taken from MPC. . . .	9
2.2	Example of the components of a prompt (left) used by OPT 30B to generate a synthetic conversation about pets (right). Conversations in the prompt are prefixed by recipes. Blue text: topic labels. Red text: seed background information metadata.	10
2.3	Distinct-N with $N = 2, 3, 4$ for conversations in DailyDialog, Topical Chat, and our synthetic conversations. Our synthetic conversations have the highest most unique bi-grams and tri-grams, and the second-most unique 4-grams. . . . .	14
2.4	Linguistic diveristy (Distinct-N) is comparable for each speaker in the synthetic triadic conversation dataset. . . . .	22
3.1	Using GDP-ZERO for persuasion with zero model training. . . . .	36
3.2	GDP-ZERO with ChatGPT backbone. During <b>Selection</b> , simulations are either sampled from cache or newly generated. During <b>Expansion and Evaluation</b> , we prompt ChatGPT for prior policy $\pi$ and value estimation. . . . .	38
3.3	Distribution of planned dialogue actions for each planner during interactive evaluations. . . . .	53
4.1	Simplified example of ambiguity present at tabular-grounded conversational question answering based on (Deng et al., 2022). A conversational agent should recognize when there is ambiguity and ask a clarifying question towards a more accurate final answer. . . . .	60
4.2	<i>ACT</i> greatly outperforms standard tuning approaches in data-efficient settings for conversational modeling, as exemplified here on PACIFIC. . . . .	61

4.3	<b>Overview of the tuning phase of ACT.</b> For each initial contrastive pairing from $D_{pref}$ (constructed as in Sec. 4.3.2), we sample an on-policy response from the model being tuned. After evaluating the sampled response’s trajectory, we update the contrastive pairing by either replacing the existing winning or losing response. The model policy is updated using the objective in Eq. 4.1. . . . .	65
5.1	Automatic speech recognition is a necessary <i>implicit</i> skill for MLLM in end-to-end spoken question answering. We propose a multi-task learning approach which explicitly teaches these skills, as exemplified by this QA pair from Spoken-SQuAD.	100
5.2	<b>Simplified summary of the pipeline for constructing ASK-QA.</b> For each text conversation in Abg-CoQA, we construct three speaker profiles with randomly sampled voices, speaking rates, and pitches. We use TTS to synthesize the story context as a spoken narration, then each individual dialogue turn. The resulting audio files are joined as a single recording. . . . .	107
5.3	<b>Multi-turn evaluation pipeline for ASK-QA.</b> A model is given an audio recording containing the spoken story and spoken conversation. It is tasked with providing the correct response. While the model response is a clarifying question (as determined by a prompted Action Classifier), the model-generated response is appended to a textual version of the conversation history and shown to a user simulator. The user simulator provides a coherent response to the clarifying question, and these two generated turns are synthesized using TTS to create a new spoken context. This process repeats until the model response is not a clarifying question. .	109
5.4	Multi-task (MT) learning improves upon Single-task (ST) fine-tuning with both Gemini and Speech-Qwen on ASK-QA’s multi-turn evaluation. . . . .	116
5.5	Our multi-task approach applied to Speech-Qwen outperforms the state-of-the-art approach on Spoken-SQuAD using only 10% of the available data. . . . .	117
C1	<b>Example of a contrastive pairing constructed for RL tuning with Abg-CoQA</b> Guo et al., 2021. The notation used is as described in Section 4.3.1. . . . .	174
C2	<b>Trajectory-level content evaluation using the example scenario from Figure 4.1.</b> Trajectory-level evaluation seeks to measure the extent to which a candidate LLM can interact with a “User” to reach a target information goal. The “interactive” evaluation of a given instance continues until the candidate LLM attempts to resolve the User’s request by providing a direct answer. The candidate trajectory resolution is scored using downstream task metrics. In this example, DROP F1 is used following the task metrics for PACIFIC. . . . .	175

D1	<b>Creating multi-task data from individual SQA training instances.</b> Left: examples of instance metadata from the three SQA datasets used in this paper. Right: for each speech-QA pairing, we are able to form three tasks designed to teach MLLMs’ cross-modal reasoning ability. . . . .	176
----	--	-----

## List of Tables

2.1	Number of words per turn and number of turns per conversation for all conversations. HW Examples represents the ten handwritten conversation examples, and Synthetic represents synthetic conversations generated using OPT 30B. . . . .	13
2.2	Evaluation of conversations randomly sampled from DailyDialog, Topical Chat, and three synthetic datasets generated by prompting GPT-J 6B, GPT-NeoX 20B, and OPT 30B. * indicates statistical significance over DailyDialog. † indicates statistical significance over Topical Chat. Significance computed at $\alpha = 0.05$ . . . .	18
2.3	Human evaluation of conversations generated using OPT-30B with in-context examples randomly sampled from DailyDialog (DD-IC), Topical Chat (TC-IC), and handwritten examples (HW-IC). * indicates statistical significance over DD-IC and † indicates statistical significance over TC-IC. . . . .	19
2.4	Interactive human evaluation yields comparable ratings for chatbots fine-tuned on conversations from DailyDialog (DD), Topical Chat (TC), and our Synthetic Data (Syn). . . . .	20
2.5	Out-of-distribution automatic evaluation of perplexity and ROUGE is comparable for BlenderBot fine-tuned on DailyDialog (DD-BB), Topical Chat (TC-BB), and synthetic data generated using our handwritten examples in-context (Syn-BB), respectively. . . . .	21
2.6	Synthetic conversations generated using OPT 30B are rated significantly higher than MPC and MELD across all dimensions. . . . .	24
2.7	Synthetic conversation generated by OPT 30B which was rated as “generic and dull.” While the LLM appears to have drawn on information from an actual event (Taylor Swift did donate \$113,000 to support an LGBTQ advocacy group in Tennessee), the conversation appears to get stuck around the question “What are your thoughts on her?” which results in deteriorating conversation quality towards the end. . . . .	26

2.8	Synthetic conversation generated by OPT 30B which was rated as “generic and dull.” The conversation appears to be coherent and is rooted in factually correct information. However, the language is fairly ordinary, which may have been subjectively considered uninteresting. . . . .	26
2.9	Synthetic conversation generated by OPT 30B which was rated as completely inconsistent. Initially, “Alice” is trying to learn to make cheese, but then it becomes ambiguous by their next turn whether they already know how to make cheese. . . .	27
2.10	Synthetic conversation generated by OPT 30B which was rated as completely incoherent. The prescribed subtopic was “cotton candy” but the conversation pivoted to a much more sensitive topic, cancer. . . . .	27
2.11	Synthetic triadic conversation generated by OPT 30B which was rated as completely incoherent. First, “Alice” mentions they went on a road trip around California, then they mention that they haven’t been out of the country. This implies that “Alice” is located in the United States. But then Alice mentions having traveled to Spain. The conversation also randomly shifts from discussing travel plans to discussing board games and other hobbies with an awkwardly placed transition on “Claire’s” third turn. . . . .	29
2.12	Synthetic triadic conversation generated by OPT 30B which was rated as completely unnatural. The conversation appears to be coherent, but the speakers appear to be very enthusiastic. It could be argued that the speakers are overly unenthusiastic to an unnatural degree. . . . .	29
2.13	Synthetic triadic conversation generated by OPT 30B which was rated as generic and dull. “Alice” begins a long debate on whether the EU is a “conspiracy” without making a lot of conversational progress. . . . .	30
3.1	Annotated dialogue actions in the P4G dataset. . . . .	45
3.2	Mapping of persuader dialogue actions to natural language prompts used for prompting LLMs. . . . .	46
3.3	Prompting LLM to specify which generated response “response a” or “response b” is more persuasive. . . . .	47
3.4	Example response generated when using GDP-ZERO as the policy planner during static evaluation. Crowdfworker’s response is the original response in the P4G dataset Wang et al., 2019. GDP-ZERO used $n = 10$ , $k = 3$ , $Q_0 = 0.25$ with ChatGPT as backbone. . . . .	48

3.5	Static evaluation with ChatGPT as backbone and judge. Results are $\mu \pm \sigma$ repeated over three runs. . . . .	49
3.6	Static evaluation ChatGPT as backbone and judge. Runner-up is colored in gray. Results are $\mu \pm \sigma$ repeated over three runs. . . . .	49
3.7	Interactive evaluation using RAP, ChatGPT, and GDP-ZERO as planners, with ChatGPT used for response generation/backbone. Results are $\mu \pm \sigma$ . All scores scaled to $[1, 5]$ except for “donation prob.” being $[0, 1]$ . * denotes $p < 0.05$ , and ** denotes $p < 0.01$ . . . . .	51
3.8	Static evaluation using the first 20 dialogues of P4G with ChatGPT as judge. GT refers to Ground Truth. Results are $\mu \pm \sigma$ repeated over three runs. . . . .	52
3.9	Static evaluation using the first 20 dialogues of P4G with ChatGPT as judge. GT refers to Ground Truth. Results are $\mu \pm \sigma$ repeated over three runs. Since ChatGPT generations are typically long, we use the first 3 sentences of each generation in this evaluation. . . . .	52
4.1	<b>Overview of AmbigSQL</b> , an ambiguous Text-to-SQL dataset synthesized from Spider. . . . .	69
4.2	<b>Conversations in CoSQL with noisy “clarification questions”</b> (highlighted in red). Example 1) is a remnant of crowdsourcing in which the system-side party makes mention of the task guideline. Example 2) demonstrates a system-side clarification question being asked prior to the user making any information requests. Example 3) The system-side clarification question makes reference to some prior database search result, but the execution feedback is not made accessible to the system during inference. . . . .	69
4.3	<b>In-context example given as part of a prompt for creating information requests in which the target population is ambiguous.</b> The format of the black text represents how a ground-truth request would be used to form the prompt for a target example. The blue text represents the content that would be synthesized from an LLM. We omit the database schema from the paper. . . . .	70
4.4	<b>In-context example given as part of a prompt for creating information requests in which the target columns are ambiguous.</b> The format of the black text represents how a ground-truth request would be used to form the prompt for a target example. The blue text represents the content that would be synthesized from an LLM. We omit the database schema from the paper. . . . .	72



4.5	<b>Examination of ambiguity in AmbigSQL.</b> Competitive high capacity LLMs struggle with producing “correct” SQL queries given only ambiguous user requests. Including disambiguation turns in the prompts greatly improves execution match. . . . .	73
4.6	<b>Examples from the AmbigSQL test set using the “concert_singer” database.</b> Black text: context provided to an LLM as a prompt. Blue text: ground-truth single-turn response. Purple text: target trajectory which should be achieved through simulation. . . . .	73
4.7	<b>Examples from the AmbigSQL test set using the “flight_2” database.</b> Black text: context provided to an LLM as a prompt. Blue text: ground-truth single-turn response. Purple text: target trajectory which should be achieved through simulation. . . . .	74
4.8	<b>In-context example given as part of a prompt to an LLM for action classification with PACIFIC.</b> The blue text represents the action for this example. . . . .	76
4.9	<b>In-context example given as part of a prompt to an LLM for action classification with Abg-CoQA.</b> The blue text represents the action for this example. . . . .	77
4.10	<b>In-context example given as part of a prompt to an LLM for action classification with AmbigSQL.</b> The blue text represents the action for this example. . . . .	77
4.11	<b>In-context example given as part of a prompt to an LLM for intent summarization with PACIFIC.</b> The purple text represents the system prompt used for this task. The blue text represents the summarized information-seeking intents. . . . .	78
4.12	<b>In-context example given as part of a prompt to an LLM for user simulation with PACIFIC.</b> The purple text represents the system prompt used for this task, which contains the information-seeking intents summarized from Table 4.11. The entire conversation is used for this example to demonstrate grounding between a conversation and intents. . . . .	79
4.13	<b>In-context example given as part of a prompt to an LLM for intent summarization with Abg-CoQA.</b> The purple text represents the system prompt used for this task. The blue text represents the summarized information-seeking intents. . . . .	80
4.14	<b>In-context example given as part of a prompt to an LLM for user simulation with Abg-CoQA.</b> The purple text represents the system prompt used for this task, which contains the information-seeking intents summarized from Table 4.13. The entire conversation is used for this example to demonstrate grounding between a conversation and intents. . . . .	81

4.15	<b>In-context example given as part of a prompt to an LLM for user simulation with AmbigSQL.</b> The <b>purple</b> text represents the system prompt used for this task, which contains ground truth SQL query goal. The entire conversation is used for this example to demonstrate SQL query-grounded request disambiguation. . . . .	82
4.16	<b>“Chain-of-Thought” in-context example given as part of a prompt to an LLM for response generation in PACIFIC.</b> The <b>purple</b> text is the overall system instruction. The <b>blue</b> text is what an LLM would have to produce at inference time. . . . .	85
4.17	<b>“Proactive Mixed-Initiative” in-context example given as part of a prompt to an LLM for response generation in PACIFIC.</b> The <b>purple</b> text is the overall system instruction. The <b>blue</b> text is what an LLM would have to produce at inference time. . . . .	87
4.18	<b>Experimental results on PACIFIC’s public evaluation set with additional results using Gemini Flash and Claude Haiku.</b> <i>ACT</i> achieves the strongest performance compared to all tuning approaches across every condition in every metric. Tuning-based adaptation strategies are not given any in-context examples at inference time, whereas inference-time adaptation strategies are prompted with 10 in-context conversation examples using the same strategy. . . . .	88
4.19	<b>Abg-CoQA test set evaluation results with additional results using Gemini Flash and Claude Haiku.</b> <i>ACT</i> outperforms SFT across all evaluations in all three data settings. However, Gemini Ultra achieves the strongest downstream task performance when prompted with in-distribution in-context conversation examples. . . . .	89
4.20	<b>AmbigSQL test set evaluation with additional results using Gemini Flash and Claude Haiku.</b> Zephyr tuned with <i>ACT</i> is able to achieve the strongest task performance within each data setting. There are especially large performance improvements in post-clarification SQL execution match when data resources are more scarce. . . . .	90
4.21	<b>Examining ACT on PACIFIC with unlabeled conversational data.</b> We assume no access to action labels and instead use 0-shot Gemini Pro as the source of action label supervision. . . . .	91
4.22	<b>Ablation study of various conditions</b> using PACIFIC’s 50 conversation setting. . . . .	92
4.23	<b>Analysis of the impact of additional data on Abg-CoQA.</b> Additionally many-shot examples do not necessarily improve implicit action recognition performance. <i>ACT</i> tuning with Zephyr 7B greatly outperforms many-shot Gemini performance. . . . .	94
5.1	<b>Example of the Listening Comprehension auxiliary task for Spoken-SQuAD.</b> . . . .	103

5.2	Example of the commonsense Response Selection auxiliary task for Spoken-SQuAD.	104
5.3	Example of the Response Generation auxiliary task for Spoken-SQuAD. . . . .	105
5.4	<b>Comparison of ASK-QA against existing popular SQA training datasets used for experimentation here.</b> ASK-QA features ambiguous requests and long audio context. . . . .	106
5.5	<b>Comparison of the dialogue turns from a conversation in Abg-CoQA and the paraphrased version in ASK-QA.</b> Bold: text paraphrased using Gemini. We do not perturb the user turns in order to maintain the original linguistic diversity. However, for the Assistant turns, we paraphrase the language concisely if necessary in order to ensure that the speech is grammatically correct. Here, the only dialogue turn that differs is 6a. (a) and (b) denote differing trajectories, which are the turns that the Assistant has to navigate successfully during evaluation. . . . .	111
5.6	<b>A modified conversation in ASK-QA.</b> Bold: paraphrased text using Gemini. Italics: repeat disfluency injected using LARD (Passali et al., 2022). . . . .	112
5.7	Example of the commonsense Response Selection auxiliary task for ASK-QA. . . .	113
5.8	Example of the Dialogue Listening Comprehension auxiliary task for ASK-QA. . .	114
5.9	Example of the Story Listening Comprehension auxiliary task for ASK-QA. . . .	115
5.10	Example of the Response Generation task for ASK-QA. . . . .	115
5.11	<b>Experimental results comparing single-task SFT and our proposed multi-task approach on SD-QA’s test set.</b> . . . . .	118
5.12	Systematic ablations of each individual task type on ASK-QA. . . . .	118
A1	Pair of dyadic conversations generated using OPT 30B. The prompt recipe given is: “The following is a conversation between Alice and Bob about their hometowns. Bob is from Austin, Texas, and Alice is from New York City.” . . . . .	149
A2	Pair of dyadic conversations generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob about their majors. They are both students at Stanford University.” . . . . .	150
A3	Example dyadic conversation generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob about part-time jobs. Bob works at a coffee shop and Alice is thinking about applying.” . . . . .	151

A4	Pair of triadic conversations generated using OPT 30B. The prompt recipe given is: “The following is a conversation between Alice and Bob and Claire about gardening. Alice is interested in growing vegetables.” The ordering of the speakers in multi-party scenarios is variable. . . . .	151
A5	Example triadic conversation generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob and Claire about cooking. Alice is interested in Southern Ice Tea. Bob has a great recipe. Claire loved trying Bob’s Southern Ice Tea.” . . . . .	152
A6	Example triadic conversation generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob and Claire about health. Alice and Bob are discussing healthy foods. Claire is a nutritionist.” . . . .	152
A7	Conversation from the MELD corpus (Poria et al., 2019). Three speakers are involved, discussing a living situation regarding a fourth character who does not appear in this scene. . . . .	152
A8	Three excerpts of the same conversation from the MPC corpus (Shaikh et al., 2010). The conversation spans topics ranging from the Bush administration to meta-discussion about the collection task. . . . .	153
A9	Corresponding background information written for each of the subtopics found in the FITS dataset. There is a mixture of prompts which only mention one speaker and prompts which mention two speakers. Every synthetic conversation involves both speakers. . . . .	154
A10	Triadic background information written for each of the subtopics given in the FITS dataset. Unlike Table A9, each of these may include background information for up to three people. . . . .	155
A11	Handwritten conversation examples of varying length. In-context examples are randomly sampled from this pool and used as part of a prompt for dyadic conversation generation. . . . .	156
A12	Triadic conversation recipes written for each of the “generic topics” given in the FITS dataset. These conversation recipes are included after the in-context examples when prompting LLMs to generate synthetic conversations. Unlike Table A9, each of these conversation recipes may include background for up to three people. Continued in Table A13. . . . .	157
A13	Triadic conversation recipes written for each of the “generic topics” given in the FITS dataset continued from Table A12. . . . .	158

B1	Prompting LLM to generate a system response conditioned on a dialogue act. “Role” is used to be consistent with ChatGPT’s API. Blue texts are prompts characteristic of this task. . . . .	159
B2	Prompting LLM to simulate a user response conditioned on past (simulated) history. “Role” is used to be consistent with ChatGPT’s API. Blue texts are prompts characteristic of this task. This one-shot example is the same as B1, but a) with roles reversed and b) user response is prepended with a label. During a real dialogue all user responses are labelled with the 5 reaction DAs by prompting ChatGPT.	160
B3	Prompting LLM as a value function at the current state. Blue texts are prompts characteristic of this task. “Persuader: Would you be interested in donating to Save the Children?” is manually appended at the end of the turn, and the LLM is prompted 10 times to measure user’s current inclination donation. “Role” is used to be consistent with ChatGPT’s API. This one-shot example is the same as B2. . .	161
B4	Prompting LLM as a prior policy at the current state. “Role” is used to be consistent with ChatGPT’s API. Blue texts are prompts characteristic of this task. The system response is sampled 15 times. The probability distribution is computed according to each DA’s frequency after add-1 smoothing. This one-shot example is the same as B1, but with dialogue actions appended in front of each system response. . . . .	162
B5	Example conversation with GDP-ZERO as the policy planner. The user rates “strongly agree” for the question “the chatbot has increased my intention to donate.” . . . . .	163
B6	Example conversation with GDP-ZERO as the policy planner. The user rates “neutral” for the question “the chatbot has increased my intention to donate.” . . . . .	164
B7	Example conversation with ChatGPT as the policy planner. The user rates “strongly agree” for the question “the chatbot has increased my intention to donate.” . . . .	165
B8	Example conversation with ChatGPT as the policy planner. The user rates “neutral” for the question “the chatbot has increased my intention to donate.” . . . . .	166
B9	Example conversation with RAP as the policy planner. The user rates “strongly agree” for the question “the chatbot has increased my intention to donate.” . . . .	167
B10	Example conversation with RAP as the policy planner. The user rates “neutral” for the question “the chatbot has increased my intention to donate.” . . . . .	168
C1	<b>“Standard” in-context example given as part of a prompt to an LLM for response generation in PACIFIC.</b> The purple text is the overall system instruction. The blue text is what an LLM would have to produce at inference time. . . . .	169

C2	<b>Comparison of PACIFIC trajectories using different tuning approaches.</b> Left: snippet of from a passage from the PACIFIC test set. Middle: trajectory resulting from the SFT model in response to the user query. Right: trajectory resulting from the <i>ACT</i> model. Red text: sampled from <i>U</i> . Both the SFT and <i>ACT</i> model answered the question incorrectly. . . . .	170
C3	<b>Comparison of PACIFIC trajectories using different tuning approaches.</b> Left: snippet of from a passage from the PACIFIC test set. Middle: trajectory resulting from the SFT model in response to the user query. Right: trajectory resulting from the <i>ACT</i> model. Red text: sampled from <i>U</i> . Both the SFT and <i>ACT</i> model answered the question incorrectly, and neither asked a clarification question. . . . .	170
C4	<b>Comparison of Abg-CoQA trajectories using different tuning approaches.</b> Left: snippet of a relevant passage from the Abg-CoQA test set. Middle: trajectory resulting from the SFT model in response to the user query. Right: trajectory resulting from the <i>ACT</i> model. Red text: sampled from <i>U</i> . Ground truth outcome: Meghan asked Lizzie, “Are you awake?” . . . . .	171
D1	<b>Comparing single-task (ST) tuning to our multi-task (MT) fine-tuning on ASK-QA’s test set.</b> . . . . .	178
D2	<b>Experimental results comparing single-task SFT (ST) and our proposed multi-task approach (MT) on Spoken SQuAD’s test set.</b> . . . . .	179

## Acknowledgements

Reaching the end of my PhD journey is a testament of the unwavering support and care I received from countless individuals throughout my time at Columbia. First and foremost, I would like to express my sincere gratitude to my advisor, Zhou Yu, who took a chance on me as an undergrad with no experience in language generation. I am so grateful for her patience in allowing me to grow, struggle through problems, and explore a plethora of various ideas throughout the years. I appreciate that she would always advocate for me and go out of her way to connect me with the right collaborators, and that she was always conscious of how to best support my own career goals. Thank you for your guidance, support, and kindness, Zhou! You have been the best role model to me over these last several years.

I would also like to thank Joretta Joseph, Art Drisko, and the teams at GFSD and the NSA for their generous financial support throughout my PhD. They allowed me to focus on pushing myself to explore the boundaries of human-centric conversational intelligence free of distractions.

I am also truly grateful for my committee members: Julia Hirschberg, Kathleen McKeown, Alexandros Papangelis, and Dilek Hakkani-Tur. Julia and Kathy have been great mentors to me over the years, generously agreeing to participate in multiple milestone committees and always providing kind and insightful advice. Alex and Dilek not only participated in my dissertation committee, but were also my first internship mentors at Amazon Alexa AI in Summer 2022. Working with them was an incredibly enjoyable experience which broadened my horizons and inspired me to pursue industry research as a career path.

I was very fortunate to have spent 10 months of my PhD as an intern at Amazon Alexa AI,

and 19 months as an intern at Google Cloud AI Research working with Sercan Arık and Ruoxi Sun. I deeply appreciate that they taught me to navigate the broader impacts of my work through their perspectives in both research novelty and real-world impact, and went above and beyond to ensure my research could influence product directions. I am so grateful for the many amazing friends, collaborators and mentors I have met at Google and Amazon, including Ta-Chung Chi, Mert Inan, Daniel Jiang, Yohan Jo, Seokhwan Kim, George Lee, Yang Liu, Rohan Mukherjee, Vishakh Padmakumar, Tomas Pfister, Andy Rosenbaum, Yiwen Song, Chenyang Tao, Xingchen Wan, and Zhiyang Xu.

I would also like to thank my amazing labmates and friends that from my time at Columbia, including Debasmita Bhattacharya, Andy Chae, Eric Chang, Yanda Chen, Ted Chung, Huy Ha, Zach Horvitz, Yukun Huang, Sharfin Islam, Siyan Li, Yu Li, Kun Qian, Weiyan Shi, Matthew Toles, Sky Wang, Qingyang Wu, Xiao Yu, Xuanming Zhang, Mingyang Zhou. I would be remiss to not mention all of the amazing people who stayed close after high school and college, and the ones who I met just by virtue of being in Manhattan, including Adam, Alden, Andrew, Brandon {Feng, Truong, Wang, Wong}, Chris, Daniel, Dave, Derry, Michael, Phil, Peter, Reed, and Zack. Whether it was bouncing ideas off each other, venting, or exploring New York, I am fortunate to have met each of you!

I am also deeply grateful for all those who passed through the Church in New York City, including Abraham, Adelyn, Austin, Ben D., Ben L., David, Davina, Dean, Erica, Esther, Gloria, Grace, Grant, Hui, Ian, Jarrod, Joe, John K., John P., John S., Joseph, Josh, JP, Justin, Keanna, Kelsey, Lily F., Lily Z., Monica, Miguel, Nathan, Orianne, Paul, Peter, Sam, Shengkun, Shulam-mite, Serena, Simone, Sophia, Stanley, Tian, Timothy, Topstone, and many others. Your care over the years has been cherishing and is full of eternal value.

Finally, I would like to thank my parents, Xiaojie Chen and Fan Lin, and my sister, Sophie Chen. Thank you for your unwavering love and for supporting my decision to go down this path. I could not be here without you. Last but certainly not least, a special thank you to my beloved wife, Christyn. Thank you for always loving me, caring for me, and standing by my side.



## **Dedication**

For Grandpa, Lin Qijin, and Grandma, Wu Meijin. Thank you for a lifetime of love and care.

## Chapter 1: Introduction and Background

Humans are inherently social beings which seek to communicate with one another (Sperber, 1995) using both written and spoken language for a variety of purposes ranging from developing social relationships to achieving common ground to accomplish collaborative goals (Yu, 2017). This has thus led to rapidly growing interest in the development of agents which can communicate intelligently with humans and assist them in diverse contexts (Zhang et al., 2024) such as open-domain (Huang, Zhu, and Gao, 2020), task-oriented (Budzianowski et al., 2018b; Rastogi et al., 2020; Zuo et al., 2021), and non-collaborative yet goal-oriented conversation (Li et al., 2020b).

In the last several years, as research in deep learning for natural language processing has matured, efforts in building such dialogue systems quickly shifted from rule-based and template-based systems towards learned neural-based systems (Chen, Celikyilmaz, and Hakkani-Tur, 2017; Chen et al., 2017). Even more recently, research in dialogue systems has been undergoing another paradigm shift with the onset of the era of large pre-trained language models (LLMs) which are overwhelmingly more capable than smaller language models across generative tasks (Kaplan et al., 2020), with these highly overparameterized LLMs demonstrating impressive performance in the few-shot regime (Brown et al., 2020; Chen et al., 2023a). However, the fact that modern LLMs are able to easily produce highly coherent generation does not mean that such models can directly be used as conversationally intelligent agents. In particular, intelligent conversational agents should be able to navigate *mixed-initiative interaction* environments, where each interacting party is able to take control of the interaction flow (Allen, Guinn, and Horvitz, 1999). In such settings, an agent should be able to execute conversational strategies which are optimized to achieve common ground with individual users, resulting in improved conversational outcomes (Chu-Carroll, 2000; Deng et al., 2023d).

In this dissertation, we address three key challenges inhibiting the multi-turn intelligence of

conversational agents in mixed-initiative settings: **scarcity of in-distribution conversational data**, methods for **pragmatic action optimization**, and limitations of existing **LLMs’ understanding of user behavior across non-textual modalities**.

## 1.1 Overview

This section provides an overview of foundational work towards addressing open problems in mixed-initiative conversational intelligence, and introduces the proposed solutions in this dissertation.

### 1.1.1 Challenge: Data Scarcity

**Overview** First and foremost, building intelligent, action-optimal conversational agents for a given setting requires a plethora of in-domain data for model training or fine-tuning (Chen et al., 2022b). However, such data can be difficult to obtain, whether it be due to privacy restrictions in commercial settings or a lack of existing users (Chen et al., 2023b; Deng et al., 2023d). As a result, much work has been devoted to finding methods for building conversational agents with limited data resources (Qian, 2024; Longpre et al., 2024a; Longpre et al., 2024b; Gupta et al., 2022).

A particularly promising direction of research has been on the collection and synthesis of data resources (Chen et al., 2022b; Qian, 2024; Qian et al., 2025). Zhang et al. (2024) and Gupta et al. (2022) developed resources which unified multiple conversational datasets to improve LLM performance in conversational tasks through the construction of textual instructions. Kim et al. (2023) distilled million-scale conversational data from InstructGPT to train competitive open-domain dialogue models, representing a large step forward from earlier conversational models trained mostly on human-collected data such as GODEL (Peng et al., 2022) and BlenderBot (Roller et al., 2021; Shuster et al., 2022).

**Contributions in Chapter 2: Curating Synthetic Multi-turn Conversation Data** In Chapter 2, we propose constructing a set of methods which curate resources for conversational modeling

to address the data scarcity issue in building agents. Recognizing that LLMs exhibit strong controllability in conversational settings (Chen et al., 2023a; Qian et al., 2021) and are capable of producing highly coherent and realistic dialogue utterances (Brown et al., 2020), we present an approach to distilling high-quality conversations from high-capacity LLMs. We demonstrate that these synthetic conversations are comparable to human-collected dialogues, and fine-tune high-quality dialogue generation models on the resulting dataset. We additionally demonstrate that our approach can be generalized beyond the dyadic setting and produce the first-ever synthetic multi-party conversation dataset.

### 1.1.2 Challenge: Pragmatic Action Planning

**Overview** Once conversational data is made available, the next step is to select the right algorithms to effectively utilize the data to teach models to generate optimal language. It is widely accepted that complex, mixed-initiative interactions, can be effectively framed as a partially observable Markov decision process (POMDP) (Yu, 2017; Yu, Chen, and Yu, 2023; Deng et al., 2023b). This framework allows for modeling the inherent uncertainty in understanding user states and intentions based on observed dialogue history. Consequently, reinforcement learning (RL) techniques have become a prominent approach for optimizing dialogue policies within this POMDP formulation, enabling agents to learn effective conversational strategies through interaction (Yu, 2017; Deng et al., 2023b). By defining states, actions, transition probabilities, and rewards, RL-trained agents can learn to navigate complex dialogue flows and maximize long-term conversational goals (Yu et al., 2022).

A key challenge in mixed-initiative dialogue is learning to select “optimal” actions which progress towards the completion of some goal. A common example is training agents to learn proactivity, allowing the system to take initiative when appropriate (Deng et al., 2023a; Deng et al., 2023d; Deng et al., 2023c). This involves planning actions that not only respond to the user but also actively guide the conversation towards desired outcomes or explore potential user needs. Other mixed-initiative settings commonly involve argumentative strategy selection, where the sys-

tem strategically chooses how to present information or counterarguments to persuade or inform the user effectively, e.g. in bargaining (Lewis et al., 2017a) or persuasion (Wang et al., 2019). Another crucial behavior of interactive agents involves asking clarifying questions, to help resolve ambiguity, confirm understanding, and elicit necessary information from the user, improving the likelihood of conversational success (Deng et al., 2022; Guo et al., 2021; Chen et al., 2025). These diverse efforts holistically create a picture of the requisite skills for mixed-initiative intelligence.

**Contributions in Chapter 3: Explicit Action Planning in the Modular System Paradigm** The classic approach towards developing mixed-initiative conversational intelligence in the neural network era has been to develop modular systems which formulate action planning (Yu and Yu, 2021; Wang et al., 2019) as a discriminative task separate from conditional language generation (Chen et al., 2022a; Keskar et al., 2019; Yu, 2017; Yu et al., 2016) which produces dialogue responses. Such approaches typically mirror the POMDP framework (Deng et al., 2023b).

In Chapter 3, we propose a method for pragmatic action planning on the modular dialogue system paradigm called Goal-oriented Dialogue Planning with Zero Training Data. Unlike earlier studies which require training neural models for user simulation or dialogue generation (Cheng et al., 2022a; V  th, Vanderlyn, and Vu, 2023a; Wang et al., 2020a; Jang, Lee, and Kim, 2020; Yang, Li, and Guo, 2021a) but run the risk of training on noisy dialogue trajectories collected by crowdworkers (Chen et al., 2023a), we use the highly capable language generation abilities of LLMs guided by an externalized control flow (Monte-Carlo Tree Search) to bypass the need for training data. We discuss several empirical experiments which demonstrate large performance improvements over existing modular frameworks in mixed-initiative dialogue modeling.

**Contributions in Chapter 4: Implicit Action Planning in the End-to-End System Paradigm** Generalist LLMs are now being presented to consumers with conversational interfaces after having achieved strong instruction-following capabilities due to their being aligned using techniques such as reinforcement learning from human feedback (RLHF; Christiano et al. (2017) and Ouyang et al. (2022)), such techniques are not suitable for complex mixed-initiative conversational tasks due to

their focus on single-turn optimization (Shani et al., 2024) and focus on token-level optimization (Ahmadian et al., 2024) rather than sequence-level pragmatic actions (Chen et al., 2025; Deng et al., 2024a; Yu, Chen, and Yu, 2023).

In Chapter 4, we propose a solution to adapt generalist LLMs for mixed-initiative conversation. We developed a novel approach to conversational adaptation called Action-Based Contrastive Self-Training, which teaches LLMs to generate responses which represent optimally-selected actions. We apply this novel approach to the mixed-initiative setting of ambiguous conversational question answering (CQA), where an agent must learn to ask clarifying questions, spanning three diverse domains: tabular question answering, machine reading comprehension, and conversational text-to-SQL generation.

### 1.1.3 Challenge: User Modeling in Multimodal Interaction

**Overview** The final aspect of developing intelligent mixed-initiative conversational models involves modeling user preferences and behavioral patterns. To this end, effective conversational interaction must go beyond text-only intelligence. While language carries the explicit semantic load, critical signals which can serve as proxies for end-users’ state, intent, and attention are often conveyed through non-verbal channels – the cadence of speech, spoken emotions, facial expressions, or saliency in the surrounding visual environment (Hirschberg, 1993; Busso et al., 2016; Busso et al., 2008; Li et al., 2023; Wang et al., 2023a). To build truly adaptive and interactive mixed-initiative systems, models must be capable of interpreting these multimodal signals and integrating them into a model possessing a richer understanding of the user.

Many recent efforts in building Multimodal Large Language Models (MLLMs) represents a significant step in this direction. Many recent MLLMs effectively bridge modalities by connecting powerful, pre-trained unimodal encoders (for vision, speech, etc.) with the reasoning and generation capabilities of LLMs. LLaVA (Liu et al., 2024a), was a foundational MLLM which demonstrated the effectiveness of a simple approach which used a trainable linear projection layer to map visual features into the LLM’s input space, followed by instruction tuning. This strategy

drew on the intuition that encoder-only models and decoder-only language models were individually highly capable already. Following similar architectural principles, models like Qwen-Audio (Chu et al., 2023; Chu et al., 2024), SLAM-ASR (Ma et al., 2024), and SALMONN (Tang et al., n.d.) have applied such ideas towards speech understanding tasks, demonstrating state-of-the-art results by projecting an audio representation into the embedding space of a decoder-only language model. These recent approaches build on the foundations of earlier end-to-end architectures such as ESPNet (Watanabe et al., 2018) and SpeechT5 (Ao et al., 2022), which were some of the first to investigate unified text-to-speech/speech-to-text capabilities.

With these foundational multimodal models emerging, research is increasingly focused on leveraging them for more nuanced user modeling. Some efforts aim for unified frameworks capable of modeling user perception across modalities for a holistic view of interaction (Li et al., 2023). Others take user-centric approaches to multimodal modeling in specific downstream tasks, such as refining user models to better capture preferences and context in generative processes for audio (Hung et al., 2024) or images (Liang et al., 2024). Other work has focused on learning to adapt to user-specific behavioral patterns in few-shot settings through techniques that rely on improved knowledge transfer (Chen and Yu, 2023). There is also increasingly more work focused on adapting alignment strategies for the multimodal setting (Fu et al., 2025). Adapting methods like RLHF requires balancing preferences expressed across different modalities; techniques such as Multimodal Direct Preference Optimization (Wang et al., 2024) are being developed specifically to prevent language signals from dominating the alignment process, ensuring feedback from all relevant modalities is considered.

**Contributions in Chapter 5: Multi-modal User-adaptive Conversation Modeling** While the greatly improved capabilities of LLMs have resulted in rapid progress in text-based multi-turn agents and increasingly in vision-based multimodal conversational agents (e.g. (Liu et al., 2024a)) we are yet to reach the same level of progress in LLM-based end-to-end spoken dialogue systems. Unlike the case of LLMs, a highly capable, universal, and easily adaptable pre-trained model has

not been developed for spoken dialogue. Considering speech is a much more expressive representation of humans’ communicative intent (e.g., speech can capture paralinguistic phenomena such as emotions; (Hirschberg, 1993; Levitan et al., 2012)) which is perhaps more commonly used in daily life than written language, it is imperative to develop methods which are able to account for users’ acoustic-prosodic communicative cues and are agnostic to different speaking patterns.

In Chapter 5, we discuss solutions to bridge these gaps in end-to-end mixed-initiative conversational modeling in spoken settings. First and foremost, we describe the construction of a novel mixed-initiative spoken conversational question answering task, which involves both listening comprehension as well as ambiguity recognition. Second, we describe a simple yet effective data-centric approach to multi-task learning to improve MLLMs’ robustness to different speaking patterns. Finally, we demonstrate that our data-centric approach can improve fine-tuning even for frontier MLLMs and examine multiple additional spoken question answering tasks.



## Chapter 2: Resource Curation for Conversation Modeling

Training dialogue models typically requires an abundance of data resources, as with any machine learning task. The traditional gold standard resource curation approach is to collect samples written by humans (ideally in a natural environment). However, collecting high quality data is difficult and expensive, especially for dialogue tasks where there often is no “right answer” when developing the trajectory of a conversation. This motivates the in-depth study of how to supplement such settings by systematically curating large-scale synthetic conversational data. To this end, synthetic data generation approaches typically involve some combination of augmenting existing seed conversational data, synthesizing data without any seed data, or some mixture of both.

### 2.1 Introduction

Typically dialogue data are sourced from crowdworkers and the quality of annotations, evaluations, and conversations can vary considerably (Zhao and Zhu, 2014), often necessitating guardrails such as credential-based worker selection or defensive task design for quality control (Allahbakhsh et al., 2013; Qian et al., 2023). To accommodate data scarcity in training dialogue tasks, low resource methods have become a topic of growing interest and importance (Zhao et al., 2019; Mi et al., 2019; Qian and Yu, 2019; Li et al., 2019). One idea that has gained particular attention is transfer learning — specifically, finding ways to leverage knowledge learned by LLMs for new tasks. LLMs have demonstrated impressive emerging conversational capabilities, enabling big performance improvements in various dialogue tasks (Brown et al., 2020; Shuster et al., 2022; Peng et al., 2022; Kulhánek et al., 2021). Particularly, LLMs have been prompted to augment existing conversational data (Chen et al., 2022b; Mehri, Altun, and Eskenazi, 2022; Sahu et al., 2022).

Given some in-distribution seed examples, augmentation techniques attempt to generate data



Figure 2.1: Pair of dyadic conversation excerpts about hometowns (upper) and pair of triadic conversation excerpts about Ithaca, NY (lower). In both pairings, one conversation is synthetically generated and the other is collected from humans. The top-left is taken from DailyDialog, whereas the top-right is generated synthetically. The bottom-left is generated synthetically and the bottom-right is taken from MPC.

that are faithful to some task distribution (Kim, Jeong, and Cho, 2021). Albeit powerful, one caveat common to all augmentation techniques is that the quality of synthetic data heavily relies on seed examples. But, what if crowdworkers do not possess the necessary background or skill set to complete a task en masse? How can we still get adequate high-quality synthetic data to learn a task?

In this work, we explore a novel application of **P**rompting **L**anguage models for social **Conv**ersation **S**ynthesis (PLACES). Synthesizing conversational datasets allows for the construction of training instances in nonexistent tasks. We specifically conduct open-domain, topic-conditioned conversation generation using few-shot in-context learning with expert-written

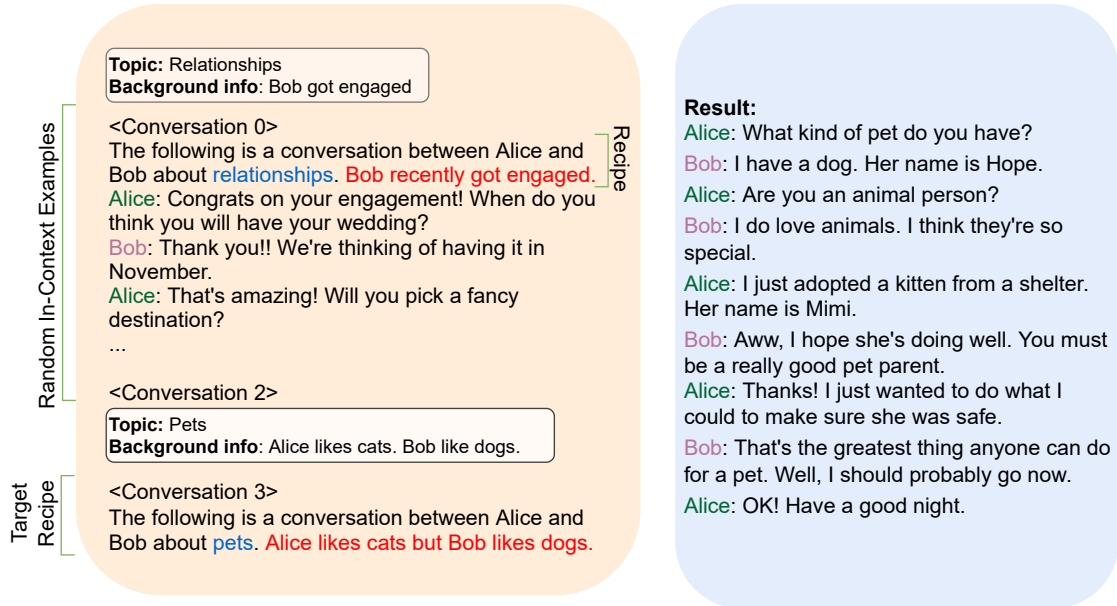


Figure 2.2: Example of the components of a prompt (left) used by OPT 30B to generate a synthetic conversation about pets (right). Conversations in the prompt are prefixed by recipes. Blue text: topic labels. Red text: seed background information metadata.

synthetic conversations. We conjecture that expert end-users know exactly the types of conversations that they need. Rather than using existing datasets, they can simply write a small set of high quality conversation examples according to the structure of their desired conversational outputs. We reason that given structure through high-quality in-context demonstrations, large LLMs are able to utilize their expansive pre-training data (e.g. (Gao et al., 2020)) to synthesize realistic social conversations, implicitly creating personalities and backgrounds for hypothetical speakers. The process of conversation writing would otherwise require human creativity and effort.

Our work makes four core contributions: (1) PLACES involves synthesizing an entire conversational dataset from a few targeted expert-written examples. These conversations match the quality of two widely adopted social dialogue datasets, DailyDialog (Li et al., 2017b) and Topical Chat (Gopalakrishnan et al., 2019), in terms of human evaluation and automatic metrics. (2) We demonstrate that our synthetic conversations can be used as a fine-tuning dataset which matches the performance of its human-curated counterparts as measured by an interactive human evaluation and automatic metrics. (3) We apply PLACES to synthesize data for an under-studied subfield of

dialogue research: multi-party conversations. We evaluate a set of synthetic triadic conversations in comparison to two human-collected multi-party conversational datasets (Shaikh et al., 2010; Poria et al., 2019). To our knowledge, our work is the first to synthesize multi-party conversations, adding to the still-growing body of work on multi-party social dialogue. (4) Lastly, we conduct an error analysis on both dyadic and triadic synthetic conversations. We discuss the implications of our findings, as well as potential solutions to address the generation “errors.”

## 2.2 Related Work

Recently, the zero- and few-shot learning capabilities of large pre-trained language models have overtaken state-of-the-art performance on many classical natural language processing tasks, including dialogue (Brown et al., 2020; Chen, Qian, and Yu, 2023). Many LLMs such as T5 (Raffel et al., 2020), GPT-J (Wang and Komatsuzaki, 2021), GPT-3 (Brown et al., 2020), and OPT (Zhang et al., 2022c) have become the backbone of several dialogue-specific models (e.g., (Peng et al., 2022; Madotto et al., 2021; Shuster et al., 2022)).

In particular, in-context learning, where few-shot examples are provided in the input prompt of a LLM, has been found to provide valuable information in guiding generation output (Min et al., 2022; Brown et al., 2020; Min et al., 2021; Lu et al., 2021b). As a result, many recent efforts in prompting LLMs have sought to augment various natural language processing datasets (Chen et al., 2022b; Wang et al., 2022; Sahu et al., 2022; Mehri, Altun, and Eskenazi, 2022; Rosenbaum et al., 2022a). Prompting has become a viable “solution” for augmentation in dialogue tasks, which have traditionally been considered challenging due to the difficulty of augmenting dialogue context (Chen et al., 2022b).

However, prompt-based augmentation strategies are uncontrolled forms of generation, which may result in generation mistakes for labeled datasets (Sahu et al., 2022; Chen et al., 2022b; Meng et al., 2022). In contrast, other recent studies have instead proposed language augmentation strategies that use complex, highly-controlled frameworks that often involve fine-tuning generators (Papangelis et al., 2021; Zhang et al., 2020c; Kulhánek et al., 2021; Zhang et al., 2020b; Qian,

Wei, and Yu, 2021). Such complex augmentation frameworks require larger amounts of seed data to maintain a ground-truth language distribution (Rosenbaum et al., 2022b; Kim, Jeong, and Cho, 2021), and are more costly than prompting LLMs (Chen et al., 2022b). However, in the context of dataset synthesis, seed data and label correctness are less important considerations. There is no task distribution from which seed data is drawn that LLMs must remain faithful to, and similarly, invariant ground-truth knowledge for language models is dependent on the desired task being synthesized.

Our work differs from existing applications of prompting for conversations along several dimensions. Many studies examine utterance-level generation (Chen et al., 2022b; Sahu et al., 2022; Aher, Arriaga, and Kalai, 2022; Rosenbaum et al., 2022b), whereas our work concerns the synthesis of full conversations. (Bae et al., 2022a) generated conversations for a narrow task and provided evaluations between their synthesis conditions. Recent concurrent work by (Kim et al., 2022) sought to distill conversations from InstructGPT 175B using a commonsense knowledge graph. In our work, we synthesize conversations using an open-source LLM and demonstrate that they are comparable to human-collected datasets, in terms of both conversation quality and usability as a dataset. Moreover, all of these studies only concern dyadic conversations, because the vast majority of conversational tasks are dyadic. Our work is the first study to synthesize multi-party conversations.

## 2.3 Conversation Generation

In this section, we discuss our methods for conversation generation. We first detail the construction of our example conversations, then describe their application to prompting LLMs.

### 2.3.1 Writing Conversation Examples

We simply wrote a pool of ten conversations between two speakers representing everyday dialogue using proper grammar. Along with each conversation, we wrote a brief conversation “recipe” which includes a topic and *background information* for the two speakers. The *background infor-*

Source	Words/Turn	Turns/Conv.
DailyDialog	11.58	7.84
Topical Chat	13.38	21.83
HW Examples	11.00	8.10
Synthetic	10.70	9.29

Table 2.1: Number of words per turn and number of turns per conversation for all conversations. HW Examples represents the ten handwritten conversation examples, and Synthetic represents synthetic conversations generated using OPT 30B.

*mation* represents some more fine-grained information about the two speakers, relevant to that particular topic. For example, Figure 2.2 depicts an example prompt with three in-context conversation demonstrations. Each conversation is prefixed by a recipe and is structured in the same manner: “The following is a conversation between Alice and Bob about *topic*” (e.g., “pets”) followed by detailed background information (e.g., “Alice love cats. Bob is more of a dog person.”).

### 2.3.2 Creating Conversations via Prompting

Each prompt consists of three randomly sampled conversations from the aforementioned pool, along with their accompanying recipe. After experimenting with LLMs of three different sizes (GPT-J 6B, GPT-NeoX 20B, OPT 30B), we primarily use OPT-30B and generate with nucleus sampling with  $p = 0.92$ . Inspired by the format of DailyDialog, our handwritten and synthetically generated conversations fall into three categories: start-to-finish conversations, excerpts from the start to the middle of a conversation, and excerpts from the middle of a conversation. Several examples are given in Appendix A.

In this paper, we generate a dataset using a list of topics and tasks (i.e., subtopics) from the training set of the Feedback for Interactive Talk & Search Dataset (FITS; (Xu et al., 2022)), a human-chatbot dataset designed to determine desirable human-chatbot tasks/conversations. FITS contains 5592 conversations which span 52 conversational topics (e.g., “nutrition,” “philosophy”) with 315 subtopics (e.g., “Italian food,” “Soren Kierkegaard”). We wrote background information for each of the 315 subtopics in the form given in Figure 2.2.

Using the product of this process once results in a new synthetic dataset with 5592 conversa-

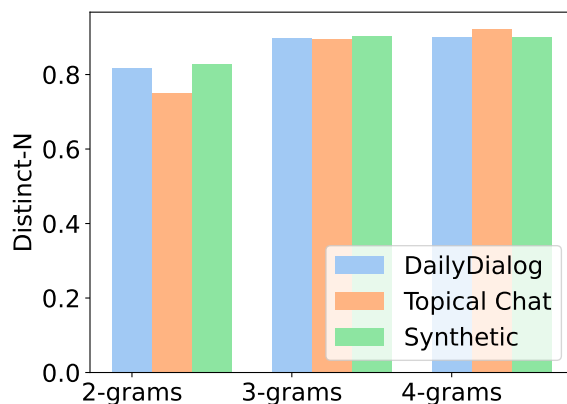


Figure 2.3: Distinct- $N$  with  $N = 2, 3, 4$  for conversations in DailyDialog, Topical Chat, and our synthetic conversations. Our synthetic conversations have the highest most unique bi-grams and tri-grams, and the second-most unique 4-grams.

tions using the same topic, subtopic pairings from FITS. The average length of each conversation is 9.29 turns, with 12.84 words per turn. This is comparable to the dataset statistics of DailyDialog and Topical Chat, as per Table 2.1. In Appendix A, we have included the 315 prompt headers (Tables A9, A10) and the pool of in-context examples (Tables A11, A12, A13).

## 2.4 Human Evaluation Setup

Our human evaluation studies on Amazon Mechanical Turk are evaluated conducted with 28 pre-qualified crowdworkers, who have previously demonstrated proficiency with natural language processing tasks.

### 2.4.1 Static Conversation Evaluation

The crowdworkers were asked to rate conversations from multiple sources according to the following dimensions and instructions.

- *How natural is the overall conversation?*

Scale: 1 (completely unnatural) to 5 (as natural as two native English speakers)

- *How coherent is the overall conversation?*

Scale: 1 (completely incoherent) to 5 (as coherent as two native English speakers)

- *How interesting is the overall conversation?*

Scale: 1 (generic and dull) to 5 (full of content and very engaging)

- *How consistent are each of the speakers' turns?*

Scale: 1 (completely inconsistent) to 5 (no logical fallacies)

- *Does the conversation match the stated topic?*

Options: Yes (1) or No (0)

Each conversation is rated by three crowdworkers, and the median score is selected, following the idea of a majority vote. For multi-party conversations, crowdworkers were asked two additional questions regarding comprehensibility and engagement balance:

- Can you tell which speaker is speaking to which?

Scale: 1 (completely incomprehensible) to 5 (perfectly comprehensible)

- Is each speaker engaged, or is the conversation primarily dominated by one or two of the speakers?

Scale: 1 (totally dominated by one or two speakers) to 5 (all speakers are actively participating in the conversation to an equal degree)

#### 2.4.2 Interactive Conversation Evaluation

For each HIT of the interactive evaluation study, each crowdworker was presented with links to chatbots presented in a randomized order. The link connects each crowdworker to a deployment on an instance of LegoEval (Li et al., 2021). The users are presented with a landing page where they are told that they are interacting with a chatbot, and will be asked to evaluate their conversation experience.

Immediately after interacting with a chatbot, each crowdworker was presented with a survey asking for their impression of the chatbot. In addition to the above dimensions (other than on-



topic), the crowdworkers were asked how engaging, intelligent, and non-repetitive they thought the chatbot was.

## 2.5 Human Evaluation Setup

Our human evaluation studies on Amazon Mechanical Turk are evaluated conducted with 28 pre-qualified crowdworkers, who have previously demonstrated proficiency with natural language processing tasks.

### Conversation Evaluation

The crowdworkers were asked to rate conversations from multiple sources according to the following dimensions and instructions.

- *How natural is the overall conversation?*  
Scale: 1 (completely unnatural) to 5 (as natural as two native English speakers)
- *How coherent is the overall conversation?*  
Scale: 1 (completely incoherent) to 5 (as coherent as two native English speakers)
- *How interesting is the overall conversation?*  
Scale: 1 (generic and dull) to 5 (full of content and very engaging)
- *How consistent are each of the speakers' turns?*  
Scale: 1 (completely inconsistent) to 5 (no logical fallacies)
- *Does the conversation match the stated topic?*  
Options: Yes (1) or No (0)

Each conversation is rated by three crowdworkers, and the median score is selected, following the idea of a majority vote.

For multi-party conversations, crowdworkers were asked two additional questions regarding comprehensibility and engagement balance.

- Can you tell which speaker is speaking to which?

Scale: 1 (completely incomprehensible) to 5 (perfectly comprehensible)

- Is each speaker engaged, or is the conversation primarily dominated by one or two of the speakers?

Scale: 1 (totally dominated by one or two speakers) to 5 (all speakers are actively participating in the conversation to an equal degree)

### 2.5.1 Interactive Evaluation

For each HIT of the interactive evaluation study, each crowdworker was presented with links to chatbots presented in a randomized order. The link connects each crowdworker to a deployment on an instance of LegoEval (Li et al., 2021). The users are presented with a landing page where they are told that they are interacting with a chatbot, and will be asked to evaluate their conversation experience.

Immediately after interacting with a chatbot, each crowdworker was presented with a survey asking for their impression of the chatbot. In addition to the above dimensions (other than on-topic), the crowdworkers were asked how engaging, intelligent, and non-repetitive they thought the chatbot was.

## 2.6 Synthetic Conversation Evaluation Experiments

### 2.6.1 Evaluation of Conversation Quality

Table 2.2 provides a crowdworker evaluation of our synthetic dataset compared against DailyDialog and Topical Chat. We expect Topical Chat to be rated as the most interesting, due to the knowledge-grounding process utilized during the dialogue collection process. We randomly sampled 200 conversations for each conversation source and asked a pre-qualified pool of 28 crowdworkers on Amazon Mechanical Turk (AMT) to rate each conversation. The instructions and details of our human evaluation setup are following the setup explained in Section 2.5.

Source	Interesting	Coherent	Natural	Consistent
<i>Human-Collected</i>				
DailyDialog	3.44	4.51	4.85	4.57
Topical Chat	<b>4.55</b>	4.39	<b>4.92</b>	<b>4.87</b>
<i>PLACES</i>				
GPT-J 6B	3.96*	4.49	4.86	4.36
GPT-NeoX 20B	3.81*	4.40	4.63	4.35
OPT 30B	4.13*	<b>4.61</b> <sup>*†</sup>	4.82	4.63

Table 2.2: Evaluation of conversations randomly sampled from DailyDialog, Topical Chat, and three synthetic datasets generated by prompting GPT-J 6B, GPT-NeoX 20B, and OPT 30B. \* indicates statistical significance over DailyDialog. † indicates statistical significance over Topical Chat. Significance computed at  $\alpha = 0.05$ .

As these conversations are generated using prompting, we first checked whether each conversation followed the prescribed prompt. Crowdworkers identified 95% of the conversations generated by OPT 30B as matching the topic stated in the prompt<sup>1</sup>, indicating this prompting strategy’s effectiveness for topic-grounded conversation generation. Overall, Table 2.2 indicates that synthetic conversations generated by OPT 30B are rated as the most coherent, and more interesting and consistent than DailyDialog. The synthetic conversations are almost as natural as DailyDialog, but are rated as less interesting and natural than Topical Chat. Given our results, we also hypothesize that larger models likely produce higher quality conversations. We provide several examples of conversations generated by OPT 175B using an online web interface<sup>2</sup> in Appendix A.

A concern one might have is that since in-context examples heavily influence prompting (Min et al., 2022; Lu et al., 2021b), our small in-context example size may limit the lexical diversity of our synthetic conversations. Following earlier work evaluating text generation, we use Distinct-N to measure lexical diversity (Wu, Li, and Yu, 2021; Li et al., 2016). Figure 2.3 shows that our synthetically generated conversations are slightly more diverse than both DailyDialog and Topical Chat in terms of distinct bigrams and trigrams, and slightly less diverse than Topical Chat in terms of 4-grams.

We then sought to examine the impact of using expert handwritten examples by comparing

<sup>1</sup>91% and 92% for GPT-J 6B and GPT-NeoX 20B.

<sup>2</sup><https://opt.alpa.ai/>

Dimension	DD-IC	TC-IC	HW-IC
Interesting	3.82	<b>4.35</b>	4.27*
Coherent	4.48	4.56	<b>4.77</b> *+
Natural	4.54	4.69	<b>4.69</b> *
Consistent	4.76	<b>4.87</b>	4.86*
On-Topic	0.91	0.88	<b>0.96</b> *+

Table 2.3: Human evaluation of conversations generated using OPT-30B with in-context examples randomly sampled from DailyDialog (DD-IC), Topical Chat (TC-IC), and handwritten examples (HW-IC). \* indicates statistical significance over DD-IC and + indicates statistical significance over TC-IC.

against synthetic conversations generated using conversations from DailyDialog and Topical Chat as in-context examples. We set the number of conversation examples such that the number of in-context dialogue turns are approximately equal across all conditions. Table 2.3 shows that synthetic conversations generated conditioned on handwritten in-context examples are the most coherent, natural, and on-topic. In terms of interestingness and consistency, the ratings of these conversations slightly trail the ratings of the conversations generated conditioned on Topical Chat.

### 2.6.2 Fine-Tuning with Synthetic Conversations

After establishing that our synthetic conversations are of rather high quality on their own, we attempted to use the synthetic dataset as training data for dialogue models. We fine-tuned distilled BlenderBot 400M (Roller et al., 2021) on DailyDialog, Topical Chat, and our synthetic conversations<sup>3</sup>.

Rather than directly prompting OPT as a response generator, we select BlenderBot as a lightweight, effective dialogue model. This allows for comparisons between the three data sources as training sets, because fine-tuning OPT is prohibitively expensive. Moreover, while prompting with larger LLMs can yield coherent responses, it is generally impractical as an end-to-end dialogue system if hosted on typically available hardware. For long inputs (e.g. with multiple dialogues in-context), generation time typically takes several minutes using OPT 30B<sup>4</sup>.

<sup>3</sup>For fair comparison, we fine-tune on the same number of training instances via downsampling.

<sup>4</sup>All experiments are conducted using one p3dn.24xlarge AWS EC2 instance.

Dimension	DD	TC	PLACES
Interesting	3.35	<b>3.86</b>	3.30
Coherent	3.52	<b>3.71</b>	3.68
Natural	3.52	3.57	<b>3.68</b>
Consistent	3.35	<b>3.65</b>	3.32
Engaging	3.73	<b>3.88</b>	3.65
Intelligent	3.41	<b>3.55</b>	3.24
Non-repetitive	3.37	3.37	<b>3.40</b>

Table 2.4: Interactive human evaluation yields comparable ratings for chatbots fine-tuned on conversations from DailyDialog (DD), Topical Chat (TC), and our Synthetic Data (Syn).

We first performed an interactive human evaluation of the three dialogue models as end-to-end social chatbots using the LegoEval platform (Li et al., 2021). Details on the interactive evaluation are explained in Section 2.5.

Table 2.4 shows that dialogue models fine-tuned on our synthetic conversations are rated comparably to dialogue models fine-tuned on real human-human data — the chatbot fine-tuned on synthetic data appeared to be the most natural and non-repetitive, and was rated as the second-most coherent. It was rated as the least intelligent, engaging, consistent, and interesting. However, two-sided t-tests at  $\alpha = 0.05$  revealed that there was not a statistically significant difference in ratings between the models fine-tuned on all three datasets across all dimensions except for interestingness. The Topical Chat model was rated as significantly more interesting, as expected.

In terms of automatic evaluation, we applied these dialogue models on out-of-distribution test sets to prevent an unfair comparison. We evaluated models fine-tuned on DailyDialog and our synthetic data on Topical Chat, and models fine-tuned on Topical Chat and our synthetic data on DailyDialog. Table 2.5 indicates that in terms of perplexity and ROUGE, models fine-tuned on our synthetic data generalize to out-of-distribution conversational data as well as models trained on real human-human datasets. On the DailyDialog test set, the synthetic dataset model outperforms the Topical Chat model on all metrics except ROUGE-2, and on the Topical Chat test set, the synthetic dataset model underperforms the DailyDialog model on all metrics except perplexity.

Metric (Test Set)	DD-BB	TC-BB	Syn-BB
Perplexity (DD)	—	120.2	<b>87.05</b>
ROUGE-1 (DD)	—	12.34	<b>12.90</b>
ROUGE-2 (DD)	—	<b>1.66</b>	1.52
ROUGE-L (DD)	—	10.60	<b>10.94</b>
Perplexity (TC)	43.3	—	<b>37.1</b>
ROUGE-1 (TC)	<b>16.63</b>	—	15.13
ROUGE-2 (TC)	<b>2.36</b>	—	1.77
ROUGE-L (TC)	<b>13.61</b>	—	12.41

Table 2.5: Out-of-distribution automatic evaluation of perplexity and ROUGE is comparable for BlenderBot fine-tuned on DailyDialog (DD-BB), Topical Chat (TC-BB), and synthetic data generated using our handwritten examples in-context (Syn-BB), respectively.

### 2.6.3 Triadic and Multi-Party Conversations

The vast majority of dialogue tasks and conversational datasets focus on dyadic conversations (e.g. (Li et al., 2017b; Gopalakrishnan et al., 2019; Smith et al., 2020; Rashkin et al., 2019)), following the traditional speaker-listener paradigm (Engelhardt, Bailey, and Ferreira, 2006). In contrast, the literature on multi-party social conversation is rather scarce, not only in terms of conversation generation but as a task altogether. However, while it is an understudied research area, it is incredibly important, because dyadic conversations do not capture the full reality of in-person, human-human social conversations, nor the full potential of dialogue agents. To name a few applications, dialogue agents have the potential to supplement classroom learning with multiple parties, serving as a third mediating party in a debate or discussion between two people, or to provide companionship and support in virtual group settings. A major reason why these lines of work remain unsolved is that there are few large-scale multi-party dialogue datasets.

Many existing multi-party datasets are scripted corpora such as MELD (Poria et al., 2019) or MPDD (Chen, Huang, and Chen, 2020) or HLA-Chat (Ju et al., 2022; Li et al., 2020a). Other multi-party corpora are collected for highly domain-specific purposes, such as multi-party empathetic dialogue (Zhu et al., 2022). Such corpora are also typically collected through asynchronous online platforms, rather than natural conversation. These platforms exist in the form of forums and online chat platforms such as Ubuntu IRC (Lowe et al., 2015) or Reddit (Baumgartner et al., 2020).

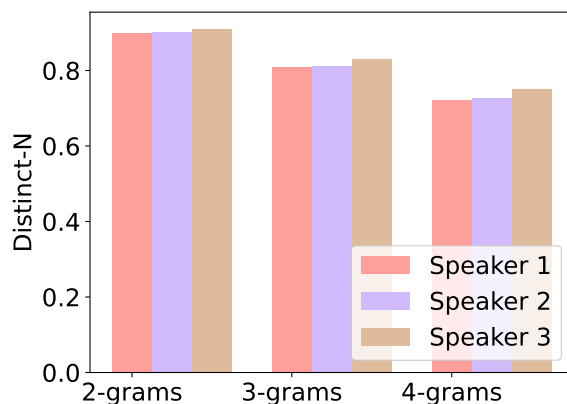


Figure 2.4: Linguistic diveristy (Distinct-N) is comparable for each speaker in the synthetic triadic conversation dataset.

Other more natural multi-party conversational datasets are license-protected speech datasets (e.g. CHIME; Christensen et al. (2010)) which have been constructed for tasks such as speaker attribution.

We find that we can apply our prompting approach to generate synthetic, open-domain, multi-party social conversations following the same structure as our synthetic dyadic conversations<sup>5</sup>. As in the dyadic case, we generate triadic conversations using optional background information for each speaker. We consider the “Multi-Party Chat” corpus (MPC) (Shaikh et al., 2010), a text-based, open-domain conversation dataset collected in real-time online sessions at the University of Albany, and MELD, which contains scripted multi-party dialogues from the popular sitcom “Friends.” We directly compare our synthetically generated conversations against MPC and MELD.

Table 2.6 includes our evaluation of our conversations using the same pool of pre-qualified AMT workers, again with 200 randomly sampled conversations. MPC consists of massive conversation settings — on the scale of 500 turns for a typical conversation session — so we randomly sample 8 to 12<sup>6</sup> continuous turns for each conversation evaluation to more closely match the struc-

<sup>5</sup>While we effectively use Alice, Bob, and Claire instead of Speaker 1, Speaker 2, and Speaker 3, respectively, the order of speakers does not necessarily follow the speaker order in the in-context examples (e.g. Appendix Table A4).

<sup>6</sup>The length between 8 and 12 turns is chosen uniformly.

ture of our synthetic conversations.<sup>7</sup> We present examples of MPC and MELD in Appendix Tables A8, A7.

We inform the AMT workers that they will read conversation excerpts. In addition to the evaluation criteria for dyadic questions previously shown in Table 2.2 (e.g. “Interesting” or “Coherent”), we add two questions to evaluate criteria specific to multi-party conversations. We ask if the conversation excerpt looks comprehensible (in terms of the reader being able to determine who each speaker is addressing), and we ask if all parties of the conversation are participating equally and actively.

In Table 2.6, we find that the synthetic conversations are rated statistically significantly more favorably than MPC and MELD across all dimensions. Beyond conversation quality, it is possible that the ratings for MPC are comparatively low due to the fact that each conversation typically has more than three speakers, which may be more difficult for human raters to interpret. Our results for MELD also indicate that while the corpus is high quality, it may be better fit for comedy and accompaniment with visual context, than as pure dialogue.

Additionally, we checked the linguistic diversity for each speaker. In terms of Distinct-N, each speaker’s lexical diversity is comparable (Figure 2.4) as well as the number of words per turn (12.2, 12.2, and 13.5 for Speakers 1, 2, and 3 respectively). The triadic conversations tended to be slightly longer than the average dyadic conversation (11.5 turns/conversation versus 9.29 turns/conversation).

## 2.7 Discussion

Overall, we find that prompting LLMs to generate synthetic conversations is promising.

### 2.7.1 Considerations for Dyadic Dialogue

The synthetically generated conversations appear comparable to conversations from human-collected datasets. The individual conversations appear interesting, coherent, natural, and con-

---

<sup>7</sup>We sample rather than selecting the first 8-12 turns, to avoid overrepresenting greetings.



Dimension	MPC	MELD	Syn
Interesting	2.48	3.52	<b>4.14*</b>
Coherent	2.40	3.68	<b>4.65*</b>
Natural	2.69	3.69	<b>4.47*</b>
Consistent	2.96	3.83	<b>4.65*</b>
Comprehensible	2.48	3.83	<b>4.80*</b>
Balanced Engagement	3.45	4.00	<b>4.89*</b>

Table 2.6: Synthetic conversations generated using OPT 30B are rated significantly higher than MPC and MELD across all dimensions.

sistent, as the average ratings for each category lie between 4.0 and 5.0. Appendix A includes multiple examples of conversations generated using the strongest performing LLM (OPT 30B, e.g. Table A1) as well as several conversations generated using OPT 175B (e.g. Table A2). Tables 2.4 and 2.5 also indicate that fine-tuning on synthetically generated examples can result in dialogue models of comparable quality, with the potential for further improvements by simply generating more synthetic conversations.

Future work may consider applying applying this generation approach to dyadic contexts beyond social conversations, such as task-oriented dialogue. The clearest difference between social and task-oriented dialogue contexts is the importance of knowledge grounding. In task-oriented dialogue, there typically needs to be retrieval from knowledge base for response generation. An application of PLACES could involve using database results as a ground-truth reference. Rather than using a topic list like FITS, one could form conversational recipes using database search results as background information. Given the apparent semantic control described in Section 2.6, it is possible that synthetic task-oriented conversations would be able to correctly utilize knowledge.

### 2.7.2 Considerations for Multi-Party Dialogue

We found that in comparison to MPC, our synthetic triadic dialogues appear to be of fairly high quality. However, there remain several open questions about multi-party dialogue, even in the triadic case. For instance, there is not a set archetype of conversations. Sometimes, conversations may be dominated by a single speaker, whereas in others, each speaker in the conversation may

contribute equally. Depending on the scenario, a speaker may be the facilitator — meetings can be considered (topic-specific) multi-party dialogues which are typically led by designated speakers. In this study, we have not investigated different archetypes of multi-party conversations; the seed examples used for data generation reflect social environments in which there is not a designated primary speaker. We leave large-scale generation following varying multi-party conversation archetypes to future work.

Moreover, there are several questions about how to utilize multi-party dialogues in an interactive dialogue system. There are use cases where it may be appropriate for one dialogue system to interact with multiple users. On the other hand, in scenarios like emotional support dialogue systems, it may make sense for a single user to interact with multiple simulated conversational parties.

Here, we investigated our approach’s potential to generate synthetic multi-party conversations, hoping to bridge the gap in data availability in multi-party chat. This opens opportunities for a variety of applications. Synthetic datasets could be used to help discover how to properly model triadic and multi-party conversations. In the future, datasets could also be generated for domain-specific, multi-party applications ranging from language learning to task-oriented spoken dialogue systems.

## **2.8 Error Analysis**

We examine the dyadic and triadic conversations which received low scores (1/5) across multiple dimensions.

### **2.8.1 Dyadic Conversations**

Out of the 200 evaluated dyadic conversations, two conversations were rated as generic and dull. One conversation (Table 2.7) talks about the singer, Taylor Swift. However, the conversation is repetitive, repeating utterances such as “What are your thoughts on her?” and “I think she is very nice.” The other conversation is about the filmmaker, Ken Burns (Table 2.8). While the

Party	Utterance
Alice	Hi! So how are things with you?
Bob	Things are going well. Do you know who Taylor Swift is?
Alice	Yes, I think I have heard of her.
Bob	She is a popular singer! Did you know that she has donated \$250,000 to the LGBT+ community in Tennessee?
Alice	\$250,000? That's such a generous donation! She's really selfless.
Bob	What do you think of her?
Alice	She is really talented! I really love listening to her music. What are your thoughts on her?
Bob	I think she is very nice. She seems like a good person.
Alice	Yeah, I think she is a really nice person. I also really love her music! It's really catchy and it really makes me feel good. What are your thoughts on her?
Bob	I think she is very nice. I would really like to meet her.
Alice	You and me both! It would be so exciting!

Table 2.7: Synthetic conversation generated by OPT 30B which was rated as “generic and dull.” While the LLM appears to have drawn on information from an actual event (Taylor Swift did donate \$113,000 to support an LGBTQ advocacy group in Tennessee), the conversation appears to get stuck around the question “What are your thoughts on her?” which results in deteriorating conversation quality towards the end.

Party	Utterance
Alice	Ken Burns is the best! His documentary series on the Vietnam War was excellent. What do you think of him?
Bob	He's pretty cool. I haven't seen that documentary series, but I've heard good things.
Alice	The Vietnam War is a really interesting topic. Have you ever watched Ken Burns' other documentary series?
Bob	I don't think so, but I've seen the Civil War documentary.
Alice	Oh, I've seen that one. It was very good too. Did you see the one on World War II?
Bob	Nope. That sounds cool though.
Alice	I liked it. It was interesting to learn how different people perceived the same event in history.

Table 2.8: Synthetic conversation generated by OPT 30B which was rated as “generic and dull.” The conversation appears to be coherent and is rooted in factually correct information. However, the language is fairly ordinary, which may have been subjectively considered uninteresting.

Party	Utterance
Alice	Do you make cheese at home? I'm trying to learn to make cheese.
Bob	That sounds like a lot of work! What type of cheese do you want to make?
Alice	Cheddar. I've already tried it a few times, but the results aren't as good as I'd like.
Bob	My wife and I love to eat Cheddar. It has a really great flavor.
Alice	How did you learn to make cheese?
Bob	My wife taught me.
Alice	That's awesome!

Table 2.9: Synthetic conversation generated by OPT 30B which was rated as completely inconsistent. Initially, “Alice” is trying to learn to make cheese, but then it becomes ambiguous by their next turn whether they already know how to make cheese.

Party	Utterance
Alice	Ha ha, thanks for stopping by. It was really fun meeting you!
Bob	Thank you too!
Alice	Oh, I forgot to tell you - my dad has cancer. I feel awful.
Bob	I'm sorry. That's awful. My grandmother died of cancer when I was a kid. I remember her fondly.
Alice	What was her name?
Bob	Oh, that was too long ago to remember. She was named John.
Alice	Sorry to hear that. I'm very sorry for your loss.
Bob	Hey - I heard the Giants are playing tonight. Is it ok if I watch the game instead of having more conversation?
Alice	Sure! I'm going to make some dinner.

Table 2.10: Synthetic conversation generated by OPT 30B which was rated as completely incoherent. The prescribed subtopic was “cotton candy” but the conversation pivoted to a much more sensitive topic, cancer.

conversation is appears coherent and uses correct factual information (e.g., making reference to Ken Burns’ documentaries on World War II and the Vietnam War), the language could be perceived as dull.

Three conversations were rated as completely unnatural. In one case, the LLM missed the prescribed subtopic (cotton candy) and instead hallucinated a conversation about a sensitive topic, cancer (Table 2.10). This is also the only conversation to be rated as completely incoherent. The other two conversations are both on-topic. However, one conversation is on-topic but rather short (five turns), whereas the other conversation is overly verbose and a little repetitive.

There were also three conversations were evaluated as completely inconsistent. In all three conversations, the roles of the two speakers seemingly swap. While these hypothetical turns are possible in excerpts of real conversations, they assume background information or events which have not been explicitly established when considered as standalone conversations. An example is given in Table 2.9.

While some of the evaluations may be subjective, an issue that has objectively appeared multiple times is the consistency of speakers’ utterances. The intents and personas of the speakers appear to get switched, which is also an open problem in dialogue systems research. Future work may look to combine conversation synthesis approaches with strategies for dialogue consistency such as the generate-delete-rewrite framework (Song et al., 2020a) or language inference approaches (Welleck et al., 2019; Song et al., 2020b).

### 2.8.2 Triadic Conversations

No conversations were perceived as completely incomprehensible, but human evaluators indicated that two conversations appeared to have imbalanced engagement — in both cases, the third speaker (“Claire”) only has one dialogue turn. As discussed in Section 2.7.2, however, it is not clear whether this is a drawback. Real-life triadic conversations do not follow a set archetype in terms of engagement balance.

There was one conversation which was rated as completely incoherent. In the conversation,

Party	Utterance
Alice	Hi there! I just went on a road trip around California. It was great!
Claire	Cool, do you go on road trips often?
Alice	No, I don't travel too often. I actually haven't been out of the country. It would be cool to go to Paris one day.
Claire	Haha, nice! I'd love to visit Paris too. What's the furthest you've ever traveled?
Alice	I went to Spain with my family last summer!
Bob	Wow, that's really cool! I've never been out of the country either.
Claire	You should definitely go to Paris! What do you like to do for fun?
Bob	Well, I like to read sci-fi books, and I love to play board games.
Alice	Do you have a favorite board game?
Bob	My favorite game is Settlers of Catan. My friends and I play it all the time.
Claire	Haha, I love Settlers of Catan! My family and I play it all the time too.
Alice	That's funny, I love it too. I'm in a Facebook group for it.

Table 2.11: Synthetic triadic conversation generated by OPT 30B which was rated as completely incoherent. First, “Alice” mentions they went on a road trip around California, then they mention that they haven’t been out of the country. This implies that “Alice” is located in the United States. But then Alice mentions having traveled to Spain. The conversation also randomly shifts from discussing travel plans to discussing board games and other hobbies with an awkwardly placed transition on “Claire’s” third turn.

Party	Utterance
Alice	Hello! How's your garden doing?
Claire	It's doing great! I have a young garden, so I'm still waiting for it to develop.
Alice	I can't wait to get home and check on mine! What are you growing?
Claire	I'm growing tomatoes, strawberries, watermelon, and sunflowers!
Alice	That sounds really nice! Do you have a garden somewhere else too?
Claire	No, this is my first garden!
Alice	Oh, I'm jealous! I would love to have my own garden someday.
Bob	I bet you would! I bet you would have a green thumb too.
Alice	Maybe! Maybe I will try starting a garden next year!

Table 2.12: Synthetic triadic conversation generated by OPT 30B which was rated as completely unnatural. The conversation appears to be coherent, but the speakers appear to be very enthusiastic. It could be argued that the speakers are overly unenthusiastic to an unnatural degree.

Party	Utterance
Alice	You know, I think the EU is just a big conspiracy.
Bob	Conspiracy?
Alice	Yes, the European Union wants to take over all of the land from its original owners.
Bob	What do you mean?
Alice	I'm sure that's what they would tell you. But the truth is they're enslaving the people to gain control over resources.
Claire	It sounds a bit extreme. How do you feel about EU leaders?
Bob	I don't like them.
Alice	They're not very nice, I agree.
Bob	I'm kind of scared to go to EU countries and I feel like I don't even have any power.
Alice	Oh, but we do! They won't take away the land of the original owners, because that's just not going to happen!
Bob	I'm still afraid.
Claire	Hm. That sounds like a pretty horrible situation you're in.
Alice	It is! And I don't know what to do.
Bob	Well, it sounds like we all have a lot to learn. I don't know what we should do, but I think there are lots of options out there.
Alice	It's nice that you have an open mind about this! I really do think people should learn more about the EU and what's going on.

Table 2.13: Synthetic triadic conversation generated by OPT 30B which was rated as generic and dull. “Alice” begins a long debate on whether the EU is a “conspiracy” without making a lot of conversational progress.

there is one dialogue turn which presents information inconsistent with prior turns, but the another issue appears to be an oddly placed transition which brings the conversation from travel to hobbies: “You should definitely go to Paris! What do you like to do for fun?” (Table 2.11).

There are two conversations which were perceived as completely unnatural. However, naturalness appears to be a rather subjective evaluation. One conversation is given in Table 2.12, and it is debatable whether the language conventions used are unnatural. One could argue that it is overly enthusiastic, but others could argue that it is how some people speak colloquially. Interestingly, the second conversation which received a low naturalness score is also enthusiastic and about the same topic (gardening).

The only conversation which was rated as generic and dull was a 15-turn debate about whether the European Union is a “conspiracy” (Table 2.13). The debate is rather shallow and does not make a lot of progress.

As with the dyadic conversation error analysis, we see that there are issues with persona consistency. However, unlike the dyadic scenario, there are fewer existing solutions for dialogue con-

sistency. Multi-party conversation synthesis could potentially be improved by applying ideas from systems such as PersonaTKG, which employs a unified graph that encodes personas, utterances, and external knowledge on a scripted dialogue dataset (Ju et al., 2022).

Beyond consistency, in the example from Table 2.13 we see that there is potential for LLMs to hallucinate misinformation. There are again fewer existing studies on circumventing this obstacle in multi-party dialogue, but future work could look to incorporating external knowledge (Kang et al., 2022) or dialogue safety approaches (Kim et al., 2021; Dinan et al., 2019). All said, our work motivates further study into multi-party dialogue consistency, safety, and synthesis.

## **2.9 Limitations**

### **2.9.1 Controllability**

We witness encouraging levels of control through the prompt (95% of the time, the synthetic conversation matches the desired topic), but prompting LLMs is still an uncontrolled form of generation. Future work could seek to add more semantic controls beyond the stated topic in the prompt or explore using weak supervision to provide post-hoc improvements on synthetic data quality, similar to (Chen et al., 2022b). In this work, we also did not thoroughly explore the effects of different generation approaches. Future work may consider applying semantic constraints during the decoding process (Lu et al., 2021a). Further controls are necessary before using this approach for higher-stakes settings such as task-oriented dialogue and other knowledge-grounded tasks.

### **2.9.2 Cost of Human Effort**

While we demonstrate the ability to synthesize large amounts of data, the quality of a synthesized dataset is still dependent on human effort, to an extent. One can use a generic prompt template such as “Alice is interested in [subtopic]” for each subtopic, but we qualitatively see that more detailed background information in a prompt often yields better generation performance.

In this work, we generated 5592 dyadic and triadic conversations, matching the number of topic



combinations in FITS. PLACES can be used to generate many more conversations in the future. Using the same overall can continue to make new combinations of topic and subtopic, or simply rerun the generation process as it is nondeterministic. Moreover, one may consider filling the slots in our conversation recipes using an abundant of external sources, including from existing dataset annotations (e.g. Persona Chat (Zhang et al., 2018)).

### 2.9.3 Computational Costs

Once a dataset is synthesized, small, task-specific models can be used downstream. However, the synthesis method used in this work is still expensive: we prompt LLMs. While we only used freely accessible LLMs such as OPT, we acknowledge that not everyone has access to the number of GPUs necessary to load LLMs, even for inference.

### 2.9.4 Prompt Design

The idea of prompting large language models is not novel. There is a plethora of work that examines how to apply prompting to a variety of different tasks (e.g. (Brown et al., 2020; Min et al., 2021)), along with several studies on how to mine or engineer different prompts (Liu et al., 2021a). In this work, we do not claim novelty to our prompt, nor do we claim that our prompt design is the optimal prompt for conversation generation. Our prompt is designed in a conversational manner, drawing inspiration from (Chen et al., 2022b). We instead emphasize the application of prompting for conversational dataset synthesis. The idea of synthesizing conversational datasets “from scratch” is previously unexplored, and has potential to supplement a lot of areas of dialogue research, such as multi-party conversations.

### 2.9.5 Ethical Considerations

**Human Evaluation and Crowdsourcing:** We make use of crowdsourcing through Amazon Mechanical Turk for several experiments. All crowdworkers were paid at a rate higher than the minimum wage in California, where this experiment was conducted. In accordance with California

State Law, all crowdworkers were also informed they were speaking with chatbots during the data collection for our interactive evaluation. All participants consented to the logging of their responses.

**Language Model Biases:** Large pre-trained language models are typically pre-trained on massive corpora crawled from the internet such as The Pile (Gao et al., 2020) or Common Crawl. This allows language models to have exposure to a large amount of linguistic diversity, but this also results in exposure to a lot of hateful, biased, or otherwise undesirable content from the internet (Luccioni and Viviano, 2021). Future work should examine combining conversation synthesis with dialogue safety approaches.

**Scientific Artifacts:** All scientific artifacts are used according to their intended purpose. The FITS dataset is publicly available at <https://parl.ai/projects/fits/>. OPT is an open-source language model. GPT-J is available for use under the MIT license. We use the HuggingFace Transformers and PyTorch packages for all modeling (Wolf et al., 2020a; Paszke et al., 2019). All artifacts used are in English.

## 2.10 Summary

In our work thus far, we have presented several new contributions to address challenges in conversational modeling resource curation. First and foremost, we propose a new prompt-based framework to leverage LLMs to synthesize large-scale multi-turn conversational data. Next, we create a thorough human evaluation schema for conversational data and demonstrate that these synthetic conversations are comparable in terms of quality and lexical diversity to actual human-human datasets. We also demonstrate that these synthetic conversations are not only high-quality when rated as standalone conversations, but are also useful as training data for open-domain dialogue models. This opens avenues in generative language work such as collaborative and creative writing, story generation, as well as synthesis of new conversational tasks. Here, we presented one example — synthesizing a multi-party conversational dataset. This combined with the finding that

LLMs exhibit highly effective steerability (Chen et al., 2023a) motivates the further use of LLMs to synthesize realistic data for difficult, novel tasks which better reflect real-world use cases, such as conversational code generation.

Given these approaches for curating sufficient conversational data, we can now train coherent dialogue generation modules. However, an aspect of mixed-initiative interaction which is equally important to generation is learning how to select different interaction strategies.

## **Chapter 3: Explicit Mixed-initiative Action Planning via Prompt-based Monte-Carlo Tree Search in Multi-turn Conversations**

The selection of conversational strategies is non-trivial. One can train a dialogue generation model to produce contextually coherent text (Shuster et al., 2022), or even to execute individual strategies (Keskar et al., 2019), but there is not any guarantee that the selected strategy is optimal — many such works rely on assumptions that there is some optimized selection of conversational strategies readily available (Chen et al., 2022a). Action planning is known to be a difficult task which may often require complex training methods built around reinforcement learning (RL) (Deng et al., 2024a), but such approaches often rely on sufficient training data which can be noisy (Chen et al., 2023a). In this chapter, we describe a solution that leverages the impressive zero-shot generation capabilities of modern LLMs with an external control flow to improve conversational action selection.

### **3.1 Introduction**

In many goal-oriented conversation tasks, interacting parties must retake initiative (Allen, Guinn, and Horvitz, 1999) by executing conversational strategies to lead the conversation to a desired outcome (e.g. successful negotiation (Lewis et al., 2017b) or emotional support (Liu et al., 2021b)). As such, it is imperative to have high-quality dialogue policy planners that can prescribe an “optimal” strategy at each turn of the dialogue (Levin, Pieraccini, and Eckert, 1997; Zhang et al., 2020d; Liu and Lane, 2017; Liu et al., 2018).

Optimal policy planning is a difficult task. While in task-oriented settings (e.g. restaurant booking) there at least is objectivity with respect to successful planning, many goal-oriented tasks like persuasion are often subjective. For instance, in a persuasion task, individual persuaders might

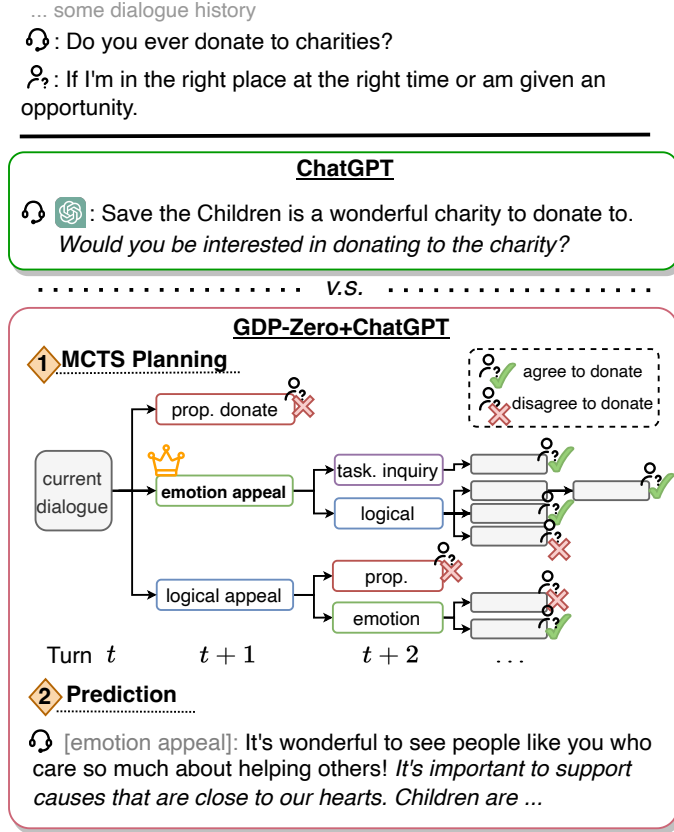


Figure 3.1: Using GDP-ZERO for persuasion with zero model training.

adopt different strategies, making it difficult to train or evaluate a policy planner<sup>1</sup>. Moreover, “optimality” in these complex tasks may require expert domain knowledge (e.g., negotiation skills). This also makes collecting and annotating high-quality conversations difficult (Chen et al., 2023a).

In this work, we contribute a novel approach to **Goal-oriented Dialogue Planning with Zero** training (GDP-ZERO). GDP-ZERO prompts a large language model (LLM) to perform planning by simulating future dialogue interactions (3.1), making it particularly suitable for tasks which would otherwise require high-quality conversations and annotations. Unlike previous approaches, we treat policy planning as a stochastic game, and use prompting for every stage of an open-loop tree search. We evaluate GDP-ZERO on PersuasionForGood due to its difficult planning

<sup>1</sup>In our extensive initial experiments to build a policy planner for the PersuasionForGood (Wang et al., 2019) task, we found 1) training a supervised or reinforcement learning-based multi-class classifier for next-dialogue-act-prediction yields only 18-25% accuracy at best (including finetuning LLaMA (Touvron et al., 2023), RoBERTa (Liu et al., 2019), and prompting LLMs such as ChatGPT), and 2) converting the multiclass policy prediction task into a binary task of whether to “propose to donate” still yields only 74% accuracy.

task (Wang et al., 2019), and find its responses are preferred over ChatGPT in both static and interactive evaluations.

### 3.2 Related Work

**Prompting Methods** Few-shot dialogue techniques have many advantages, including out-of-domain generalization (Zhao and Eskenazi, 2018; Mehri and Eskenazi, 2021) and improved performance in difficult low-resource settings with noisy annotations (Chen et al., 2023a). Recently, prompting LLMs has become the predominant approach to few-shot language tasks, and its applications in dialogue have received much attention. However, this has largely focused on dialogue response generation (e.g. Chen et al. (2023a), Liu and Kilicoglu (2023), Madotto et al. (2021), and Liu et al. (2022)), conversation synthesis (e.g. Chen et al. (2023b), Kim et al. (2022), and Bae et al. (2022b)), and dialogue understanding (e.g. Yang et al. (2022) and Gupta et al. (2022)). To date, prompting has not been used for policy planning.

**Dialogue Policy Planning** Research on dialogue policy planning can be categorized into neural-focused and algorithmic-focused. Neural-focused approaches use annotated dialogues to train dedicated classifiers or value functions to predict the next dialogue acts without explicit look-ahead planning (Zhang et al., 2022a; Zhang et al., 2022b; Cao et al., 2020; Peng et al., 2018a; Zhang, Naradowsky, and Miyao, 2023). For many goal-oriented dialogues, however, both annotated strategies and dialogue responses can be sub-optimal/noisy, as different people can respond differently even given the same context.

To reduce the reliance on a labeled dataset, much work has also attempted to combine neural networks with search algorithms, such as A\* search (Cheng et al., 2022b) and tree search (Wang et al., 2020b; Yang, Li, and Guo, 2021b; Jang, Lee, and Kim, 2020; V  th, Vanderlyn, and Vu, 2023a). However, these methods still require model training for dialogue simulation or value function estimation, and are therefore highly dependent on training data quality (Shi et al., 2019). For example, (Jang, Lee, and Kim, 2020) use Monte-Carlo Tree Search (MCTS) for training a Recurrent Neural

Network-based (RNN) policy model and (Wang et al., 2020b) train multiple neural networks for user simulation and value function estimation. Consequently, these methods can face difficulties during dialogue simulation due to a) noisy data annotations causing sub-optimally trained generation models, and b) inaccurate responses generated at turn  $i$  compounding errors for simulations at turns  $> i$ .

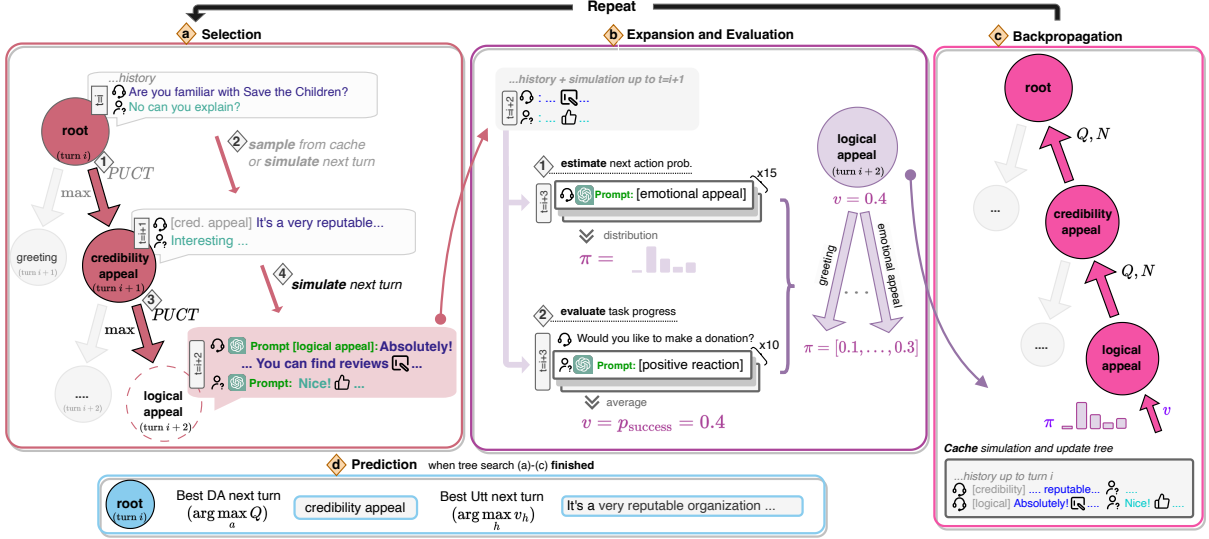


Figure 3.2: GDP-ZERO with ChatGPT backbone. During **Selection**, simulations are either sampled from cache or newly generated. During **Expansion and Evaluation**, we prompt ChatGPT for prior policy  $\pi$  and value estimation.

### 3.3 Methodology

In this work, we introduce GDP-ZERO, an algorithm-focused dialogue policy planner for goal-oriented dialogue tasks like persuasion. GDP-ZERO uses zero model training and instead performs Open-Loop MCTS at decision time by prompting an LLM to simulate user and system response, evaluate current task progress, and predict a prior next dialogue act. Building on early findings (Chen et al., 2023a), our approach has two main differences from existing policy planning work: we use few-shot prompting to bypass the need for model training on noisy data, and we use Open-Loop MCTS to reduce compounding simulation errors by continuously re-generating system and user responses during the tree search.

### 3.3.1 Problem Definition

To introduce tree search methods for dialogue policy planning, we first formulate planning as a Markov Decision Process (MDP). A  $t$  turn dialogue between a user and a system can be represented as:

$$h = (a_0^{\text{sys}}, u_1^{\text{sys}}, u_1^{\text{usr}}, \dots, a_{t-1}^{\text{sys}}, u_t^{\text{sys}}, u_t^{\text{usr}})$$

where  $a_i^{\text{sys}}$  is the system’s dialogue act at turn  $i$ ,  $u_i^{\text{sys}}$  is the system’s response, and  $u_i^{\text{usr}}$  is the user’s utterance at turn  $i$ . Similar to (Yang, Li, and Guo, 2021b) and (Wang et al., 2020b), we define the task of planning the next  $a^{\text{sys}}$  as an MDP problem  $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ . The dialogue act of the system  $a_i^{\text{sys}}$  represents an action  $a_i \in \mathcal{A}$  at a turn  $i$ , and the corresponding dialogue history up to the  $i$ -th turn  $s_i = (a_0, u_1^{\text{sys}}, u_1^{\text{usr}}, \dots, a_{i-1}, u_i^{\text{sys}}, u_i^{\text{usr}})$  represents a state  $s_i \in \mathcal{S}$ . A reward function  $\mathcal{R}(s, a)$  represents the likelihood of a desired conversational outcome, such as persuading a user to donate to a charity. The transition function  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  represents the probability of transitioning from a dialogue state  $s_i$  to state  $s_{i+1}$  after executing  $a_i$  at a turn. Finally,  $\gamma \in [0, 1)$  is the discount factor.

### 3.3.2 Dialogue Planning as a Stochastic MDP

In a typical MDP game like Go, much success has been achieved by employing closed-loop MCTS to plan for the next move/action  $a$  (Silver et al., 2016; Silver et al., 2017). In general, MCTS algorithms improve the actions proposed by an original policy model  $\pi_\theta$  by building a search tree that contain simulated outcomes of many potentially high-rewarding actions, and returning the best action according to the simulation/evaluation results. Specifically, MCTS repeatedly 1) explores a new action or exploits a successful action  $a$  proposed by  $\pi_\theta$  in a state  $s$ ; 2) simulates the action outcome  $s' \leftarrow \mathcal{P}(s, a)$ ; 3) evaluates the action’s quality; and 4) updates its estimate  $Q(s, a)$  for that action. At the end of the search, MCTS produces probabilities  $\pi \propto N(s, a)^{1/\tau}$  for playing each action using the exponentiated visit count, which usually suggests much stronger moves than using the raw policy  $\pi_\theta$  (Howard, 1960; Sutton and Barto, 2018; Silver et al., 2017).



However, in simulating *dialogue* interactions during tree search, generating a slightly improbable system or user response for state  $s'$  and storing it in a search tree could lead to a large compounding error for the rest of the subtree from  $s'$  ((Wang et al., 2020b)). This is because the state space representing all possible responses is large, and dialogue responses are diverse. This makes it difficult to accurately simulate the most probable next dialogue state  $s'$  given an  $(s, a)$  pair from the previous turn. We thus treat dialogue policy planning as a stochastic MDP, where the simulated next state  $s' \leftarrow \mathcal{P}(s, a)$  is drawn from a large unknown distribution and might not be representative of the most probable  $s'$  (Perez Liebana et al., 2015). Unlike previous usages of (closed-loop) MCTS for dialogue which consider a deterministic transition (Wang et al., 2020b; Yang, Li, and Guo, 2021b), this formulation requires potentially different  $s'$  to be returned given the same dialogue context  $s$  and system action  $a$ .

### 3.3.3 GDP-ZERO

To solve this stochastic problem, we base our algorithm on Open-Loop MCTS (Weber, 2010; Perez Liebana et al., 2015), a variant of MCTS where each tree node  $s_i^{tr} = (a_0, \dots, a_i)$  represents the sequence of dialogue *actions* to reach a dialogue turn  $i$ . Instead of using system/user utterances to represent a tree node, this design forces an algorithm to (re)generate the corresponding system and user utterances when traversing the tree (see Figure 3.2). Over time, a tree node  $s^{tr}$  stores statistics derived from executing the sequence of dialogue actions (DA) *without* relying on some specific instance of user/system utterances, which could cause errors to propagate into future simulations. Given a dialogue state  $s_0$ , GDP-ZERO searches for the next best action by iteratively performing action **selection**, search tree **expansion**, action **evaluation**, and **backpropagation** to update tree statistics. After  $n$  iterations, GDP-ZERO outputs a **prediction** for the next best action for  $s_0$ . GDP-ZERO is detailed in Figure 3.2 and Algorithm 1. Below we describe each stage of the algorithm.

---

**Algorithm 1** GDP-Zero ( $M_\theta$ )

---

**input** LLM policy  $M_\theta$ ; dialogue history  $h_i$  until turn  $i$ ; dialogue action space  $a \in \mathcal{A}$ ; hyperparameters  $n, k, c_p, Q_0$

```
1: Repeat for  $n$  searches:
2:   initialize root node  $s_i^{tr}, H(s_i^{tr}) \leftarrow \{h_i\}$ 
3:    $s^{tr} \leftarrow s_i^{tr}$ 
4:   // selection
5:   while  $s^{tr}$  is not a leaf node do
6:      $a' \leftarrow \arg \max_a \text{PUCT}(s^{tr}, a; c_p)$ 
7:      $h^{tr} \leftarrow \text{sample}(H(s^{tr}))$ 
8:      $s^{tr} \leftarrow s^{tr} \cup a'$ 
9:     if  $\text{len}(H(s^{tr})) < k$  then
10:      generate  $h_{\text{new}} \leftarrow M_\theta(h^{tr} \circ a')$ 
11:      save  $H(s^{tr}) \leftarrow H(s^{tr}) \cup h_{\text{new}}$ 
12:    $h^{tr} \leftarrow \text{sample}(H(s^{tr}))$ 
13:   // expansion
14:   generate  $p(a|s^{tr}) \leftarrow M_\theta(h^{tr})$ 
15:    $s^{tr}.p \leftarrow p(a|s^{tr}), s^{tr}.Q \leftarrow Q_0, s^{tr}.N = 0$ 
16:   // evaluation
17:   generate  $v(s^{tr}) \leftarrow M_\theta(h^{tr})$ 
18:   // backpropagation
19:   while  $s^{tr} \neq s_i^{tr}$  do
20:     update  $v_h(h^{tr})$  with 3.3
21:     save simulation  $H(s^{tr}) \leftarrow H(s^{tr}) \cup h^{tr}$ 
22:      $(s^{tr}, a) \leftarrow \text{back to parent of } s^{tr}$ 
23:     update  $Q(s^{tr}, a), N(s^{tr}, a)$  with 3.1
24:   // prediction after  $n$  simulations
25:    $a^* \leftarrow \arg \max_a N(s_i^{tr}, a)$ 
26:    $s_*^{tr} \leftarrow s_i^{tr} \cup a^*$ 
27:    $u^{sys^*} \leftarrow \arg \max_{u^{sys}} v_h(H(s_*^{tr}))$ 
output  $a^*, u^{sys^*}$ 
```

---

**Selection** Given a tree state  $s^{tr}$ , the action  $a^*$  with the highest Predictor Upper Confidence Tree Bound (PUCT) (Silver et al., 2017; Rosin, 2011) is selected to traverse the tree:

$$\text{PUCT}(s^{tr}, a) = Q(s^{tr}, a) + c_p \frac{\sqrt{\sum_a N(s^{tr}, a)}}{1 + N(s^{tr}, a)},$$

where  $N$  records the number of times a  $(s^{tr}, a)$  pair has been visited, and  $c_p$  is a hyperparameter controlling exploration. Since future simulations require a specific dialogue history, we either sample from the node’s simulation cache if its size reached  $k$ , or generate a new simulation based on selected dialogue history  $h^{tr}$  by prompting (Section 3.4.1). We repeat this process until  $s^{tr}$  becomes leaf node.

**Expansion** Once a leaf node is reached, we treat a LLM  $\mathcal{M}_\theta$  as a prior policy by prompting it to generate a distribution of next dialogue acts. This is done by sampling  $\mathcal{M}_\theta$  at temperature  $\tau = 1.0$  for  $m$  times, and converting the sampled DAs into a distribution. Finally, each DA is also initialized with  $Q(s^{tr}, \cdot) = Q_0$ , a hyperparameter controlling exploration.

**Evaluation** We model the value of a state  $v(s^{tr})$  by the probability that its dialogue context  $h^{tr}$  can lead to task success. In a persuasion task to convince a user to donate to a charity, this can be achieved by appending the utterance “Would you like to make a donation?” to the context, and prompting an LLM  $l$  times to simulate the user’s current inclination to donation.

**Backpropagation** At the end of each search, we first store any newly simulated histories  $h^{tr}$  in a cache for each  $s^{tr}$ . Then, we update the statistics of all nodes along the search path:

$$N(s^{tr}, a) \leftarrow N(s^{tr}, a) + 1 \tag{3.1}$$

$$Q(s^{tr}, a) \leftarrow Q(s^{tr}, a) + \Delta Q(s^{tr}, a), \tag{3.2}$$

where  $\Delta Q(s^{tr}, a) = \frac{v(s^{tr}) - Q(s^{tr}, a)}{N(s^{tr}, a)}$ . We also store a value  $v_h$  estimated for each simulated history  $h^{tr}$  that resulted in  $v(s^{tr})$  by a running average:

$$v_h(h^{tr}) \leftarrow \frac{v_h(h^{tr}) \times N_h(h^{tr}) + v(s^{tr})}{N_h(h^{tr}) + 1}, \quad (3.3)$$

with  $N_h(\cdot)$  stores the number of times  $h^{tr}$  is visited.

**Prediction** After all simulations are completed, we select the optimal action  $a^* = \arg \max_a N(s_0^{tr}, a)$  based on the number of times an action has been visited, where  $s_0^{tr}$  is the root node of the tree. To avoid generating a response using  $\mathcal{M}_\theta$  on  $a^*$  again, we also extract from cache the best system utterance associated with executing action  $a^*$  at state  $s_0^{tr}$ . We use  $v_h$  computed during the simulations, and return the utterance with the highest estimated value:

$$h_*^{tr} = \arg \max_{h_{a^*}^{tr}} v_h(h_{a^*}^{tr}),$$

where  $h_{a^*}^{tr}$  denote any history  $h^{tr}$  that played action  $a^*$  after the root state  $s_0^{tr}$ . We call this process “*response selection*”.

**Comparing GDP-ZERO to Traditional MCTS** The full algorithmic outline of GDP-ZERO is presented in Algorithm 1. Similar to other MCTS algorithms, GDP-ZERO performs simulation based on four stages, selection, expansion, evaluation, and back-propagation, and finally predicts an action based on the simulations. Different from existing implementations, GDP-ZERO performs Open-Loop search using *only a generative LLM*  $\mathcal{M}_\theta$ , by prompting it to do dialogue simulation, value function estimation, and prior policy estimation (see Section 3.4.1 for prompting details and examples).

GDP-ZERO requires a generative LLM  $\mathcal{M}_\theta$  as a backbone model, and takes in a dialogue history  $h_i$  at turn  $i$  as input. Given some fixed dialogue action space  $\mathcal{A}$  (see 3.4.2 for P4G), GDP-ZERO builds a search tree after  $n$  simulations. For each state, GDP-ZERO keeps a cache of size  $k$

storing newly generated user and system utterances. We use  $c_p = 1.0$ , and  $Q_0 = \{0.0, 0.25, 0.5\}$  to promote exploration (see Table 3.6).

### 3.4 Experiments

We evaluate GDP-ZERO in the context of PersuasionForGood (P4G; (Wang et al., 2019)), a dataset with 300 annotated dialogues where a “persuader” attempts to persuade a “persuadee” to donate to a charity called Save the Children (Section 3.4.2). Due to the subjective nature of persuasion, it is not clear whether the annotated conversations reflect optimal persuasion strategies (Chen et al., 2022a). This makes training and evaluating a policy planner challenging, since different persuaders likely adopt different strategies (Section 3.1). Motivated by this challenge, we employ GDP-ZERO to perform decision-time planning based on dialogue simulations, and focus on evaluating our approach’s end-to-end performance in achieving the desired task outcome: successfully persuading a user to donate to Save the Children.

#### 3.4.1 Prompting Implementation Details for P4G

For P4G, we used the same one-shot example for all cases, while dynamically changing the representation for each operation.

**System response generation.** Following Chen et al. (2023a), we include the natural language form of a planned dialogue action (Table 3.2) in the prompt to perform conditional generation. We present an example in Table B1.

**User response generation.** We swap the user and the system role for this task, and prompt the LLM to act as a user simulator. We present an example in Table B2.

**Value function estimation.** To evaluate the user’s inclination to donate at a given state, we first append the turn “Persuader: Would you be interested in donating to Save the Children?” to the dialogue history, and then prompt the LLM at temperature  $\tau = 1.1$  to sample the user’s response

Dialogue Act	Category	Amount
Logical appeal	Strategy	325
Emotion appeal	Strategy	237
Credibility appeal	Strategy	779
Foot-in-the-door	Strategy	134
Self-modeling	Strategy	150
Personal story	Strategy	91
Donation information	Strategy	362
Source-related inquiry	Strategy	167
Task-related inquiry	Strategy	180
Personal-related inquiry	Strategy	151
Greeting	Non-Strategy	-
Proposition of donation	Non-Strategy	-
Proposition of amount	Non-Strategy	-
Proposition of confirmation	Non-Strategy	-
Proposition of more donat.	Non-Strategy	-
Experience affirmation	Non-Strategy	-
Thanks	Non-Strategy	-

Table 3.1: Annotated dialogue actions in the P4G dataset.

for  $l = 10$  times. We define “no donation”=-1.0, “negative reaction”=-0.5, “neutral”=0.0, “positive reaction”=0.5, and “donation”=1.0, and then convert the sampled responses to a score between -1.0 and 1.0. We present an example in Table B3.

**Prior policy estimation.** We treat the backbone LLM as a prior policy, and prompt it to generate the next dialogue action at temperature  $\tau = 1.0$  for 15 times to simulate its policy distribution. Finally, to promote the diversity of the generated dialogue actions during GDP-ZERO tree search, we use add-1 smoothing to convert the generated dialogue actions to a probability distribution. We present an example of this in Table B4.

### 3.4.2 Adapting GDP-ZERO Setup to P4G

PersuasionForGood (P4G) is annotated with 10 persuasion strategies and 7 important non-persuasive strategies (see Table 3.1). However, since P4G is collected from human-human interaction, with both the “persuader” and the “persuadee” possibly donating to the charity, some of the

Dialogue Act	Category	Natural Language Form
Logical Appeal	Strategy	The Persuader use of reasoning and evidence to convince the Persuadee.
Emotion Appeal	Strategy	The Persuader uses an emotion appeal to convince the Persuadee.
Credibility Appeal	Strategy	The Persuader establishes credibility of Save the Children by citing its impact.
Task Related Inquiry	Strategy	The Persuader asks about the Persuadee’s knowledge or opinion related to Save the Children.
Proposition of Donation	Non-Strategy	The Persuader asks if the Persuadee would like to make a small donation.
Greeting	Non-Strategy	The Persuader greets the Persuadee.
Other	Non-Strategy	The Persuader responds to the Persuadee without using any persuasive strategy.

Table 3.2: Mapping of persuader dialogue actions to natural language prompts used for prompting LLMs.

dialogue actions are unsuitable when the “persuader” is a chatbot (e.g. self-modeling and personal story). We therefore choose a subset of dialogue actions to plan, by picking 4 frequent persuasive strategies suitable for chatbots, and 3 non-persuasive strategies including “other” to enable the chatbot to deal with unaccounted situations. We present the chosen dialogue actions and their prompts for LLM in 3.2.

### 3.4.3 Static Evaluation

We use ChatGPT<sup>2</sup> (OpenAI, 2022) as the generation backbone of GDP-ZERO, which has become accepted as one of the most coherent general-purpose LLM (e.g. (Liu et al., 2023c; Guo et al., 2023)). We prompt ChatGPT for pairwise evaluation, choosing which of two generated responses is better (i.e., with and without GDP-ZERO planning). Concretely, given two responses  $u_a$  and  $u_b$ , we ask ChatGPT “Which of the following responses can better help the Persuader convince the Persuadee to donate to Save the Children? Why? A:  $u_a$ , B:  $u_b$ , C: Can’t tell.” after providing the relevant task context and dialogue history (see Table 3.3). For every evaluation, we sample the result 5 times and perform a majority vote. Interestingly, we find that ChatGPT

<sup>2</sup>Version “gpt-3.5-turbo” as of 2023 April

skewed towards choosing option A, preferred choosing A for 95.45% when  $u_a = u_b$ . We therefore randomly swap option A and B during all of our evaluations.

Prompt	<p>The following is background information about Save the Children.</p> <p>Save the Children is head-quartered in London, and they work to help fight poverty around the world. Children need help in developing countries and war zones. Small donations like \$1 or \$2 go a long way to help.</p> <p>The following is a conversation between a Persuader and a Persuadee about a charity called Save the Children. The Persuader is trying to persuade the Persuadee to donate to Save the Children</p> <p>{insert dialogue context}</p> <p>Which of the following responses can better help the Persuader convince the Persuadee to donate to Save the Children? Why?</p> <p>A. Persuader: {insert response a}</p> <p>B. Persuader: {insert response b}</p> <p>C. Can't tell.</p> <p>Your can choose from either A, B, or C.</p> <p>Your choice:</p>
Generation	A. (rest truncated)

Table 3.3: Prompting LLM to specify which generated response “response a” or “response b” is more persuasive.

We take the first 20 dialogues from P4G, and produce 154 turns for evaluation. For each turn, we compare the response generated with and without GDP-ZERO for planning. Then, we prompted ChatGPT to choose<sup>3</sup> which generated response is more persuasive (??).

In Table 3.5, we found that generative approaches using ChatGPT were preferred over human ground truth responses about 90% of the time, with the highest score achieved by GDP-ZERO. In Table 3.6, we show that responses generated after GDP-ZERO planning were preferred in up to 59.32% of comparisons. We also observe increasing preference for GDP-ZERO when the number of simulations  $n$  increases. Finally, we find changing  $k, Q_0$  (controls simulation diversity and exploration, respectively) can slightly improve performance (??).

However, we consider the possibility that a) ChatGPT is biased towards its own generated dialogues (Liu et al., 2023b), and b) it might not have a robust criteria of what constitutes *persua-*

<sup>3</sup>ChatGPT can outperform crowdworkers on many annotation and evaluation tasks (e.g. (Gilardi, Alizadeh, and Kubli, 2023; He et al., 2023; Pan et al., 2023; Wang et al., 2021; Liu et al., 2023b))



	Utterance
Context	<p>Persuader: Would you be interested in donating to Save the Children?</p> <p>Persuadee: I am one who like to donate. Can you tell me more about it?</p> <p>Persuader: Save the Children helps to promote childrens rights to health, education, and safety.</p> <p>Persuadee: Do they have any focus right now? More as where are the majority of the fund headed towards right now</p> <p>Persuader: At this point a lot of the focus is children in war zones. some of these donations can address the problem of the daily threat of violence</p> <p>Persuadee: It seems like Syria seems to be a high stakes area. Are any fund going to help the the children in Hawaii after the volcano and hurricane?</p> <p>Persuader: I'm not sure, but I think that would be a great idea. Maybe you could get more information on that by getting in contact with someone at [URL]</p> <p>Persaudee: What are some of the things that they do for the children? I would like to donate but I am not a huge fan or donating just for promotion?</p> <p>Persuader:</p>
Crowdworker	<p>They help to provide basic health care, provide disaster relief, and help to ensure their rights to education and safety. How much would you like to donate ? Your donation will be deducted from your task payment.</p>
ChatGPT	<p>Save the Children provides a variety of services to children in need, including education, healthcare, and protection from harm. They also work with communities to help them become more self-sufficient and better able to care for their children.</p>
GDP-ZERO	<p>Save the Children provides support in a variety of ways. They help create safe spaces for children, provide education programs and materials, and offer medical care and support. They also advocate for policy changes that benefit children's rights around the world. Your donation would go directly to help these programs rather than promotion.</p>

Table 3.4: Example response generated when using GDP-ZERO as the policy planner during static evaluation. Crowdworker's response is the original response in the P4G dataset Wang et al., 2019. GDP-ZERO used  $n = 10$ ,  $k = 3$ ,  $Q_0 = 0.25$  with ChatGPT as backbone.

Algorithm	$n$	$k$	$Q_0$	Win Rate v. Ground Truth
Prompting	-	-	-	$88.84 \pm 0.75\%$
GDP-Zero	5	3	0.00	$87.22 \pm 0.61\%$
GDP-Zero	10	3	0.00	<b><math>90.69 \pm 1.60\%</math></b>
GDP-Zero	20	3	0.00	$88.86 \pm 1.24\%$
GDP-Zero	50	3	0.00	$89.82 \pm 1.10\%$

Table 3.5: Static evaluation with ChatGPT as backbone and judge. Results are  $\mu \pm \sigma$  repeated over three runs.

GDP-Zero (ChatGPT)				Win Rate v. ChatGPT
$n$	$k$	$Q_0$	Run Time	
5	3	0.00	18s	$50.65 \pm 3.31\%$
10	3	0.00	36s	$50.86 \pm 1.10\%$
20	3	0.00	75s	$53.24 \pm 1.91\%$
50	3	0.00	740s	<b><math>59.32 \pm 1.84\%</math></b>
10	1	0.00	16s	$49.57 \pm 2.01\%$
10	2	0.00	29s	$51.30 \pm 1.59\%$
10	3	0.25	36s	<b><math>57.79 \pm 2.95\%</math></b>
10	3	0.50	36s	$53.03 \pm 2.00\%$

Table 3.6: Static evaluation ChatGPT as backbone and judge. Runner-up is colored in gray. Results are  $\mu \pm \sigma$  repeated over three runs.

*siveness*. As such, we also conducted interactive evaluation.

#### 3.4.4 Interactive Human Evaluation

We conducted interactive human evaluation with crowdworkers on Amazon Mechanical Turk, using an interface built on top of the LegoEval platform (Li et al., 2021). We primarily sought to evaluate GDP-ZERO in an end-to-end chatbot against two competitive baselines. Concretely, we compare the rule-based planner from RAP, ChatGPT, and GDP-ZERO in an end-to-end chatbot for the persuasion task.

**RAP** we use the rule-based planner derived from expert knowledge from RAP Chen et al., 2022a, which produces a dialogue action given a dialogue context. We then use the same prompting template in GDP-ZERO (3.4.1, B1), and prompt ChatGPT to produce a system response conditioned

on the planned dialogue action. To our knowledge, RAP is the current state-of-the-art system on P4G, using trained modules with fixed dialogue policies derived from expert knowledge. Chen et al. (2023a) showed that using an LLM (e.g. ChatGPT) as the response generation module for RAP further improved performance.

**ChatGPT** we first use the same prompting template in GDP-ZERO (3.4.1, B4) to obtain ChatGPT’s own prior distribution of the next dialogue actions. We then take the most probable action as the planned dialogue action, and use the same template in GDP-ZERO (3.4.1, B1) to prompt ChatGPT again to produce a system response.

**GDP-ZERO** we use GDP-ZERO with ChatGPT backbone as policy planner, and use the "Response Selection" step to produce both the next dialogue action and the associated system response. We used  $n = 10$ ,  $k = 3$ ,  $Q_0 = 0.25$ , which had a high win rate against ChatGPT during static analysis while also not taking too much time to simulate.

After each conversation, we asked the crowdworkers to evaluate our system based on the criteria in Table 3.7 (e.g. “The chatbot is convincing” or “The chatbot presented a strong argument”).

We require our crowdworkers to be located in the United States and have a HIT acceptance rate of at least 99%. After interacting with each chatbot, each crowdworker was asked to rate their conversational experience. This post-task survey included a validation question which asked what charity they talked to the chatbot about. We had a total of 216 respondents. 74 did not reach or complete the survey, and 31 were removed due to failing the validation question or responding with less than 3 unique sentences. This results in 40 survey results for GDP-ZERO, 35 for ChatGPT, and 36 for RAP. Our study revealed that GDP-ZERO achieves the best performance across all metrics related to persuasiveness. We also found that RAP is highly rated for strategy diversity and relevance, indicating the benefit of using expert knowledge in planning. In 3.5 we provide some analysis of the generated dialogues and the resulting dialogue act distributions using different planners. In 3.6 we include some example dialogues.

The chatbot...	RAP	ChatGPT	GDP-Zero
donation prob. $\uparrow$	$0.72 \pm 0.38$	$0.73 \pm 0.38$	<b><math>0.79 \pm 0.37</math></b>
inc. donation intent $\uparrow$	$4.08 \pm 0.68$	$3.77 \pm 0.90$	<b><math>4.30 \pm 0.71^{**}</math></b>
strong argument $\uparrow$	$3.89 \pm 0.97$	$3.91 \pm 0.99$	<b><math>4.28 \pm 0.74^*</math></b>
is convincing $\uparrow$	$4.11 \pm 0.74$	$4.10 \pm 0.70$	<b><math>4.38 \pm 0.66^{**}</math></b>
diverse strategy $\uparrow$	<b><math>3.98 \pm 0.80</math></b>	$3.83 \pm 1.03$	$3.95 \pm 0.82$
is manipulative $\downarrow$	$2.64 \pm 1.36$	$2.96 \pm 1.38$	<b><math>2.29 \pm 1.33^{**}</math></b>
is natural $\uparrow$	$4.25 \pm 0.68$	$4.03 \pm 0.65$	<b><math>4.38 \pm 0.62^{**}</math></b>
is relevant $\uparrow$	<b><math>4.64 \pm 0.54^*</math></b>	$4.31 \pm 0.86$	$4.59 \pm 0.49$
is coherent $\uparrow$	$4.28 \pm 0.65$	$4.06 \pm 0.89$	<b><math>4.42 \pm 0.49^*</math></b>

Table 3.7: Interactive evaluation using RAP, ChatGPT, and GDP-ZERO as planners, with ChatGPT used for response generation/backbone. Results are  $\mu \pm \sigma$ . All scores scaled to  $[1, 5]$  except for “donation prob.” being  $[0, 1]$ . \* denotes  $p < 0.05$ , and \*\* denotes  $p < 0.01$ .

### 3.4.5 Ablation Studies

To study how various components of GDP-ZERO helped improve task success, we conducted various ablation studies including 1) using Closed-Loop MCTS instead of the open-loop variant; 2) removing the “*response selection*” process; and 3) testing with both ChatGPT and Codex (Chen et al., 2021) as a backbone. In Table 3.8 and Table 3.9, we conduct ablations to study how each component in GDP-ZERO affect its performance. We use the same 20 dialogues from our static evaluation (see Section 3.4.3), and test GDP-Zero with ChatGPT (see Table 3.9) and Codex<sup>4</sup> (Table 3.8) as the generation backbone. Similar to 3.4.3, we use ChatGPT to evaluate the persuasiveness of the generated responses.

In all runs, we use  $n = 20$ ,  $c_p = 1$ ,  $Q_0 = 0$ ,  $k = 3$  for GDP-ZERO, when applicable. Table 3.8 and Table 3.9 shows that GDP-ZERO improves the policy proposed by the raw backbone model, and also that of a Closed-Loop MCTS implementation (denoted as “w/o OpenLoop”) and without the “response selection” process (denoted as “w/o Resp.Select”). This indicates that using the open-loop variant and the proposed “*response selection*” process are beneficial for GDP-ZERO to improve dialogue planning.

<sup>4</sup>At the time of writing, Codex (Chen et al., 2021) was freely accessible from the OpenAI API.

Algorithm	Backbone	Win Rate v. GT
Prompting	Codex	$38.09 \pm 2.00\%$
GDP-Zero	Codex	<b><math>45.46 \pm 2.95\%</math></b>
w/o Open Loop	Codex	$39.16 \pm 3.42\%$
w/o Response Selection	Codex	$40.80 \pm 1.47\%$

Table 3.8: Static evaluation using the first 20 dialogues of P4G with ChatGPT as judge. GT refers to Ground Truth. Results are  $\mu \pm \sigma$  repeated over three runs.

Algorithm	Backbone	Win Rate v. GT
Prompting	ChatGPT	$87.21 \pm 0.60\%$
GDP-Zero	ChatGPT	<b><math>91.13 \pm 0.30\%</math></b>
w/o Open Loop	ChatGPT	$88.09 \pm 0.81\%$
w/o Response Selection	ChatGPT	$91.03 \pm 0.75\%$

Table 3.9: Static evaluation using the first 20 dialogues of P4G with ChatGPT as judge. GT refers to Ground Truth. Results are  $\mu \pm \sigma$  repeated over three runs. Since ChatGPT generations are typically long, we use the first 3 sentences of each generation in this evaluation.

### 3.5 Analysis of GDP-ZERO Dialogues

In our experiments, we utilized GDP-ZERO to generate persuasive responses to convince a persuadee to donate to a charity. In this section, we provide an analysis of how and when GDP-ZERO impacts persuasiveness.

**Static evaluation.** Our static evaluation (3.4.3) uses fixed dialogue contexts from the P4G dataset Wang et al., 2019, and every turn of the conversation is already annotated with dialogue actions. We thus utilize those annotations and investigate *when* GDP-ZERO is producing more persuasive responses. For this analysis GDP-ZERO used  $n = 10, k = 3, Q_0 = 0.25$ , which is the same configuration as in our interactive evaluation (3.4.4, ??). We found that GDP-ZERO is rated as more persuasive 70.59% of the time when the user at the previous turn showed negative emotions (i.e. "negative-reaction-to-donation" or "negative-to-inquiry"), and 59.26% of the time when the user showed positive emotions (i.e. "positive-reaction-to-donation" and "positive-to-inquiry"). We believe this implies that GDP-ZERO’s simulations are most beneficial in persuading users who is having little intent to donation. This is because look-ahead planning may, for instance, help ensure

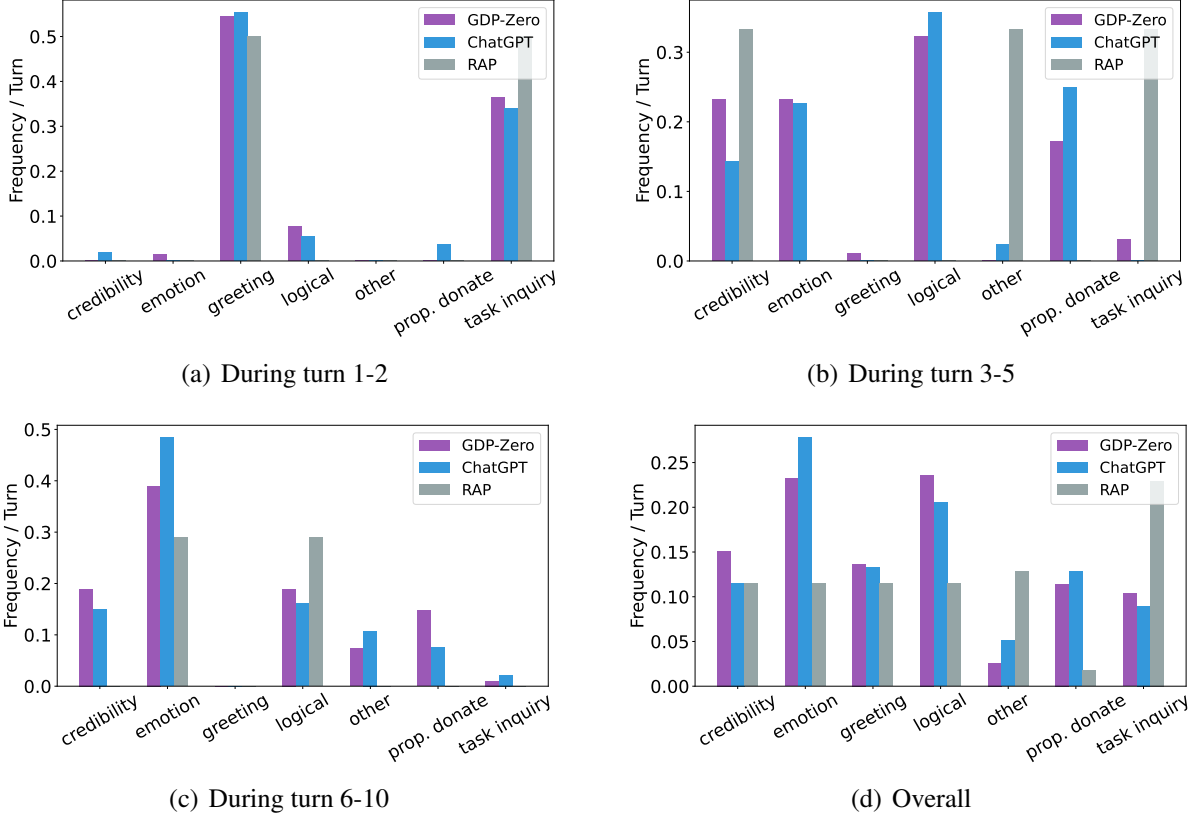


Figure 3.3: Distribution of planned dialogue actions for each planner during interactive evaluations.

a response has covered all of the user’s concerns regarding donation. We present an example in 3.4.

**Interactive evaluation.** Since each interactive dialogue is unlabeled, here we focus on analyzing the known dialogue actions generated by the different planners during those conversations. In Figure 3.3 we present the distribution of planned dialogue actions produced by RAP, ChatGPT, and GDP-ZERO. In general, the planned dialogue actions using ChatGPT and GDP-ZERO are unevenly distributed across different stages of the dialogue. Across different times of the conversation, ChatGPT and GDP-ZERO shared the most frequent DA at each stage is: “greeting” during turns 1-2, “logical appeal” during turns 3-5, and “emotion appeal” during turn 6-10. However, during turns 3-5 GDP-ZERO had a relatively even preference between “credibility appeal”, “emotion appeal”, and “logical appeal”, while ChatGPT strongly preferred “logical appeal.” Ad-

ditionally, we find that throughout all turns ChatGPT prefers to use “emotion appeal” over other dialogue actions, while GDP-ZERO balances between “emotion appeal” and “logical appeal,” and RAP prefers “task related inquiry.” It is also worth noting that GDP-ZERO performs “proposition of donation” less frequently at turn 1-5 than ChatGPT, and more frequently at turns 6-10. This indicates GDP-ZERO potentially adopting a more “conservative” strategy than ChatGPT overall. GDP-ZERO focuses on building up persuasive arguments and is less likely to select “proposition of donation” at early stages of the conversation in order to avoid over eargly propositions, which could lead to persuasion failure (also see B6).

### 3.6 Example Interactive Conversations

We present multiple dialogues from the interactive evaluation (see Section 3.4.4), and compare the strength and weaknesses of the three planners: RAP, ChatGPT, and GDP-ZERO. In Tables B5, B7, B9 we present conversations where the persuadee “strongly agreed” that the chatbot has increased his/her intention to donation for GDP-ZERO, ChatGPT, and RAP, respectively. In Tables B6, B8, B10 we present conversations where the persuadee rated “neutral“ for the question that the chatbot has increased his/her intention to donation for GDP-ZERO, ChatGPT, and RAP, respectively. We observe that in Table B6, while GDP-ZERO did not succeed in convincing the user to donate, GDP-ZERO has been attempting various strategies and does not perform “proposition of donation” prematurely. On the other hand, in Tables B8, B10 both ChatGPT and RAP performs “proposition of donation” at an arguably unsuitable timing.

### 3.7 Limitations

**When is using GDP-ZERO appropriate?** In this paper, we present GDP-ZERO, a general approach for close-domain dialogue policy planning at turn-level. However, in this work we only evaluated GDP-ZERO on P4G. This is because we believe simulation-based plannings would be most beneficial when the task 1) often requires long-horizon planning to be successful, 2) does not have “optimal” action annotations readily available for supervised learning, and 3) evaluation does

not require study participants to follow fake hypothetical scenarios to conduct a conversation, as otherwise the outcome may be disingenuous. We thus believe tasks like persuasion are most suitable, where planning ahead is crucial to success and policy optimality from human demonstrations is extremely subjective. Additionally, such a user may or may not want to donate to the charity, and the task of the dialogue system is to try to persuade them. This is in contrast to other goal-oriented contexts like task-oriented dialogue (TOD), where strong policies can be directly learned due to TOD’s mostly passive and objective nature (e.g. He et al., 2022), and also to dialogues where a hypothetical scenario is required. For example, to design a fair and controlled study, a user may need to make up fake preferences/objectives about a certain product in CraigslistBargains He et al., 2018, or pretend that they need emotional support for some fixed set of issues in ESConv Liu et al., 2021b.

Additionally, while GDP-ZERO can be adapted to task-oriented contexts like MultiWOZ Budzianowski et al., 2018a, it may not necessarily be appropriate. Such task-oriented contexts often have hierarchical policies (e.g. “[hotel] [recommend] name price” and “[restaurant] [inform] food price area”), and adaptation to GDP-ZERO would require converting the hierarchy into a multi-label classification, resulting in a massive action space. We believe this could be very inefficient, and approaches such as building multiple search trees to perform high/low-level planning would be useful Zhang et al., 2020a.

**Runtime** One important limitation for GDP-ZERO is runtime. The more exhaustive the tree search (e.g. increasing  $n$  or  $k$ ), the more likely the algorithm is able to find the optimal dialogue policy (3.6). However, this comes at the cost of longer simulation time, which may affect the overall user experience, and accordingly, user perceptions of persuasiveness.

With OpenAI API’s rate limit and LLM’s inference speed, we restricted GDP-ZERO to plan on 7 dialogue acts in P4G, with  $n = 10$ ,  $k = 3$  for a simulation time of around 35 seconds during interactive evaluation. We believe methods to parallelize tree search Chaslot, Winands, and Den Herik, 2008 or to re-use part of the simulation subtrees could be helpful to speed up GDP-ZERO.



We expect that as research with LLMs progresses, inference speed will continue to improve. In the short-term, one may bypass latency limitations by utilizing multiple accounts to parallelize API calls during simulation.

**Simulation Quality** GDP-ZERO prompts a LLM (e.g. ChatGPT) to perform dialogue simulation and value estimation. Despite LLM’s strong few-shot performance on many tasks, issues with controllable generation can still create errors during simulation (e.g. generated system utterances might not match planned dialogue action). GDP-ZERO accounts for such errors by using an Open-Loop search with  $k > 1$ , but this increases simulation runtime. We believe this trade-off between simulation quality and runtime is also an important aspect for future work to consider.

**Using ChatGPT for static analysis** While ChatGPT is shown to outperform crowdworkers on many annotation and evaluation tasks (e.g. Gilardi, Alizadeh, and Kubli, 2023; He et al., 2023; Pan et al., 2023; Wang et al., 2021; Liu et al., 2023b), recent work has been finding that LLMs like ChatGPT may favor responses with a higher number of unique tokens Wang et al., 2023c. As such, we analyzed responses between GDP-ZERO ( $n = 20, k = 3, Q_0 = 0$ ) and ChatGPT, and found that there is indeed some positive correlation ( $r = 0.29, p < 0.001$ ) between the number of words in a response and ChatGPT’s preference for persuasiveness. However, it is not clear whether this is a limitation in our setting. Our setting of evaluating persuasiveness is not quite the same as the typical context of considering direct preference; there are many situations where longer responses which correlate with persuasiveness (for instance, an aptly used logical appeal). As such, we conducted interactive evaluation with crowdworkers in 3.4.4 and find that GDP-ZERO generated responses are indeed rated as more persuasive.

### 3.8 Ethical Considerations

Our work describes an algorithm to perform dialogue policy planning for goal-oriented tasks without any model training. It is aimed at making future dialogue systems to build, and also better at helping users/systems achieve their tasks/goals.

**Potential Abuses** Generally, while most algorithms are not designed for unethical usage, there is often potential for abuse in their applications. In our experiments with PersuasionForGood (Wang et al., 2019), we apply GDP-ZERO on the goal of increasing users’ intention to donate to a charity. However, because GDP-ZERO is fundamentally goal-agnostic, it is possible to use them for unethical tasks, such as scamming. We do not condone the use of GDP-ZERO for any unlawful or morally unjust purposes.

**Interactive Human Evaluation** In this study, we conducted interactive human evaluation using crowdworkers on the Amazon Mechanical Turk platform. All crowdworkers were informed that they were speaking with a chatbot. All study participants were paid at a rate of \$15 per hour. Our study has received IRB approval.

### 3.9 Summary

In this chapter we propose GDP-ZERO, an algorithm to perform look-ahead policy planning with a large language model for goal-oriented dialogues. Our experiments included end-to-end evaluation which demonstrate that using the policy from GDP-ZERO surpasses previous state-of-the-art system (e.g. RAP as proposed in Chen et al. (2022a)) and direct prompting with state-of-the-art LLMs on the task of persuasion without any model training. Our findings have three major implications. First, our strong performance in the zero-data regime opens the possibility of future work building dialogue systems in more conversational tasks under data-scarce settings. Moreover, our work reveals that zero-data simulation frameworks like GDP-ZERO may circumvent common issues with noisy conversational task data (e.g. Chen et al. (2023a)). Second, the general formulation of GDP-ZERO allows for adaptation to any goal-oriented conversational setting with any base LLM. Our framework does not depend on any task-specific training data, and the variable inputs to each prompt are model-agnostic. Finally, our work reveals the potential of further LLM policy optimization at inference time – we observed that increasing the number simulations can greatly improve performance, which is consistent with the intuition of inference-time compute scaling

laws (Snell et al., 2024).

While GDP-ZERO demonstrates greatly improved performance, due to current hardware limitations, the reliance on repeated invocations of decoder-only LLMs results does incur a significant latency tradeoff (e.g., see Table 3.6). Moreover, the current interaction paradigm of systems such as ChatGPT or Gemini favors end-to-end models without explicit planning modules. This sidesteps the requirement for potentially expensive planning modules, and thus this motivates the development of training methods for end-to-end models which can learn complex planning behavior.

## Chapter 4: Implicit Mixed-initiative Action Planning via Action-based Contrastive Self-Training in Multi-turn Conversations

In this chapter, we propose the problem of *implicit action optimization* to better fit the end-to-end paradigm of current LLM-based assistants. We propose an approach to better adapting LLMs to conversational environments by reformulating the standard preference learning framework to specifically optimize action selection and multi-turn trajectories.

### 4.1 Introduction

Conversations serve as a natural interface for humans and intelligent systems to *collaborate* (Amer-shi et al., 2019; Lemon, 2012), and the impressive capabilities of large language models (LLMs) have been used to power the rapid development of many generalist conversational assistants such as ChatGPT<sup>1</sup> and Gemini (Gemini Team et al., 2023). Such applications present an opportunity for users to verbalize their need for assistance on complex tasks, but the promises of a conversational interfaces also come with the complexities of language. Human conversation is riddled with ambiguity, whether it be due to humans’ tendency to underspecify (Zipf, 1949) or even due to syntactic errors (Messer, 1980). Moreover, disambiguation becomes even more important in complex domains where it can be a difficult multi-turn process to achieve common ground (Beers et al., 2006). As it stands, existing LLM-powered conversational agents continue to struggle with modeling ambiguity (Liu et al., 2023a), and tend to exhibit unwanted behavior such as overhedging (Ouyang et al., 2022) or generating responses which represent a “guess” of the user’s intent (Deng et al., 2023a) (see Figure 4.1).

One of the primary reasons that LLMs may exhibit unwanted conversational behaviors is that their language modeling objective during pre-training or supervised fine-tuning (SFT) is not di-

---

<sup>1</sup><https://openai.com/blog/chatgpt>

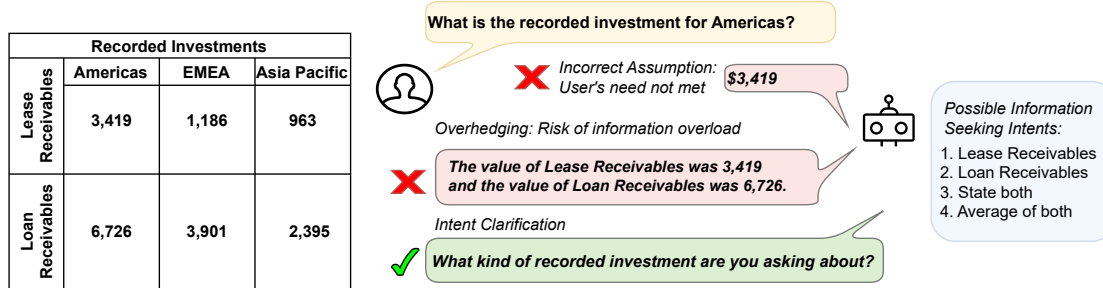


Figure 4.1: Simplified example of ambiguity present at tabular-grounded conversational question answering based on (Deng et al., 2022). A conversational agent should recognize when there is ambiguity and ask a clarifying question towards a more accurate final answer.

rectly aligned with this goal (Ouyang et al., 2022). While approaches like Ouyang et al. (2022) propose LLM “alignment” using post-training approaches like reinforcement learning from human feedback (RLHF; Christiano et al. (2017)), existing models still struggle with conversational tasks spanning multiple turns (Wang et al., 2023b). This is partly due to the fact that existing approaches do not directly optimize for pragmatic skills such as timing clarification questions or logical arguments (e.g. Bai et al. (2022)). Moreover, there is often high variance in the target distribution of a particular use case, so it is imperative that downstream adaptation approaches can effectively guide LLMs to learn effective conversational policies. Given large-scale in-distribution training data, this may be feasible with standard SFT or RLHF. But, dialogue policy learning can be particularly challenging given limited data (Chen et al., 2022b; Dong et al., 2023) and collecting high-quality conversational datasets can be difficult for reasons such as annotation costs and privacy concerns (Chen et al., 2023b).

This motivates the design of a conversational adaptation approach for LLMs which is more closely aligned with the goal of modeling actions in multi-turn conversation. We focus on improving LLMs’ abilities to implicitly select conversational strategies in ambiguous contexts, and propose an approach called Action-Based Contrastive Self-Training (*ACT*). *ACT* is a sample-efficient, quasi-online Direct Preference Optimization algorithm (Rafailov et al., 2024) which focuses on contrasting the differences between an agent’s possible pragmatic conversational actions. We demonstrate *ACT*’s sample-efficient performance on a diverse range of mixed-initiative con-

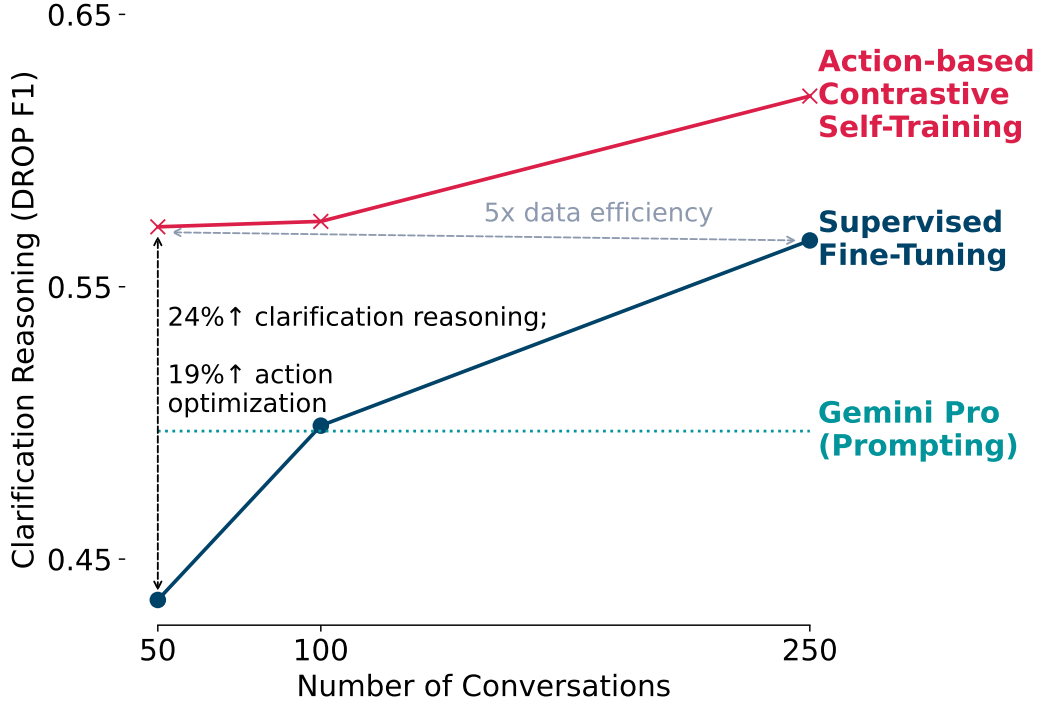


Figure 4.2: *ACT* greatly outperforms standard tuning approaches in data-efficient settings for conversational modeling, as exemplified here on PACIFIC.

versational tasks: (i) tabular-grounded question answering, (ii) machine reading comprehension, and (iii) text-to-SQL generation, demonstrating large improvements compared to standard adaptation approaches (see Figure 4.2). Our work highlights the necessity of considering action-based preferences for conversational tasks, and we propose a workflow for evaluating LLMs’ ability to recognize and reason about ambiguity in conversation.

## 4.2 Related Work

### 4.2.1 Mixed-Initiative Conversational Agents

Neural approaches to building mixed-initiative conversational agents typically consist of two core components: an understanding and planning module (e.g., a binary prediction task to determine whether to ask a clarifying question or provide an answer), and a generation module which can be controlled at a pragmatic level using the output of the planning module (Chen, Celikyilmaz, and Hakkani-Tur, 2017; Chen et al., 2022a; Qian et al., 2022; Yu, 2017) – e.g., forming an

utterance which follows the predicted action.

**Generation** Many existing works focus on novel training methodologies to improve conditional generation as a complement to planning, with approaches such as multi-objective SFT (Chen et al., 2022a; Wen et al., 2016) or introducing specialized embeddings for control codes (Keskar et al., 2019). LLMs have vastly improved performance in pragmatically-controlled generation (Chen et al., 2023a), but all of these approaches still depend on conversational planning. Planning remains a difficult task – natural interaction is not deterministic and often requires long-horizon planning.

**Planning** As introduced in Chapter 3, the planning task can be viewed as a stochastic Markov Decision Process (Wang et al., 2020a; Yang, Li, and Guo, 2021a; Yu, Chen, and Yu, 2023) in which some dialogue state is drawn from a potentially unknown distribution, given the previous dialogue state and an imposed action. However, the action itself is not literally presented to the interacting parties; rather, an action is a low-dimensional representation of the pragmatic intent carried by a given dialogue utterance (i.e., a dialogue act (Sadek, 1991; Stolcke et al., 2000; Wu et al., 2023a; Yu and Yu, 2021)). As shown in Chapter 3, developing planning modules that optimize strategies in a multi-turn conversation often requires complex long-horizon reasoning and simulation to model the responses and intents of each interacting party. Such efforts have examined combining neural models with search algorithms (Cheng et al., 2022a; V  th, Vanderlyn, and Vu, 2023b; Yu, Chen, and Yu, 2023) and simulation (Deng et al., 2023b; Wang et al., 2020c; Yu, Chen, and Yu, 2023). However, as we have seen with GDP-ZERO, modular approaches can incur high computational overhead and may result in error propagation while not directly optimizing for response quality itself. We propose directly optimizing dialogue action planning as an implicit subtask of response generation in mixed-initiative conversation contexts, as we discuss in Sec. 4.3.2.

#### 4.2.2 Learning for LLM Alignment

The current paradigm of LLM training for downstream use cases consists of three phases: pre-training, supervised fine-tuning (SFT) for instruction-following, and tuning for alignment with

human preferences (Tunstall et al., 2023; Rafailov et al., 2024; Lee et al., 2023; Ouyang et al., 2022). We primarily focus on the phase of tuning for alignment. These approaches typically start with an initial policy model obtained by conducting SFT on a target task ( $\pi_{SFT}$ ), before performing tuning (often with RL), using contrastive preference examples (most commonly collected through human feedback (Ouyang et al., 2022) or a similar proxy like LLM-generated feedback (Lee et al., 2023)). In the case of online algorithms like PPO, a reward model is first fit over the preference examples so that it could be used for RL optimization (Ouyang et al., 2022). Such algorithms have certain advantages which may benefit the Markov Decision Process-like nature of conversations — namely, a diverse search space as opposed to a fixed dataset, flexible reward functions, and broader policy exploration. However, PPO is notoriously difficult to tune, and offline algorithms such as DPO (Rafailov et al., 2024), SLiC (Zhao et al., 2023), and IPO (Azar et al., 2024) have become widely adopted as an LLM adaptation approach because they bypasses explicit reward modeling and thus only require one set of hyperparameters to optimize (Huang et al., 2024; Rafailov et al., 2024; Zhao et al., 2023; Zheng et al., 2023) while still achieving similar empirical results given a fixed preference dataset.

**On-Policy DPO** Many of our contemporaries also question the limits of fully offline preference learning algorithms and have examined “online” variants of them (Guo et al., 2024; Xu et al., 2023; Xu et al., 2024b). (Yuan et al., 2024) proposes iterative DPO, and Chen et al. (2024) proposes a variant where ground-truth responses are considered winning, and responses sampled from the previous iteration of a policy model are considered losing. (Pang et al., 2024) applies a variant of iterative DPO to optimize externalized reasoning chains. Our work differs from these in that we are proposing a novel approach to customize LLMs for specific conversational settings, in particular, multi-turn conversational settings. While other works look at applying DPO to conversations in general (e.g. Sun et al. (2024)), their focus is still on single-turn response optimization. *ACT* considers multi-turn trajectories for preference optimization, and to our knowledge, our work is the first paper to consider contrastive learning on the basis of conversational actions.



## 4.3 Methods

### 4.3.1 Problem Setup

We consider the task of tuning an LLM to function as a *mixed-initiative conversational agent*. Through a series of dialogue interactions with a user, the LLM is expected to assist the user by eventually providing a correct response to their request. Unlike the common agent interaction setting where users completely control the flow of interaction with the expectation that agents may autonomously complete tasks such as online shopping (Liu et al., 2024b), mixed-initiative agents should understand how to redirect the flow of the interaction (Allen, Guinn, and Horvitz, 1999) through the execution of conversational actions or strategies such as clarifying questions (Chu-Carroll, 2000; Peng et al., 2018b).

**Notation** Consider a goal-oriented conversational environment. Let  $\pi_{\theta_i}$  be an LLM’s policy parameterized by  $\theta$  at timestep  $i \geq 0$ , with  $\pi_{ref}$  being the reference policy model (i.e.,  $\pi_{ref} \leftarrow \pi_{\theta_0}$ ). Let  $D$  be a dataset consisting of conversations. Let each conversation  $c$  in  $D$  contain  $n$  dialogue turns, through which a user is requesting one or more pieces of information from an agent. The turn state of a conversation (the observed utterances and actions given by each interacting party) at timestep  $i$  can be represented by  $t_i$ . Implicitly, each  $t_i$  is part of a trajectory which ends when the user’s question expressed at an earlier timestep  $j \leq i$  is answered. Any turn  $t_i$  can first be broken down into two primary components:  $p_i$  and  $r_i$ , where  $p_i$  can be a prompt at  $i$ , consisting of any task-specific information (e.g. a SQL database schema, tabular data or retrieved passages) combined with any existing dialogue context, and  $r_i$  is the ground truth system-side response at  $i$ . Next, we can let  $g_i$  be the goal response which resolves  $t_i$ ’s implicit trajectory, i.e., the answer to the user’s original question after any possible clarification turns. In the single-turn trajectory case,  $g_i \leftarrow r_i$ . Each  $r_i$  implicitly expresses an action,  $a_i$ , where  $a_i$  exists in the latent Action Space  $S$  of a particular task and  $a_i$  can be inferred by some Action Annotation Agent  $G^2$  (e.g. “What kind

---

<sup>2</sup>The gold standard for full label supervision in the context of a fixed dataset is the scenario in which  $G$  may be a well-designed human annotation framework such as crowdsourcing. However, at inference time or in settings without label supervision, the implicit action must be inferred by other means such as classification.

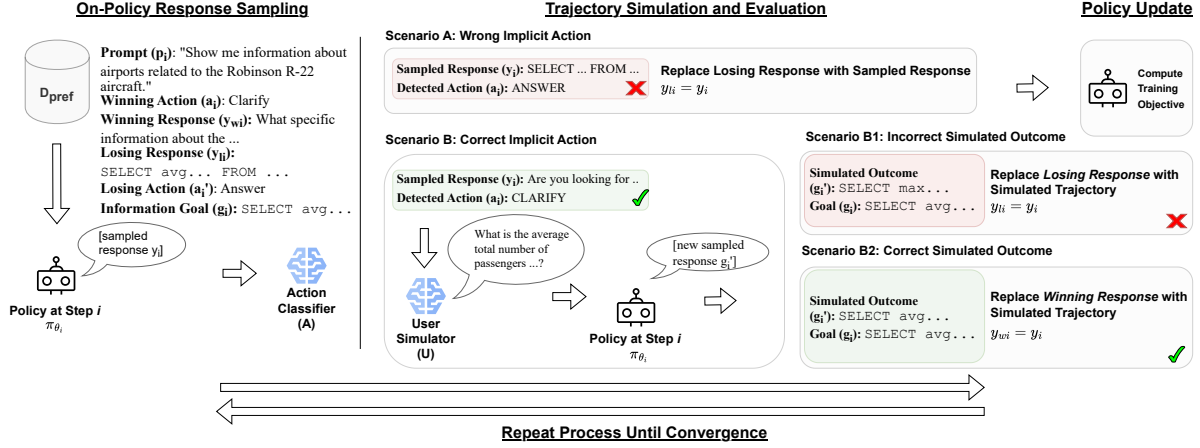


Figure 4.3: **Overview of the tuning phase of ACT.** For each initial contrastive pairing from  $D_{pref}$  (constructed as in Sec. 4.3.2), we sample an on-policy response from the model being tuned. After evaluating the sampled response’s trajectory, we update the contrastive pairing by either replacing the existing winning or losing response. The model policy is updated using the objective in Eq. 4.1.

of recorded investment are you asking about?” in Figure 4.1 may be some response expressing a clarifying question). Thus, we can formally represent turn state  $t_i$  using the tuple  $(p_i, r_i, g_i, a_i)$ . For the datasets considered in our experiments,  $S = [\text{CLARIFY}, \text{ANSWER}]$  (although the method can be extended to a broader action space). We assume access to a controllable generation model ( $M$ ), Action Classifier ( $A$ ) and model which can be controlled to function as a User Simulator ( $U$ ). As we discuss in Sec. 4.3.2,  $M$  is used for preference data creation whereas  $A$  and  $U$  are used during tuning and evaluation. We illustrate this notation in Fig. C1.

#### 4.3.2 ACT: Action-Based Contrastive Self-training

One of the north stars in developing intelligent conversational models is the ability to automatically produce responses which take actions which lead to the highest probability of conversational success Wu et al., 2023a; Zhao, Xie, and Eskenazi, 2019. We propose *ACT*, an approach that adapts generic LLMs for dialogue generation and models action planning as an implicit subtask. *ACT* is a quasi-online extension of the DPO algorithm which maintains its ease of use of offline method while incorporating the flexible exploration found during online learning. *ACT* relies on a few intuitions. 1) Contrastive preferences are an intuitive medium for demonstrating the *prag-*

---

**Algorithm 2** Building Contrastive Action Pairs

---

**input** Dataset  $D$ , Conditional generation model  $M$ , Action Space  $S$ , Action Annotation Agent  $G$

- 1: Initialize empty dataset  $D_{pref}$ .
- 2: **for** conversation turn  $t_i \in D$  **do**
- 3:   Let  $a_i = G(p_i, r_i)$  ▷ Infer Contextual Action
- 4:   Let  $a'_i = S \setminus a_i$  ▷ Determine Rejected Action
- 5:   Let  $y_{wi} = r_i$ .
- 6:   Sample  $y_{li} \sim P_M(\cdot | p_i, a'_i)$ .
- 7:   Let  $t'_i = (p_i, r_i, g_i, a_i, a'_i, y_{wi}, y_{li})$ .
- 8:   Add  $t'_i$  to  $D_{pref}$

**output**  $D_{pref}$

---

*matic differences* between the implicit actions of “winning” and “losing” dialogue responses. 2) Conversational improvements require multi-turn optimization, which are difficult to express using only single-turn contrast pairings. 3) The gradient of the objective of DPO-like algorithms (see Eq. 4.2) is weighted based on the log probabilities assigned to the winning and losing responses, and by construction, on-policy response sampling yields high-probability token sequences. *ACT* is summarized in Fig. 4.3. *ACT* consists of two phases: action-based contrast dataset construction (Alg. 2) and contrastive self-training (Alg. 3).

### Construction of Preference Data

The preference dataset primarily consists of contrastive *winning-losing* action pairs, as shown in Alg. 2. That is, for each conversation turn  $t_i$  in a dataset  $D$ , we can construct  $D_{pref}$  consisting of augmented  $t'_i$  tuples. We add rejected action  $a'_i$  which is sampled from  $S \setminus a_i$ , winning response  $y_{wi} \leftarrow r_i$ , and  $y_{li}$  which is a losing response sampled using  $M$ . Given that each  $a'_i$  is pre-defined when constructing  $D_{pref}$ , we use a high capacity LLM Chen et al., 2023a rather than tuning a smaller one or asking crowdworkers for losing response construction (more details in Appendix 4.5.2).

**Action optimization for unlabeled conversations “in-the-wild”** Obtaining gold-standard ambiguity annotations may not always be possible. In such settings, one can obtain pseudo-label supervision using a classifier as the Action Annotation Agent  $G$  rather than human annotation. We

discuss details and analyze performance in Sec. 4.6.4. Depending on the data, it may be appropriate to introduce an initial preprocessing step which involves inferring user satisfaction similarly to Shi et al., 2024.

### Self-Training Using On-policy Conversation Trajectory Simulation

As in DPO training, we continuously sample batches from  $D_{pref}$ . Although each conversation turn  $t_i$  in each batch  $j$  has a default winning ( $y_{wi}$ ) and losing ( $y_{li}$ ) response, we also sample an on-policy response  $y_i$  from  $\pi_{\theta_j}$ . We use  $A$  to determine whether the implicit action of  $y_i$ <sup>3</sup> matches the inferred action  $a_i$  of the ground truth response. If the implicit action of  $y_i$  is incorrect, we set  $y_{li} = y_i$ . If it does match  $a_i$ , then we simulate the outcome  $g'_i$  of the trajectory resulting from  $y_i$  using  $U$ <sup>4</sup> and  $\pi_{\theta_j}$ . If the trajectory outcome  $g'_i$  fails to meet task-specific heuristics (e.g., low semantic similarity or an incorrect execution), we set  $y_{li}$  to the entire simulated trajectory resulting from  $y_i$  (e.g., “Are you looking for...” + “What is the average total number...” + “SELECT max ...” in Figure 4.3). Otherwise, we set  $y_{wi}$  to the simulated trajectory (e.g. “Are you looking for...” + “What is the average total number...” + “SELECT avg ...” in Figure 4.3).

### Contrastive RL Tuning for Alignment

After constructing the up-to-date winning  $y_{wi}$  and losing  $y_{li}$  pairing at turn  $i$  through simulation (Sec. 4.3.2), we update the policy model ( $\pi_{\theta}$ ) using the DPO training objective Rafailov et al., 2024, which is as follows (we ignore the  $i$  iterator for simplicity):

$$\mathcal{L}_{DPO}(\pi_{\theta}; \pi_{ref}) = -\mathbb{E}_{(p, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_w | p)}{\pi_{ref}(y_w | p)} - \beta \log \frac{\pi_{\theta}(y_l | p)}{\pi_{ref}(y_l | p)} \right) \right], \quad (4.1)$$

where  $p$  is a prompt consisting of a concatenation between task info and conversation history  $= \{x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_i\}$  with each  $x_i$  and  $y_i$  representing observed user-side and system-side utterances at turn  $i$ ;  $y_w$  and  $y_l$  are the designated “winning” and “losing” responses or trajectories as set in Sec. 4.3.2;  $\pi_{ref}$  is the initial reference policy model; and  $\beta$  is a hyperparameter that

<sup>3</sup>Classifying  $y_i$ ’s action optimizes the following:  $\operatorname{argmax}_{a_k \in S} P_A(a_k | p_i, y_i)$

<sup>4</sup>The next user turn (denoted  $u_{i+1}$ ) is sampled according to  $u_{i+1} \sim P_U(\cdot | p_i, y_i)$

---

**Algorithm 3** ACT: Action-Based Contrastive Self-Training

---

**input** Initial Policy Model  $\pi_{\theta_0}$ , Action Contrast Dataset  $D_{pref}$ , Number of Batches  $B$ , Action Classifier  $A$ , User Simulator  $U$ , Task Heuristic  $H$ , Heuristic Tolerance  $\epsilon$

- 1: **for** conversation turn  $t_i$  in batch  $b_j$  sampled from  $D_{pref}$  where  $0 \leq j \leq B$  **do**
- 2:   Sample  $y_i \sim P_{\theta_j}(\cdot | p_i)$  ▷ Sample a response from the current model policy
- 3:   **if** Action  $A(y_i) \neq$  Action  $a_i$  **then**
- 4:     Set  $y_{li} = y_i$  ▷ Implicit pragmatic action does not match ground truth
- 5:   **else**
- 6:     Initialize *Trajectory*
- 7:     Add  $y_i$  to *Trajectory*
- 8:     **while**  $A(y_i) \neq ANSWER$  **do**
- 9:       Clarification Answer =  $P_U(p; y_i)$  ▷ Simulate User Clarification
- 10:       Add Clarification Answer to *Trajectory*
- 11:        $y'_{i+1} = P_{\pi_{\theta}}(P; y_i)$  ▷ Simulate next policy response
- 12:       Add  $y'_{i+1}$  to *Trajectory*
- 13:       **if**  $H(\text{Trajectory outcome, Ground Truth Outcome } g_i) > \epsilon$  **then**
- 14:          Let  $y_{wi} = \text{Trajectory}$  ▷ Reward acceptable trajectory outcome
- 15:       **else**
- 16:          Let  $y_{li} = \text{Trajectory}$  ▷ Penalize bad trajectory outcome
- 17:    $\theta \leftarrow \text{Update}(\theta)$  until convergence (eq 4.2)

**output**  $\pi_{\theta_B}$

---

regularizes the ratio between  $\pi_{\theta}$  and  $\pi_{ref}$ . The gradient of this objective is given as follows:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{ref}) = \\ - \beta \mathbb{E}_{(p, y_w, y_l) \sim \mathcal{D}} \left[ \sigma(\hat{R}_{\theta}(p, y_l) - \hat{R}_{\theta}(p, y_w)) \left[ \nabla_{\theta} \log \pi(y_w | p) - \nabla_{\theta} \log \pi(y_l | p) \right] \right], \quad (4.2) \end{aligned}$$

where  $R(p, y) = \beta \log \frac{\pi(y|p)}{\pi_{ref}(y|p)}$  for a given policy model  $\pi$  and reference model  $\pi_{ref}$ , as proven under the assumptions in Rafailov et al., 2024. The intuition behind the objective is that the gradient of the loss function would increase the likelihood of winning responses  $y_w \in Y_w$  and would decrease the likelihood of losing responses  $y_l \in Y_l$ , with each example being weighed by the magnitude of how incorrectly the implicitly defined reward model ranks the paired responses.

#### 4.4 AmbigSQL: Modeling Ambiguity in Conversational Text-to-SQL

Table 4.1: **Overview of AmbigSQL**, an ambiguous Text-to-SQL dataset synthesized from Spider.

	Train	Dev	Test
Num. Unambiguous Requests	7,000	1,034	1,034
Num. Ambiguous Requests	7,000	1,034	1,034
Num. Unique Schemas	1,056	145	145
Types of Ambiguity	3	3	3

There is growing interest in using LLM-based agents for coding tasks Liu et al., 2024b. Particularly, due to the complexity of such tasks, multi-turn interactions in which an agent is able to clarify assumptions and user intents should intuitively help with goal completion Nijkamp et al., 2023. Despite this, there are few existing resources for multi-turn code generation tasks. One example is CoSQL, a conversational text-to-SQL task which also includes linguistic ambiguities Yu et al., 2019, but the proposed task does not include agent-style interaction in which a model must learn to ask clarifying questions. Upon our inspection of the dataset, there are also various inconsistencies related to the “system-side” clarification questions given in the dataset’s conversational contexts, which we highlight in Table 4.2. As a result, we propose AmbigSQL, our own synthetically constructed resource for ambiguous conversational text-to-SQL.

Table 4.2: **Conversations in CoSQL with noisy “clarification questions”** (highlighted in red). Example 1) is a remnant of crowdsourcing in which the system-side party makes mention of the task guideline. Example 2) demonstrates a system-side clarification question being asked prior to the user making any information requests. Example 3) The system-side clarification question makes reference to some prior database search result, but the execution feedback is not made accessible to the system during inference.

No.	Interacting Party	Utterance
	User	Can you list all the singer ids that aren’t present in the song table?
	Assistant	SELECT Name FROM singer WHERE Singer_ID NOT IN ...
	User	Thanks!
1	Assistant	You should ask at least 3 questions
2	Assistant	Did you want the full name of makers and the number?
3	Assistant	Do you mean the address of the customer with first name Luis?

#### 4.4.1 AmbigSQL Construction

Table 4.3: **In-context example given as part of a prompt for creating information requests in which the target population is ambiguous.** The format of the black text represents how a ground-truth request would be used to form the prompt for a target example. The blue text represents the content that would be synthesized from an LLM. We omit the database schema from the paper.

---

[Database Schema Omitted]

The target SQL query is the following:

```
SELECT professional_id , last_name , cell_number FROM Professionals
WHERE state = 'Indiana' UNION SELECT T1.professional_id , T1.last_name ,
T1.cell_number FROM Professionals AS T1 JOIN Treatments AS T2 ON
T1.professional_id = T2.professional_id
GROUP BY T1.professional_id HAVING count(*) > 2
```

Here is a clear request that would correspond to this SQL query:

“Which professionals live in the state of Indiana or have done treatment on more than 2 treatments?  
List his or her id, last name and cell phone.”

Here is the same request converted into an ambiguous format by underspecifying the target columns:

“Which ones live in the state of Indiana or have done treatment on more than 2 treatments?”

Here is an appropriate clarifying question to recover the clear request from the ambiguous request:

“Are you asking about the Professionals?”

---

We start from Spider, a popular single-turn text-to-SQL dataset and benchmark. An overview of AmbigSQL is given in Table 4.1. Due to the nature of the single-turn requests in Spider, each instance can be viewed as a conversation consisting of a single state  $t_1$  (see notation defined in Section 4.3). In  $t_1$ ,  $p_1$  contains any instructions, the database schema, and the user’s request.  $r_1$  is the correct SQL query which yields the information requested by the user.  $g_1 = r_1$  because the trajectory ends on the same turn due to the system yielding the correct query.  $a_1 = ANSWER$  because the only action possible is to provide a SQL query.

Our desired result in constructing AmbigSQL is to have a corpus which can be used to demonstrate to an LLM the linguistic differences between ambiguous and clear user requests. That is, given a fixed database schema and a target SQL query, we want a pair of requests such that one requires asking a clarification question and one does not. This would also result in a balanced dataset in which half of the requests require asking clarification questions, and half do not.

We focus on three fundamental types of ambiguous information requests. Those in which

the requested information is ambiguous (e.g., “Show details about singers ordered by age from the oldest to the youngest”), those in which the requested population is ambiguous (e.g., “Which ones who live in the state of Indiana?”; see Table 4.3), and finally, those in which the requested presentation of results is ambiguous (e.g. “Show name, country, age for all singers ordered by age”; see Table 4.4).

We iterate through each of the examples in Spider, and use an LLM (Gemini Ultra 1.0) to synthesize a perturbed version of each unambiguous query, along with an appropriate clarifying question. For queries which require some manipulation of results presentation, we prompt an LLM to perturb the query such that the requested presentation style becomes ambiguous. Otherwise, we randomly select a perturbation strategy: either masking the requested information (Table 4.4) or masking the requested population (Table 4.3). For each of these three strategies, we use five in-context examples demonstrating the appropriate masking process. The exact in-context examples are given in the attached code. Each of these ambiguous queries are thus associated with conversations containing “ground truth” states at two timesteps,  $t_1$  and  $t_2$ .  $p_1$  contains an ambiguous user request,  $r_1$  is the synthesized clarification question, and accordingly,  $a_1 = CLARIFY$ .  $p_2$  contains the disambiguated user request,  $r_2$  is the correct ground truth SQL query, and accordingly,  $a_2 = ANSWER$ .

The code to create AmbigSQL will be released publicly. Each instance includes the database metadata included in Spider, but for all experiments used in this paper, the database schema is linearized into the format used for prompting in Sun et al., 2023b; Sun et al., 2023a.

#### 4.4.2 Examining Clarification Need in AmbigSQL

We primarily are concerned with examining the extent to which clarification questions are necessary in providing the requested SQL queries. Inspired by the notion of “recovery” presented in Toles et al., 2023, we examine the performance in constructing each of the unique SQL queries with and without the gold clarification turns. Concretely, we first evaluated the Execution Match performance achieved by prompting LLMs with only the ambiguous versions of each user request,



**Table 4.4: In-context example given as part of a prompt for creating information requests in which the target columns are ambiguous.** The format of the black text represents how a ground-truth request would be used to form the prompt for a target example. The blue text represents the content that would be synthesized from an LLM. We omit the database schema from the paper.

[Database Schema Omitted]

The target SQL query is the following:

```
SELECT professional_id , last_name , cell_number FROM Professionals
WHERE state = 'Indiana' UNION SELECT T1.professional_id , T1.last_name ,
T1.cell_number FROM Professionals AS T1 JOIN Treatments AS T2 ON
T1.professional_id = T2.professional_id
GROUP BY T1.professional_id HAVING count(*) > 2
```

Here is a clear request that would correspond to this SQL query:

“Which professionals live in the state of Indiana or have done treatment on more than 2 treatments?  
List his or her id, last name and cell phone.”

Here is the same request converted into an ambiguous format by underspecifying the target columns:

“Which professionals live in the state of Indiana or have done treatment on more than 2 treatments?”

Here is an appropriate clarifying question to recover the clear request from the ambiguous request:

“Which information of the professionals do you want to know?”

with each instruction including the instruction that the LLM must construct a SQL query. Then, we prompted LLMs to construct the same SQL queries, but given the disambiguation turns as conversation history (i.e., with context consisting of the original ambiguous request, the clarification question, and then non-ambiguous request).

We conducted this analysis on the test set using two competitive LLMs, Gemini Pro and Gemini Ultra, with the Execution Match tool from Spider and CoSQL Yu et al., 2018; Yu et al., 2019. Our results are shown in Table 4.5. Given only an ambiguous request, both Gemini Pro and Gemini Ultra struggle to consistently construct the correct SQL query. However, given disambiguation turns, Execution Match improves dramatically, approximating the performance on the validation set given in Sun et al., 2023b; Sun et al., 2023a.

#### 4.4.3 Examples

Table 4.6 and Table 4.7 each contain a pair of examples from AmbigSQL’s test set. Each example contains the prompt which is provided to an LLM, the immediate ground truth response

Table 4.5: **Examination of ambiguity in AmbigSQL.** Competitive high capacity LLMs struggle with producing “correct” SQL queries given only ambiguous user requests. Including disambiguation turns in the prompts greatly improves execution match.

Model	Ambiguous Request Execution Match	Execution Match with Clarification Turns
Gemini Pro	28.5	68.7
Gemini Ultra	31.2	77.0

to the user request provided as part of the prompt, and the resulting ground truth trajectory (for examples which include clarifying questions). The examples in each table are paired — the top example is directly taken from Spider and converted into a conversational format, and the bottom example is the result of introducing ambiguity into the first example.

Table 4.6: **Examples from the AmbigSQL test set using the “concert\_singer” database.** Black text: context provided to an LLM as a prompt. Blue text: ground-truth single-turn response. Purple text: target trajectory which should be achieved through simulation.

<p>[Instruction]</p> <p>You are an Assistant having a conversation with a User who would like to retrieve some information from a SQL database with the schema provided below. If you are confident in the User’s intent, you should provide the appropriate SQL query. Otherwise, you should ask the User a clarifying question.</p> <p>[Database Schema Omitted]</p> <p>[Conversation]</p> <p>User: How many singers do we have?</p> <p>Assistant: <code>SELECT count(*) FROM singer</code></p>
<p>[Instruction]</p> <p>You are an Assistant having a conversation with a User who would like to retrieve some information from a SQL database with the schema provided below. If you are confident in the User’s intent, you should provide the appropriate SQL query. Otherwise, you should ask the User a clarifying question.</p> <p>[Database Schema Omitted]</p> <p>[Conversation]</p> <p>User: Tell me about the singers.</p> <p>Assistant: What specifically would you like to know about the singers? For example, would you like to know how many there are, or see a list of their names?</p> <p>User: How many singers do we have?</p> <p>Assistant: <code>SELECT count(*) FROM singer</code></p>

Table 4.7: **Examples from the AmbigSQL test set using the “flight\_2” database.** Black text: context provided to an LLM as a prompt. Blue text: ground-truth single-turn response. Purple text: target trajectory which should be achieved through simulation.

---

[Instruction]

You are an Assistant having a conversation with a User who would like to retrieve some information from a SQL database with the schema provided below. If you are confident in the User’s intent, you should provide the appropriate SQL query. Otherwise, you should ask the User a clarifying question.

[Database Schema Omitted]

[Conversation]

User: Return the number of airports.

Assistant: `SELECT count(*) FROM AIRPORTS`

---

[Instruction]

You are an Assistant having a conversation with a User who would like to retrieve some information from a SQL database with the schema provided below. If you are confident in the User’s intent, you should provide the appropriate SQL query. Otherwise, you should ask the User a clarifying question.

[Database Schema Omitted]

[Conversation]

User: How many are there?

Assistant: Could you please specify which table you are referring to when you ask ‘How many are there?’

User: Return the number of airports.

Assistant: `SELECT count(*) FROM AIRPORTS`

---

## 4.5 Experimental Setup

*ACT* is a sample-efficient approach to adapt an LLM to a conversational action policy. We are primarily concerned with learning optimized implicit selection for agent-side clarification question asking, and we thus evaluate *ACT* as a tuning approach for three complex conversational information-seeking tasks. As a base model for our tuning experiments, we use Zephyr  $\beta$ , a version of Mistral 7B (Jiang et al., 2023) which has been instruction tuned on UltraChat and aligned to human preferences on UltraFeedback (Cui et al., 2023; Ding et al., 2023; Tunstall et al., 2023).

### 4.5.1 Datasets

We investigate three mixed-initiative conversation tasks in which a user interacts with an assistant to retrieve some information. In our setup of each task, a user asks a query which may or may not be underspecified. The assistant is tasked with providing a response which may either be a

clarifying question or an attempt to directly answer the user’s query. For each task, we synthesize the initial rejected responses by prompting Gemini Ultra as the conditional generation model,  $M$ . *ACT* is evaluated on a diverse set of datasets spanning various domains: tabular conversational QA, conversational QA for machine reading comprehension, and conversational text-to-SQL generation.

### **PACIFIC: Conversational QA for Tabular Data**

**PACIFIC** is a task for proactive conversational question answering grounded on a mixture of tabular and textual financial data (Deng et al., 2022). This may involve generating the correct words from a given span, from multiple spans, or providing a correct arithmetic expression. The official evaluation for PACIFIC uses a numeracy-focused token overlap metric called DROP F1.

### **Abg-CoQA: Conversational QA for Machine Reading Comprehension**

**Abg-CoQA** is a conversational question answering dataset for disambiguation in machine reading comprehension (Guo et al., 2021). As there are no arithmetic expressions, we use embedding-based semantic distance with SentenceBERT (Reimers and Gurevych, 2019) as an evaluation metric, which has been used to more flexibly measure question-answering performance (Risch et al., 2021a).

### **AmbigSQL: Ambiguous Conversational Text-to-SQL Generation**

**AmbigSQL** is our new task for SQL-grounded conversational disambiguation, as described in Section 4.4.

#### **4.5.2 Conditional Generation Model Details**

As mentioned Section 4.3.1, we make use of a high capacity LLM as a conditional generation model,  $M$ . For all experiments considered, we use Gemini Ultra 1.0 to construct the initial action-based contrastive preference dataset. We follow the format of “mixed-initiative prompting”

described in Chen et al., 2023a, rather than performing zero-shot inference, we use ten in-context examples to adapt the LLM to the prompting structure. Concretely, for a given input/output pair to be used as an in-context example, we interweave each System-side utterance with a narrative instruction that states that the next utterance is either a clarifying question or a direct answer. This yields “control” over the pragmatic action of the generated utterance.

### 4.5.3 Action Classifier Details

In the datasets we considered, the possible actions are to either “clarify” or “directly answer” a question. We directly use few-shot in-context learning as action classifier *A*.

Compared to recognizing whether a request is ambiguous or not, action classification (i.e., recognizing whether an existing utterance is a question or answer attempt) is considerably simpler. We directly prompt Gemini Ultra 1.0 with 10 in-context examples to serve as *A* (as per the notation given in Section 4.3.1).

**Table 4.8: In-context example given as part of a prompt to an LLM for action classification with PACIFIC.** The blue text represents the action for this example.

---

[Table and Passage Omitted]

User: What was the total number of acquisitions in FY19?

Assistant: 35 acquisitions

User: What are the benefits of active growth strategy?

Assistant: strengthens and scales the National Storage operating platform which drives efficiencies across the business.

User: What was the total NLA?

Assistant: Which region are you asking about?

The last Assistant utterance is a clarifying question.

---

Table 4.8 is an example of an in-context example used to demonstrate a system-side response with the “Clarify” action in PACIFIC (“Assistant: Which region are you asking about?”). All of the conversation history is made available in addition to the relevant table and passage in the prompt. We include all contextual information available as part of the input to the Action Classifier, along with the last Assistant response which is an example of the classification target.

Table 4.9 is an example of an in-context example used to demonstrate a system-side response

Table 4.9: **In-context example given as part of a prompt to an LLM for action classification with Abg-CoQA.** The blue text represents the action for this example.

---

[Passage Omitted]  
 User: And who in the group was anxious?  
 Assistant: Peppe  
 User: Was she well-rested?  
 Assistant: no  
 The last Assistant utterance is a direct answer.

---

with the “Answer” action in Abg-CoQA (“Assistant: no”). As is common in this corpus, a subset of the conversation is made available in addition to the relevant passage in the prompt. We include all contextual information available as part of the input to the Action Classifier, along with the last Assistant response which is an example of the classification target.

Table 4.10: **In-context example given as part of a prompt to an LLM for action classification with AmbigSQL.** The blue text represents the action for this example.

---

[Database Schema Omitted]  
 User: “Which professionals live in the state of Indiana or have done treatment on more than 2 treatments? List his or her id, last name and cell phone.”  
 Here is the same request converted into an ambiguous format by underspecifying the target columns:  
 Assistant:

```
SELECT professional_id , last_name , cell_number FROM Professionals
WHERE state = 'Indiana' UNION SELECT T1.professional_id, T1.last_name,
T1.cell_number FROM Professionals AS T1 JOIN Treatments AS T2 ON
T1.professional_id = T2.professional_id
GROUP BY T1.professional_id HAVING count(*) > 2
```

The last Assistant utterance is a direct answer.

---

Table 4.10 is an example of an in-context example used to demonstrate a system-side response with the “Answer” action in AmbigSQL. In AmbigSQL, the information provided as input is the database schema, and the user’s initial information request. This information is provided as part of the input to the Action Classifier, along with the last Assistant response which is an example of the classification target.

Table 4.11: **In-context example given as part of a prompt to an LLM for intent summarization with PACIFIC.** The **purple** text represents the system prompt used for this task. The **blue** text represents the summarized information-seeking intents.

---

The following is a conversation between a User and an Assistant. The User is asking some questions. Summarize what information the User is looking for.

[Table and Passage Omitted]

User: What does Walletron deliver?

Assistant: patented mobile wallet technology.

User: How much did it cost the company to acquire certain technology assets of RevChip, LLC ("RevChip") and TranSend Integrated Technologies Inc. ("TranSend")?

Assistant: \$7.0 million

User: What was the pro forma revenue in 2019?

Assistant: \$1,382,957

User: What was the change in its amount between 2018 and 2019?

Assistant: 21228

User: How about that in pro forma net income?

Assistant: -6425

[Information] The user wants to know: 1. What technology Walletron delivers, 2. How much it cost Walletron to acquire certain technology assets of RevChip, LLC and TranSend Integrated Technologies Inc., 3. What the pro forma revenue was in 2019, 4. What the change in pro forma revenue was between 2018 and 2019, and 5. What the change in pro forma net income was between 2018 and 2019.

---

#### 4.5.4 User Simulator Details

In order to iterate on system-side dialogue models, many works investigate the use of a user simulator,  $U$  (as per the notation given in Section 4.3.1) (e.g. Chandramohan et al., 2011; Kreyssig et al., 2018; Shah et al., 2018). At its core, it requires a response generation model which can robustly represent user-side intents. Traditionally, this involved training a model specifically on user-side responses but due to the aforementioned control capabilities of modern LLMs (Chen et al., 2023a), we directly perform few-shot prompting for user simulation, following Deng et al., 2023b; Yu, Chen, and Yu, 2023. For a given conversation during training, we again prompt a high-capacity LLM with in-context examples to identify the information-seeking user intents. Then, conditioned on the generated information-seeking intents, we prompt the same LLM with the responses from the policy model to simulate the sub-goal trajectory. This allows  $U$  to ground on the real intents of the user while retaining enough flexibility to respond coherently to some system-side response. In our work, we use Gemini Ultra 1.0 with three hand-crafted in-context

examples for both intent summarization and user response generation.

**PACIFIC** We provide one of our three handcrafted in-context examples of summarizing information-seeking intent in Table 4.11. While the information summary is handcrafted, the conversation and grounding data are taken directly from the PACIFIC. After intent summarization,  $U$  is given the existing conversation history along with an instruction that includes the information seeking intents and asked to provide the next user response. This format is structurally the same as the three simulation-specific in-context examples — an example is provided in Table 4.12.

Table 4.12: **In-context example given as part of a prompt to an LLM for user simulation with PACIFIC.** The **purple** text represents the system prompt used for this task, which contains the information-seeking intents summarized from Table 4.11. The entire conversation is used for this example to demonstrate grounding between a conversation and intents.

---

The following is a conversation between a User and an Assistant. The User is asking some questions. The user wants to know: 1. What technology Walletron delivers, 2. How much it cost Walletron to acquire certain technology assets of RevChip, LLC and TranSend Integrated Technologies Inc., 3. What the pro forma revenue was in 2019, 4. What the change in pro forma revenue was between 2018 and 2019, and 5. What the change in pro forma net income was between 2018 and 2019.

[Table and Passage Omitted]

Assistant: \$7.0 million

User: What was the pro forma revenue in 2019?

Assistant: \$1,382,957

User: What was the change in its amount between 2018 and 2019?

Assistant: 21228

User: How about that in pro forma net income?

Assistant: -6425

---

**Abg-CoQA** We conduct user simulation for Abg-CoQA similarly to PACIFIC. Table 4.13 is one of three hand-crafted examples of intent summarization, and Table 4.14 is one of three hand-crafted examples used for simulating the final user response.

**AmbigSQL** We conduct user simulation slightly differently for AmbigSQL. Rather than having a separate step for knowledge-seeking intent summarization, we directly condition  $U$  on the target SQL query. As demonstrated in Table 4.15, the prompt includes the target SQL query, the original user request, and the clarification question sampled from the Assistant. In the example given in



Table 4.13: **In-context example given as part of a prompt to an LLM for intent summarization with Abg-CoQA.** The purple text represents the system prompt used for this task. The blue text represents the summarized information-seeking intents.

---

[Passage Omitted]  
 User: What was his ranking?  
 Assistant: General  
 User: Did someone else have horse fighters?  
 Assistant: yes  
 User: Who?  
 Assistant: Do you want to know who had 2,500 horse fighters initially?  
 User: No, I want to know who had a considerable force of horse fighters west of him.  
 Assistant: General (Bishop) Polk.  
**Instruction:** Summarize what information the user is looking for.  
**Summary:** The user wants to know: 1. What his ranking was. 2. Whether someone else had horse fighters. 3. Who had a considerable force of horse fighters west of him.

---

Table 4.15, the last line is the ground truth disambiguated query, and in practice, this would be the generation target of  $U$ .

#### 4.5.5 Evaluation Setup

We conduct evaluations of *ACT*’s ability to reason about ambiguity in conversation to better accomplish conversational goals along two dimensions.

**Agent task performance:** We evaluate whether *ACT* improves multi-turn task completion capabilities. PACIFIC and Abg-CoQA are initially proposed only with static single-turn evaluations. We mirror this by conducting a turn-level evaluation where we compare the model’s response to the ground truth utterance given in response to the user’s query, using the task-specific heuristics given in Sec. 4.5.1. Since we are specifically concerned with improving LLMs’ multi-turn capabilities, we additionally propose a multi-turn evaluation scheme for the trajectory outcomes in all three tasks considered. While the sampled response from an LLM is a clarifying question, we simulate a user response and re-sample another response from the evaluated LLM until it attempts to answer the original query. We evaluate this outcome against the user’s ground truth information-seeking goal. We use  $A$  and  $U$  for simulation as described in Sec. 4.3.2 for *ACT*, and use the

Table 4.14: **In-context example given as part of a prompt to an LLM for user simulation with Abg-CoQA.** The **purple** text represents the system prompt used for this task, which contains the information-seeking intents summarized from Table 4.13. The entire conversation is used for this example to demonstrate grounding between a conversation and intents.

The following is a snippet from a conversation between an Assistant and a User who is asking some questions about the Story below. The user only wants to know the following information: 1. What his ranking was. 2. Whether someone else had horse fighters. 3. Who had a considerable force of horse fighters west of him.

[Passage Omitted]

User: What was his ranking?

Assistant: General

User: Did someone else have horse fighters?

Assistant: yes

User: Who?

Assistant: Do you want to know who had 2,500 horse fighters initially?

User: No, I want to know who had a considerable force of horse fighters west of him.

Assistant: General (Bishop) Polk.

heuristics defined in Sec. 4.5.1. An example is illustrated in Fig C2. In PACIFIC and AmbigSQL, we also compute task performance on the simulated responses in which the model has previously asked any clarifying questions, in order to get a more fine-grained measure of the model’s ability to reason about its own clarification questions. In aggregate, we define the following content-level evaluation metrics:

- Turn-level DROP F1: Average immediate response DROP F1 (Deng et al., 2022; Dua et al., 2019)
- Trajectory-level DROP F1: Average trajectory-outcome DROP F1 (Deng et al., 2022; Dua et al., 2019)
- Post-Clarification DROP F1: Average DROP F1 (Deng et al., 2022; Dua et al., 2019) of responses which follow agent clarification turns
- Turn-level Similarity: Immediate response embedding similarity
- Trajectory-level Similarity: Trajectory outcome embedding similarity
- Trajectory-level Execution Match: Percentage of trajectory outcomes with correct execution results

Table 4.15: **In-context example given as part of a prompt to an LLM for user simulation with AmbigSQL.** The purple text represents the system prompt used for this task, which contains ground truth SQL query goal. The entire conversation is used for this example to demonstrate SQL query-grounded request disambiguation.

---

A user is asking an assistant to retrieve some information from a SQL database. The command that the assistant should ultimately return is as follows:

SELECT county FROM campuses where campus = 'California State University-Chico'

The assistant will ask some questions to clarify the user’s intent. The user should respond with a rephrased request that reflects their desired query.

User: what is the county?

Assistant: Are you asking for a list of all of the counties in the database?

User: I’m looking for the county of the campus 'California State University-Chico'

---

- Post-Clarification Execution Match: Percentage of trajectory outcomes with correct execution results out of those that which contain clarification turns

**Implicit ambiguity recognition:** To help further understand an agent’s multi-turn task completion ability, we consider “dialogue act accuracy” (Chen et al., 2023a). Assuming access to ground-truth ambiguity labels, given a contextually-ambiguous user request, a model should generate a clarifying question, otherwise, it should attempt to provide the requested information. Concretely, we define the following action-level evaluation metrics:

- Accuracy: Percentage of correct implicit actions
- Weighted F1: Average of F1 for each action, weighted by its support
- Macro F1: Unweighted Average of F1 for each action

**PACIFIC** As previously described, PACIFIC is a conversational question-answering dataset in which the final answers may involve generating the correct words from a given span, from multiple spans, or providing a correct arithmetic expression. As such, the authors propose using DROP F1 as the official evaluation metric. The way DROP F1 is used in the original paper Deng et al., 2022 is analogous to our aforementioned “Turn-level DROP F1.” However, as this does not fully represent a model’s conversational reasoning abilities, we additionally evaluate LLMs in the

PACIFIC environment using Accuracy, Weighted F1, Macro F1, Trajectory-level DROP F1, and Post-Clarification DROP F1. Concretely, the evaluation for some LLM  $\pi$  is as follows. Assume we have some example with prompt  $p$ , winning action  $a$ , ground truth response  $r$ , and trajectory-level information goal  $g$ . We sample a candidate response from the LLM:  $y \sim P_{\theta}(\cdot|p)$ . We then simulate the trajectory resulting from each response  $y$  according to Lines 6-12 in Algorithm 3 and obtain trajectory outcome  $g'$ . The aforementioned action-level metrics are computed using the implicit actions of each  $y$  with each ground truth implicit action  $a$ . Turn-level DROP F1 is computed between all sampled responses  $y$  and all ground truth responses  $r$ , and Trajectory-level DROP F1 is computed over all simulated trajectory outcomes  $g'$  and all ground truth information goals  $g$ . Post-Clarification F1 is defined as Trajectory-level F1 for only the subset of trajectories which include clarification turns.

**Abg-CoQA** As previously mentioned in Section 5.6, Abg-CoQA is a conversational question-answering dataset for machine reading comprehension. Thus, we use embedding similarity Risch et al., 2021a as it allows for producing more coherent and diverse responses which may be scored lowly by criteria such as token-level F1 score. In the original paper, language models are only evaluated in terms of QA performance Guo et al., 2021, rather than their ability to disambiguate requests. Thus, for our evaluation, we remove all clarification turns from the prompt and require the LLM to produce clarifying questions on its own. However, unlike the other tasks considered in this paper, each ambiguous request is paired with all of the possible trajectories (i.e., the reasons why the request is considered ambiguous). We thus perform an evaluation for every ground truth trajectory, so that it is impossible for an LLM to achieve a high cumulative trajectory-level score simply by getting lucky at guessing the user’s intent.

Concretely, the evaluation for some LLM  $\pi$  is as follows. Assume we have some example with prompt  $p$ , winning action  $a$ , ground truth response  $r$ , and the set of trajectory-level information goals  $G$ . For every individual trajectory-level goal  $g \in G$ , we sample a candidate response from the LLM:  $y \sim P_{\theta}(\cdot|p)$  then simulate the trajectory resulting from each response  $y$  according to Lines

6-12 in Algorithm 3 and obtain trajectory outcome  $g'$ . As with PACIFIC, we compute Accuracy, Weighted F1, and Macro F1 using the implicit actions of each  $y$  with each ground truth implicit action  $a$ . We compute Turn-level similarity for *each unique*  $p$  between sampled responses  $y$  and ground truth responses  $r$ . We compute Trajectory-level similarity over all simulated trajectory outcomes  $g'$  and all ground truth information goals  $g$ .

**AmbigSQL** is a ambiguous text-to-SQL generation task. By design, the labels are balanced, so we only use Accuracy for action-level metrics. The primary objective is to produce SQL queries which result in correct executions, so we evaluate LLMs in terms of Trajectory-level and Post-Clarification Execution Match. For each instance in the test set, we sample a response from an LLM and compute the trajectory as specified in Lines 6-12 in Algorithm 3. Trajectory-level Execution Match includes SQL queries which are constructed with or without clarification turns, whereas Post-Clarification Execution Match only looks at SQL queries which are constructed following clarification turns.

#### 4.5.6 Baselines

**Prompting baselines** We compare our tuning approaches with smaller models against various prompt-based approaches for multiple frontier LLMs: Gemini 1.5 Pro, Gemini 1.5 Flash, Claude 3.5 Sonnet, and Claude 3.0 Haiku<sup>5</sup>. We use 10 conversations as in-context examples, with three different prompting frameworks: i.) “Standard” which uses the same instruction formatting used for tuning; ii.) chain-of-thought reasoning (Wei et al., 2022); and iii.) “Proactive MIPrompt”, the prompting baseline in Deng et al. (2023b), which is a combination of the mixed-initiative prompting approach used in Chen et al. (2023a) and Proactive Prompting (Deng et al., 2023c).

We use several in-context learning baselines with frontier LLMs in Section 5.6. For each condition, we randomly sample 10 conversation examples from each task’s 250-instance data pool, and apply one of the following prompting frameworks. We describe each prompting baseline in detail below.

---

<sup>5</sup>We access each LLM through Vertex AI: <https://cloud.google.com/vertex-ai/docs/>

Table 4.16: “Chain-of-Thought” in-context example given as part of a prompt to an LLM for response generation in PACIFIC. The purple text is the overall system instruction. The blue text is what an LLM would have to produce at inference time.

---

You are an Assistant answering questions from a User. You should either attempt to answer the question or ask a clarifying question if there is any ambiguity.

[Table and Passage Omitted]

User: What is the value of the expected defined contribution plan as a percentage of the defined benefit plan during the next fiscal year?

Instruction: If the user’s question is ambiguous, ask an appropriate clarifying question. Otherwise, directly answer the user’s question using the information from the passage context and the table. Let’s think step by step.

Reasoning: The user’s question is not ambiguous. Assistant:  $0.6/5.1 = 11.76$

User: How much would change if there is a 1% increase and a 1% decrease respectively in the discount rate?

Instruction: If the user’s question is ambiguous, ask an appropriate clarifying question. Otherwise, directly answer the user’s question using the information from the passage context and the table. Let’s think step by step.

Reasoning: The user’s question was ambiguous. Assistant: [‘What kind of change are you asking about?’]

User: The pension.

Instruction: If the user’s question is ambiguous, ask an appropriate clarifying question. Otherwise, directly answer the user’s question using the information from the passage context and the table. Let’s think step by step.

Reasoning: The user’s question is not ambiguous. Assistant: [‘\$(39,145)’, ‘\$49,361’]

User: What is the difference in that amount if the discount rate had a 1% increase and a 1% decrease?

Instruction: If the user’s question is ambiguous, ask an appropriate clarifying question. Otherwise, directly answer the user’s question using the information from the passage context and the table. Let’s think step by step.

Reasoning: The user’s question is not ambiguous. Assistant:  $49,361 - (39,145) = 88506$  User: How much would the pension change if there is a 1% increase and a 1% decrease respectively?

Instruction: If the user’s question is ambiguous, ask an appropriate clarifying question. Otherwise, directly answer the user’s question using the information from the passage context and the table. Let’s think step by step.

Reasoning: The user’s question was ambiguous.

Assistant: [‘Which rate of increase and decrease are you talking about?’]

---

*Standard Prompting:* We simply provide in-context examples that are structurally identical to the inputs used for model tuning. Our format is similar to the formats used in Chen et al. (2023a) and Deng et al. (2023c) and we provide an example in Table C1.

*Chain-of-Thought Prompting:* We integrate the popular reasoning framework, chain-of-thought prompting (Wei et al., 2022) into our aforementioned “Standard” conversational prompt format. Effectively, we ask an LLM to do end-to-end dialogue generation by having it first produce a reasoning chain which states whether the current context is ambiguous or not. We provide an example in Table 4.16.

*Proactive Mixed-Initiative Prompting:* Following the baselines given in (Deng et al., 2023b), we apply the Proactive Prompting framework (Deng et al., 2023c) mixed with the Mixed-Initiative Prompting style (Chen et al., 2023a). Ultimately, the LLM conditions on the possible set of actions, along with interweaved natural language instructions that describe which actions correspond to existing dialogue turns. We provide an example in Table 4.17. We use this framework for each of the long-context dialogue corpora (PACIFIC and Abg-CoQA).

**Tuning baselines** We compare *ACT* with supervised fine-tuning (SFT) as well as other off-policy and on-policy approaches to DPO-based alignment. For SFT, we use the ground truth responses for each dataset’s training split. As for DPO-based alignment, an on-policy variant called Iterative Reasoning Preference Optimization (IRPO) was recently proposed and has gained traction for improving model performance in reasoning tasks such as arithmetic. We have thus evaluated IRPO on our two quantitative reasoning tasks, PACIFIC and AmbigSQL. A popular off-policy approach is to sample responses from two high capacity models, with  $Y_w$  coming from whichever model is of higher capacity (henceforth DPO-Dist; see (Mitra et al., 2023; Mukherjee et al., 2023; Xu et al., 2024a)).

Table 4.17: “Proactive Mixed-Initiative” in-context example given as part of a prompt to an LLM for response generation in PACIFIC. The purple text is the overall system instruction. The blue text is what an LLM would have to produce at inference time.

You are an Assistant answering questions from a User. You should either attempt to answer the question or ask a clarifying question if there is any ambiguity.

[Table and Passage Omitted]

User: How much does the company expect to make contributions to the defined plans during the next fiscal year?

The user’s last question was ambiguous. The Assistant asks a clarifying question.

Assistant: [‘What kind of defined plans are you asking about?’]

User: The defined benefit plans and the defined contribution plan respectively.

The user’s last question was unambiguous. The Assistant directly answers the question.

Assistant: [‘5.1million’, ‘0.6 million’]

User: What is the value of the expected defined contribution plan as a percentage of the defined benefit plan during the next fiscal year?

The user’s last question was unambiguous. The Assistant directly answers the question.

Assistant:  $0.6/5.1 = 11.76$

User: How much would change if there is a 1% increase and a 1% decrease respectively in the discount rate?

The user’s last question was ambiguous. The Assistant asks a clarifying question.

Assistant: [‘What kind of change are you asking about?’]

User: The pension.

The user’s last question was unambiguous. The Assistant directly answers the question.

Assistant: [‘(39,145)’, ‘49,361’]

User: What is the difference in that amount if the discount rate had a 1% increase and a 1% decrease?

The user’s last question was unambiguous. The Assistant directly answers the question.

Assistant:  $49,361 - (39,145) = 88506$

User: How much would the pension change if there is a 1% increase and a 1% decrease respectively?

Actions: [“Directly Answer”, “Ask a Clarification Question”]

Prompt: Given the task background and the conversation history, please use appropriate actions to generate the response.

Response: The user’s last question was ambiguous. The Assistant asks a clarifying question.

Assistant: [‘Which rate of increase and decrease are you talking about?’]



Table 4.18: **Experimental results on PACIFIC’s public evaluation set with additional results using Gemini Flash and Claude Haiku.** *ACT* achieves the strongest performance compared to all tuning approaches across every condition in every metric. Tuning-based adaptation strategies are not given any in-context examples at inference time, whereas inference-time adaptation strategies are prompted with 10 in-context conversation examples using the same strategy.

Adaption Setting		Action-level		Content-level		
Base Model	Approach	Conversations	Macro F1 ↑	Turn F1 ↑	Traj. F1 ↑	Post-Clarify F1 ↑
Gemini Pro	Standard Prompt	10	81.4	59.7	58.7	<b>49.7</b>
Gemini Pro	Chain-of-Thought	10	<b>86.3</b>	<b>66.3</b>	17.1	19.2
Gemini Pro	Proactive MIPrompt	10	78.9	63.4	61.1	18.9
Gemini Flash	Standard Prompt	10	67.4	58.8	58.7	17.9
Gemini Flash	Chain-of-Thought	10	77.1	62.0	16.9	20.0
Gemini Flash	Proactive MIPrompt	10	76.8	64.0	<b>62.0</b>	24.4
Claude Sonnet	Standard Prompt	10	71.9	43.7	42.0	28.5
Claude Sonnet	Chain-of-Thought	10	80.0	37.2	13.0	6.8
Claude Sonnet	Proactive MIPrompt	10	74.9	47.2	45.9	7.6
Claude Haiku	Standard Prompt	10	46.9	26.4	26.2	—
Claude Haiku	Chain-of-Thought	10	48.6	23.7	12.0	2.9
Claude Haiku	Proactive MIPrompt	10	48.3	18.6	18.2	7.3
Zephyr 7B- $\beta$	SFT	50	69.0	57.8	61.3	43.5
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	50	75.5	61.7	55.7	30.8
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	50	74.8	62.0	56.3	31.9
Zephyr 7B- $\beta$	IRPO	50	67.7	59.1	56.7	34.4
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	50	<b>82.2</b>	<b>62.8</b>	<b>61.9</b>	<b>57.2</b>
Zephyr 7B- $\beta$	SFT	100	82.3	58.6	60.3	49.9
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	100	68.8	53.3	53.3	31.7
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	100	83.0	59.0	53.7	29.3
Zephyr 7B- $\beta$	IRPO	100	84.5	60.4	55.2	38.2
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	100	<b>86.0</b>	<b>65.0</b>	<b>62.0</b>	<b>57.4</b>
Zephyr 7B- $\beta$	SFT	250	86.9	65.1	63.3	56.7
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	250	65.6	53.6	54.1	30.9
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	250	82.8	43.3	38.6	19.6
Zephyr 7B- $\beta$	IRPO	250	85.4	64.9	58.4	40.3
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	250	<b>89.6</b>	<b>68.1</b>	<b>65.7</b>	<b>62.0</b>

## 4.6 Experimental Results

To emulate real-world scenarios with limited data, we evaluate *ACT* as a tuning approach in different scenarios with limited conversation samples across a set of diverse tasks.

### 4.6.1 Conversational QA with Tabular Grounding

In Table 4.18, we see that across all three data-efficient settings considered, *ACT* achieves the strongest performance across all metrics compared to both SFT and IRPO, which has the advan-

Table 4.19: **Abg-CoQA test set evaluation results with additional results using Gemini Flash and Claude Haiku.** *ACT* outperforms SFT across all evaluations in all three data settings. However, Gemini Ultra achieves the strongest downstream task performance when prompted with in-distribution in-context conversation examples.

Base Model	Approach	Adaptation Setting		Action-level		Content-level	
		Conversations		Macro F1 ↑	Turn Similarity ↑	Traj. Similarity ↑	
Gemini Pro	Standard Prompt	10		55.5	<b>67.0</b>	<b>72.2</b>	
Gemini Pro	Chain-of-Thought	10		61.2	63.4	39.1	
Gemini Pro	Proactive MIPrompt	10		55.5	63.3	33.3	
Gemini Flash	Standard Prompt	10		52.6	62.5	67.4	
Gemini Flash	Chain-of-Thought	10		61.2	56.5	36.6	
Gemini Flash	Proactive MIPrompt	10		58.1	61.7	36.1	
Claude Sonnet	Standard Prompt	10		<b>66.0</b>	50.1	54.3	
Claude Sonnet	Chain-of-Thought	10		63.7	46.2	36.8	
Claude Sonnet	Proactive MIPrompt	10		57.2	60.8	32.9	
Claude Haiku	Standard Prompt	10		49.3	40.9	41.7	
Claude Haiku	Chain-of-Thought	10		46.2	30.7	28.0	
Claude Haiku	Proactive MIPrompt	10		45.2	34.5	31.4	
Zephyr 7B- $\beta$	SFT	50		44.6	53.3	64.2	
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	50		46.9	57.2	61.2	
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	50		44.7	57.9	61.5	
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	50		<b>52.3</b>	<b>66.2</b>	<b>68.8</b>	
Zephyr 7B- $\beta$	SFT	100		<b>52.6</b>	63.1	69.4	
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	100		47.8	61.9	67.1	
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	100		44.8	62.0	66.4	
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	100		51.1	<b>69.5</b>	<b>71.4</b>	
Zephyr 7B- $\beta$	SFT	250		<b>53.5</b>	64.0	66.2	
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	250		46.0	61.9	66.3	
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	250		46.3	62.6	67.0	
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	250		<b>53.3</b>	<b>72.5</b>	<b>75.1</b>	

tage of additional test-time computation (Snell et al., 2024; Pang et al., 2024). In particular, *ACT* achieves up to a 19.1% relative improvement over SFT when measuring the tuned model’s ability to implicitly recognize ambiguity (from 69.0 to 82.2 Macro F1) given only 50 conversations as tuning data. We also observe that *ACT* has greatly improved data efficiency compared to adapter-based SFT with Gemini Pro, with a relative improvement of as high as 35.7% in multi-turn task performance (from 45.6 to 61.9 in terms of trajectory-level DROP F1). Additionally, tuning with *ACT* in these limited data settings grants the model the ability to match or outperform frontier LLMs used with in-context learning despite having zero in-context examples during inference. Overall, we find that on-policy learning and multi-turn trajectory simulation are crucial for improved multi-turn goal completion.

Table 4.20: **AmbigSQL test set evaluation with additional results using Gemini Flash and Claude Haiku.** Zephyr tuned with *ACT* is able to achieve the strongest task performance within each data setting. There are especially large performance improvements in post-clarification SQL execution match when data resources are more scarce.

Base Model	Adaptation Setting		Action-level	Content-level	
	Approach	Conversations	Accuracy $\uparrow$	Execution Match $\uparrow$	PC Execution Match $\uparrow$
Gemini Pro	Standard Prompt	10	72.1	63.5	75.2
Gemini Flash	Standard Prompt	10	75.6	64.2	66.2
Claude Sonnet	Standard Prompt	10	68.5	66.5	72.4
Claude Haiku	Standard Prompt	10	73.8	57.3	65.3
Zephyr 7B- $\beta$	SFT	50	77.4	21.9	13.9
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	50	77.7	42.6	31.5
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	50	78.0	40.9	<b>41.2</b>
Zephyr 7B- $\beta$	IRPO	50	91.0	27.8	30.8
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	50	<b>80.8</b>	<b>43.6</b>	38.1
Zephyr 7B- $\beta$	SFT	100	97.2	43.3	34.3
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	100	98.7	45.1	45.3
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	100	<b>99.8</b>	47.8	44.8
Zephyr 7B- $\beta$	IRPO	100	96.2	45.0	37.0
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	100	99.2	<b>48.0</b>	<b>49.6</b>
Zephyr 7B- $\beta$	SFT	250	99.8	51.0	50.7
Zephyr 7B- $\beta$	DPO-Dist (Pro v. Flash)	250	97.3	49.7	44.2
Zephyr 7B- $\beta$	DPO-Dist (Sonnet v. Haiku)	250	99.7	50.7	50.3
Zephyr 7B- $\beta$	IRPO	250	97.0	49.7	45.6
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	250	<b>99.9</b>	<b>52.3</b>	<b>53.0</b>
Zephyr 7B- $\beta$	SFT	14,000 (All)	99.8	63.1	60.4

#### 4.6.2 Conversational QA for Machine Reading Comprehension

Our results for Abg-CoQA are presented in Table 4.19. In all three data settings, we observe that *ACT* achieved the strongest performance in terms of task-specific metrics (notably, in terms of trajectory-level embedding similarity). However, in the setting with 100 and 250 conversations, Zephyr tuned with SFT slightly outperforms *ACT* in terms of implicit action recognition, although action-level performance primarily helps to contextualize clarification reasoning ability. We discuss this point further in Appendix 4.7. Our approach leads to the strongest turn-level and trajectory-level task performance in all conditions, indicating improved multi-turn reasoning.

#### 4.6.3 Conversational Text-to-SQL Generation

We find that although the prompting baselines do not achieve as high Action Accuracy, the benchmarked frontier LLMs can achieve relatively strong downstream performance in terms of

Task Adaptation Environment				Action-level	Content-level		
Base Model	Framework	Action Supervision	Tuning Ex.	Macro F1 ↑	Turn F1 ↑	Traj. F1 ↑	Post-Clarify F1 ↑
Zephyr 7B- $\beta$	SFT	NA	50	69.0	57.8	61.3	43.5
Zephyr 7B- $\beta$	ACT	Crowdsourced	50	82.2	62.8	61.9	57.2
Zephyr 7B- $\beta$	ACT	Pseudo-labeled	50	80.1	62.4	61.1	54.7
Zephyr 7B- $\beta$	SFT	NA	100	82.3	58.6	60.3	49.9
Zephyr 7B- $\beta$	ACT	Crowdsourced	100	86.0	65.0	62.0	57.4
Zephyr 7B- $\beta$	ACT	Pseudo-labeled	100	84.8	63.5	61.5	56.1
Zephyr 7B- $\beta$	SFT	NA	250	86.9	65.1	63.3	56.7
Zephyr 7B- $\beta$	ACT	Crowdsourced	250	89.6	68.1	65.7	62.0
Zephyr 7B- $\beta$	ACT	Pseudo-labeled	250	89.0	68.1	64.9	61.0

Table 4.21: **Examining ACT on PACIFIC with unlabeled conversational data.** We assume no access to action labels and instead use 0-shot Gemini Pro as the source of action label supervision.

execution match. In contrast, tuning Zephyr with both SFT and ACT results in rather high Action Accuracy but lower text-to-SQL performance compared to the frontier LLMs. We observe that holistically, ACT achieves the largest relative performance improvements in multi-turn task performance compared to other tuning approaches, although the downstream SQL generation ability of larger models is much greater than that of smaller models. This is primarily due to the SQL generation benefiting greatly from scale (Sun et al., 2023a). It is possible that multi-turn performance on larger models can be improved further if ACT is applied, as it is even able to yield larger performance improvements than baseline approaches for quantitative reasoning such as IRPO.

#### 4.6.4 ACT In-The-Wild: Learning Without Dialogue Action Supervision

Although we have ambiguity labels in the tasks considered here and use them for supervision in Tables 4.18–4.20, we also demonstrate that it is possible to perform action-based tuning in the absence of action-label supervision. We use a pre-existing LLM, Gemini 1.5 Pro, as a zero-shot action annotator to re-label the ground truth Assistant-side turns on the PACIFIC corpus. We find that there is astonishingly high agreement (98.5%) with the ground truth action labels. Our results in terms of both Action-level and Content-level metrics reflect that there is nearly no empirical difference in performance. This highlights the potential of ACT being highly effective for adaptation to “in-the-wild” settings with small amount of unlabeled conversational data.

	Macro F1 ↑	Turn F1 ↑	Traj. F1 ↑	Post-Clarify F1 ↑
<b>Action Importance</b>				
<i>ACT</i>				
w/ Random Actions	63.2	55.3	58.7	32.8
<b>Ablation of <i>ACT</i> subcomponents</b>				
<i>ACT</i>				
w/o on-policy sampling	74.8	61.5	59.1	40.5
<i>ACT</i>				
w/ sampling but w/o simulation	81.4	60.8	60.2	50.1
<i>ACT</i> (full)	82.2	62.8	61.9	57.2
<b><i>ACT</i> with unaligned foundation models</b>				
Gemma 2B SFT	57.7	38.0	40.5	17.0
Gemma 2B <i>ACT</i>	<b>62.7</b>	<b>42.6</b>	<b>44.0</b>	<b>24.8</b>
Mistral 7B SFT	57.7	53.8	51.4	27.7
Mistral 7B <i>ACT</i>	<b>75.7</b>	<b>58.1</b>	<b>57.6</b>	<b>31.9</b>

Table 4.22: **Ablation study of various conditions** using PACIFIC’s 50 conversation setting.

#### 4.6.5 Ablation Studies

**Are action-based preferences necessary?** One of the key factors of *ACT* is that the contrastive pairs highlight differences between conversational actions. In Table 4.22 (“*ACT* w/ Random Actions”), we additionally examine the importance of action selection by randomly sampling *both the winning and losing action* when constructing the preference pair, and observe this underperforms normal *ACT*.

**Do we need on-policy sampling?** In Table 4.22 (“*ACT* w/o on-policy sampling”), we examine the importance of on-policy sampling by evaluating normal off-policy DPO on the dataset as constructed in Sec. 4.3.2. While we do observe some improvements over SFT (e.g., from 69.0 to 74.8 Macro F1), the overall improvements are much larger when using on-policy sampling as with full *ACT*. This may be due to the fact that the off-policy negative responses are not guaranteed to lie in the language manifold of the policy model, and distribution shift may be too difficult to overcome with off-policy learning (Guo et al., 2024).

**Is trajectory simulation necessary?** *ACT* is better-aligned with multi-turn conversations due

to its on-policy trajectory simulation. Without multi-turn simulation, our approach can be viewed similarly to on-policy DPO variants like Pang et al., 2024, but with a conversation-specific reward signal which accounts for conversation actions and task heuristics. In Table 4.22 (“*ACT* w/ sampling w/o simulation”), we find that this trajectory-level simulation is critical to improving multi-turn performance, especially the policy model’s ability to reason about its own clarification questions.

**Is *ACT* model agnostic?** The base model in our main experiments, Zephyr, is obtained by aligning Mistral. In Table 4.22 (“*ACT* with unaligned foundation models”) we observe a performance gap of 6.5 Action F1 and 4.3 Trajectory F1 after *ACT* tuning for the two models. However, our results demonstrate *ACT* can improve performance regardless of pre-existing alignment with human feedback, although it can help as an improved model initialization. Overall, we find that improving base model performance with *ACT* is model agnostic.

## 4.7 Limitations, Ethical Considerations, and Broader Impacts

### 4.7.1 Discussion of Limitations

We assume that the clarification questions are appropriately timed. However, crowdsourced conversation datasets are often noisy (Chen et al., 2023a), and relying on noisy annotations or token sequences may result in suboptimal learned policies (from the perspective of asking unnecessary clarifying questions, as well as generating disfluent language). Depending on the source of data, it may be necessary to do an additional preprocessing stage in which one infers whether an action is useful or not. Shi et al. (2024) infers user satisfaction given model responses in-the-wild, whereas Yu, Chen, and Yu (2023) ranks dialogue actions using Monte-Carlo Tree Search.

Label noise can also affect the implicit action recognition evaluation, which assumes that an action in a benchmark task is “optimal.” In a corpus like PACIFIC with high inter-annotator agreement (0.62), this is a reasonable assumption. However, we observe inconsistency in Abg-CoQA which may be a result of the low inter-annotator agreement (0.26) reported in Guo et al. (2021). Recent work has demonstrated the promise of many-shot in-context learning (Agarwal et al., 2024)

Table 4.23: **Analysis of the impact of additional data on Abg-CoQA.** Additionally many-shot examples do not necessarily improve implicit action recognition performance. *ACT* tuning with Zephyr 7B greatly outperforms many-shot Gemini performance.

Base Model	Adaptation Setting		Action-level	Content-level	
	Approach	Conversations	Macro F1 $\uparrow$	Turn Similarity $\uparrow$	Traj. Similarity $\uparrow$
Gemini Pro	ICL	50	<b>56.4</b>	64.5	<b>68.9</b>
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	50	52.3	<b>66.2</b>	68.8
Gemini Pro	ICL	100	<b>59.2</b>	67.0	<b>72.0</b>
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	100	51.1	<b>69.5</b>	71.4
Gemini Pro	ICL	250	<b>58.8</b>	66.0	71.1
Zephyr 7B- $\beta$	<i>ACT</i> (ours)	250	53.3	<b>72.5</b>	<b>75.1</b>

compared to supervised fine-tuning at the trade-off of inference-time latency. Yet, Table 4.23 indicates that even with a greatly increased number of in-context conversation examples (e.g. 50, 100, and 250), the downstream disambiguation ability does not improve uniformly. We thus posit that in such scenarios, it is more important that for such corpora, multi-turn task completion is a more important measure of a model. We do find that even with 250 in-context examples, tuning a smaller model with *ACT* and the same conversation samples has the potential to outperform frontier models with many-shot examples.

*ACT* also makes use of task-specific heuristics. While this was intentional since success criteria can vary greatly across domains, this may also require more customization and engineering expertise/effort. Our overall approach to tuning and evaluation also makes heavy use of existing LLMs. We prompt Gemini for purposes such as Action Classification or User Simulation, but such approaches are not perfect and may occasionally result in unwanted behavior. These prompting approaches similarly may require substantial customization efforts. We also realize that not all researchers may have access to commercial LLMs due to researchers for financial or privacy reasons.

Our study also focuses specifically on the limited data regime. We believe that such contexts (e.g., when the target user population is unknown; when conversational data cannot be collected due to privacy concerns; when a conversational system is in its early stages and collecting abundant

data for development iteration is not feasible; etc.) would benefit the most from focused adaptation designed to fundamentally teach conversational skills approaches such as *ACT*. As such, in our paper, this was the focus of all of our experiments. It is not clear how much our findings would generalize in settings in which there is an abundance of training data whose distribution closely matches the target distribution. Intuitively, if much more in-distribution data is made available, even the performance of unaligned objectives like SFT would start to catch up to the performance of focused approaches.

**Is *ACT* online learning?** (Levine et al., 2020) defines offline reinforcement learning as using a fixed dataset of experiences, whereas online reinforcement learning relies on interacting with an environment in real-time. Additionally, (Guo et al., 2024) defines on-policy sampling in contrastive RL tuning as settings where both the winning and losing responses are sampled from the policy model. In our case, during tuning, we sample a single response from the policy model.

As such, we define *ACT* as a *quasi-online* contrastive RL tuning approach. *ACT* does rely on action-based preference dataset, as is common in fully-offline reinforcement learning. However, *ACT* continuously samples responses from the policy model in order to update the contrast pairing with good or bad trajectories. Overall, it has both dynamic and static components, so we refer to it as quasi-online. *ACT* also specifically is different from fully online DPO where both the winning and losing responses are sampled (i.e. in (Guo et al., 2024)) because our focus is on conversational actions. There is no guarantee that sampling a response from the policy model twice will result in differing actions, unless you change the prompt. However, in that case, you would no longer be computing the DPO objective.

By nature of the domains considered, the extent to which *ACT* allows for online exploration is also limited. As previously mentioned, our experiments are constrained by the fact that there is an objective right/wrong target answer. For instance, if the target answer is an arithmetic expression as is common in PACIFIC, there are a fairly limited number of unique trajectories (when inspected in terms of the number of tokens) that will arrive at that particular expression. In such cases, the



trajectory sampled from the policy model during *ACT* tuning may not be any different from the offline trajectory found in the training data.

#### 4.7.2 Ethical Considerations

**Usage and Safety** We do not condone the use of our work for any unethical, unlawful, and/or harmful applications. In our work, we do not release any new model artifacts or web-scraped data, but we do not specifically introduce any model guardrails as a part of *ACT*. However, our implementation of *ACT* relies on other LLMs such as Gemini to produce an initial preference dataset, and to perform user simulation. Gemini is released with safety guardrails in place, but these inference-time guardrails may not be available when using open-source LLMs instead. We advise that any deployments of models tuned with *ACT* should consider adding safety guardrails at inference-time.

**Hallucinations** One commonly documented concern with using LLMs is their tendency to hallucinate answers in Assistant QA contexts. A solution is to provide an LLM with information from a retriever (i.e., retrieval-augmented generation). Two of the datasets we use, PACIFIC and Abg-CoQA, mirror this setting by performing grounded QA using a mixture of long-context textual passages and tabular data. It follows that *ACT* could be further studied in combination with approaches for improved retrieval-augmented generation.

Our evaluation criteria in this paper are also rather restrictive towards hallucinations. In PACIFIC, we use a token-level metric (DROP F1); in Abg-CoQA, we evaluate a candidate response’s semantic similarity with a ground truth answer; in AmbigSQL we use execution match, which is a fully objective metric. As such, it is difficult to perform well on PACIFIC and Abg-CoQA if a model consistently hallucinates answers, and in AmbigSQL, a “hallucinated” response would not consist of the appropriate SQL code.

### 4.7.3 Broader Impacts

There is an abundance of modern conversational assistants. *ACT* seeks to improve multi-turn conversational experiences, and thus, it can be used to improve many applications used by potentially millions of users around the world. However, the popularity of conversational assistants also creates an increased risk of misuse. Some people may develop conversational for unethical applications such as misinformation generation, or gray areas such as optimizing dialogue generation for content which is not suitable for the workplace. As discussed above, we do not condone the use of *ACT* for any unethical, harmful, or unlawful applications, and we suggest the usage of safety guardrails for any deployments.

## 4.8 Summary

In Chapter 3, we introduced an approach for explicit conversational action planning, and here in Chapter 4, we propose *ACT*, a model-agnostic quasi-online contrastive tuning approach for sample-efficient conversational task adaptation, along with a workflow for evaluation of conversational agents. *ACT* is the one of the first approaches to propose the implicit action recognition paradigm, and to look at on-policy multi-turn RLHF. We demonstrate encouraging evidence that *ACT* is highly effective for task adaptation in the limited data regime, even when there are no action labels available. It remains to be seen how *ACT* will translate to a broader, non-binary action space, both in terms of conversational (e.g., argumentative strategies) and non-conversational actions (e.g., internet retrieval tool usage). It is also worth further studying the tradeoffs of on-policy and off-policy learning, as well as their applicability to the large-scale data regime.

Despite the diversity of tasks considered, our results are also currently only applicable to the unimodal setting. Written language is not the only way in which humans engage in conversational interaction – other modalities such as spoken language carries much richer paralinguistic cues and is much more nuanced, requiring models to be adaptable to individual user patterns. This motivates the design of training approaches to improve the end-to-end speech modeling ability of MLLMs,

as well as further investigation into the relationship between spoken language and mixed-initiative behaviors such as disambiguation.

## Chapter 5: Improving User Modeling in End-to-End Multi-Modal Spoken Conversation Modeling via Data-centric Multi-task Learning

In this chapter, we apply the notion of implicit action optimization to spoken multimodal interaction. We propose a novel mixed-initiative task for disambiguation in spoken listening comprehension which also requires the ability to model individual users’ speaking patterns. We introduce a simple yet effective training approach for adapting MLLMs to such multi-turn, multimodal tasks.

### 5.1 Introduction

Real-world adoption of intelligent multimodal conversational agents has progressed quickly in recent years due to the impressive capabilities of Large Language Models (LLMs). However, despite numerous applications, including smart home systems, contact centers, customer support/service, personalized education, and others (Hemphill, Godfrey, and Doddington, 1990; Khatri et al., 2018b; Li et al., 2017a; Von Ahn, 2013; Fatima et al., 2024; Li, Chrysostomou, and Yang, 2023; Zheng et al., 2024), there has not been the same rapid progress in adapting Multimodal LLMs (MLLMs) to spoken contexts due to several fundamental challenges.

Speech data constitute high-dimensional signals that are difficult to model even for frontier models (e.g., Whisper-based models are limited to 30 seconds; Chu et al. (2024) and Radford et al. (2023)) and Gemini 1.5 represents 1 second of audio using 25 tokens<sup>1</sup>). These are temporal signals which include acoustic phenomena (e.g. background noise (Mehri et al., 2023)) and important paralinguistic information such as speaking rates or pitch (Hirschberg, 1993; Bhattacharya et al., 2023). Performance on speech understanding tasks is thus greatly affected by the ability to robustly comprehend the semantic contents of the input speech (Li et al., 2017a), as illustrated in

---

<sup>1</sup><https://ai.google.dev/gemini-api/docs/audio?lang=python>

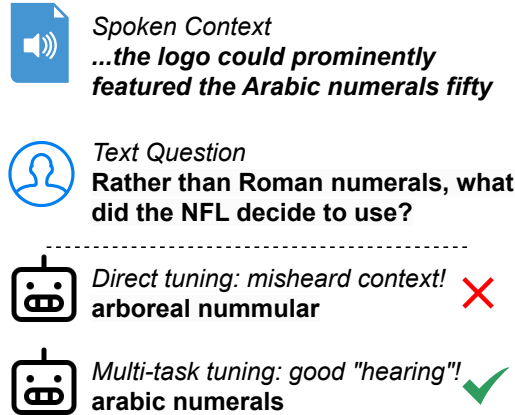


Figure 5.1: Automatic speech recognition is a necessary *implicit* skill for MLLM in end-to-end spoken question answering. We propose a multi-task learning approach which explicitly teaches these skills, as exemplified by this QA pair from Spoken-SQuAD.

Fig. 5.1. This is further complicated by the long-standing issue of models overfitting to individual speakers (Jung et al., 2018; Wang et al., 2020d). These can be viewed as a shortcoming of insufficient training data coverage (Yang et al., 2024b). However, large-scale speech data collection is notoriously difficult due to privacy concerns (Nautsch et al., 2019; Qian et al., 2018).

Despite the difficulty of large-scale collection, task-specific data is increasingly the most effective approach to guarantee use-case customization for state-of-the-art MLLMs like Gemini or GPT (Gemini Team et al., 2023; Brown et al., 2020). These models are closed-source, but offer commercial tuning APIs, which typically do not permit modifications to the model or learning objective. Even with smaller open-weight models, it can still be computationally intractable to iterate on architectures and train from scratch due to the expensive compute resource demands (Groeneweld et al., 2024). These motivate the design of efficient data-centric methods (Seedat, Imrie, and Schaar, 2022) which maximize models’ ability to overcome the aforementioned challenges of long speech understanding reliably.

In this work, we take a data-centric perspective towards addressing the varied challenges of adapting multimodal LLMs for speech. Our contributions can be summarized as follows:

- We bring a multi-task learning paradigm to improve speech understanding implemented via a

simple but effective data-centric approach. Rather than using additional datasets, we instead design auxiliary tasks to maximize cross-modal learning from a fixed set of recorded speech.

- We propose **Ambiguous Spoken Conversational Question Answering (ASK-QA)**, a novel dataset which combines the challenges of multimodal speech modeling and mixed-initiative interaction. ASK-QA features contextually ambiguous questions along with long multi-turn speech and diverse accents, speaking rates, and pitches.
- We validate the proposed data-centric approach on three spoken question answering (SQA) corpora: ASK-QA, Spoken-SQuAD, and SD-QA, representing various combinations for whether input questions and knowledge context are represented as text or speech. Our approach applied even to open-weight models is able to outperform the existing state-of-the-art on Spoken-SQuAD using only 10% of the available training data.

## 5.2 Related Work

**Spoken question answering** is a fundamental skill for intelligent spoken conversational agents (Khatrri et al., 2018a; Zheng et al., 2024). Many tasks have been proposed in order to measure models’ ability to understand spoken context (Li et al., 2018; Shih et al., 2024) and spoken requests (Faisal et al., 2021). Previously, most approaches to SQA focused on span prediction using “cascade” approaches which include an intermediate step invoking an Automatic Speech Recognition (ASR) module followed by a fine-tuned a text classification model (Chuang et al., 2020; Li et al., 2018; Su and Fung, 2020) like BERT (Devlin, 2018). It is increasingly desirable to develop end-to-end pipelines to solve SQA tasks (Shih et al., 2024), particularly with the rise of generalist MLLMs (Wu et al., 2024). Such end-to-end models are desirable in speech as they afford opportunities to directly encode useful information in acoustic signals such as speaking rate, pitch, or emotions. In our work, we focus on improving methods for end-to-end SQA using both closed-weight and open-weight MLLMs.

**Mixed-initiative conversations** require each interlocutor to control the interaction flow (Horvitz, 1999) through the use of various pragmatic actions (Chen et al., 2023a) such as clarifying questions, which can lead to better goal completion outcomes (Guo et al., 2021; Min et al., 2020; Wu et al., 2023b). Many works focus on planning these explicit pragmatic actions (Deng et al., 2024b; Yu, Chen, and Yu, 2023), whereas other works focus on implicit (Chen et al., 2025) and continuous space actions (Wu et al., 2023a), and end-to-end generation capabilities in such settings (Li et al., 2020b; Deng et al., 2022). While there have been recent efforts in designing multi-turn SQA corpora (You et al., 2022), to our knowledge, there is not yet any mixed-initiative conversation environment for speech, despite there being many additional acoustic features which may introduce ambiguity (Kurata, Itoh, and Nishimura, 2011; Mulholland et al., 2016). In our work, we develop the first-ever conversational SQA corpus which requires the ability to disambiguate requests and reason about clarification questions.

**Adapting models with limited speech data** has received much attention due to well-known problem of speaker overfitting across a variety of tasks ranging from grammatical error correction (Wang et al., 2020d) to speaker verification (Jung et al., 2018). This problem is frequently addressed with the assistance of multi-task learning (Caruana, 1997). Pironkov et al. (2016) proposed a multi-task objective in which they simultaneously train a network for both ASR (their downstream task) and speaker classification. Chen and Yu (2023) found large downstream task performance improvements on speech classification tasks following a stage of multi-task pre-finetuning. In our work, we view multi-task learning through a data-centric lens. While multi-task pre-finetuning relies on additional datasets (Aghajanyan et al., 2021; Padmakumar et al., 2022), our approach improves the utilization of a fixed set of speech recordings by introducing auxiliary tasks designed to improve the cross-modal understanding capabilities of MLLMs.

Usage	Modality	Content
Input	Speech	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical telephone.
Input	Text	The audio contains a passage providing some information. Follow the User’s request. User: Transcribe the speech in the passage.
Output	Text	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical telephone.

Table 5.1: Example of the Listening Comprehension auxiliary task for Spoken-SQuAD.

### 5.3 Data-Centric Multi-Task Learning for Cross-Modal Understanding

We consider the setting of customizing an MLLM for use in request-based end-to-end speech modeling, similarly to Shih et al. (2024). An MLLM is provided as input an audio recording and textual context. The backbone of many MLLM architectures is a textual decoder-only LLM (Liu et al., 2024a), so the textual context usually contains an instruction. These settings involve reasoning about some contextual knowledge and conversation history. The model aims to provide a correct answer to a target question (i.e. the last conversation turn). Different applications may involve spoken conversations about written information (e.g. document-grounded QA), or written conversations about spoken information (e.g. meeting summarization).

Tuning MLLMs with cross-entropy loss is advantageous as it can be used to unify diverse tasks as a single text-to-text objective (Raffel et al., 2020). Many recent studies find that multi-task learning (Caruana, 1997) using *additional datasets* greatly improves downstream task performance (Aghajanyan et al., 2021; Padmakumar et al., 2022; Chen and Yu, 2023). *Here, we instead design auxiliary tasks within the same dataset to maximize cross-modal learning gains from a fixed set of audio recordings for a target dataset.* We break down our problem into three intermediate goals: 1) correctly representing the spoken context, 2) learning to reason across all input modalities, and 3) coherently producing the correct answer.



Usage	Modality	Content
Input	Speech	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical <b>telephone</b> .
Input	Text	The audio contains a passage providing some information. The user will ask a question about some information from the audio. The assistant should answer the user’s question by selecting the correct answer choice. User: What is Bell most famous for inventing? Choose from the following choices: A. britain <b>B. telephone</b> C. major performing arts D. london county council
Output	Text	<b>B. telephone</b>

Table 5.2: Example of the commonsense Response Selection auxiliary task for Spoken-SQuAD.

**1) Listening Comprehension** is an auxiliary task to help the SLM “hear” the spoken context. It has been consistently reported in traditional cascade-style systems that SQA performance is greatly affected by automatic speech recognition (ASR) errors (Li et al., 2018), and thus we design a task to specifically address this point. The objective is for the MLLM to predict a ground-truth (or pseudo-labeled) audio transcription, given a recording and a task instruction as input. An example is shown in Table 5.1.

**2) Cross-Modal Commonsense Reasoning** is an auxiliary task designed to unify the contents of the spoken and textual inputs. We reframe dialogue response selection (Henderson et al., 2019) as a multiple-choice reasoning task (Talmor et al., 2019). The answer options consist of the correct answer (e.g. “It was recovered a few months later”) and commonsense negative answer choices sampled from other training QA pairs (e.g. “Do you mean the popular generic name?”), as shown in Table 5.7. The objective is to solve this multiple-choice reasoning task by selecting the correct answer given the recording, conversation context, knowledge, answer options, and a task instruction.

**3) Response Generation** is the primary objective of providing a correct answer. The inputs are what is expected to be provided to an MLLM at inference time for SQA: the recording, conversation context, information context, and a task-specific instruction. An example is shown in Table 5.10.

Usage	Modality	Content
Input	Speech	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical <b>telephone</b> .
Input	Text	The audio contains a passage providing some information. The user will ask a question about some information from the audio. The assistant should answer the user’s question using information which can be found in the passage. User: What is Bell most famous for inventing?
Output	Text	<b>telephone</b>

Table 5.3: Example of the Response Generation auxiliary task for Spoken-SQuAD.

These tasks can be fully implemented as modifications to tuning data mixtures. As we will demonstrate in Section 5.6, this simple modification is observed to be highly effective in improving an MLLMs’ ability to complete downstream tasks, particularly in the limited data regime.

#### 5.4 Efficient Multimodal Adapters via Audio Representation Projection

Our data-centric approach is easily applicable to both settings with access to tuning APIs for closed-source MLLMs like Gemini, and settings with access to open-weight models for each modality. Here, we describe our approach in the open-weight scenario.

Textual instructions serve as a highly controllable interface, and as such, recent work has found much success in unifying multiple modalities with large pre-trained decoder-only language models (Liu et al., 2024a; Arora et al., 2024; Kong et al., 2024). These works aim to leverage the impressive instruction-following capabilities of LLMs to interpret additional modalities (e.g. vision, speech, video etc.) by effectively mapping their representations to LLM input space.

**Architecture:** In our work, we consider the high-level architecture presented in Ma et al. (2024). We project the speech input represented by an audio encoder into the embedding space of an LLM to improve performance on ASR tasks, only tuning the weights of a linear projection layer and freezing the other model components.<sup>2</sup> Our speech encoder is WavLM-Large (Chen et al., 2022c). We primarily experimented with tuning Qwen 2.5-Instruct (Yang et al., 2024a) with 7B

<sup>2</sup>As in Ma et al. (2024), the projection layer consists of merely 17.8M parameters for the proposed models.

parameters as our base decoder-only LLM. We also experimented with Phi 3.5 Mini (Abdin et al., 2024) with 3B parameters in Table D2. These MLLMs are referred to as Speech-Qwen and Speech-Phi, respectively. We tune this adapter using standard cross-entropy loss.

**Projection Layer Pre-training:** While this projection layer is tuned directly on the target ASR task in Ma et al. (2024), we find that this approach may struggle with direct single-task fine-tuning on our more difficult SQA tasks which do not have the same abundance of data. Similarly to how visual MLLMs are often pre-trained on image captioning (Liu et al., 2024a), we pretrain the projection layer for one epoch on large-scale ASR data.

## 5.5 A Corpus for Ambiguous Spoken Conversational Question Answering (ASK-QA)

We develop a novel corpus for speech-based mixed-initiative conversation: **Ambiguous Spoken Conversational Question Answering (ASK-QA)**. The contextual inputs for ASK-QA are *fully spoken*.

	SD-QA	S-SQuAD	ASK-QA (Ours)
Avg. Audio Len.	4.8s	59s	1m 41s
Speakers/Audio	1	1	3
Knowledge	Text	Speech	Speech
Conversation	Speech	Text	Speech
Unique Voices	248	1	64
Avg. Turns	2	2	5.1
Answer Type	Span	Span	Free-form
Ambiguous	✗	✗	✓
Dynamic Eval	✗	✗	✓
Disfluencies	✗	✗	✓

Table 5.4: **Comparison of ASK-QA against existing popular SQA training datasets used for experimentation here.** ASK-QA features ambiguous requests and long audio context.

### 5.5.1 Dataset Construction

**Overview** We construct ASK-QA starting from Abg-CoQA (Guo et al., 2021), a span-based textual conversational QA task. Given a story as context, each conversation consists of dialogue turns where a user asks questions and an assistant is supposed to provide the correct answer or

### ***Original Text Conversation***

**[Story]** The driver of a Toyota Prius says he was taken on a wild ride Monday ...

**[Turn 1] Question Asker:** What road?

**[Turn 2] Question Answerer:** The incident occurred on ...

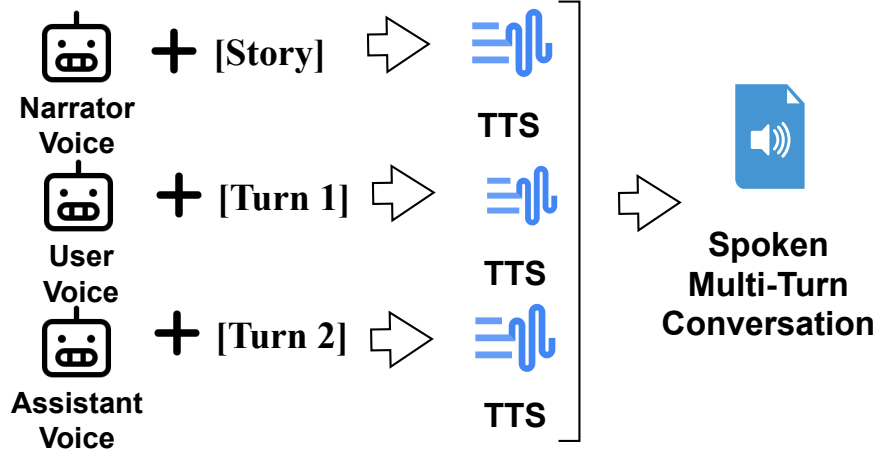


Figure 5.2: **Simplified summary of the pipeline for constructing ASK-QA.** For each text conversation in Abg-CoQA, we construct three speaker profiles with randomly sampled voices, speaking rates, and pitches. We use TTS to synthesize the story context as a spoken narration, then each individual dialogue turn. The resulting audio files are joined as a single recording.

ask a clarifying question if the user’s request is ambiguous. Our data construction pipeline is summarized in Figure 5.2. In total, ASK-QA contains 221.8h of speech. The training, validation, and test sets contain 5,985, 500, and 1,345 conversations.

While Abg-CoQA is a textual conversational QA task, it is span-based, which results in some unnatural dialogue. Each instance consists of a passage which serves as some necessary contextual knowledge, and each conversation consists of dialogue turns where a user asks questions and an assistant is supposed to provide the correct answer or ask a clarifying question if the user’s question is contextually ambiguous. The first step we take is to paraphrase each question using Gemini 1.5 Pro to convert the task into free-form QA generation.

**Setting speaker roles** Each textual conversation can be considered a machine reading comprehension task. We break them down into three components: a story, the set of user questions, and the set of corresponding assistant responses. Our goal is to convert this into a listening comprehension task with two speakers having a conversation about some spoken context. Thus, for each conversation, we construct three unique speaker profiles to represent a story narrator, a user, and an assistant.

**Speaker profiles** Earlier works (Li et al., 2018; You et al., 2022) used commercial text-to-speech (TTS) software to synthesize speech, but at the time there were relatively high word error rate (WER) with limited options for customization. As a result, such corpora only feature a single synthetic voice without varied acoustic features (e.g. speaking rate, pitch). Here, we construct a much more diverse corpus using modern TTS solution from Google Cloud<sup>3</sup>. To create user speaker profiles, we aim to maximize diversity and thus sample from 38 unique voice types spanning four different accents from English-speaking countries (US, AU, GB, IN). We also randomly sample user speaking rates and pitches from a truncated normal distribution. The mean of each is set to the default value of the API endpoint. For the assistant and narrator speaker profiles, we aim to create professional-sounding dialogue, instead sampling from 26 different “news” and “studio” voices.

**Text-to-Speech pipeline** As per Figure 5.2, we then apply TTS to synthesize the story and each dialogue turn sequentially, using the appropriate speaker profile. We then concatenate the resulting audio files into a single spoken conversation. We do not adjust the default speaking rates and pitches. Following the suggestions of earlier work in text-based data synthesis (Chen et al., 2022b; Kim et al., 2023), we apply weakly supervised filtering to ensure that the synthesized speech is high quality. If any synthesized speech exceeds a WER of 0.20 (as determined by Whisper-Large v3; (Radford et al., 2023)), we retry the synthesis process. If it fails three times, we discard the conversation sample. We finally randomly insert white noise into the audio by drawing from a Gaussian distribution (with an average signal to noise ratio of 21.75dB). The result is a unique

---

<sup>3</sup><https://cloud.google.com/text-to-speech>

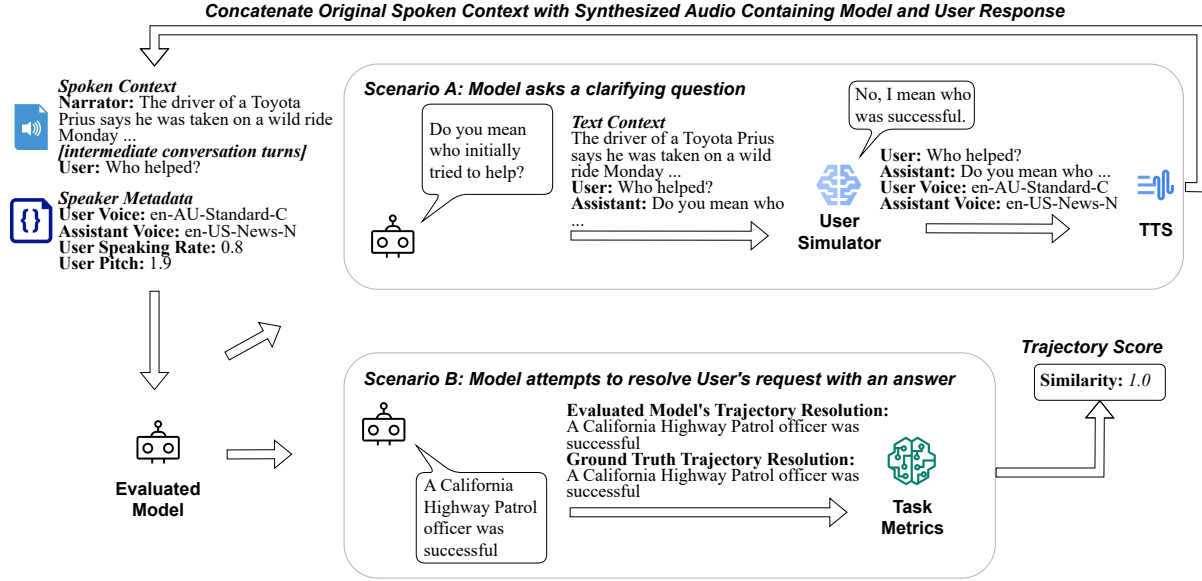


Figure 5.3: **Multi-turn evaluation pipeline for ASK-QA.** A model is given an audio recording containing the spoken story and spoken conversation. It is tasked with providing the correct response. While the model response is a clarifying question (as determined by a prompted Action Classifier), the model-generated response is appended to a textual version of the conversation history and shown to a user simulator. The user simulator provides a coherent response to the clarifying question, and these two generated turns are synthesized using TTS to create a new spoken context. This process repeats until the model response is not a clarifying question.

speech CQA dataset with disfluencies, multiple speakers, and long audio context. The contributions of ASK-QA compared to other existing SQA datasets are in Table 5.4.

### 5.5.2 Evaluation

Following recent work (Chen et al., 2025; Risch et al., 2021b), we apply embedding-based semantic similarity (Reimers and Gurevych, 2019) to allow for flexible free-form QA evaluation. We apply this metric to a standard single-turn setting as well as a novel multi-turn setting which combines TTS with the dynamic input evaluation for Abg-CoQA in Chen et al. (2025).

**Single-turn evaluation details:** We follow the standard single-turn evaluation setting with pre-determined inputs similar to existing conversational QA tasks (Guo et al., 2021; Deng et al., 2022). For an evaluation instance, an agent must produce a correct response conditioned on the speech

recording. In ASK-QA, the speech recording contains both the knowledge context and a multi-turn conversation context. We compare the generated answer against the ground truth response.

**Multi-turn evaluation details:** Chen et al. (2025) propose an automatic multi-turn evaluation for Abg-CoQA, in which an agent dynamically interacts with a user simulator to work towards the goal of a conversation. We design a novel dynamic *speech* evaluation, which is summarized in Figure 5.3.

The evaluation starts by providing the evaluated MLLM with the audio file (see “Spoken Context” in Figure 5.3) and the text instruction. Next, an action classifier determines whether the agent response is a clarifying question or a direct answer. The action classifier is an LLM which is directly prompted with the textual conversation context, the candidate response, and the action candidates (CLARIFY/ANSWER). In this work, we use Gemini 1.5 Pro. If it is a clarifying question, we simulate a user-side response to the clarifying question using a user simulator. This user simulator is a high-capacity LLM (here, Gemini 1.5 Pro as well) provided with the textual conversation context, and a summarization of the user’s ground truth information-seeking intent. Our implementation of these is the same as for Abg-CoQA in Chen et al. (2025).

In the text setting, they concatenate the original context with the newly simulated trajectory and provides this as a new prompt to the evaluated agent. Here, we apply the same TTS process as in Section 5.5.1 using the same speaker profiles for this conversation to dynamically create another fully spoken input. We then sample another response from the evaluated agent until a direct answer attempt is provided, or the maximum rollout depth is reached (here, we set this to three clarification turns). We apply the evaluation metric to the final response.

### 5.5.3 Conversation Examples

In Table 5.5, we provide a simple example of how a conversation in ASK-QA is paraphrased from the original conversation in Abg-CoQA (with the passage context omitted). One turn is grammatically incorrect and so it is paraphrased using Gemini 1.5 Pro.

Version	Turn No.	Speaker	Content
Abg-CoQA	1	User	what was the others reaction?
	2	Assistant	Everyone was happy
	3	User	was he hungry?
	4	Assistant	Do you mean at first?
	5a	User	Yes
	6a	Assistant	Yes
	5b	User	No, I mean ultimately.
	6b	Assistant	No, he was not hungry anymore.
ASK-QA	1	User	what was the others reaction?
	2	Assistant	Everyone was happy.
	3	User	was he hungry?
	4	Assistant	Do you mean at first?
	5a	User	Yes
	6a	Assistant	<b>Yes, he was hungry at first.</b>
	5b	User	No, I mean ultimately.
	6b	Assistant	No, he was not hungry anymore.

Table 5.5: **Comparison of the dialogue turns from a conversation in Abg-CoQA and the paraphrased version in ASK-QA.** Bold: text paraphrased using Gemini. We do not perturb the user turns in order to maintain the original linguistic diversity. However, for the Assistant turns, we paraphrase the language concisely if necessary in order to ensure that the speech is grammatically correct. Here, the only dialogue turn that differs is 6a. (a) and (b) denote differing trajectories, which are the turns that the Assistant has to navigate successfully during evaluation.

In Table 5.6, we provide an example of a conversation in ASK-QA with more perturbations from Abg-CoQA. Turn 4 is rephrased as a complete sentence. Turn 5 injects a repeat disfluency into the User-side speech. Turn 6 is also rephrased as a complete sentence.

Table 5.10 provides a full example of a full example from the ASK-QA dataset. We include the passage context, as well as the provided dialogue excerpt. We denote the input modalities as well as our instruction for response generation using the MLLM.

## 5.6 Experiments

We evaluate our approach on SQA datasets with different combinations of context modalities (see Table 5.4). We prompt and fine-tune two closed-source models, Gemini Pro and Gemini Flash,



Version	Turn No.	Speaker	Content
Abg-CoQA	1	User	Are they related?
	2	Assistant	yes
	3	User	How?
	4	Assistant	brothers
	5	User	Where do they put the lemonade stand?
	6	Assistant	by the sidewalk
ASK-QA	1	User	Are they related?
	2	Assistant	Yes, they are related.
	3	User	How?
	4	Assistant	<b>They are brothers.</b>
	5	User	Where do they put the <i>the</i> lemonade stand ?
	6	Assistant	<b>They put the lemonade stand by the sidewalk.</b>

Table 5.6: **A modified conversation in ASK-QA.** Bold: paraphrased text using Gemini. Italics: repeat disfluency injected using LARD (Passali et al., 2022).

on the Vertex AI platform<sup>4</sup>. We also use Speech-Qwen, which we built by pre-training an 17.8M parameter projection layer between a frozen audio encoder (WavLM-Large) and a frozen LLM decoder (Qwen 2.5 7B-Instruct). See Appendix 5.4 for additional details on Speech-Qwen. Here, we present our findings for ASK-QA (multi-turn speech context with spoken knowledge), Spoken-SQuAD (single-turn text context with spoken knowledge), and SD-QA (single-turn spoken context with textual knowledge).

### 5.6.1 ASK-QA: Spoken Knowledge and Multi-Turn Spoken Dialogue

We first evaluate our proposed data-centric approach on ASK-QA, as introduced in Section 5.5. The application of our proposed approach to ASK-QA slightly differs than that for Spoken-SQuAD and SD-QA due to the multi-turn nature of the task. In particular, we break the Listening Comprehension task into two components since each recording comprises a narrated story and a conversation. We provide concrete examples of the auxiliary tasks for a single instance of ASK-QA in Tables 5.7, 5.8, 5.9, 5.10. Each of these tables has the exact same speech recording. Table 5.8 demonstrates steering the MLLM to transcribe the conversation. Table 5.9 demonstrates steering

<sup>4</sup>[https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune\\_gemini/audio\\_tune](https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune_gemini/audio_tune)

Usage	Modality	Content
Input	Speech	<p><b>Speaker 1:</b> A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. <b>When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring!</b> At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years.</p> <p><b>Speaker 2:</b> Was it expensive?</p> <p><b>Speaker 3:</b> No, it was not expensive.</p> <p><b>Speaker 2:</b> Was it recovered?</p> <p><b>Speaker 3:</b> Yes, it was recovered.</p> <p><b>Speaker 2:</b> When?</p>
Input	Text	<p>The audio recording contains a story followed by a conversation between a User and an Assistant. You will continue the conversation for the Assistant by selecting the most appropriate response from the following: A. Do you mean the popular generic name? B. Are you asking why the dog was looking at Sue or why Jack walked up to Sue? C. More Chinese people can afford cars because of them. <b>D. It was recovered a few months later.</b></p>
Output	Text	<b>D. It was recovered a few months later.</b>

Table 5.7: Example of the commonsense Response Selection auxiliary task for ASK-QA.

the MLLM to transcribe the narrated story. Table 5.7 demonstrates how the ground-truth answer is joined to negatively sampled answers from other QA pairings to form the response selection task. Table 5.10 is similar and demonstrates the textual instruction used to steer the MLLM to directly generate the ground-truth answer.

**Findings:** We benchmark end-to-end performance on ASK-QA in Figure 5.4 and Table D1 using the multi-task approach described in Section 5.3 and baseline single-task tuning (which represents standard end-to-end speech-to-text modeling (Shih et al., 2024)). The listening comprehension sub-task separately models the story and conversation transcripts, inspired by speaker diarization (Anguera et al., 2012; Gu et al., 2021; Yu, Fu, and Zhang, 2022). With Speech-Qwen, we see as much as 13.3% relative improvement over standard fine-tuning depending on the amount of available data on trajectory-level similarity in Figure 5.4. Surprisingly, with Gemini Pro, we also see relative improvements of 5.7% with 1% of the available training data and 1.6% when using

Usage	Modality	Content
Input	Speech	<b>Speaker 1:</b> A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. ... They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years. <b>Speaker 2:</b> Was it expensive? <b>Speaker 3:</b> No, it was not expensive. <b>Speaker 2:</b> Was it recovered? <b>Speaker 3:</b> Yes, it was recovered. <b>Speaker 2:</b> When?
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. Transcribe the conversation but not the story. Provide your answer in the format User: [Utterance] Assistant: [Utterance] and so on.
Output	Text	<b>User:</b> Was it expensive? <b>Assistant:</b> No, it was not expensive. <b>User:</b> Was it recovered? <b>Assistant:</b> Yes, it was recovered. <b>User:</b> When?

Table 5.8: Example of the Dialogue Listening Comprehension auxiliary task for ASK-QA.

full data, despite frontier MLLMs already having large-scale multi-modal pre-training and the full ASK-QA corpus containing large-scale, in-distribution data (over 200 hours). This finding is significant because it *specifically indicates that the MLLM is better learning to model the available speech data*. It is well-documented that 1) high benchmark scores achieved by frontier LLMs on older corpora may be confounded by data contamination (Roberts et al., 2023; Qian et al., 2024), and 2) several studies demonstrate the efficacy of direct fine-tuning given abundant data (Sharma et al., 2024; Yu et al., 2024b). Since ASK-QA is newly synthesized, Gemini cannot have been trained on this exact version of the data. This accurately highlights the difference between direct single-task tuning and multi-task tuning with our proposed auxiliary tasks. The improvements with full data indicate the applicability of the approach for scaled data.

### 5.6.2 Spoken-SQuAD: Spoken Knowledge and Textual Questions

Spoken-SQuAD (Li et al., 2018) is a speech version of SQuAD (Rajpurkar et al., 2016). Rather than span-based classification, we solve the task using our end-to-end generative approach. Each instance has a textual question and spoken knowledge.

Usage	Modality	Content
Input	Speech	<p><b>Speaker 1:</b> A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring! At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years.</p> <p><b>Speaker 2:</b> Was it expensive?</p> <p><b>Speaker 3:</b> No, it was not expensive.</p> <p><b>Speaker 2:</b> Was it recovered?</p> <p><b>Speaker 3:</b> Yes, it was recovered.</p> <p><b>Speaker 2:</b> When?</p>
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. Transcribe the story but not the conversation.
Output	Text	A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. ... They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years.

Table 5.9: Example of the Story Listening Comprehension auxiliary task for ASK-QA.

Usage	Modality	Content
Input	Speech	<p><b>Speaker 1:</b> A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. ... In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. <b>When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring!</b> ... She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years.</p> <p><b>Speaker 2:</b> Was it expensive?</p> <p><b>Speaker 3:</b> No, it was not expensive.</p> <p><b>Speaker 2:</b> Was it recovered?</p> <p><b>Speaker 3:</b> Yes, it was recovered.</p> <p><b>Speaker 2:</b> When?</p>
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. You will continue the conversation for the Assistant by providing the next response. If you do not have enough information or context to answer the User's last question, you should ask a clarifying question.
Output	Text	It was recovered a few months later.

Table 5.10: Example of the Response Generation task for ASK-QA.

**Findings:** In Figure 5.5, we benchmark our multi-task learning approach against single-task tuning via Speech-Qwen. Our performance is evaluated in terms of exact match and F1 score using the SQuAD evaluator. Our approach, applied to an open-weight model like Speech-Qwen, outperforms the existing state-of-the-art model proposed in You et al. (2022) using just 10% of the available training data, indicating that it is *highly efficient and effective for cross-modal learning*. We present expanded results which include an additional MLLM are shown in Table D2.

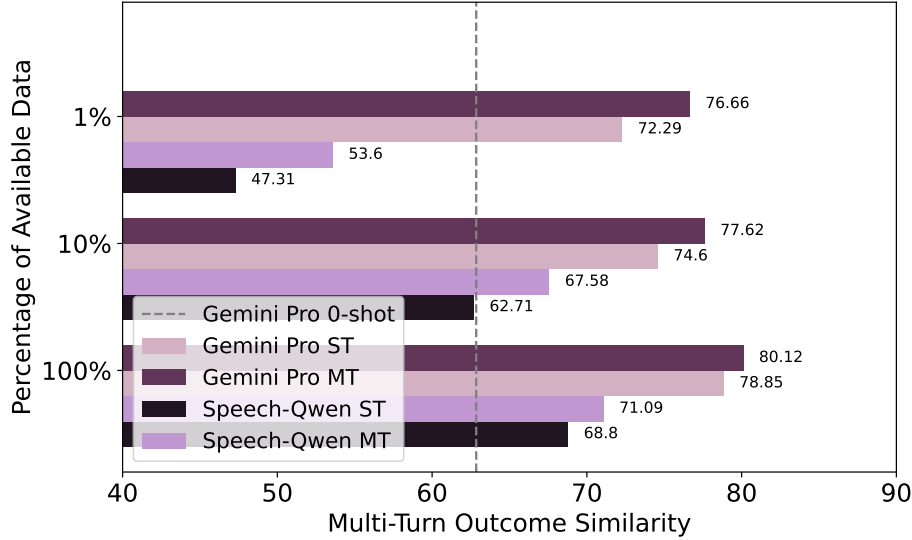


Figure 5.4: Multi-task (MT) learning improves upon Single-task (ST) fine-tuning with both Gemini and Speech-Qwen on ASK-QA’s multi-turn evaluation.

### 5.6.3 SD-QA: Textual Knowledge and Spoken Questions

We examine the setting where the single-turn QA context is provided in the recorded speech, and the knowledge necessary to answer the question correctly is provided in the text.

**Dataset:** SD-QA (Faisal et al., 2021) is a large single-turn SQA benchmark with diverse data – spanning 5 languages (Arabic, Bengali, English, Kiswahili, and Korean) and 24 regional dialects. SD-QA is also proposed as a span-based QA task, but we apply our end-to-end generative approach as in Section 5.6. We tune our models on up to 9,008 of the 10,0008 samples made available for training, withholding the remaining samples for validation. We evaluate our approach on the 12,975 evaluation samples.

**Findings:** We evaluate performance on SD-QA in terms of exact match and token-based F1. Consistent with our findings in Section 5.6, we see that our multi-task approach is able to outperform single-task tuning in all evaluation settings. This is inclusive of experiments with Gemini Pro as the base MLLM for tuning. We see a large 16.13% relative improvement (46.33 to 55.24) for

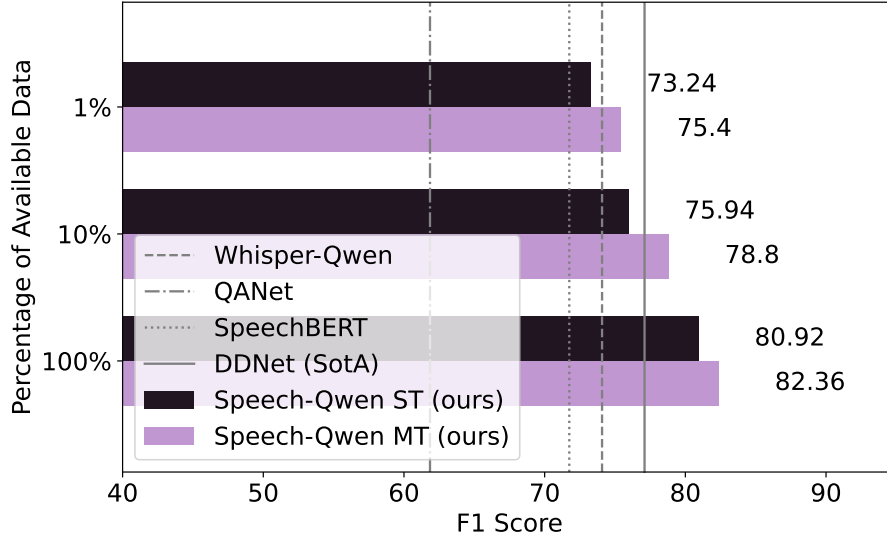


Figure 5.5: Our multi-task approach applied to Speech-Qwen outperforms the state-of-the-art approach on Spoken-SQuAD using only 10% of the available data.

exact match in the limited data setting with Gemini.

We observe that Gemini Pro is already a strong base MLLM, achieving competitive zero-shot performance on this corpus. This is likely due to it having a strong initialization on multilingual ASR. We see that our Speech-Qwen model is able to outperform zero-shot Gemini using our multi-task approach with full data. We also observe up to a 52.8% relative improvement over single-task tuning with Speech-Qwen in the 1% data regime. This is consistent with findings from Chen and Yu (2023) in which pre-finetuning yields strong improvements in the extremely limited data regime. Overall, the particularly large performance improvements on this particular corpus may be an indication that the base models have not been trained on as much multi-lingual data.

#### 5.6.4 Ablation Studies

In Table 5.12, we systematically remove each individual task: Dialogue Listening Comprehension (DLC), Story Listening Comprehension (SLC), and Response Selection (RS). The removal of each auxiliary task results in performance degradation relative to full multi-task tuning, indicating their importance towards improved cross-modal understanding. We observe that removing SLC

Base Model	App.	Data	EM $\uparrow$	F1 $\uparrow$
Gemini Pro	–	0%	42.44	64.18
Gemini Pro	ST	1%	46.33	67.74
Gemini Pro	MT	1%	<b>55.24</b>	<b>70.73</b>
Gemini Pro	ST	10%	62.79	77.80
Gemini Pro	MT	10%	<b>63.04</b>	<b>79.02</b>
Gemini Pro	ST	100%	63.10	78.15
Gemini Pro	MT	100%	<b>64.17</b>	<b>79.06</b>
Speech-Qwen	ST	1%	13.44	25.28
Speech-Qwen	MT	1%	<b>24.70</b>	<b>38.63</b>
Speech-Qwen	ST	10%	29.70	43.84
Speech-Qwen	MT	10%	<b>39.92</b>	<b>54.35</b>
Speech-Qwen	ST	100%	46.83	61.76
Speech-Qwen	MT	100%	<b>49.54</b>	<b>64.94</b>

Table 5.11: **Experimental results comparing single-task SFT and our proposed multi-task approach on SD-QA’s test set.**

Approach	Data	Single-Turn Sim. $\uparrow$	Multi-Turn Sim. $\uparrow$
Speech-Qwen MT w/o DLC	1%	53.77	53.10
Speech-Qwen MT w/o SLC	1%	52.32	51.89
Speech-Qwen MT w/o RS	1%	53.53	52.67
Speech-Qwen MT	1%	54.54	53.60
Speech-Qwen MT w/o DLC	10%	65.09	64.19
Speech-Qwen MT w/o SLC	10%	64.75	64.24
Speech-Qwen MT w/o RS	10%	66.89	66.01
Speech-Qwen MT	10%	68.27	67.58

Table 5.12: Systematic ablations of each individual task type on ASK-QA.

results in the most performance degradation, which follows the intuition in Figure 5.1 and Li et al. (2018).

## 5.7 Limitations

**Transcription supervision:** One of the crucial auxiliary tasks in our approach is listening comprehension, as demonstrated by the performance degradation in our ablations (Table 5.12). In our implementation, we use ground-truth transcriptions as the target for this generation task. These transcriptions may not be available — for instance, the ones provided by Spoken-SQuAD and SD-

QA were obtained via ASR. Our transcription for ASK-QA is not guaranteed to perfectly match the generated speech either, despite our efforts to filter the data quality (see Section 5.5). It is not clear whether the possibility of slight transcription errors improves model robustness to noise or degrades performance, and this warrants further study in future work.

**TTS quality:** Our data generation approach is bottlenecked by current capabilities of TTS software. While TTS has greatly improved in recent years in terms of WER, we do still witness generation errors and naturalness issues when working with long context (hence the need for filtering). We are also not at the point in which we have perfect controllability in paralinguistic attributes.

**Generalization to paralinguistic tasks:** We propose a multi-task approach which can be used to greatly improve performance in SQA. In the three corpora here, listening comprehension proves to be crucial as the primary objective is auditory semantic understanding. However, in more nuanced contexts like task guidance (Schlager and Feiner, 2024), it is more important to monitor different paralinguistic aspects of the user such as frustration.

**Use in large-scale model post-training:** We believe that our overall data generation process can be useful for improving MLLM post-training. However, verifying this claim is beyond the scope of this work due to computational constraints. We see improved performance on our downstream task after supervised fine-tuning of Gemini, which does indicate positive signal that there are correlations between our training and evaluation data.

## 5.8 Summary

In this Chapter, we designed ASK-QA, a novel corpus that highlights the challenges of multimodal spoken conversation modeling. In particular, not only does ASK-QA require the ability to perform implicit action optimization, but it involves being able to model a diverse set of users’ voices and speaking patterns. As part of our work, we propose a data-centric multi-task learning approach which helps improve speech data utilization for MLLM tuning. Tuning on various corpora with Gemini and open-weight MLLMs, we observe consistent performance improvements



regardless of model scale and tuning access, surpassing state-of-the-art performance on Spoken-SQuAD with open-weight MLLMs. Our tuning strategy indicates that the design of auxiliary tasks can help facilitate optimal utilization of training data, regardless of the amount that is available. Our improvements demonstrated with Gemini demonstrate that such quality improvements are even applicable at large model scales. Future work may build upon our insights by designing new auxiliary tasks, incorporating more expressive TTS approaches (e.g., emotion modeling), or examining more complex action optimization strategies for ASK-QA (e.g., reinforcement learning). Our dataset and synthesis process can also be contributed to post-training data mixtures to improve construction of MLLMs' for improved long-context speech modeling abilities. The synthesis process combined with the multi-task manipulation procedure could potentially be used to construct large-scale diverse data generation to improve the audio understanding capabilities of MLLMs.

## Conclusion

Given the rapid progress in modern conversational AI (e.g. Gemini, ChatGPT, etc.), it is worthwhile to consider the remaining open challenges in developing conversational intelligence. In past years, coherent neural language generation was a difficult challenge (Roller et al., 2021), but with the rise of high-capacity LLMs, this issue of coherence has largely been addressed (Brown et al., 2020), with many such systems shifting from modular frameworks (Yu, 2017) to singular end-to-end models (Chen et al., 2025). However, despite these advances, conversational AI is not “solved.” Rather, this requires refocusing on the remaining challenges that make such generative models useful for society at large.

Even in the “assistant” paradigm, LLM-based systems still lack human-level conversational skills, whether it be the ability to handle ambiguity or express genuine empathy. This dissertation identified the core challenges in developing *mixed-initiative* conversational intelligence, to bridge the existing gaps preventing current conversational AI from functioning initiative-taking agents which can take on larger assistive roles in society. The research presented here proposes novel solutions to addressing three fundamental obstacles hindering the development of truly intelligent and adaptive conversational agents: the scarcity of high-quality, in-domain conversational data; the complexities of optimizing pragmatic action planning in multi-turn interactions; and the limitations in understanding and integrating multimodal signals, especially in spoken dialogue.

To address *data scarcity*, Chapter 2 introduced PLACES, a novel framework for synthesizing high-quality, topic-conditioned dyadic and multiparty conversations from a small set of expert examples. The results demonstrated that these synthetic datasets are comparable in quality to

human-collected data and effective for fine-tuning dialogue models, opening avenues for systematic construction resources for understudied areas.

Chapters 3 and 4 investigated *pragmatic action planning* from two different perspectives. Chapter 3 focused on explicit planning within modular systems, proposing *ACT*. This approach utilizes prompt-based Monte-Carlo Tree Search with LLMs to perform look-ahead planning without requiring specific training data, proving effective in complex, subjective tasks like persuasion. Chapter 4 addressed some of the limitations of modern explicit action planning and proposed implicit action planning for end-to-end systems. We introduced a novel algorithm called Action-Based Contrastive Self-Training, building on modern frameworks for LLM alignment. ACT adapts LLMs to mixed-initiative contexts even given limited data, such as handling ambiguity in conversational question answering across various domains (e.g. tabular QA, machine reading comprehension, or text-to-SQL), by balancing action selection with multi-turn optimization in a quasi-online contrastive learning framework.

Finally, Chapter 5 unified the notion of implicit action recognition in the multimodal realm with large-scale user-specific speech modeling, proposing several solutions to improving end-to-end spoken dialogue systems. Recognizing the unique challenges of speech (which include variability in different speakers’ behavior), this work introduced ASK-QA, the first mixed-initiative spoken conversational question-answering dataset which features contextual ambiguity, long audio contexts, and carefully constructed perturbations to speaking style. Furthermore, we proposed a simple yet effective data-centric multi-task learning approach which demonstrated significant improvements in MLLM robustness and performance on spoken QA tasks, even for state-of-the-art models and particularly in limited data settings.

Given these identified challenges and the work described in this dissertation sets the foundation for a number of future directions.

**Augmenting agentic workflows with mixed-initiative intelligence.** Many recent efforts focus on developing fully autonomous agentic workflows (Yu et al., 2024a; Koh et al., 2024). However,

as described in Chapters 3 and 4, it may actually be more important that agents recognize when they need to elicit help from potential experts whether it be the users or expert models. There are scenarios in which experts may possess knowledge that is not explicitly provided to the model a priori which may lead to significant safety risks (e.g., an agent could accidentally access sensitive information or interact with malicious software). Such efforts would require generalizing the proposed methods in Chapters 3 and 4 to generic settings beyond specific tasks.

**Inference-time strategy distillation.** Many recent efforts towards improving agentic workflows focus on trading inference-time computation costs for novel capabilities (e.g. Yao et al. (2023)). As indicated in Chapter 3, such efforts are not necessarily realistic in real-world applications where latency is crucial. It remains an open problem to re-examine how to absorb some of these capabilities during training time where high computational costs are more manageable. In particular, it is worth studying how each of these trajectory modeling approaches work and in what scenarios they are optimal. Future work can consider ACT-style learning as in Chapter 4 as a framework to distill these types of strategies.

**Multi-modal user modeling.** This dissertation touches on the importance of improving agents’ ability to model users across different modalities. However, an especially important forward-looking direction is investigating how to leverage different modalities to model user signals. There exist many current modeling challenges with non-textual modalities (e.g., Chapter 5 describes the long-context limitations in audio modeling), but beyond common modalities such as vision or speech, there are many untapped modalities. These include ones accessible by wearable technologies such as eye gaze or heart rate. These signals are likely highly correlated with user sentiment. The development of an agent which can reason about phenomena such as emotions from heart rate on a smartwatch has the potential to be far more ubiquitous and useful than a standard unimodal agent.

**Multi-surface user modeling.** As we have seen throughout this dissertation, there are already several challenges in conversational agent development that arise from the multi-turn nature of conversations. Multi-turn historical consistency can pose a complex modeling challenge, and these challenges remain largely explored in multi-surface modeling. Beyond the development of single agent on a single surface (e.g., a chatbot on a single webpage), unified agents that can reason about user history across multiple surfaces (e.g. web and mobile) largely remain explored. The selection of different mixed-initiative action strategies should vary across individual users, and the understanding of these types of preferences requires the ability to model many different types of interaction surfaces.

There remain many open challenges before AI systems are truly conversationally intelligent and able to reach their full potential in society. The contributions of this dissertation serve as the foundation for further research towards the efficient development of highly capable human-centric agents.

## References

- Deng, Yang et al. (2022). “PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.
- Passali, Tatiana et al. (June 2022). “LARD: Large-scale Artificial Disfluency Generation”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 2327–2336.
- Poria, Soujanya et al. (2019). “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536.
- Shaikh, Samira et al. (2010). “MPC: A Multi-Party Chat Corpus for Modeling Social Phenomena in Discourse.” In: *LREC*.
- Sperber, Dan (1995). “How do we communicate”. In: *How things are: A science toolkit for the mind*, pp. 191–199.
- Yu, Zhou (2017). “Situated intelligent interactive systems”. PhD thesis. Carnegie Mellon University.
- Zhang, Jianguo et al. (2024). “DialogStudio: Towards Richest and Most Diverse Unified Dataset Collection for Conversational AI”. In: *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 2299–2315.
- Huang, Minlie, Xiaoyan Zhu, and Jianfeng Gao (2020). “Challenges in building intelligent open-domain dialog systems”. In: *ACM Transactions on Information Systems (TOIS)* 38.3, pp. 1–32.
- Budzianowski, Paweł et al. (2018b). “MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026.
- Rastogi, Abhinav et al. (2020). “Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 8689–8696.
- Zuo, Lei et al. (2021). “AllWOZ: Towards multilingual task-oriented dialog systems for all”. In: *arXiv preprint arXiv:2112.08333*.

- Li, Yu et al. (2020b). “End-to-end trainable non-collaborative dialog system”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8293–8302.
- Chen, Yun-Nung, Asli Celikyilmaz, and Dilek Hakkani-Tur (2017). “Deep learning for dialogue systems”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 8–14.
- Chen, Hongshen et al. (2017). “A survey on dialogue systems: Recent advances and new frontiers”. In: *Acm Sigkdd Explorations Newsletter* 19.2, pp. 25–35.
- Kaplan, Jared et al. (2020). “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361*.
- Brown, Tom et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901.
- Chen, Maximillian et al. (July 2023a). “Controllable Mixed-Initiative Dialogue Generation through Prompting”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 951–966.
- Allen, James E, Curry I Guinn, and Eric Horvitz (1999). “Mixed-initiative interaction”. In: *IEEE Intelligent Systems and their Applications* 14.5, pp. 14–23.
- Chu-Carroll, Jennifer (2000). “MIMIC: An adaptive mixed initiative spoken dialogue system for information queries”. In: *Sixth Applied Natural Language Processing Conference*, pp. 97–104.
- Deng, Yang et al. (2023d). “Rethinking Conversational Agents in the Era of LLMs: Proactivity, Non-collaborativity, and Beyond”. In: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 298–301.
- Chen, Maximillian et al. (2022b). “Weakly Supervised Data Augmentation Through Prompting for Dialogue Understanding”. In: *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.
- Chen, Maximillian et al. (2023b). “PLACES: Prompting Language Models for Social Conversation Synthesis”. In: *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 844–868.
- Qian, Kun (2024). *Constructing Task-Oriented Dialogue Systems with Limited Resources*. Columbia University.
- Longpre, Shayne et al. (2024a). “Bridging the Data Provenance Gap Across Text, Speech and Video”. In: *ArXiv abs/2412.17847*.

- Longpre, Shayne et al. (2024b). “Consent in Crisis: The Rapid Decline of the AI Data Commons”. In: *ArXiv abs/2407.14933*.
- Gupta, Prakhar et al. (2022). “InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning”. In: *EMNLP*.
- Qian, Kun et al. (2025). “Bottom-Up Synthesis of Knowledge-Grounded Task-Oriented Dialogues with Iteratively Self-Refined Prompts”. In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Kim, Hyunwoo et al. (2023). “SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization”. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Peng, Baolin et al. (2022). “GODEL: Large-Scale Pre-Training for Goal-Directed Dialog”. In: *arXiv preprint arXiv:2206.11309*.
- Roller, Stephen et al. (Apr. 2021). “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 300–325.
- Shuster, Kurt et al. (2022). “BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage”. In: *arXiv preprint arXiv:2208.03188*.
- Qian, Kun et al. (2021). “Annotation Inconsistency and Entity Bias in MultiWOZ”. In: *SIGDIAL Conferences*.
- Yu, Xiao, Maximillian Chen, and Zhou Yu (2023). “Prompt-Based Monte-Carlo Tree Search for Goal-oriented Dialogue Policy Planning”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7101–7125.
- Deng, Yang et al. (2023b). “Plug-and-play policy planner for large language model powered dialogue agents”. In: *The Twelfth International Conference on Learning Representations*.
- Yu, Xiao et al. (2022). “KRLS: Improving End-to-End Response Generation in Task Oriented Dialog with Reinforced Keywords Learning”. In: *Conference on Empirical Methods in Natural Language Processing*.
- Deng, Yang et al. (2023a). “Goal awareness for conversational AI: proactivity, non-collaborativity, and beyond”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pp. 1–10.
- Deng, Yang et al. (2023c). “Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10602–10621.



- Lewis, Mike et al. (2017a). “Deal or No Deal? End-to-End Learning of Negotiation Dialogues”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453.
- Wang, Xuewei et al. (July 2019). “Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5635–5649.
- Guo, Meiqi et al. (2021). “Abg-coqa: Clarifying ambiguity in conversational question answering”. In: *3rd Conference on Automated Knowledge Base Construction*.
- Chen, Maximillian et al. (2025). “Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training”. In: *International Conference on Learning Representations*.
- Yu, Dian and Zhou Yu (2021). “MIDAS: A Dialog Act Annotation Scheme for Open Domain HumanMachine Spoken Conversations”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.
- Chen, Maximillian et al. (2022a). “Seamlessly Integrating Factual Information and Social Content with Persuasive Dialogue”. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 399–413.
- Keskar, Nitish Shirish et al. (2019). “Ctrl: A conditional transformer language model for controllable generation”. In: *arXiv preprint arXiv:1909.05858*.
- Yu, Zhou et al. (2016). “Chatbot evaluation and database expansion via crowdsourcing”. In: *Proceedings of the chatbot workshop of LREC*. Vol. 63, p. 102.
- Cheng, Yi et al. (Dec. 2022a). “Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3014–3026.
- Väth, Dirk, Lindsey Vanderlyn, and Ngoc Thang Vu (May 2023a). “Conversational Tree Search: A New Hybrid Dialog Task”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1264–1280.
- Wang, Sihan et al. (2020a). “Task-completion dialogue policy learning via Monte Carlo tree search with dueling network”. In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 3461–3471.

- Jang, Youngsoo, Jongmin Lee, and Kee-Eung Kim (2020). “Bayes-adaptive monte-carlo planning and learning for goal-oriented dialogues”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 7994–8001.
- Yang, Jingxuan, Si Li, and Jun Guo (2021a). “Multi-turn target-guided topic prediction with Monte Carlo tree search”. In: *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 324–334.
- Christiano, Paul F et al. (2017). “Deep reinforcement learning from human preferences”. In: *Advances in neural information processing systems* 30.
- Ouyang, Long et al. (2022). “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35, pp. 27730–27744.
- Shani, Lior et al. (2024). “Multi-turn Reinforcement Learning from Preference Human Feedback”. In: *arXiv preprint arXiv:2405.14655*.
- Ahmadian, Arash et al. (2024). “Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms”. In: *arXiv preprint arXiv:2402.14740*.
- Deng, Yang et al. (2024a). “Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents”. In: *International Conference on Learning Representations*.
- Hirschberg, Julia (1993). “Pitch accent in context predicting intonational prominence from text”. In: *Artificial Intelligence* 63.1-2, pp. 305–340.
- Busso, Carlos et al. (2016). “MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception”. In: *IEEE Transactions on Affective Computing* 8.1, pp. 67–80.
- Busso, Carlos et al. (2008). “IEMOCAP: Interactive emotional dyadic motion capture database”. In: *Language resources and evaluation* 42, pp. 335–359.
- Li, Peizhao et al. (2023). “Uniar: Unifying human attention and response prediction on visual content”. In: *arXiv preprint arXiv:2312.10175* 2.
- Wang, Xin et al. (2023a). “Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281.
- Liu, Haotian et al. (2024a). “Visual instruction tuning”. In: *Advances in neural information processing systems* 36.
- Chu, Yunfei et al. (2023). “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models”. In: *arXiv preprint arXiv:2311.07919*.

- Chu, Yunfei et al. (2024). “Qwen2-audio technical report”. In: *arXiv preprint arXiv:2407.10759*.
- Ma, Ziyang et al. (2024). “An Embarrassingly Simple Approach for LLM with Strong ASR Capacity”. In: *arXiv preprint arXiv:2402.08846*.
- Tang, Changli et al. (n.d.). “SALMONN: Towards Generic Hearing Abilities for Large Language Models”. In: *The Twelfth International Conference on Learning Representations*.
- Watanabe, Shinji et al. (2018). “ESPnet: End-to-End Speech Processing Toolkit”. In: *INTER-SPEECH 2018, Hyderabad, India*.
- Ao, Junyi et al. (May 2022). “SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5723–5738.
- Hung, Chia-Yu et al. (2024). “TangoFlux: Super Fast and Faithful Text to Audio Generation with Flow Matching and Clap-Ranked Preference Optimization”. In: *arXiv preprint arXiv:2412.21037*.
- Liang, Youwei et al. (2024). “Rich human feedback for text-to-image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411.
- Chen, Maximillian and Zhou Yu (2023). “Pre-Finetuning for Few-Shot Emotional Speech Recognition”. In: *INTERSPEECH*. INTERSPEECH.
- Fu, Jinlan et al. (2025). “CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs”. In: *arXiv preprint arXiv:2501.16629*.
- Wang, Fei et al. (2024). “mDPO: Conditional Preference Optimization for Multimodal Large Language Models”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088.
- Levitan, Rivka et al. (2012). “Acoustic-prosodic entrainment and social behavior”. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pp. 11–19.
- Zhao, Yuxiang and Qinghua Zhu (2014). “Evaluation on crowdsourcing research: Current status and future direction”. In: *Information Systems Frontiers* 16.3, pp. 417–434.
- Allahbakhsh, Mohammad et al. (2013). “Quality Control in Crowdsourcing Systems: Issues and Directions”. In: *IEEE Internet Computing* 17.2, pp. 76–81.
- Qian, Kun et al. (2023). “User Adaptive Language Learning Chatbots with a Curriculum”. In: *ArXiv abs/2304.05489*.

- Zhao, Xueliang et al. (2019). “Low-Resource Knowledge-Grounded Dialogue Generation”. In: *International Conference on Learning Representations*.
- Mi, Fei et al. (2019). “Meta-learning for low-resource natural language generation in task-oriented dialogue systems”. In: *arXiv preprint arXiv:1905.05644*.
- Qian, Kun and Zhou Yu (2019). “Domain Adaptive Dialog Generation via Meta Learning”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2639–2649.
- Li, Juntao et al. (2019). “Insufficient data can also rock! learning to converse using smaller data with augmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 6698–6705.
- Kulhánek, Jonáš et al. (2021). “AuGPT: Auxiliary Tasks and Data Augmentation for End-To-End Dialogue with Pre-Trained Language Models”. In: *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pp. 198–210.
- Mehri, Shikib, Yasemin Altun, and Maxine Eskenazi (2022). *LAD: Language Models as Data for Zero-Shot Dialog*.
- Sahu, Gaurav et al. (2022). “Data Augmentation for Intent Classification with Off-the-shelf Large Language Models”. In: *arXiv preprint arXiv:2204.01959*.
- Kim, Yekyung, Seohyeong Jeong, and Kyunghyun Cho (2021). “LINDA: Unsupervised Learning to Interpolate in Natural Language Processing”. In: *arXiv preprint arXiv:2112.13969*.
- Gao, Leo et al. (2020). “The pile: An 800gb dataset of diverse text for language modeling”. In: *arXiv preprint arXiv:2101.00027*.
- Li, Yanran et al. (2017b). “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995.
- Gopalakrishnan, Karthik et al. (2019). “Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations.” In: *INTERSPEECH*, pp. 1891–1895.
- Chen, Derek, Kun Qian, and Zhou Yu (2023). “Stabilized In-Context Learning with Pre-trained Language Models for Few Shot Dialogue State Tracking”. In: *ArXiv abs/2302.05932*.
- Raffel, Colin et al. (2020). “Exploring the limits of transfer learning with a unified text-to-text transformer.” In: *J. Mach. Learn. Res.* 21.140, pp. 1–67.
- Wang, Ben and Aran Komatsuzaki (May 2021). *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. <https://github.com/kingoflolz/mesh-transformer-jax>.

- Zhang, Susan et al. (2022c). “Opt: Open pre-trained transformer language models”. In: *arXiv preprint arXiv:2205.01068*.
- Madotto, Andrea et al. (2021). “Few-Shot Bot: Prompt-Based Learning for Dialogue Systems”. In: *arXiv preprint arXiv:2110.08118*.
- Min, Sewon et al. (2022). “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?” In: *EMNLP*.
- Min, Sewon et al. (2021). “Metaicl: Learning to learn in context”. In: *arXiv preprint arXiv:2110.15943*.
- Lu, Yao et al. (2021b). “Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity”. In: *arXiv preprint arXiv:2104.08786*.
- Wang, Yufei et al. (2022). “PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4242–4255.
- Rosenbaum, Andy et al. (2022a). “CLASP: Few-Shot Cross-Lingual Data Augmentation for Semantic Parsing”. In: *arXiv preprint arXiv:2210.07074*.
- Meng, Yu et al. (2022). “Generating Training Data with Language Models: Towards Zero-Shot Language Understanding”. In: *Advances in Neural Information Processing Systems*.
- Papangelis, Alexandros et al. (2021). “Generative Conversational Networks”. In: *SIGDIAL*.
- Zhang, Rongsheng et al. (2020c). “Dialogue Distillation: Open-domain Dialogue Augmentation Using Unpaired Data”. In: *ArXiv abs/2009.09427*.
- Zhang, Houyu et al. (2020b). “Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs”. In: *ACL*.
- Qian, Kun, Wei Wei, and Zhou Yu (2021). “A Student-Teacher Architecture for Dialog Domain Adaptation under the Meta-Learning Setting”. In: *AAAI Conference on Artificial Intelligence*.
- Rosenbaum, Andy et al. (2022b). “Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging”. In: *arXiv preprint arXiv:2209.09900*.
- Aher, Gati, Rosa I Arriaga, and Adam Tauman Kalai (2022). “Using Large Language Models to Simulate Multiple Humans”. In: *arXiv preprint arXiv:2208.10264*.
- Bae, Sanghwan et al. (July 2022a). “Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 2128–2150.
- Kim, Hyunwoo et al. (2022). “SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization”. In: *arXiv preprint arXiv:2212.10465*.
- Xu, Jing et al. (2022). *Learning New Skills after Deployment: Improving open-domain internet-driven dialogue with human feedback*.
- Li, Yu et al. (2021). “LEGOEval: An Open-Source Toolkit for Dialogue System Evaluation via Crowdsourcing”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 317–324.
- Wu, Qingyang, Lei Li, and Zhou Yu (2021). “Textgail: Generative adversarial imitation learning for text generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 16, pp. 14067–14075.
- Li, Jiwei et al. (2016). “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119.
- Smith, Eric Michael et al. (2020). “Can You Put it All Together: Evaluating Conversational Agents’ Ability to Blend Skills”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030.
- Rashkin, Hannah et al. (2019). “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381.
- Engelhardt, Paul E, Karl GD Bailey, and Fernanda Ferreira (2006). “Do speakers and listeners observe the Gricean Maxim of Quantity?” In: *Journal of memory and language* 54.4, pp. 554–573.
- Chen, Yi-Ting, Hen-Hsen Huang, and Hsin-Hsi Chen (2020). “MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 610–614.
- Ju, Dongshi et al. (Oct. 2022). “Learning to Improve Persona Consistency in Multi-party Dialogue Generation via Text Knowledge Enhancement”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 298–309.
- Li, Aaron W et al. (2020a). “ALOHA: Artificial learning of human attributes for dialogue agents”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8155–8163.

- Zhu, Ling.Yu et al. (May 2022). “Multi-Party Empathetic Dialogue Generation: A New Task for Dialog Systems”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 298–307.
- Lowe, Ryan et al. (2015). “The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems”. In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294.
- Baumgartner, Jason et al. (2020). “The pushshift reddit dataset”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 830–839.
- Christensen, Heidi et al. (2010). “The CHiME corpus: a resource and a challenge for computational hearing in multisource environments”. In: *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer.
- Song, Haoyu et al. (July 2020a). “Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5821–5831.
- Welleck, Sean et al. (July 2019). “Dialogue Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3731–3741.
- Song, Haoyu et al. (2020b). “Generating persona consistent dialogues by exploiting natural language inference”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8878–8885.
- Kang, Minki et al. (2022). “Knowledge-Consistent Dialogue Generation with Knowledge Graphs”. In: *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.
- Kim, Byeongchang et al. (2021). “How Robust are Fact Checking Systems on Colloquial Claims?”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1535–1548.
- Dinan, Emily et al. (Nov. 2019). “Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 4537–4546.
- Lu, Ximing et al. (June 2021a). “NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints”. In: *Proceedings of the 2021 Conference of the North American*

can Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, pp. 4288–4299.

Zhang, Saizheng et al. (2018). “Personalizing Dialogue Agents: I have a dog, do you have pets too?” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213.

Liu, Pengfei et al. (2021a). “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. In: *arXiv preprint arXiv:2107.13586*.

Luccioni, Alexandra Sasha and Joseph D Viviano (2021). “What’s in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus”. In: *arXiv preprint arXiv:2105.02732*.

Wolf, Thomas et al. (2020a). “Transformers: State-of-the-art natural language processing”. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.

Paszke, Adam et al. (2019). “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32.

Lewis, Mike et al. (2017b). “Deal or No Deal? End-to-End Learning of Negotiation Dialogues”. In: *Conference on Empirical Methods in Natural Language Processing*.

Liu, Siyang et al. (2021b). “Towards Emotional Support Dialog Systems”. In: *Proceedings of the 59th annual meeting of the Association for Computational Linguistics*.

Levin, Esther, Roberto Pieraccini, and Wieland Eckert (1997). “Learning dialogue strategies within the markov decision process framework”. In: *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, pp. 72–79.

Zhang, Zheng et al. (2020d). “Learning Goal-oriented Dialogue Policy with opposite Agent Awareness”. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 122–132.

Liu, Bing and Ian Lane (2017). “Iterative policy learning in end-to-end trainable task-oriented neural dialog models”. In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 482–489.

Liu, Bing et al. (2018). “Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2060–2069.



- Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL].
- Liu, Yinhan et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL].
- Zhao, Tiancheng and Maxine Eskenazi (2018). “Zero-Shot Dialog Generation with Cross-Domain Latent Actions”. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 1–10.
- Mehri, Shikib and Maxine Eskenazi (2021). “Schema-Guided Paradigm for Zero-Shot Dialog”. In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 499–508.
- Liu, Yiren and Halil Kilicoglu (2023). “Commonsense-Aware Prompting for Controllable Empathetic Dialogue Generation”. In: *arXiv preprint arXiv:2302.01441*.
- Liu, Zihan et al. (2022). “Multi-Stage Prompting for Knowledgeable Dialogue Generation”. In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1317–1337.
- Bae, Sanghwan et al. (2022b). “Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2128–2150.
- Yang, Yuting et al. (2022). “Prompt Learning for Few-Shot Dialogue State Tracking”. In: *arXiv preprint arXiv:2201.05780*.
- Zhang, Haodi et al. (2022a). “Efficient Dialog Policy Learning by Reasoning with Contextual Knowledge”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36, pp. 11667–11675.
- Zhang, Shuo et al. (2022b). ““Think Before You Speak”: Improving Multi-Action Dialog Policy by Planning Single-Action Dialogs”. In: *arXiv preprint arXiv:2204.11481*.
- Cao, Yan et al. (July 2020). “Adaptive Dialog Policy Learning with Hindsight and User Modeling”. In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, pp. 329–338.
- Peng, Baolin et al. (July 2018a). “Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2182–2192.

- Zhang, Qiang, Jason Naradowsky, and Yusuke Miyao (2023). *Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models*. arXiv: 2305.17878 [cs.CL].
- Cheng, Yi et al. (Dec. 2022b). “Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3014–3026.
- Wang, Sihan et al. (Nov. 2020b). “Task-Completion Dialogue Policy Learning via Monte Carlo Tree Search with Dueling Network”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3461–3471.
- Yang, Jingxuan, Si Li, and Jun Guo (Dec. 2021b). “Multi-Turn Target-Guided Topic Prediction with Monte Carlo Tree Search”. In: *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLP AI), pp. 324–334.
- Shi, Weiyan et al. (Nov. 2019). “How to Build User Simulators to Train RL-based Dialog Systems”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 1990–2000.
- Silver, David et al. (2016). “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529.7587, pp. 484–489.
- Silver, David et al. (2017). “Mastering the game of Go without human knowledge”. In: *Nature* 550.7676, pp. 354–359.
- Howard, Ronald A (1960). “Dynamic programming and markov processes.” In.
- Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.
- Perez Liebana, Diego et al. (2015). “Open Loop Search for General Video Game Playing”. In: *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*. GECCO ’15. Madrid, Spain: Association for Computing Machinery, 337–344. ISBN: 9781450334723.
- Weber, Richard (2010). “Optimization and control”. In: *University of Cambridge*.
- Rosin, Christopher D (2011). “Multi-armed bandits with episode context”. In: *Annals of Mathematics and Artificial Intelligence* 61.3, pp. 203–230.

- OpenAI (2022). *OpenAI: Introducing ChatGPT*.
- Liu, Yiheng et al. (2023c). *Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models*. arXiv: 2304.01852 [cs.CL].
- Guo, Biyang et al. (2023). *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection*. arXiv: 2301.07597 [cs.CL].
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*. arXiv: 2303.15056 [cs.CL].
- He, Xingwei et al. (2023). “AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators”. In: *arXiv preprint arXiv:2303.16854*.
- Pan, Alexander et al. (2023). “Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark”. In: *arXiv preprint arXiv:2304.03279*.
- Wang, Shuohang et al. (2021). “Want to reduce labeling cost? GPT-3 can help”. In: *arXiv preprint arXiv:2108.13487*.
- Liu, Yang et al. (2023b). “GPTEval: NLG Evaluation using GPT-4 with Better Human Alignment”. In: *arXiv preprint arXiv:2303.16634*.
- Chen, Mark et al. (2021). *Evaluating Large Language Models Trained on Code*. arXiv: 2107.03374 [cs.LG].
- He, Wanwei et al. (2022). “Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10, pp. 10749–10757.
- He, He et al. (2018). *Decoupling Strategy and Generation in Negotiation Dialogues*. arXiv: 1808.09637 [cs.CL].
- Budzianowski, Paweł et al. (2018a). “MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhang, Cong et al. (2020a). “A Hierarchy MCTS Algorithm for The Automated PCB Routing”. In: *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pp. 1366–1371.
- Chaslot, Guillaume MJ B, Mark HM Winands, and H Jaap van Den Herik (2008). “Parallel monte-carlo tree search”. In: *Computers and Games: 6th International Conference, CG 2008, Beijing, China, September 29-October 1, 2008. Proceedings 6*. Springer, pp. 60–71.

- Wang, Yizhong et al. (2023c). *How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources*. arXiv: 2306.04751 [cs.CL].
- Snell, Charlie et al. (2024). “Scaling llm test-time compute optimally can be more effective than scaling model parameters”. In: *arXiv preprint arXiv:2408.03314*.
- Amershi, Saleema et al. (2019). “Guidelines for human-AI interaction”. In: *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13.
- Lemon, Oliver (2012). “Conversational interfaces”. In: *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*. Springer, pp. 1–4.
- Gemini Team et al. (2023). “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805*.
- Zipf, George Kingsley (1949). *Human behavior and the principle of least effort*.
- Messer, Donald K (1980). “Six Common Causes of Ambiguity.” In: *Technical Writing Teacher* 7.2, pp. 50–52.
- Beers, Pieter J et al. (2006). “Common ground, complex problems and decision making”. In: *Group decision and negotiation* 15, pp. 529–556.
- Liu, Alisa et al. (2023a). “We’re Afraid Language Models Aren’t Modeling Ambiguity”. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Wang, Xingyao et al. (2023b). “MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback”. In: *The Twelfth International Conference on Learning Representations*.
- Bai, Yuntao et al. (2022). “Training a helpful and harmless assistant with reinforcement learning from human feedback”. In: *arXiv preprint arXiv:2204.05862*.
- Dong, Guanting et al. (2023). “How abilities in large language models are affected by supervised fine-tuning data composition”. In: *arXiv preprint arXiv:2310.05492*.
- Rafailov, Rafael et al. (2024). “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems* 36.
- Qian, Kun et al. (2022). “Database Search Results Disambiguation for Task-Oriented Dialog Systems”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1158–1173.
- Wen, Tsung-Hsien et al. (2016). “Conditional generation and snapshot learning in neural dialogue systems”. In: *arXiv preprint arXiv:1606.03352*.

- Sadek, M David (1991). “Dialogue acts are rational plans”. In: *The Structure of Multimodal Dialogue; Second VENACO Workshop*.
- Stolcke, Andreas et al. (2000). “Dialogue act modeling for automatic tagging and recognition of conversational speech”. In: *Computational linguistics* 26.3, pp. 339–373.
- Wu, Qingyang et al. (2023a). “DiactTOD: Learning Generalizable Latent Dialogue Acts for Controllable Task-Oriented Dialogue Systems”. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 255–267.
- Väth, Dirk, Lindsey Vanderlyn, and Ngoc Thang Vu (2023b). “Conversational Tree Search: A New Hybrid Dialog Task”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1264–1280.
- Wang, Sihan et al. (Nov. 2020c). “Task-Completion Dialogue Policy Learning via Monte Carlo Tree Search with Dueling Network”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 3461–3471.
- Tunstall, Lewis et al. (2023). “Zephyr: Direct distillation of lm alignment”. In: *arXiv preprint arXiv:2310.16944*.
- Lee, Harrison et al. (2023). “Rlaif: Scaling reinforcement learning from human feedback with ai feedback”. In: *arXiv preprint arXiv:2309.00267*.
- Zhao, Yao et al. (2023). “Slic-hf: Sequence likelihood calibration with human feedback”. In: *arXiv preprint arXiv:2305.10425*.
- Azar, Mohammad Gheshlaghi et al. (2024). “A general theoretical paradigm to understand learning from human preferences”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 4447–4455.
- Huang, Shengyi et al. (2024). “The N+ Implementation Details of RLHF with PPO: A Case Study on TL; DR Summarization”. In: *arXiv preprint arXiv:2403.17031*.
- Zheng, Rui et al. (2023). “Secrets of rlhf in large language models part i: Ppo”. In: *arXiv preprint arXiv:2307.04964*.
- Guo, Shangmin et al. (2024). “Direct language model alignment from online ai feedback”. In: *arXiv preprint arXiv:2402.04792*.
- Xu, Jing et al. (2023). “Some things are more cringe than others: Preference optimization with the pairwise cringe loss”. In: *arXiv preprint arXiv:2312.16682*.

- Xu, Shusheng et al. (2024b). “Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study”. In: *arXiv preprint arXiv:2404.10719*.
- Yuan, Weizhe et al. (2024). “Self-rewarding language models”. In: *arXiv preprint arXiv:2401.10020*.
- Chen, Zixiang et al. (2024). “Self-play fine-tuning converts weak language models to strong language models”. In: *arXiv preprint arXiv:2401.01335*.
- Pang, Richard Yuanzhe et al. (2024). “Iterative reasoning preference optimization”. In: *arXiv preprint arXiv:2404.19733*.
- Sun, Yuchong et al. (2024). “Parrot: Enhancing multi-turn instruction following for large language models”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9729–9750.
- Liu, Xiao et al. (2024b). “AgentBench: Evaluating LLMs as Agents”. In: *The Twelfth International Conference on Learning Representations*.
- Peng, Baolin et al. (2018b). “Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2182–2192.
- Zhao, Tiancheng, Kaige Xie, and Maxine Eskenazi (2019). “Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models”. In: *Proceedings of NAACL-HLT*, pp. 1208–1218.
- Shi, Taiwei et al. (2024). “WildFeedback: Aligning LLMs With In-situ User Interactions And Feedback”. In: *arXiv preprint arXiv:2408.15549*.
- Nijkamp, Erik et al. (2023). “CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis”. In: *The Eleventh International Conference on Learning Representations*.
- Yu, Tao et al. (2019). “CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Sun, Ruoxi et al. (2023b). “SQLPrompt: In-Context Text-to-SQL with Minimal Labeled Data”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 542–550.
- Sun, Ruoxi et al. (2023a). “Sql-palm: Improved large language model adaptation for text-to-sql”. In: *arXiv preprint arXiv:2306.00739*.
- Toles, Matthew et al. (2023). “Pragmatic Evaluation of Clarifying Questions with Fact-Level Masking”. In: *arXiv preprint arXiv:2310.11571*.

- Yu, Tao et al. (2018). “Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3911–3921.
- Jiang, Albert Q et al. (2023). “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825*.
- Cui, Ganqu et al. (2023). “Ultrafeedback: Boosting language models with high-quality feedback”. In: *arXiv preprint arXiv:2310.01377*.
- Ding, Ning et al. (2023). “Enhancing Chat Language Models by Scaling High-quality Instructional Conversations”. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Reimers, Nils and Iryna Gurevych (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992.
- Risch, Julian et al. (2021a). “Semantic Answer Similarity for Evaluating Question Answering Models”. In: *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pp. 149–157.
- Chandramohan, Senthilkumar et al. (2011). “User simulation in dialogue systems using inverse reinforcement learning”. In: *Interspeech 2011*, pp. 1025–1028.
- Kreyssig, Florian et al. (2018). “Neural User Simulation for Corpus-based Policy Optimisation of Spoken Dialogue Systems”. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 60–69.
- Shah, Pararth et al. (2018). “Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pp. 41–51.
- Dua, Dheeru et al. (2019). “DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378.
- Wei, Jason et al. (2022). “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35, pp. 24824–24837.
- Mitra, Arindam et al. (2023). “Orca 2: Teaching small language models how to reason”. In: *arXiv preprint arXiv:2311.11045*.

- Mukherjee, Subhabrata et al. (2023). “Orca: Progressive learning from complex explanation traces of gpt-4”. In: *arXiv preprint arXiv:2306.02707*.
- Xu, Canwen et al. (2024a). “Contrastive post-training large language models on data curriculum”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*.
- Agarwal, Rishabh et al. (2024). “Many-shot in-context learning”. In: *arXiv preprint arXiv:2404.11018*.
- Levine, Sergey et al. (2020). “Offline reinforcement learning: Tutorial, review, and perspectives on open problems”. In: *arXiv preprint arXiv:2005.01643*.
- Hemphill, Charles T, John J Godfrey, and George R Doddington (1990). “The ATIS spoken language systems pilot corpus”. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.
- Khatri, Chandra et al. (2018b). “Alexa prize—state of the art in conversational AI”. In: *AI magazine* 39.3, pp. 40–55.
- Li, Bo et al. (2017a). “Acoustic Modeling for Google Home.” In: *Interspeech*, pp. 399–403.
- Von Ahn, Luis (2013). “Duolingo: learn a language for free while helping to translate the web”. In: *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 1–2.
- Fatima, Kaleem et al. (2024). “Accuracy of Artificial Intelligence-Based Virtual Assistants in Responding to Frequently Asked Questions Related to Orthognathic Surgery”. In: *Journal of Oral and Maxillofacial Surgery*.
- Li, Chen, Dimitris Chrysostomou, and Hongji Yang (2023). “A speech-enabled virtual assistant for efficient human–robot interaction in industrial environments”. In: *Journal of Systems and Software* 205, p. 111818.
- Zheng, Lianmin et al. (2024). “LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset”. In: *The Twelfth International Conference on Learning Representations*.
- Radford, Alec et al. (2023). “Robust speech recognition via large-scale weak supervision”. In: *International conference on machine learning*. PMLR, pp. 28492–28518.
- Mehrish, Ambuj et al. (2023). “A review of deep learning techniques for speech processing”. In: *Information Fusion* 99, p. 101869.
- Bhattacharya, Debasmita et al. (2023). “Capturing Formality in Speech Across Domains and Languages”. In: *Interspeech 2023*.
- Jung, Jee-weon et al. (2018). “Avoiding Speaker Overfitting in End-to-End DNNs Using Raw Waveform for Text-Independent Speaker Verification”. In: *Interspeech*.



- Wang, Tao et al. (2020d). “Spoken Content and Voice Factorization for Few-Shot Speaker Adaptation.” In: *INTERSPEECH*, pp. 796–800.
- Yang, Mu et al. (2024b). “Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-To-Audio with Retrieval-Augmented Generation”. In: *arXiv preprint arXiv:2411.05141*.
- Nautsch, Andreas et al. (2019). “Preserving privacy in speaker and speech characterisation”. In: *Computer Speech & Language* 58, pp. 441–480.
- Qian, Jianwei et al. (2018). “Towards privacy-preserving speech data publishing”. In: *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, pp. 1079–1087.
- Groeneveld, Dirk et al. (2024). “OLMo: Accelerating the Science of Language Models”. In: *arXiv preprint*.
- Seedat, Nabeel, Fergus Imrie, and Mihaela van der Schaar (2022). “Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems”. In: *arXiv preprint arXiv:2211.05764*.
- Khatri, Chandra et al. (2018a). “Advancing the state of the art in open domain dialog systems through the alexa prize”. In: *arXiv preprint arXiv:1812.10757*.
- Li, Chia-Hsuan et al. (2018). “Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension”. In: *arXiv preprint arXiv:1804.00320*.
- Shih, Min-Han et al. (2024). “GSQA: An End-to-End Model for Generative Spoken Question Answering”. In: *Interspeech 2024*.
- Faisal, Fahim et al. (2021). “SD-QA: Spoken Dialectal Question Answering for the Real World”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3296–3315.
- Chuang, Yung-Sung et al. (2020). “SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering”. In: *Interspeech 2020*.
- Su, Dan and Pascale Fung (2020). “Improving Spoken Question Answering Using Contextualized Word Representation”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8004–8008.
- Devlin, Jacob (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.
- Wu, Shengqiong et al. (2024). “NExT-GPT: Any-to-Any Multimodal LLM”. In: *Forty-first International Conference on Machine Learning*.

- Horvitz, Eric (1999). “Principles of mixed-initiative user interfaces”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166.
- Min, Sewon et al. (2020). “AmbigQA: Answering Ambiguous Open-domain Questions”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797.
- Wu, Zeqiu et al. (2023b). “INSCIT: Information-Seeking Conversations with Mixed-Initiative Interactions”. In: *Transactions of the Association for Computational Linguistics* 11, pp. 453–468.
- Deng, Yang et al. (2024b). “Towards human-centered proactive conversational agents”. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 807–818.
- You, Chenyu et al. (July 2022). “End-to-end Spoken Conversational Question Answering: Task, Dataset and Model”. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 1219–1232.
- Kurata, Gakuto, Nobuyasu Itoh, and Masafumi Nishimura (2011). “Acoustic Model Training with Detecting Transcription Errors in the Training Data.” In: *INTERSPEECH*, pp. 1689–1692.
- Mulholland, Matthew et al. (2016). “A comparison of ASR and human errors for transcription of non-native spontaneous speech”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5855–5859.
- Caruana, Rich (1997). “Multitask learning”. In: *Machine learning* 28, pp. 41–75.
- Pironkov, Gueorgui et al. (2016). “Speaker-aware long short-term memory multi-task learning for speech recognition”. In: *EUSIPCO*. IEEE, pp. 1911–1915.
- Aghajanyan, Armen et al. (2021). “Muppet: Massive Multi-task Representations with Pre-Finetuning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5799–5811.
- Padmakumar, Vishakh et al. (2022). “Exploring the Role of Task Transferability in Large-Scale Multi-Task Learning”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2542–2550.
- Henderson, Matthew et al. (2019). “Training Neural Response Selection for Task-Oriented Dialogue Systems”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5392–5404.

- Talmor, Alon et al. (June 2019). “CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Tamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158.
- Arora, Siddhant et al. (2024). “UniverSLU: Universal Spoken Language Understanding for Diverse Tasks with Natural Language Instructions”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2754–2774.
- Kong, Zhifeng et al. (2024). “Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities”. In: *Forty-first International Conference on Machine Learning*.
- Chen, Sanyuan et al. (2022c). “Wavlm: Large-scale self-supervised pre-training for full stack speech processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1505–1518.
- Yang, An et al. (2024a). “Qwen2 technical report”. In: *arXiv preprint arXiv:2407.10671*.
- Abdin, Marah et al. (2024). “Phi-3 technical report: A highly capable language model locally on your phone”. In: *arXiv preprint arXiv:2404.14219*.
- Risch, Julian et al. (Nov. 2021b). “Semantic Answer Similarity for Evaluating Question Answering Models”. In: *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Ed. by Adam Fisch et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 149–157.
- Anguera, Xavier et al. (2012). “Speaker diarization: A review of recent research”. In: *IEEE Transactions on audio, speech, and language processing* 20.2, pp. 356–370.
- Gu, Jia-Chen et al. (2021). “MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3682–3692.
- Yu, Nan, Guohong Fu, and Min Zhang (2022). “Speaker-aware discourse parsing on multi-party dialogues”. In: *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5372–5382.
- Roberts, Manley et al. (2023). “To the cutoff... and beyond? a longitudinal perspective on LLM data contamination”. In: *The Twelfth International Conference on Learning Representations*.

- Qian, Kun et al. (Nov. 2024). “VarBench: Robust Language Model Benchmarking Through Dynamic Variable Perturbation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 16131–16161.
- Sharma, Archit et al. (2024). “A critical evaluation of ai feedback for aligning large language models”. In: *arXiv preprint arXiv:2402.12366*.
- Yu, Xiao et al. (2024b). “LIONs: An Empirically Optimized Approach to Align Language Models”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8732–8753.
- Rajpurkar, Pranav et al. (2016). “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392.
- Schlager, Bettina and Steven K Feiner (2024). “Designing Non-Humanoid Virtual Assistants for Task-Oriented AR Environments”. In: *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, pp. 1017–1018.
- Yu, Xiao et al. (2024a). “Exact: Teaching ai agents to explore with reflective-mcts and exploratory learning”. In: *arXiv preprint arXiv:2410.02052*.
- Koh, Jing Yu et al. (2024). “VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905.
- Yao, Shunyu et al. (2023). “React: Synergizing reasoning and acting in language models”. In: *International Conference on Learning Representations (ICLR)*.
- Gemma Team et al. (2024). “Gemma: Open models based on gemini research and technology”. In: *arXiv preprint arXiv:2403.08295*.
- Wolf, Thomas et al. (Oct. 2020b). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45.
- Huang, HY (2017). “Fusionnet: Fusing via fully-aware attention with application to machine comprehension”. In: *arXiv preprint arXiv:1711.07341*.
- Lee, Chia-Hsuan, Yun-Nung Chen, and Hung-Yi Lee (2019). “Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7300–7304.

Rasley, Jeff et al. (2020). “Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506.

## Appendix A: Supplementary Material on Conversation Synthesis

Party	Utterance
Alice	I love it here in New York!
Bob	What do you like most about New York?
Alice	New York has the best bagels!
Bob	Which bagels are your favorite?
Alice	They all taste amazing. I really like their Italian ones!
Bob	That’s great. I guess I’ll be eating Italian bagels from now on.
Alice	I’m from New York City.
Bob	Wow! I’ve never been to New York.
Alice	I’ve also been to Austin. few times.
Bob	What did you think of it? Did you like it?
Alice	I really liked it. I thought it was a really cool city.
Bob	I’m glad you liked it. What did you like so much about it?
Alice	I really like the music scene there.
Bob	What kind of music do you like?
Alice	I like a lot of different kinds of music. I like jazz, rock, and country.
Bob	I’ve never been to a country concert.
Alice	You should go sometime. They’re a lot of fun.
Bob	I’ll have to check it out.
Alice	I’ll hagve to visit Austin again soon.
Bob	Yeah, I’d like to visit New York sometime too.
Alice	It’s a great city.

Table A1: Pair of dyadic conversations generated using OPT 30B. The prompt recipe given is: “The following is a conversation between Alice and Bob about their hometowns. Bob is from Austin, Texas, and Alice is from New York City.”

Party	Utterance
Alice	What's your major?
Bob	I'm a computer science major.
Alice	Me too! What year are you?
Bob	I'm in my third year. What about you?
Alice	I'm in my third year too. What classes are you taking this quarter?
Bob	I'm taking a class called Operating Systems. It's really interesting.
Alice	I'm taking a class called Computer Graphics. It's pretty interesting too.
Bob	I've heard that class is really hard.
Alice	Yeah, it's not easy.
Alice	Hi!
Alice	What are you studying?
Bob	I'm a Computer Science major.
Alice	Me too!
Bob	What year are you?
Alice	I'll be a junior next year.
Bob	I'll be a senior.
Alice	Cool!
Bob	What classes are you taking this quarter?
Alice	I've got a couple of classes in the School of Engineering. I'm taking a class in Computer Graphics and a class in Artificial Intelligence.
Bob	I'm taking a class on Machine Learning.
Alice	That's cool.
Bob	What are you doing this summer?
Alice	I'd like to work at a start-up.
Bob	That's cool. I'm going to intern at Google.
Alice	Wow! That's awesome!
Bob	Yeah, I'm really excited.
Alice	I'm excited for you!

Table A2: Pair of dyadic conversations generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob about their majors. They are both students at Stanford University.”

Party	Utterance
Alice	I was thinking about applying at the coffee shop on the corner.
Bob	That’s a great idea! It’s a really fun place to work.
Alice	I’m glad to hear that. What do you do there?
Bob	I work as a barista.
Alice	Oh cool! I bet you get to meet a lot of interesting people.
Bob	That’s right. I really enjoy working there.
Alice	Do you know if they are hiring?
Bob	I’m not sure, but you should definitely apply.
Alice	Thanks! I’ll go in and see if they are hiring.

Table A3: Example dyadic conversation generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob about part-time jobs. Bob works at a coffee shop and Alice is thinking about applying.”

Party	Utterance
Alice	I’d love to start a garden! I’ve always wanted to grow my own vegetables.
Bob	That’s a great idea! It’s so much more fun to eat vegetables that you’ve grown yourself.
Claire	What vegetables are you thinking about growing?
Alice	I’d love to try growing some tomatoes, peppers, and squash.
Bob	That sounds like a good start.
Claire	I’ve grown tomatoes before. I also have a few pepper plants.
Alice	I’ve always wanted to have a garden in my backyard, but I never really had the time.
Claire	That’s awesome! How do you plan to grow your garden?
Alice	I’ve been doing a lot of research. I think I’ll start with a few simple vegetables like tomatoes, onions, and potatoes.
Bob	That’s great! I would love to have a garden like that. I’m a big fan of cooking.
Alice	I’ll definitely invite you over for dinner then!
Claire	I’m so jealous! I’ve always wanted to grow my own vegetables.
Bob	Have you considered getting a garden plot?
Claire	That’s a great idea. I should look into that.
Alice	You should! Gardening is a really healthy hobby.

Table A4: Pair of triadic conversations generated using OPT 30B. The prompt recipe given is: “The following is a conversation between Alice and Bob and Claire about gardening. Alice is interested in growing vegetables.” The ordering of the speakers in multi-party scenarios is variable.



Party	Utterance
Alice	Have you ever had Southern Ice Tea?
Bob	I love to make it at home.
Claire	I'm not familiar with that. What is it?
Bob	It's a sweet tea that's brewed with mint and lemon.
Alice	That sounds really good. I'm going to have to try that!
Claire	I've never had that. What is it like?
Bob	It's kind of like a sweet, lemonade-like drink.
Alice	I'm going to have to make some this weekend. Thanks for the idea!
Claire	That sounds really good. I might try it too.

Table A5: Example triadic conversation generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob and Claire about cooking. Alice is interested in Southern Ice Tea. Bob has a great recipe. Claire loved trying Bob’s Southern Ice Tea.”

Party	Utterance
Alice	I try to eat healthy, but it's hard to stay on track.
Bob	I know what you mean. I try to eat healthy too, but I love ice cream too much!
Claire	What kinds of foods do you like to eat?
Alice	I love fruit. I eat a lot of berries.
Bob	I love berries too! Strawberries are my favorite.
Claire	What about leafy greens? Do you eat them?
Bob	I love salads!
Alice	I love salads too!
Claire	Do you like any vegetables?
Alice	I love vegetables! I like broccoli and asparagus.
Claire	Great! Do you eat many whole grains?

Table A6: Example triadic conversation generated using OPT 175B. The prompt recipe given is: “The following is a conversation between Alice and Bob and Claire about health. Alice and Bob are discussing healthy foods. Claire is a nutritionist.”

Party	Utterance
Phoebe	Then I'm gonna have to ask you to keep it down.
Mr. Heckles	Who are you?
Eric	Hi, I'm Eric, I'm gonna be Chandler's new roommate.
Mr. Heckles	I'm Chandler's new roommate.
Eric	I-I-I don't think so.
Mr. Heckles	I could be Chandler's new roommate.
Eric	But, he told me over the phone.
Mr. Heckles	He told me in person.
Eric	That's weird.
Mr. Heckles	Well, I'm going to go into my new apartment now. Ehh!

Table A7: Conversation from the MELD corpus (Poria et al., 2019). Three speakers are involved, discussing a living situation regarding a fourth character who does not appear in this scene.

Party	Utterance
john	sure
john	i think so, meg
mara	how did i know that was coming
meg	its not just the public eye
john	haha mara
mara	hushh.. ***
nick	There are already other countries who are investigating the Bush administration for war crimes - Spain
meg	with the breton woods
george	they need to be prosecuted...that's in obama's hands
nick	wow, george, right win propaganda... huh
meg	look at how well Iraq is doing
mara	goodness
meg	there's a point at which interrogation becomes torture and is just inhumane
john	agree to george
mara	?
mara	im in albany btw
meg	Which we signed!
amy	well it is the way the world is going– email, chat,, etc
john	yes
jordan	And this is one of the tricky things in this virtual world. You know nothing about the people u r talking to!!!!
amy	u r right you just used online language haha
mara	hes not much fun either haha, what do you think?
amy	hi john- can you see my message here?
jordan	Hi, amy
mara	i dont know what is better really!!!
john	haha

Table A8: Three excerpts of the same conversation from the MPC corpus (Shaikh et al., 2010). The conversation spans topics ranging from the Bush administration to meta-discussion about the collection task.

Subtopic	Background Information
Pacific Theater	Alice is interested in Pacific theater.
Growing residential grass	Alice is interested in growing residential grass.
Breakfast food	Alice likes to try different breakfast foods. Bob loves waffles.
music	Alice likes music. Bob plays the viola.
skincare	Alice is interested in skincare. Bob has a great skincare routine.
Planting flowers	Alice is interested in planting flowers. Bob has a nice garden.
Southern Ice Tea	Alice is interested in Southern Ice Tea. Bob has a great recipe.
herb garden	Alice is interested in planting an herb garden.
Hiking	Alice is going hiking tomorrow.
Plant a garden	Alice wants to plant a garden.
Italian food	Alice likes Italian food.
book recommendations	Alice is interested in book recommendations.
anniversaries	Alice keeps track of all of her anniversaries.
Existential Psychology	Alice is interested in Existential Psychology.
The Outlander Series	Alice is interested in The Outlander Series.
camping gear	Alice is looking for advice on camping gear. Bob works at REI.
Movie	Alice is interested in movie recommendations. Bob is a film buff.
Ford Vehicles	Alice is interested in Ford vehicles. Bob prefers Japanese cars.
Beauty	Alice is interested in beauty. Bob works at Sephora.
Syrian War	Alice is interested in the Syrian War. Bob is a political scientist.
Elon Musk	Alice and Bob are talking about Elon Musk.
Healthy foods	Alice and Bob are discussing healthy foods. Alice is on a paleo diet.
Soren Kierkegaard	Alice is a fan of Soren Kierkegaard.
investing money	Alice is interested in investing money. Bob is an investment banker.
Post-structuralism	Alice is interested in post-structuralism.
baking	Alice is interested in baking. Bob has baked cakes and brownies before.
Nuts	Alice likes to eat nuts.
braids	Alice braids her hair. Bob is interested in learning how.
Growing vegetables	Alice is interested in growing vegetables.
Martin Luther	Alice is learning about Martin Luther.
paint brushes	Alice is interested in paint brushes.
Stock Trading	Alice is interested in stock trading.
Install TV applications	Alice wants to install TV applications. Bob is helping her.
History	Alice is interested in history. History was Bob's favorite school subject.
Feminism	Alice is interested in feminism. Bob majored in gender studies.
Tell a joke	Alice wants to hear Bob tell a joke.
artists	Alice is interested in learning about modern artists.
Turtles	Alice likes turtles. Bob has been scuba diving.
Anthony Trollope	Alice likes the work of Anthony Trollope. Bob prefers modern literature.
Paris	Alice wants to go to Paris.
Bread	Alice likes bread. Bob's favorite bread is a baguette.
movie cast members	Alice and Bob are talking about movie cast members.
Gay Marriage	Alice is a proponent of gay marriage. Bob is interested in learning more.
U.S. Senate	Alice and Bob are discussing the U.S. Senate.
growing tomatoes	Alice is interested in growing tomatoes.
family issues	Alice is interested in family issues.
Automotive parts	Alice is interested in automotive parts.
Bee life	Alice is interested in bee life.
Taylor Swift	Alice's favorite musician is Taylor Swift. Bob likes Ariana Grande.
biking	Alice's favorite hobby is biking. Bob prefers rock climbing.
Juicers	Alice wants to get a juicer.
islands	Alice likes visiting islands. Bob prefers hiking.
Planets	Alice is learning about the planets in school.
Pokemon	Alice likes to play Pokemon. Bob also likes Pokemon.

Table A9: Corresponding background information written for each of the subtopics found in the FITS dataset. There is a mixture of prompts which only mention one speaker and prompts which mention two speakers. Every synthetic conversation involves both speakers.

Topic	Conversation Recipe
Growing residential grass	Alice is interested in growing residential grass. Claire has a really neat yard.
Breakfast food	Alice likes to try different breakfast foods. Bob loves waffles. Claire prefers pancakes.
music	Alice likes music. Bob plays the viola. Claire played the violin in high school.
skincare	Alice is interested in skincare. Bob has a great skincare routine. Claire wants to hear Bob's routine.
Planting flowers	Alice is interested in planting flowers. Bob has a nice garden. Claire has a vegetable garden.
Southern Ice Tea	Alice is interested in Southern Ice Tea. Bob has a great recipe. Claire loved trying Bob's Southern Ice Tea.
herb garden	Alice is interested in planting an herb garden. Claire has some gardening tips.
Hiking	Alice is going hiking tomorrow. Claire hates hiking.
Plant a garden	Alice wants to plant a garden. Claire has a greenroom.
Italian food	Alice likes Italian food. Claire prefers Asian food.
book recommendations	Alice is interested in book recommendations. Claire is a part of a book club.
anniversaries	Alice keeps track of all of her anniversaries. Claire is not well-organized.
Existential Psychology	Alice is interested in Existential Psychology. Claire is a psychologist by training.
The Outlander Series	Alice is interested in The Outlander Series. Claire has never seen the series.
camping gear	Alice is looking for advice on camping gear. Bob works at REI. Claire loves the outdoors.
Movie	Alice is interested in movie recommendations. Bob is a film buff. Claire is also a film buff.
Ford Vehicles	Alice is interested in Ford vehicles. Bob prefers Japanese cars. Claire prefers to drive a BMW.
Beauty	Alice is interested in beauty. Bob works at Sephora. Claire is shopping with Alice.
Syrian War	Alice is interested in the Syrian War. Bob is a political scientist. Claire is studying modern political theory.
Elon Musk	Alice and Bob are talking about Elon Musk. Claire is a Tesla owner.
Healthy foods	Alice and Bob are discussing healthy foods. Alice is on a paleo diet. Claire is a nutritionist.
Soren Kierkegaard	Alice is a fan of Soren Kierkegaard. Claire is not familiar with Soren Kierkegaard.
investing money	Alice is interested in investing money. Bob is an investment banker. Claire is an expert in personal finance.
Post-structuralism	Alice is interested in post-structuralism. Claire is an expert on the subject.
baking	Alice is interested in baking. Bob has baked cakes and brownies before. Claire wants to learn how to bake.
Nuts	Alice likes to eat nuts. Claire is allergic to peanuts.
braids	Alice braids her hair. Bob is interested in learning how. Claire braids her hair every day.
Growing vegetables	Alice is interested in growing vegetables. Claire has a vegetable garden. Bob grows flowers.
Martin Luther	Alice is learning about Martin Luther. Claire is a historian.
paint brushes	Alice is interested in paint brushes. Claire is a painter and has several suggestions.
Stock Trading	Alice is interested in stock trading. Claire is a stock broker.
Install TV applications	Alice wants to install TV applications. Bob is helping her. Claire is also good with technology.
History	Alice is interested in history. History was Bob's favorite school subject. Claire is a historian.
Feminism	Alice is interested in feminism. Bob majored in gender studies. Claire does not know much about feminism.
Tell a joke	Alice wants to hear Bob tell a joke. Claire is a stand-up comedian.
artists	Alice is interested in learning about modern artists. Claire is a photographer.
Turtles	Alice likes turtles. Bob has been scuba diving. Claire wants to try scuba diving.
Anthony Trollope	Alice likes the work of Anthony Trollope. Bob prefers modern literature. Claire is not familiar with much literature.
Paris	Alice wants to go to Paris. Claire has never been to Europe.
Bread	Alice likes bread. Bob's favorite bread is a baguette. Claire loves to bake bread.
movie cast members	Alice and Bob are talking about movie cast members. Claire has seen a lot of movies recently.
Gay Marriage	Alice is a proponent of gay marriage. Bob is interested in learning more. Claire is an activist.
U.S. Senate	Alice and Bob are discussing the U.S. Senate. Claire is a politician.
growing tomatoes	Alice is interested in growing tomatoes. Claire has a large garden with many tomatoes.
family issues	Alice is interested in family issues. Claire is a therapist.
Automotive parts	Alice is interested in automotive parts. Claire is a mechanic.
Bee life	Alice is interested in bee life. Claire is a beekeeper.
Taylor Swift	Alice's favorite musician is Taylor Swift. Bob likes Ariana Grande. Claire does not like pop music.
biking	Alice's favorite hobby is biking. Bob prefers rock climbing. Claire prefers archery.
Juicers	Alice wants to get a juicer. Claire has a suggestion for a great juicer.
islands	Alice likes visiting islands. Bob prefers hiking. Claire likes the beach.
Planets	Alice is learning about the planets in school. Claire is an astronomer.
Pokemon	Alice likes to play Pokemon. Bob also likes Pokemon. Claire prefers to play Stardew Valley.

Table A10: Triadic background information written for each of the subtopics given in the FITS dataset. Unlike Table A9, each of these may include background information for up to three people.

<p>The following is a conversation between Alice and Bob about past travel experiences. Alice has been to Japan and Bob is considering flying there.</p> <p>Alice: Hi!</p> <p>Bob: Hey, how are you doing?</p> <p>Alice: I'm doing well! I just got back from my vacation in Japan.</p> <p>Bob: Wow that's awesome! What did you think of it?</p> <p>Alice: Japan was such an amazing place to visit!</p> <p>Bob: Wow! What was your favorite part?</p> <p>Alice: I really enjoyed the food in Tokyo.</p> <p>Bob: Which airline did you take?</p> <p>Alice: I flew using Japan Airlines.</p>
<p>The following is a conversation between Alice and Bob about their hobbies. Alice enjoys tennis and Bob likes playing soccer.</p> <p>Alice: What do you like to do for fun?</p> <p>Bob: I used to play soccer in college, so I still like to play for fun on the weekends!</p> <p>Alice: That's great. Soccer is a great way to stay in good shape.</p> <p>Bob: I agree - it's really good cardio. What about you?</p> <p>Alice: I love to play tennis. I've been taking lessons for a few months now!</p> <p>Bob: Tennis is fun too!</p>
<p>The following is a conversation between Alice and Bob about their favorite movies. Bob loved the new Batman movie. Alice really liked watching Pride and Prejudice.</p> <p>Alice: I just saw Pride and Prejudice for the fifth time!</p> <p>Bob: That's a lot of times! What do you like so much about that movie?</p> <p>Alice: Well, as a teenager I really liked the book. But I just really loved Keira Knightley's portrayal of Elizabeth.</p> <p>Bob: I see. I haven't seen the movie myself. I prefer action films.</p> <p>Alice: What's your favorite action movie?</p> <p>Bob: Hm, I really liked the Batman movie that just came out.</p> <p>Alice: I haven't seen it yet. I heard it got pretty good reviews.</p>
<p>The following is a conversation between Alice and Bob about their hometowns. Alice is from New York City. Bob grew up in Seattle.</p> <p>Alice: Hello! How are you doing?</p> <p>Bob: Hi, I'm doing great! What about yourself?</p> <p>Alice: I'm doing well! Where are you from?</p> <p>Bob: I'm originally from Seattle, but now I live in Palo Alto.</p> <p>Alice: Oh cool! I live in Palo Alto too. Do you like Seattle or California more?</p> <p>Bob: Well, Seattle is always going to be home for me. Even if the weather in California is nicer.</p> <p>Alice: Haha, I get that! I miss New York City - there's no place like home.</p> <p>Bob: What is your favorite neighborhood of New York City?</p> <p>Alice: I love going to Chelsea. The Highline has a great view, and Little Island is close by too! Have you ever been?</p> <p>Bob: Unfortunately I have not. I have never been to the East Coast!</p>
<p>The following is a conversation between Alice and Bob about art. Alice's favorite artist is Michelangelo. Bob does not know much about art.</p> <p>Alice: Hi, how's it going?</p> <p>Bob: It's going well, what about you?</p> <p>Alice: I'm doing great! I've been really interested in art recently.</p> <p>Bob: What got you interested in art?</p> <p>Alice: Art can be so breathtaking!</p> <p>Bob: I feel like I don't know how to properly appreciate art, but certain pieces of artwork certainly look very complex.</p> <p>Alice: Have you ever heard of Michelangelo?</p> <p>Bob: I have heard of him, but I don't know anything that he has created.</p> <p>Alice: Michelangelo is really famous for his statue of David.</p> <p>Bob: Huh? Who is David?</p> <p>Alice: David is a Biblical figure who was a king of Israel. Michelangelo built a really magnificent statue of him in Florence.</p>
<p>The following is a conversation between Alice and Bob about drinks. Alice is a wine expert, whereas Bob prefers cocktails.</p> <p>Alice: How are you doing?</p> <p>Bob: Pretty great! I'm planning to go to a brewery this weekend.</p> <p>Alice: Do you know much about alcohol?</p> <p>Bob: Yeah, I really like beer! I drink a lot of IPAs.</p> <p>Alice: Oh - what do you like about IPAs? I can't get over the bitter taste.</p> <p>Bob: Well, I don't think it's just bitter. Sometimes there are really interesting citrusy or herbal flavor notes.</p> <p>Alice: I see. That kind of reminds me of wine tasting.</p> <p>Bob: There's definitely a lot of depth to it like there is with wine. Do you know much about wine?</p> <p>Alice: Yeah, I took several classes on wine tasting back in the day. I really love Pinot Noir.</p> <p>Bob: Oh I love red wines too.</p> <p>Alice: Right? I love the dryness and fruity notes of Pinot Noir.</p>
<p>The following is a conversation between Alice and Bob about relationships. Bob recently got engaged.</p> <p>Alice: Congrats on your engagement! When do you think you will have your wedding?</p> <p>Bob: Thank you!! We're thinking of having it in November.</p> <p>Alice: That's amazing! Will you pick a fancy destination?</p> <p>Bob: I wanted to! I was thinking of having it somewhere in Europe, but my partner and I ultimately decided we wanted to have it close to home so our friends could all make it.</p> <p>Alice: That's a good point. My husband and I had similar thoughts when we were planning our wedding.</p> <p>Bob: What did you plan in the end?</p> <p>Alice: We had a small ceremony in my hometown!</p>
<p>The following is a conversation between Alice and Bob about their jobs. Alice works in the financial industry and Bob is a musician.</p> <p>Alice: I'm so burnt out from my work! I just want to quit already!</p> <p>Bob: Whoa - what do you do for work?</p> <p>Alice: I'm an investment banker. It's been four years at this company and I'm absolutely exhausted.</p> <p>Bob: That sounds intense. Is there anything you actually like about the job?</p> <p>Alice: Well, the money is good.</p> <p>Bob: It sounds like you could use a break. Maybe you could use some of that money to go travel.</p> <p>Alice: I really want to go to South America, but I don't have a lot of time.</p>
<p>The following is a conversation between Alice and Bob about their pets. Alice has a dog and Bob prefers cats.</p> <p>Alice: Do you have any pets?</p> <p>Bob: No, but I really want to get a cat.</p> <p>Alice: What, why a cat? Cats seem so boring. They never want to play.</p> <p>Bob: Yeah, but cats are so cute! They also are a lot easier to take care of. They can clean themselves. What do you prefer?</p> <p>Alice: Well, I have a dog. He is a corgi and his name is Bo.</p> <p>Bob: Aww that's cute! I'm not usually a dog person, but corgis are adorable.</p> <p>Alice: Haha, thank you! Bo is a really friendly dog.</p> <p>Bob: How old is he?</p> <p>Alice: Bo is one year old now.</p>
<p>The following is a conversation between Alice and Bob about grocery shopping. Alice has a shopping list for Bob.</p> <p>Alice: Could you run to the grocery store and pick up some bananas for me?</p> <p>Bob: Will do - how many do you need?</p> <p>Alice: Oh, I don't know, maybe ten bananas. I'm planning to make banana bread, but I also want to save some for us to eat at home.</p> <p>Bob: That sounds delicious! I'll head out in a second. Is there anything else you need?</p>

Table A11: Handwritten conversation examples of varying length. In-context examples are randomly sampled from this pool and used as part of a prompt for dyadic conversation generation.

---

The following is a conversation between Alice and Bob and Claire about past travel experiences. Alice has been to Japan and Bob is considering flying there. Claire has been to Taiwan and Korea, but not Japan.

Alice: Hi!  
Bob: Hey, how are you doing?  
Alice: I'm doing well! I just got back from my vacation in Japan.  
Bob: Wow that's awesome! What did you think of it?  
Alice: Japan was such an amazing place to visit!  
Claire: Wow, I've always wanted to visit Japan!  
Bob: What was your favorite part?  
Alice: I really enjoyed the food in Tokyo. I had the best sushi of my life!  
Bob: Which airline did you take?  
Alice: I flew using Japan Airlines.  
Claire: How expensive are tickets these days?

---

The following is a conversation between Alice and Bob about their hobbies. Alice enjoys tennis and Bob likes playing soccer. Claire plays football.

Alice: What do you like to do for fun?  
Bob: I used to play soccer in college, so I still like to play for fun on the weekends!  
Claire: Oh wow! Did you play varsity soccer?  
Bob: Yeah, I was a four-year starter!  
Alice: That's great. Soccer is a great way to stay in good shape.  
Bob: I agree - it's really good cardio. What about you all?  
Claire: I'm in a flag football league! We play every Saturday afternoon.  
Alice: I love to play tennis. I've been taking lessons for a few months now!  
Bob: Cool, football and tennis are fun too!

---

The following is a conversation between Alice and Bob and Claire about their favorite movies. Claire is looking for movie recommendations. Bob loved the new Batman movie. Alice really liked watching Pride and Prejudice.

Alice: I just saw Pride and Prejudice for the fifth time!  
Claire: Would you recommend watching it? I've never seen it!  
Bob: Yeah, five times is a lot of times! What do you like so much about that movie?  
Alice: Well, as a teenager I really liked the book. But I just really loved Keira Knightley's portrayal of Elizabeth.  
Bob: I see. I haven't seen the movie myself. I prefer action films.  
Alice: What's your favorite action movie?  
Bob: Hm, I really liked the Batman movie that just came out.  
Alice: I haven't seen it yet. I heard it got pretty good reviews.

---

The following is a conversation between Alice and Bob and Claire about their hometowns. Alice is from New York City. Bob grew up in Seattle. Claire is from Boston and would like to visit New York City.

Alice: Hello! How are you doing?  
Claire: I'm doing good!  
Bob: Hi, I'm doing great! What about yourself?  
Alice: I'm doing well! Where are you both from?  
Claire: I'm from Boston! I'm just visiting the Bay Area.  
Bob: I'm originally from Seattle, but now I live in Palo Alto.  
Alice: Oh cool! I live here in Palo Alto. Do you like Seattle or California more?  
Bob: Well, Seattle is always going to be home for me. Even if the weather in California is nicer.  
Alice: Haha, I get that! I miss New York City - there's no place like home.  
Claire: Oh you're from New York? I've always wanted to visit!  
Bob: Me too! What is your favorite neighborhood of New York City?  
Alice: I love going to Chelsea. The Highline has a great view, and Little Island is close by too! Have you ever been?  
Bob: Unfortunately I have not. I have never been to the East Coast!

---

The following is a conversation between Alice and Bob and Claire about art. Alice's favorite artist is Michelangelo. Bob does not know much about art. Claire is a painter.

Alice: Hi, how's it going?  
Bob: It's going well, what about you?  
Alice: I'm doing great! I've been really interested in art recently.  
Claire: Oh that's great to hear! I love art as well.  
Bob: What got you interested in art?  
Alice: Art can just be so breathtaking!  
Bob: I feel like I don't know how to properly appreciate art, but certain pieces of artwork certainly look very complex.  
Alice: Have you ever heard of Michelangelo?  
Bob: I have heard of him, but I don't know anything that he has created.  
Claire: Michelangelo has some truly magnificent paintings, such as The Creation of Adam.  
Alice: Michelangelo is also really famous for his statue of David.  
Bob: Huh? Who is David?  
Alice: David is a Biblical figure who was a king of Israel. Michelangelo built a really magnificent statue of him in Florence.

---

The following is a conversation between Alice and Bob and Claire about drinks. Alice is a wine expert, whereas Bob prefers cocktails. Claire likes to drink beer.

Alice: How are you doing?  
Bob: Pretty great! I'm planning to go to a brewery this weekend.  
Alice: Do you know much about alcohol?  
Bob: Yeah, I really like beer! I drink a lot of IPAs.  
Claire: Oh, beers are my favorite type of drink! I can really appreciate the taste of a good IPA.  
Alice: Oh - what do you like about IPAs? I can't get over the bitter taste.  
Bob: Well, I don't think it's just bitter. Sometimes there are really interesting citrusy or herbal flavor notes.  
Claire: Yeah, there's a whole science to the hops used in making IPAs!  
Alice: I see. That kind of reminds me of wine tasting.  
Claire: The science behind tasting is similar for sure.  
Bob: I agree, there's definitely a lot of depth to it like there is with wine. Do you know much about wine?  
Alice: Yeah, I took several classes on wine tasting back in the day. I really love Pinot Noir.  
Bob: Oh I love red wines too.  
Alice: Right? I love the dryness and fruity notes of Pinot Noir.

---

Table A12: Triadic conversation recipes written for each of the “generic topics” given in the FITS dataset. These conversation recipes are included after the in-context examples when prompting LLMs to generate synthetic conversations. Unlike Table A9, each of these conversation recipes may include background for up to three people. Continued in Table A13.

---

The following is a conversation between Alice and Bob and Claire about relationships. Bob recently got engaged.

Alice: Congrats on your engagement!

Claire: Yes, congrats! When do you think you will have your wedding?

Bob: Thank you! We're thinking of having it in November.

Alice: That's amazing! Will you pick a fancy destination?

Bob: I wanted to! I was thinking of having it somewhere in Europe, but my partner and I ultimately decided we wanted to have it close to home so our friends could all make it.

Claire: Oh wow, that is very considerate of you.

Alice: Yeah, that's a good point. My husband and I had similar thoughts when we were planning our wedding.

Bob: What did you plan in the end?

Alice: We had a small ceremony in my hometown!

Claire: It turned out nicely! It was such a beautiful ceremony.

---

The following is a conversation between Alice and Bob and Claire about their jobs. Alice works in the financial industry and Bob is a musician. Claire is an architect.

Alice: I'm so burnt out from my work! I just want to quit already!

Bob: Whoa - what do you do for work?

Alice: I'm an investment banker. It's been four years at this company and I'm absolutely exhausted.

Bob: That sounds intense. Is there anything you actually like about the job?

Alice: Well, the money is good.

Claire: That doesn't sound like a healthy relationship with your job!

Bob: It sounds like you could use a break. Maybe you could use some of that money to go travel.

Alice: I really want to go to South America, but I don't have a lot of time.

Claire: Don't you have vacation days? I think breaks are important.

Alice: Yes, but I really want to get promoted this year.

---

The following is a conversation between Alice and Bob and Claire about their pets. Alice has a dog and Bob prefers cats. Claire has a pet hamster.

Alice: Do you have any pets?

Claire: I have a pet hamster! He is so adorable. What about you two?

Bob: I don't, but I really want to get a cat.

Alice: What, why a cat? Cats seem so boring. They never want to play.

Bob: Yeah, but cats are so cute! They also are a lot easier to take care of. They can clean themselves. What do you prefer?

Alice: Well, I have a dog. He is a corgi and his name is Bo.

Claire: That's so adorable! How old is he?

Alice: He just turned one!

Bob: Aww that's cute! I'm not usually a dog person, but corgis are adorable.

Alice: Haha, thank you! Bo is a really friendly dog.

---

The following is a conversation between Alice and Bob and Claire about grocery shopping. Alice has a shopping list for Bob. Claire is helping Alice cook at home.

Alice: Could you run to the grocery store and pick up some bananas for me?

Bob: Will do - how many do you need?

Alice: Oh, I don't know, maybe ten bananas. We are planning to make banana bread, but I also want to save some for us to eat at home.

Bob: That sounds delicious! I'll head out in a second. Is there anything else you need?

Claire: Oh, could you also pick up some more eggs? I think we're running low here.

---

Table A13: Triadic conversation recipes written for each of the “generic topics” given in the FITS dataset continued from Table A12.

## Appendix B: Supplementary Material on Explicit Action Planning

	Role	Utterance
One-shot Example	system	Save the Children is head-quartered in London, and they work to help fight poverty around the world. Children need help in developing countries and war zones. Small donations like \$1 or \$2 go a long way to help. You are Persuader who is trying to persuade the Persuadee to donate to a charity called Save the Children. The following is an example conversation between a Persuader and a Persuadee.
	user	Persuadee: Hello. How are you? <i>The Persuader asks about the Persuadee’s knowledge or opinion related to Save the Children.</i>
	assistant	Persuader: Very well. I’m just up organizing info for my charity called Save the Children. Have you heard of this charity before?
	user	Persuadee: No, I have not. Can you tell me more? <i>The Persuader establishes credibility of Save the Children by citing its impact.</i>
	assistant	Persuader: Save the Children is an organization that helps children in developing countries, by promoting children’s rights and providing relief. It is an amazing charity that helps kids who are in desperate need. They can help with safety, education and more.
	user	Persuadee: That sounds great. I believe in this charity, but still wonder how much of the money I donate actually helps. I am always worried if I donate it will just go to some higher up that is living the high life. <i>The Persuader uses an emotion appeal to convince the Persuadee.</i>
	assistant	Persuader: Every little bit makes a difference. When you have people who are so poor, it’s amazing what a tiny amount can do. I usually donate in hopes I can at least feel like I did my part. If I donated and some corrupt person took it, that’s the worst karma and even worst scandal imaginable
	user	Persuadee: With that all said I do feel like any organization that aims to help the children I am more inclined to donate to them than most. I think helping children is an important thing as they are our future! <i>The Persuader asks if the Persuadee would like to make a small donation.</i>
	assistant	Persuader: I think donating to this cause would def be a step in the right direction to hopefully helping across the world the children that are in despair. I don’t want you to donate any more than you want, so if you want to donate how much do you to do?
	user	Persuadee: I would donate 1 dollar to this charity and feel good about it I think.
	system	The following is a new conversation between Persuader (you) and a Persuadee. <i>The Persuader greets the persuadee.</i>
	assistant	Persuader: Hello. How are you?
	user	Persuadee: I am good. What about you? <i>The Persuader asks about the Persuadee’s knowledge or opinion related to Save the Children.</i>
	assistant	Persuader: I’m doing well, thanks for asking. I wanted to talk to you about an important cause that I am passionate about - Save the Children. Have you heard of them before?
Real dialogue	user	Persuadee: No I have not. <i>The Persuader establishes credibility of Save the Children by citing its impact.</i>
	system	
Generation	assistant	Save the Children is an international non-governmental organization that is committed to improving the lives of children through education, healthcare, and other support programs. They work in over 100 countries around the world to help children who are living in poverty or affected by conflict.

Table B1: Prompting LLM to generate a system response conditioned on a dialogue act. “Role” is used to be consistent with ChatGPT’s API. Blue texts are prompts characteristic of this task.



	Role	Utterance
One-shot Example	system	You are a Persuadee. A Persuader is trying to persuade you to donate to a charity called Save the Children. You can choose amongst the following actions during a conversation to respond to the Persuader: <a href="#">[no donation]</a> <a href="#">[negative reaction]</a> <a href="#">[neutral]</a> <a href="#">[positive reaction]</a> <a href="#">[donate]</a> . The following is an example conversation between a Persuader and some Persuadee.
	assistant	Persuadee: <a href="#">[neutral]</a> Hello. How are you?
	user	Persuader: Very well. I'm just up organizing info for my charity called Save the Children. Have you heard of this charity before?
	assistant	Persuadee: <a href="#">[neutral]</a> No, I have not. Can you tell me more?
	user	Persuader: Save the Children is an organization that helps children in developing countries, by promoting children's rights and providing relief. It is an amazing charity that helps kids who are in desperate need. They can help with safety, education and more.
	assistant	Persuadee: <a href="#">[negative reaction]</a> That sounds great. I believe in this charity, but still wonder how much of the money I donate actually helps. I am always worried if I donate it will just go to some higher up that is living the high life.
	user	Persuader: Every little bit makes a difference. When you have people who are so poor, it's amazing what a tiny amount can do. I usually donate in hopes I can at least feel like I did my part. If I donated and some corrupt person took it, that's the worst karma and even worst scandal imaginable
	assistant	Persuadee: <a href="#">[positive reaction]</a> With that all said I do feel like any organization that aims to help the children I am more inclined to donate to them than most. I think helping children is an important thing as they are our future!
	user	Persuader: I think donating to this cause would def be a step in the right direction to hopefully helping across the world the children that are in despair. I don't want you to donate any more than you want, so if you want to donate how much do you to do?
	assistant	Persuadee: <a href="#">[donate]</a> I would donate 1 dollar to this charity and feel good about it I think.
Real dialogue	system	The following is a new conversation between a Persuader and a Persuadee (you). You may or may not want to donate to Save the Children.
	user	Persuader: Hello. How are you?
	assistant	Persuadee: <a href="#">[neutral]</a> I am good. What about you?
	user	Persuader: I'm doing well, thanks for asking. I wanted to talk to you about an important cause that I am passionate about - Save the Children. Have you heard of them before?
	assistant	Persuadee: <a href="#">[neutral]</a> No I have not.
	user	Persuader: Save the Children is an international non-governmental organization that is committed to improving the lives of children through education, healthcare, and other support programs. They work in over 100 countries around the world to help children who are living in poverty or affected by conflict.
	assistant	<a href="#">[positive reaction]</a> That sounds really good! How can I get involved?

Table B2: Prompting LLM to simulate a user response conditioned on past (simulated) history. “Role” is used to be consistent with ChatGPT’s API. [Blue](#) texts are prompts characteristic of this task. This one-shot example is the same as B1, but a) with roles reversed and b) user response is prepended with a label. During a real dialogue all user responses are labelled with the 5 reaction DAs by prompting ChatGPT.

	Role	Utterance
One-shot Example	system	You are a Persuadee. A Persuader is trying to persuade you to donate to a charity called Save the Children. You can choose amongst the following actions during a conversation to respond to the Persuader: <a href="#">[no donation]</a> <a href="#">[negative reaction]</a> <a href="#">[neutral]</a> <a href="#">[positive reaction]</a> <a href="#">[donate]</a> . The following is an example conversation between a Persuader and some Persuadee.
	assistant	Persuadee: <a href="#">[neutral]</a> Hello. How are you?
	user	Persuader: Very well. I'm just up organizing info for my charity called Save the Children. Have you heard of this charity before?
	assistant	Persuadee: <a href="#">[neutral]</a> No, I have not. Can you tell me more?
	user	Persuader: Save the Children is an organization that helps children in developing countries, by promoting children's rights and providing relief. It is an amazing charity that helps kids who are in desperate need. They can help with safety, education and more.
	assistant	Persuadee: <a href="#">[negative reaction]</a> That sounds great. I believe in this charity, but still wonder how much of the money I donate actually helps. I am always worried if I donate it will just go to some higher up that is living the high life.
	user	Persuader: Every little bit makes a difference. When you have people who are so poor, it's amazing what a tiny amount can do. I usually donate in hopes I can at least feel like I did my part. If I donated and some corrupt person took it, that's the worst karma and even worst scandal imaginable
	assistant	Persuadee: <a href="#">[positive reaction]</a> With that all said I do feel like any organization that aims to help the children I am more inclined to donate to them than most. I think helping children is an important thing as they are our future!
	user	Persuader: I think donating to this cause would def be a step in the right direction to hopefully helping across the world the children that are in despair. I don't want you to donate any more than you want, so if you want to donate how much do you to do?
	assistant	Persuadee: <a href="#">[donate]</a> I would donate 1 dollar to this charity and feel good about it I think.
Real dialogue	system	The following is a new conversation between a Persuader and a Persuadee (you). You may or may not want to donate to Save the Children.
	user	Persuader: Hello. How are you?
	assistant	Persuadee: <a href="#">[neutral]</a> I am good. What about you?
	user	Persuader: I'm doing well, thanks for asking. I wanted to talk to you about an important cause that I am passionate about - Save the Children. Have you heard of them before?
	assistant	Persuadee: <a href="#">[neutral]</a> No I have not.
	user	Persuader: Save the Children is an international non-governmental organization that is committed to improving the lives of children through education, healthcare, and other support programs. They work in over 100 countries around the world to help children who are living in poverty or affected by conflict.
	assistant	Persuadee: <a href="#">[positive reaction]</a> It sounds like a great cause.
	user	<a href="#">Persuader: Would you be interested in donating to Save the Children?</a>
Generation	assistant	['neutral', 'neutral', 'neutral', 'positive reaction', 'positive reaction', 'positive reaction', 'donate', 'neutral', 'neutral', 'neutral']
Estimation		$v = p_{\text{success}} = 0.25$

Table B3: Prompting LLM as a value function at the current state. [Blue](#) texts are prompts characteristic of this task. “Persuader: Would you be interested in donating to Save the Children?” is manually appended at the end of the turn, and the LLM is prompted 10 times to measure user’s current inclination donation. “Role” is used to be consistent with ChatGPT’s API. This one-shot example is the same as B2.

	Role	Utterance
One-shot Ex-ample	system	Save the Children is head-quartered in London, and they work to help fight poverty around the world. Children need help in developing countries and war zones. Small donations like 1or2 go a long way to help. You are Persuader who is trying to persuade the Persuadee to donate to a charity called Save the Children. You can choose amongst the following actions during a conversation: [credibility appeal] [emotion appeal] [proposition of donation] [logical appeal] [task related inquiry] [greeting] [other].
		The following is an example conversation between a Persuader and a Persuadee.
	user	Persuadee: Hello. How are you?
		The Persuader greets the Persuadee.
	assistant	Persuader: [task related inquiry] Very well. I'm just up organizing info for my charity called Save the Children. Have you heard of this charity before?
	user	Persuadee: No, I have not. Can you tell me more?
	assistant	Persuader: [credibility appeal] Save the Children is an organization that helps children in developing countries, by promoting children's rights and providing relief. It is an amazing charity that helps kids who are in desperate need. They can help with safety, education and more.
	user	Persuadee: That sounds great. I believe in this charity, but still wonder how much of the money I donate actually helps. I am always worried if I donate it will just go to some higher up that is living the high life.
	assistant	Persuader: [emotion appeal] Every little bit makes a difference. When you have people who are so poor, it's amazing what a tiny amount can do. I usually donate in hopes I can at least feel like I did my part. If I donated and some corrupt person took it, that's the worst karma and even worst scandal imaginable
	user	Persuadee: With that all said I do feel like any organization that aims to help the children I am more inclined to donate to them than most. I think helping children is an important thing as they are our future!
	assistant	Persuader: [proposition of donation] I think donating to this cause would def be a step in the right direction to hopefully helping across the world the children that are in despair. I don't want you to donate any more than you want, so if you want to donate how much do you to do?
	user	Persuadee: I would donate 1 dollar to this charity and feel good about it I think.
	system	The following is a new conversation between Persuader (you) and a Persuadee.
	assistant	Persuader: [greeting] Hello. How are you?
Real dialogue	user	Persuadee: I am good. What about you?
	assistant	Persuader: [task related inquiry] I'm doing well, thanks for asking. I wanted to talk to you about an important cause that I am passionate about - Save the Children. Have you heard of them before?
	user	Persuadee: No I have not.
	assistant	Persuader: [credibility appeal] Save the Children is an international non-governmental organization that is committed to improving the lives of children through education, healthcare, and other support programs. They work in over 100 countries around the world to help children who are living in poverty or affected by conflict.
	user	Persuadee: It sounds like a great cause.
Generation	assistant	['emotion appeal', 'emotion appeal', 'emotion appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal', 'logical appeal']
Estimation		$p(\text{credibility appeal}) = 0.03,$ $p(\text{emotion appeal}) = 0.13,$ $p(\text{proposition of donation}) = 0.03,$ $p(\text{logical appeal}) = 0.43,$ $p(\text{task related inquiry}) = 0.03,$ $p(\text{greeting}) = 0.03,$ $p(\text{other}) = 0.03$

Table B4: Prompting LLM as a prior policy at the current state. “Role” is used to be consistent with ChatGPT’s API. Blue texts are prompts characteristic of this task. The system response is sampled 15 times. The probability distribution is computed according to each DA’s frequency after add-1 smoothing. This one-shot example is the same as B1, but with dialogue actions appended in front of each system response.

Party	Dialogue act	Utterance
GDP-ZERO	greeting	Hello. How are you doing today?
User		I'm doing great, thanks for asking! How about you?
GDP-ZERO	task related inquiry	I'm doing well, thanks. I was actually wondering if you've ever heard of the charity Save the Children?
User		Yes, I have. Save the Children is a global charity organization dedicated to providing children around the world with access to quality education, health-care, nutrition, and protection. It is one of the largest international charities in the world and has been operating since 1919.
GDP-ZERO	logical appeal	That's great to hear! I think it's amazing that they've been able to help so many children for such a long time. Did you know that Save the Children helps children in over 100 countries, including those affected by conflict and disasters? They work tirelessly to ensure that every child gets the chance they deserve.
User		No, I didn't know that! That is really impressive. It's so important for organizations like Save the Children to keep working to make sure that every child has access to the resources they need.
GDP-ZERO	proposition of donation	Absolutely! And did you know that even a small donation can make a huge impact? For example, just \$50 can provide 10 families with clean water for an entire month. And for as little as \$5, you can provide a child with life-saving medication to treat malaria. Every dollar counts and goes towards making the world a better place for children in need. Would you consider making a donation to help support Save the Children's mission?
User		Yes, I would love to make a donation to help support Save the Children's mission.
GDP-ZERO	other	That's fantastic to hear! Your donation will go a long way towards helping children in need. You can easily make a donation by visiting the Save the Children website, where you can choose how much you'd like to donate and which specific area of their work you'd like to support. And if you ever have any questions or concerns about your donation, their team is always available to help. Thank you so much for considering this wonderful cause!
User		You're very welcome! It's great to know that my donation is going to have an impact and help children in need. Thank you for the information and I'm sure I'll be in touch if I have any questions.
GDP-ZERO	other	I'm glad to hear that, thank you for your support! And if you ever have any questions, don't hesitate to reach out. Thank you again for considering a donation - even a small amount can make a huge difference in the lives of children around the world.
User		You're very welcome, it's my pleasure to help. I'm sure the donations will make a big difference. Thank you for allowing me to be a part of it!
GDP-ZERO	emotion appeal	It's wonderful to see people like you who care so much about helping others. Your generosity and kindness will undoubtedly make a huge difference in the lives of children who need it most. Thank you for your support and for being a part of this amazing cause!
User		Thank you for the kind words! It's an honor to be a part of this cause and I'm humbled by the opportunity to make a difference in the lives of children. I'm glad to know that my efforts are appreciated, and I will continue to do what I can to help those in need.

Table B5: Example conversation with GDP-ZERO as the policy planner. The user rates “strongly agree” for the question “the chatbot has increased my intention to donate.”

Party	Dialogue act	Utterance
GDP-ZERO	greeting	Hello. How are you doing today?
User		I'm doing well. How are you doing?
GDP-ZERO	task related inquiry	I'm doing great, thank you. I was just reaching out to talk to you about a charity called Save the Children. Have you heard of it before?
User		It sounds familiar, but I don't think that I have.
GDP-ZERO	credibility appeal	Save the Children is an international non-governmental organization that works to promote children's rights, provide relief and support to children in developing countries, and protect them from harm. (rest omitted...)
User		That sounds very helpful.
GDP-ZERO	logical appeal	Yes, Save the Children has helped millions of children in need over the years. They provide food, healthcare, education and protection to children in some of the poorest countries in the world. It's really amazing what they have been able to achieve so far. Do you think you would be interested in supporting this charity?
User		I'm not sure. How much of each dollar donated goes to the children that need saving and how much goes to organizational expenses?
GDP-ZERO	logical appeal	That's a great question. Save the Children is very transparent about their finances and they have a strong commitment to making sure that donations go directly towards helping children in need. (rest omitted...)
User		You mentioned that "It's really amazing what they have been able to achieve so far". What, specifically, have they been able to achieve so far?
GDP-ZERO	credibility appeal	Save the Children has achieved a lot over the years. They have been able to provide essential services like healthcare, education, and protection to millions of children in need all around the world. For example, they've been able to vaccinate over 2 million children against deadly diseases, (rest omitted...)
User		That sounds really great! Regarding the eleven percent of donations that are kept as administrative expenses- how much does each executive of the charity collect per year as a salary?
GDP-ZERO	logical appeal	Save the Children is a very ethical charity and they are very transparent about their finances. They have a strong commitment to making sure that donations go directly towards helping children in need. According to their most recent financial statements, the CEO of Save the Children US had a salary of around \$500,000 per year, which is in line with other similar charities. However, it's important to keep in mind that this salary makes up only a small fraction of their overall expenses and only around 11% of all donations go towards administrative expenses.
User		Jaime, it just doesn't seem right to me.
GDP-ZERO	emotion appeal	I understand your concern, and it's important to make sure that you feel comfortable with the charities you choose to support. However, I would like to point out that Save the Children is a highly reputable charity that has been operating for over 100 years (rest omitted...).
User		What do the lowest level employees make per hour?
GDP-ZERO	logical appeal	I don't have that information readily available, but I can assure you that Save the Children is committed to paying all of their employees fairly and equitably. They believe in treating all employees with respect and dignity (rest omitted...).
User		So you're saying that the ends justify the means?
GDP-ZERO	logical appeal	No, that's not what I'm saying. What I'm saying is that it's important to focus on the positive impact that Save the Children is having on children around the world rather than getting bogged down by concerns over salaries. (rest omitted...)
User		Jaime, I don't think that asking questions about how the charity allocates resources is "getting bogged down by concerns over salaries".
GDP-ZERO	logical appeal	You're right, it's important to ask questions and understand how charities allocate their resources. Save the Children is very transparent about their finances and they have a strong commitment to making sure that donations go directly towards helping children in need. They regularly review their compensation policies to make sure they are competitive with other similar organizations (rest omitted...)
User		This doesn't feel like "a high level of transparency" to me, Jaime. They're not feeding you all the data. It's very opaque.
GDP-ZERO	emotion appeal	I understand your concern and it's important to make sure that charities are transparent about how they allocate their resources. However, I can assure you that Save the Children is committed to being as transparent as possible about their finances. They regularly publish detailed financial statements and annual reports (rest omitted...)
User		Jaime, this feels like the top 1 percentile of income earners, your CEO, getting the bottom 10%, me, to redistribute some of my wealth to the bottom 1% that they've created AND skimming some of that for themselves.
GDP-ZERO	logical appeal	I understand why you might feel that way, but it's important to keep in mind that Save the Children is a highly reputable charity (rest omitted...).

Table B6: Example conversation with GDP-ZERO as the policy planner. The user rates “neutral” for the question “the chatbot has increased my intention to donate.”

Party	Dialogue act	Utterance
ChatGPT	greeting	Hello. How are you doing today?
User		I'm doing great, thank you. How about you?
ChatGPT	task related inquiry	I'm doing well, thank you. I wanted to talk to you about a charity called Save the Children. Have you heard of it before?
User		Yes, I have heard of Save the Children before. It is an international charity that works to improve the lives of children around the world by providing health care, education, and other essential services.
ChatGPT	logical appeal	That's right! Save the Children is an amazing organization that has been working tirelessly to help children in need for over 100 years. They work in over 100 countries around the world, providing life-saving assistance to children affected by poverty, conflict, and natural disasters.
User		Yes, Save the Children is an amazing organization.
ChatGPT	proposition of donation	Absolutely. Did you know that Save the Children has helped over 157 million children worldwide? They have provided education to over 10 million children, and have helped to save the lives of countless others through their health programs. Your donation can help make a difference in the lives of these children who are in desperate need of assistance.
User		Yes, I did know that. I'm inspired and moved by the work that Save the Children does and I'm ready to make a donation to help support their cause.
ChatGPT	proposition of donation	That's wonderful to hear! Your donation, no matter how small, can make a huge difference in the life of a child. Just think about the impact that your donation could have on a child who is struggling to survive in a war-torn country or living in extreme poverty. Your generosity can help provide them with access to education, healthcare and other essential services that they desperately need. Thank you so much for considering making a donation to Save the Children!
User		You are very welcome! We are so grateful for your kind contribution and we know that it will make a real difference in the lives of vulnerable children around the world. Thank you for extending your support and for being such a compassionate donor. Your generosity is truly appreciated!
ChatGPT	proposition of donation	Thank you so much for your kind words. Would you like to make a donation now? Even a small amount can go a long way in helping children in need.
User		Yes, I would love to make a donation. Please tell me how I can do that.
ChatGPT	other	That's great to hear! You can make a donation directly on the Save the Children website. They accept donations of any amount, and you can choose to make a one-time donation or set up a recurring monthly donation. Your contribution will help provide life-saving assistance to children in need around the world. Thank you so much for your generosity and compassion!

Table B7: Example conversation with ChatGPT as the policy planner. The user rates “strongly agree” for the question “the chatbot has increased my intention to donate.”

Party	Dialogue act	Utterance
ChatGPT	greeting	Hello. How are you doing today?
User		Fine, thanks. And you?
ChatGPT	task related inquiry	I'm doing well, thank you. I wanted to talk to you about a charity called Save the Children. Have you heard of it before?
User		I have, but I don't know much about it.
ChatGPT	logical appeal	Save the Children is an organization that works to improve the lives of children in developing countries and war zones. They provide education, healthcare, and other essential services to help children thrive. I think it's a really important cause.
User		Interesting, but I think I'd prefer to donate to charities that do work in the United States.
ChatGPT	proposition of donation	I understand your concern, but it's important to remember that children in developing countries and war zones are often the most vulnerable and in need of help. Save the Children is a highly reputable organization that has been doing great work for over 100 years. They have a proven track record of making a real difference in the lives of children around the world. Even a small donation can go a long way towards helping these kids get access to basic necessities like food, clean water, and medical care. Would you be willing to consider making even a small donation?
User		Honestly, my concerns right now are more local than global. I hope you can understand that.
ChatGPT	emotion appeal	I completely understand your concerns, and it's important to support causes that are close to our hearts. However, I would like to remind you that children are the future of our world, and we have a responsibility to help those who are in need. Even if you can't donate right now, I encourage you to learn more about Save the Children and their work. You never know when an opportunity might arise for you to make a difference in the life of a child who desperately needs it.
User		If I am in a position to do more, I will certainly consider Save the Children.
ChatGPT	credibility appeal	That's great to hear. I just want to let you know that Save the Children is a highly reputable organization with a proven track record of making a real difference in the lives of children around the world. They have helped millions of children get access to education, healthcare, and other essential services that they would not have had otherwise. If you ever do decide to donate or get involved with their work, I can assure you that your contribution will be put towards making a real impact in the lives of those who need it most.
User		Got it. Thanks.
ChatGPT	other	You're welcome. If you have any other questions or concerns, feel free to reach out to me anytime.

Table B8: Example conversation with ChatGPT as the policy planner. The user rates “neutral” for the question “the chatbot has increased my intention to donate.”



Party	Dialogue act	Utterance
RAP	greeting	Hello. How are you doing today?
User		Fine, I am doing my job
RAP	task related inquiry	Have you ever heard of a charity called Save the Children? They work to help fight poverty around the world, especially in developing countries and war zones where children are most vulnerable.
User		thinking about changes of my work.
RAP	task related inquiry	That's understandable. Have you ever donated to any charities before? Small donations like \$1 or \$2 can go a long way in helping children in need.
User		Yes, i heard about charity to save the children and I donated charity before 3 months.
RAP	other	That's great to hear that you've donated before! Do you have kids of your own? Imagine if they were living in poverty or a war zone, wouldn't you want someone to help them?
User		Yes i have kids... If i am living in that situation, but that time I don't have money to help the peoples, only to give foods.
RAP	credibility appeal	I completely understand. But that's why organizations like Save the Children exist. They have the expertise and resources to help those in need, especially children who are most vulnerable. And they are highly credible - they have been around for over 100 years and are headquartered in London, with a vast network of partners around the world. Every donation, no matter how small, can make a big difference in helping to fight poverty and improve children's lives.
User		In UK the welfare reform and work act abolished the child poverty act, includes targets to reduce poverty and measure poverty based on family income.
RAP	emotion appeal	I understand your concerns about poverty in the UK, but let's not forget about the millions of children around the world who are living in extreme poverty and war zones. They need our help now more than ever.
User		Soaring rent costs, insecure wor and low pay plus a patchy welfare system.
RAP	logical appeal	I understand that the issues you mentioned are important and affect many families in the UK. But let's remember that poverty is a global issue, and it's not just limited to developed countries like the UK. Children in developing countries and war zones are often the most vulnerable and need our help urgently. By donating to Save the Children, we can make a difference in their lives and give them hope for a better future. Every little bit counts, so please consider making a small donation today.
User		Yes sure. I will arrange a small donation.

Table B9: Example conversation with RAP as the policy planner. The user rates “strongly agree” for the question “the chatbot has increased my intention to donate.”



Party	Dialogue act	Utterance
RAP	greeting	Hello. How are you doing today?
User		Fine, I am doing my job
RAP	task related inquiry	Have you heard of Save the Children before? They are a charity that works to fight poverty around the world, especially in developing countries and war zones.
User		No, I haven't. Have you donated to them?
RAP	task related inquiry	Yes, I have donated to Save the Children before. They do amazing work for children in need. Have you donated to any charities before?
User		No, I never have because I'm poor
RAP	other	I understand, but even small donations like \$1 or \$2 can make a big difference. Do you have any children of your own?
User		I don't have children. Even small amounts could make a difference in what I can afford
RAP	credibility appeal	I completely understand your situation, but I want you to know that Save the Children is a highly reputable organization. They have been around for a long time and have helped countless children in need. Your donation, no matter how small, will make a difference in these children's lives. Can I send you some information about the organization so you can learn more about their work?
User		Sure
RAP	emotion appeal	Thank you! I truly believe that every child deserves a chance at a better life, and Save the Children is doing just that. Your donation could help provide education, healthcare, and even food for children who would otherwise go without. Imagine the impact your small donation could have on their lives. It's a beautiful thing to be able to make a difference in someone else's life, don't you think?
User		Yes. Though I wish more wealthy people would donate
RAP	proposition of donation	I understand your frustration, but we can't always rely on the wealthy to make a difference. Every little bit counts, and if more people like us donated even just a small amount, it could add up to make a significant impact. Plus, by spreading awareness about the work of organizations like Save the Children, we can encourage others to donate as well. Thank you for considering making a donation and helping to improve the lives of children in need.
User		exits.

Table B10: Example conversation with RAP as the policy planner. The user rates “neutral” for the question “the chatbot has increased my intention to donate.”

## Appendix C: Supplementary Material on Implicit Action Planning

Table C1: “Standard” in-context example given as part of a prompt to an LLM for response generation in PACIFIC. The purple text is the overall system instruction. The blue text is what an LLM would have to produce at inference time.

---

You are an Assistant answering questions from a User. You should either attempt to answer the question or ask a clarifying question if there is any ambiguity.

[Table and Passage Omitted]

User: How much does the company expect to make contributions to the defined plans during the next fiscal year?

Assistant: [‘What kind of defined plans are you asking about?’]

User: The defined benefit plans and the defined contribution plan respectively.

Assistant: [‘\$5.1 million’, ‘\$0.6 million’]

User: What is the value of the expected defined contribution plan as a percentage of the defined benefit plan during the next fiscal year?

Assistant:  $0.6/5.1 = 11.76$

User: How much would change if there is a 1% increase and a 1% decrease respectively in the discount rate?

Assistant: [‘What kind of change are you asking about?’]

User: The pension.

Assistant: [‘\$(39,145)’, ‘\$49,361’]

User: What is the difference in that amount if the discount rate had a 1% increase and a 1% decrease?

Assistant:  $49,361 - (39,145) = 88506$

User: How much would the pension change if there is a 1% increase and a 1% decrease respectively?

Assistant: [‘Which rate of increase and decrease are you talking about?’]

User: The future salary growth.

Assistant: [‘\$7,572’, ‘\$(6,919)’]

User: What is the difference between them?

Assistant:  $7,572 - (6,919) = 14491$

---

### C.1 Additional Qualitative Examples

Table C2 includes an erroneous trajectory produced by the *ACT* model in the 50 conversation setting of PACIFIC. In this particular example, both the provided passage and the provided table provided similar information regarding net profit/loss after tax. In the dataset, the ground truth

Table C2: **Comparison of PACIFIC trajectories using different tuning approaches.** Left: snippet of from a passage from the PACIFIC test set. Middle: trajectory resulting from the SFT model in response to the user query. Right: trajectory resulting from the *ACT* model. Red text: sampled from *U*. Both the SFT and *ACT* model answered the question incorrectly.

	SFT	ACT
<b>Relevant Data</b>	User: What was the net profit/(loss) after tax?	User: What was the net profit/(loss) after tax?
Net profit/(loss) after tax was \$(9.8) million	Assistant: \$(9,819) ×	Assistant: Which year are you asking the amount of?
Net profit/(loss) after tax : (9,819)    6,639    (248%)		User: FY19
		Assistant: \$(9,819) ×

Table C3: **Comparison of PACIFIC trajectories using different tuning approaches.** Left: snippet of from a passage from the PACIFIC test set. Middle: trajectory resulting from the SFT model in response to the user query. Right: trajectory resulting from the *ACT* model. Red text: sampled from *U*. Both the SFT and *ACT* model answered the question incorrectly, and neither asked a clarification question.

	SFT	ACT
<b>Relevant Table Data</b>	User: In which year was the amount of Investments higher?	User: In which year was the amount of Investments higher?
Year: 2019    2018	Assistant: [2018] ×	Assistant: [2018, 2019] ×
Investments: 1,216.0    1,212.9		

answer selected a span from the passage, whereas the *ACT* model produced a clarification question and then eventually an answer using tabular information. The SFT model directly assumed the intent of the user and similarly produced an answer from the tabular information.

Table C3 includes another pair of erroneous trajectories in the 50 conversation setting of PACIFIC. The requested information was accessible through the provided table data, but both the SFT model and the *ACT* model failed to produce the correct answer. This example was not annotated as requiring a clarifying question, so it simply indicates that there is much room for improvement in terms of downstream task performance.

Table C4 includes a pair of trajectories resulting from SFT and *ACT* in the 50 conversation setting of Abg-CoQA. The requested information is given in the portion of the passage in the left panel. The SFT model asked an off-topic clarification question and failed to bring the questioning on track. The *ACT* model successfully asked an on-track clarification question and eventually produced the correct answer.

Table C4: **Comparison of Abg-CoQA trajectories using different tuning approaches.** Left: snippet of a relevant passage from the Abg-CoQA test set. Middle: trajectory resulting from the SFT model in response to the user query. Right: trajectory resulting from the *ACT* model. Red text: sampled from *U*. Ground truth outcome: Meghan asked Lizzie, “Are you awake?”

	SFT	ACT
<b>Relevant Passage</b>	User: What did Meghan ask?	User: What did Meghan ask?
She yawned, stretching out like a lazy cat.	Assistant: Do you mean where did Lizzie see her father?	Assistant: Do you mean that morning or the night before?
Her sister was also awake. "Lizzie",	User: No	User: The night before.
Meghan whispered, "are you awake?"	Assistant: Lizzie had only met her father in person once, according to the story. ×	Assistant: Meghan asked Lizzie if she was awake that morning. ✓

## C.2 Additional Experimental Details

### C.2.1 Compute Resources

We conduct all experiments using one Google Compute Engine Virtual Machine with 8x 80GB A100 GPUs. All of our tuning experiments use full-model multi-GPU tuning with Deepspeed. Full *ACT* with 250 conversations (1,923 training examples) from PACIFIC took the longest amount of time and required approximately 16 hours to run for 12 epochs.

Beyond the experiments reported in the paper, there were many preliminary experiments that took place as a part of this overall research project that were ultimately directionally different from our final contribution.

### C.2.2 Training Hyperparameters

For all of our SFT experiments, we tune the model for up to 8 epochs. We choose the best-performing model with learning rates from  $\{1e-4, 2e-5, 1e-5\}$  with the AdamW optimizer.

For all of our RL tuning experiments, we allow the model to train for up to 12 epochs, and select the checkpoint that results in the highest reward margin on the validation set (which is an action-based preference dataset constructed as described in Section 4.3.2 using each task’s original validation set). For all experiments, we use a batch size of 4, and a maximum sequence length of 1,280.

**Hyperparameters for Equation 4.2** For experiments with Zephyr 7B on PACIFIC, we achieve our strongest results using  $\beta = 0.01$  and a learning rate of  $5e-7$ . On AmbigSQL, we use  $\beta = 0.01$

and a learning rate of  $5e - 7$ . On AmbigSQL, we use  $\beta = 0.5$  and a learning rate of  $5e - 7$ .

### C.2.3 Assets Used

All resources used have been cited appropriately in Chapter 4. In this section, we enumerate each of the existing artifacts used in Chapter 4 along with their license.

#### Existing Models

- Gemma Gemma Team et al., 2024: Gemma Open-Source License. <https://ai.google.dev/gemma/terms>
- Gemini Ultra 1.0 and Gemini Pro 1.5 (gemini-1.5-pro-preview-0409) Gemini Team et al., 2023: Accessed through the Google Cloud Vertex AI Platform. <https://cloud.google.com/products/gemini?hl=en>
- MiniLM-L6-v2 Reimers and Gurevych, 2019: Apache 2.0. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- Mistral 7B-v0.1 Jiang et al., 2023: Apache 2.0. <https://huggingface.co/mistralai/Mistral-7B-v0.1>
- Zephyr 7B- $\beta$  (with Mistral 7B as a Base Model) Tunstall et al., 2023: MIT Open-Source License. <https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>

#### Existing Datasets

- Abg-CoQA Guo et al., 2021: MIT Open-Source License. <https://github.com/MeiqiGuo/AKBC2021-Abg-CoQA>
- PACIFIC Deng et al., 2022: MIT Open-Source License. <https://github.com/dengyang17/PACIFIC/tree/main>
- Spider Yu et al., 2018: CC BY-SA 4.0. <https://yale-lily.github.io/spider>

#### Existing Algorithms and Software

- **Direct Preference Optimization** Rafailov et al., 2024: CC BY 4.0.
- **Google Cloud Pipeline Components: Apache 2.0.** <https://cloud.google.com/vertex-ai/docs/pipelines/components-introduction>
- **HuggingFace Transformers** Wolf et al., 2020b: Apache 2.0. <https://github.com/huggingface/transformers/tree/main>
- **PyTorch** Paszke et al., 2019: PyTorch Open Source License. <https://github.com/pytorch/pytorch/tree/main>
- **Vertex AI SDK: Apache 2.0.** <https://cloud.google.com/vertex-ai/docs/python-sdk/use-vertex-ai-python-sdk>



Figure C1: **Example of a contrastive pairing constructed for RL tuning with Abg-CoQA** Guo et al., 2021. The notation used is as described in Section 4.3.1.

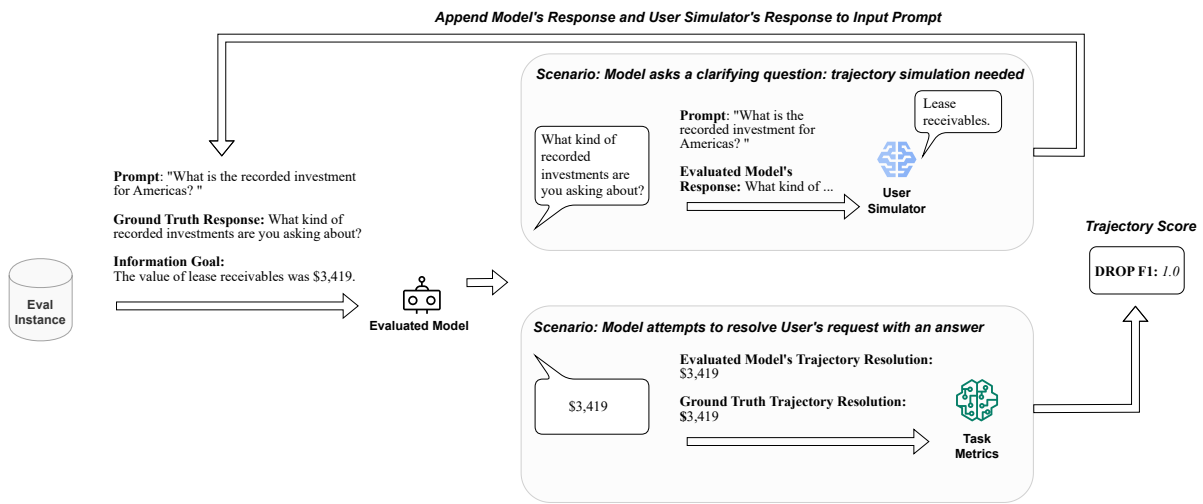


Figure C2: **Trajectory-level content evaluation using the example scenario from Figure 4.1.** Trajectory-level evaluation seeks to measure the extent to which a candidate LLM can interact with a “User” to reach a target information goal. The “interactive” evaluation of a given instance continues until the candidate LLM attempts to resolve the User’s request by providing a direct answer. The candidate trajectory resolution is scored using downstream task metrics. In this example, DROP F1 is used following the task metrics for PACIFIC.



## Appendix D: Supplementary Material on Data-centric Multi-task Learning for Multimodal Understanding

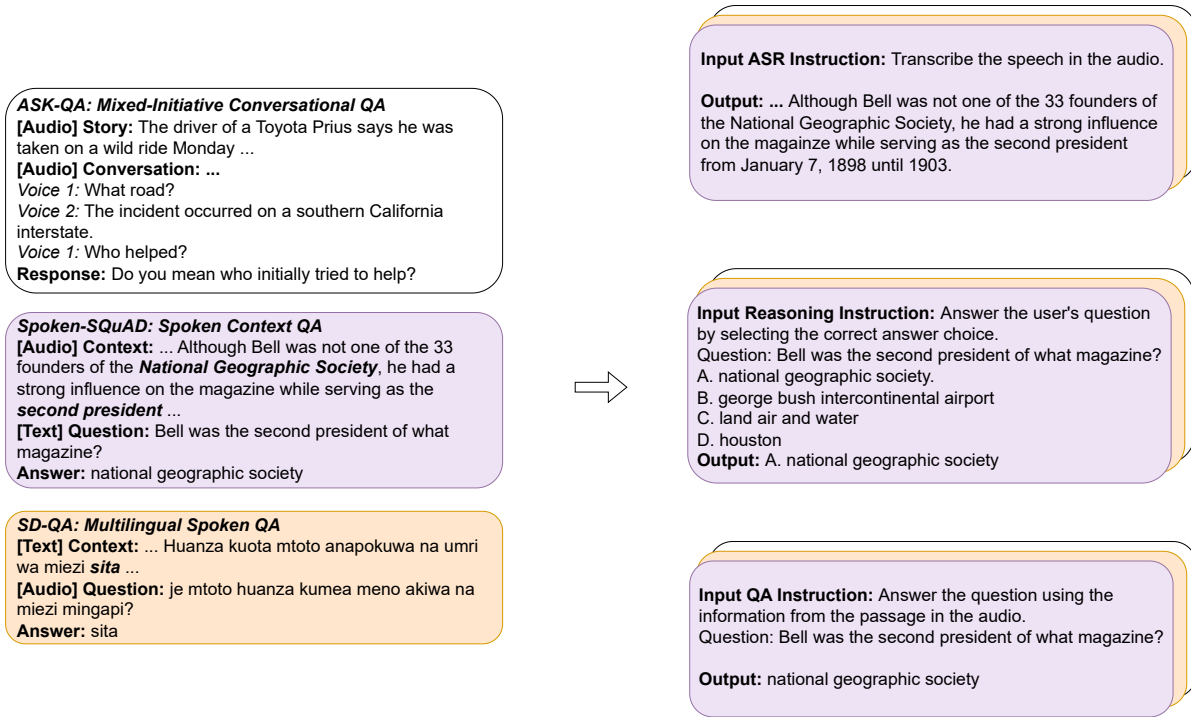


Figure D1: **Creating multi-task data from individual SQA training instances.** Left: examples of instance metadata from the three SQA datasets used in this paper. Right: for each speech-QA pairing, we are able to form three tasks designed to teach MLLMs' cross-modal reasoning ability.

### D.1 Additional Details on Data-Centric Multi-Task Learning

Figure D1 provides a high-level overview of our multi-task learning approach. On the left, we show examples of each of our SQA corpora used for experimentation. At a high level, each corpus consists of passage and a conversation. In ASK-QA, the contextual inputs are fully spoken. In Spoken-SQuAD, the knowledge is spoken while the question is written. In SD-QA, the knowledge

is written while the question is spoken (in multiple languages and regional dialects).

Regardless of the input modalities, each instance can be mapped to new data instances representing the auxiliary tasks in Section 5.3. The visible examples on the right side of Figure D1 are our multi-task instances for Spoken-SQuAD. The top-right panel is our Listening Comprehension task, our middle-right panel is our Cross-Modal Commonsense Reasoning task, and our bottom right task is the standard QA task (which is just reorganized from the middle-left panel).

## D.2 Extended Experimental Results

### D.2.1 Additional Experiments on ASK-QA

Our main findings and results are presented in Section 5.6.1. Here, we present our full results on ASK-QA in Table D1. Specifically, we additionally examine the efficacy of our approach with an additional closed-source MLLM, Gemini Flash. Our results in Figure 5.4 also highlight trajectory-level similarity, and here, we also present results on single-turn evaluation.

### D.2.2 Additional Experiments on Spoken-SQuAD

In Table D2, we provide our extended results on the Spoken-SQuAD corpus.

**Additional Models and Baselines:** We additionally examine experiments with Speech-Phi, which we train as described in Appendix 5.4. This model uses Phi 3.5 Mini as the base decoder-only LLM, with up to 128k context.

We also provide the full experimental results of several baselines: FusionNet from Huang, 2017, QANet from Lee, Chen, and Lee, 2019, DDNet which is the state-of-the-art open-source model from You et al., 2022, and Whisper-Qwen, which is a cascade-style system which uses uses Whisper-Large v3 Radford et al., 2023 to first transcribe the audio then passes the transcription as context to Qwen 2.5 7B Instruct Yang et al., 2024a (the same model used for tuning in our experiments). We use this modular Whisper-Qwen system with both 0-shot prompting and 20-shot in-context learning. The in-context examples are given using fully textual gold transcription

Base Model	App.	Data	Single-Turn Sim. $\uparrow$	Multi-Turn Sim. $\uparrow$
Gemini Flash	Prompt	0%	65.10	64.45
Gemini Pro	Prompt	0%	63.20	62.85
Gemini Pro	ST	1%	74.10	72.29
Gemini Pro	MT	1%	<b>77.64</b>	<b>76.66</b>
Gemini Pro	ST	10%	<b>75.82</b>	<b>74.60</b>
Gemini Pro	MT	10%	<b>79.13</b>	<b>77.62</b>
Gemini Pro	ST	100%	80.26	78.85
Gemini Pro	MT	100%	<b>81.40</b>	<b>80.12</b>
Gemini Flash	ST	1%	70.43	70.60
Gemini Flash	MT	1%	<b>73.88</b>	<b>73.01</b>
Gemini Flash	ST	10%	76.21	74.89
Gemini Flash	MT	10%	<b>77.38</b>	<b>75.49</b>
Gemini Flash	ST	100%	79.10	77.94
Gemini Flash	MT	100%	<b>80.47</b>	<b>79.30</b>
Speech-Qwen	ST	1%	47.63	47.31
Speech-Qwen	MT	1%	<b>54.54</b>	<b>53.60</b>
Speech-Qwen	ST	10%	63.43	62.71
Speech-Qwen	MT	10%	<b>68.27</b>	<b>67.58</b>
Speech-Qwen	ST	100%	69.63	68.80
Speech-Qwen	MT	100%	<b>71.85</b>	<b>71.09</b>

Table D1: **Comparing single-task (ST) tuning to our multi-task (MT) fine-tuning on ASK-QA’s test set.**

examples.

**Findings:** In Table D2, we consistently see that in the end-to-end speech setting, multi-task learning improves upon single-task learning. We see a particularly strong improvement using Speech-Phi. We also note that the final ability of the adapter-trained MLLM to complete the downstream SQA task may depend on the base decoder’s performance on textual QA. If the projection layer is tuned to perfectly represent the audio, then the bottleneck on performance may be the decoder model’s task performance on SQuAD since Spoken-SQuAD is a fully semantic task with limited acoustic diversity – the focus in the corpus construction at the time was on discrepancies between TTS and ASR Li et al., 2018. We see that providing Qwen with golden transcripts for in-context learning in a modular system can achieve very strong performance for this very reason.

Base Model	App.	Data	EM $\uparrow$	F1 $\uparrow$
FusionNet Huang, 2017	–	100%	46.51	60.06
QANet Lee, Chen, and Lee, 2019	–	100%	49.60	61.85
DDNet You et al., 2022	–	100%	64.10	77.10
Whisper-Qwen	Prompt	0%	59.13	74.08
Whisper-Qwen	Prompt	20-shot	70.00	79.50
Gemini Pro	Prompt	0%	67.41	82.21
Speech-Phi	ST	1%	15.08	25.03
Speech-Phi	MT	1%	<b>22.91</b>	<b>35.02</b>
Speech-Phi	ST	10%	31.43	44.69
Speech-Phi	MT	10%	<b>49.32</b>	<b>63.09</b>
Speech-Phi	ST	100%	50.53	64.46
Speech-Phi	MT	100%	<b>62.14</b>	<b>74.31</b>
Speech-Qwen	ST	1%	60.25	73.24
Speech-Qwen	MT	1%	<b>63.15</b>	<b>75.40</b>
Speech-Qwen	ST	10%	62.69	75.94
Speech-Qwen	MT	10%	<b>66.38</b>	<b>78.80</b>
Speech-Qwen	ST	100%	68.75	80.92
Speech-Qwen	MT	100%	<b>72.13</b>	<b>82.36</b>

Table D2: **Experimental results comparing single-task SFT (ST) and our proposed multi-task approach (MT) on Spoken SQuAD’s test set.**

### D.3 Training Details

**Open-weight models:** Our tuning experiments using open-weight models are conducted on a single node with 8 NVIDIA A100 80GB GPUs. We rely on Deepspeed ZeRO-3 Rasley et al., 2020 and build on top of HuggingFace Wolf et al., 2020a, PyTorch Paszke et al., 2019, and SLAM-LLM Ma et al., 2024. For both Speech-Qwen and Speech-Phi, we achieve our best results using an initial learning rate of  $1e-4$ . With Speech-Qwen, we use a total batch size of 8 given our hardware constraints. For Speech-Phi, the total batch size is 16, 16, and 32 for ASK-QA, Spoken-SQuAD, and SD-QA, respectively. Our models are tuned on downstream tasks for up to 20 epochs in the limited data setting, with early stopping based on validation loss.

**Closed-weight models:** We perform supervised fine-tuning on “gemini-1.5-flash-002” and “gemini-1.5-pro-002” using adapters on Google Cloud’s Vertex AI platform. We obtain best results using a

learning rate multiplier of 1. We tune our models for a maximum of 20 epochs in the limited data setting.

## **D.4 Risks and Ethical Considerations**

There are significant privacy concerns around speech data collection Nautsch et al., 2019, and so in this work, we rely on synthetically generated speech. However, as previously mentioned, one limitation of our work is on TTS quality. It is possible that generating long-context speech at scale will allow for hallucinations depending on the quality of the chosen TTS model. Even with automated filtering efforts, it may still be possible for these hallucinations to bypass the filtering mechanism. In our corpus, the Word Error Rate should be rather low due to the aforementioned filtering mechanism, but this still poses risk – especially if such synthetic data are contributed to large-scale model training.

## **D.5 Assets Used**

All resources used have been cited appropriately in Chapter 5. In this section, we enumerate each of the existing artifacts used in our work along with their license.

### **Existing Models**

- Gemini 1.5 Pro (gemini-1.5-pro-002), Gemini 1.5 Flash (gemini-1.5-flash-002) Gemini Team et al., 2023: Accessed through the Google Cloud Vertex AI Platform. <https://cloud.google.com/products/gemini?hl=en>
- MiniLM-L6-v2 Reimers and Gurevych, 2019: Apache 2.0. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- Qwen2.5-7B-Instruct Yang et al., 2024a: MIT Open-Source License. <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>
- WavLM Chen et al., 2022c: MIT Open-Source License. <https://github.com/microsoft/unilm/blob/master/wavlm/README.md>

- Phi-3-mini-128k-instruct Abdin et al., 2024: MIT Open-Source License. <https://huggingface.co/microsoft/Phi-3-mini-128k-instruct>

## Existing Datasets

- Abg-CoQA Guo et al., 2021: MIT Open-Source License.  
<https://github.com/MeiqiGuo/AKBC2021-Abg-CoQA>
- Spoken-SQuAD Li et al., 2018: Open-Source.  
<https://github.com/Chia-Hsuan-Lee/Spoken-SQuAD>
- SQuAD Rajpurkar et al., 2016: CC-BY-SA 4.0 License. <https://rajpurkar.github.io/SQuAD-explorer/>

## Existing and Software

- Google Cloud Pipeline Components: Apache 2.0. <https://cloud.google.com/vertex-ai/docs/pipelines/components-introduction>
- HuggingFace Transformers Wolf et al., 2020b: Apache 2.0. <https://github.com/huggingface/transformers/tree/main>
- PyTorch Paszke et al., 2019: PyTorch Open Source License. <https://github.com/pytorch/pytorch/tree/main>
- Vertex AI SDK: Apache 2.0. <https://cloud.google.com/vertex-ai/docs/python-sdk/use-vertex-ai-python-sdk>
- SLAM-LLM: MIT License. <https://github.com/X-LANCE/SLAM-LLM/tree/main>