ProMISe: A Proactive Multi-turn Dialogue Dataset for Information-seeking Intent Resolution

Yash Parag Butala[♠†], **Siddhant Garg**[★↑], **Pratyay Banerjee**[♣], **Amita Misra**[♣]
[♠]Carnegie Mellon University [★]Meta AI [♣]Amazon Alexa AI
ypb@cs.cmu.edu
sidgarg@meta.com
{pratyay, misrami}@amazon.com

Abstract

Users of AI-based virtual assistants and search systems encounter challenges in articulating their intents while seeking information on unfamiliar topics, possibly due to complexity of the user's intent or the lack of meta-information on the topic. We posit that an iterative suggested question-answering (SQA) conversation can improve the trade-off between the satisfaction of the user's intent while keeping the information exchange natural and cognitive load of the interaction minimal on the users. In this paper, we evaluate a novel setting ProMISe by means of a sequence of interactions between a user, having a predefined information-seeking intent, and an agent that generates a set of SQA pairs at each step to aid the user to get closer to their intent. We simulate this two-player setting to create a multi-turn conversational dataset of SQAs and user choices (1025 dialogues comprising 4453 turns and 17812 SQAs) using human-feedback, chain-of-thought prompting and web-retrieval augmented large language models. We evaluate the quality of the SQs in the dataset on attributes such as diversity, specificity, grounding, etc, and benchmark the performance of different language models for the task of replicating user behavior.

1 Introduction

Users of AI-based virtual assistants and search systems such as Google Search, Alexa, Bing, etc. often face challenges in effectively satisfying their information-seeking intents, especially on unfamiliar topics. This stems from a combination of (i) the inability of the user to formulate the appropriate question(s) for the agent owing to the complexity of the intent, (ii) the user lacking meta-information on an unfamiliar topic that is required to phrase the appropriate question(s) to the agent, and (iii) the agent's response being long, complicated and cognitively challenging for the user to process.

[†]Work done during internship at Amazon Alexa AI
[↑]Work completed at Amazon Alexa AI

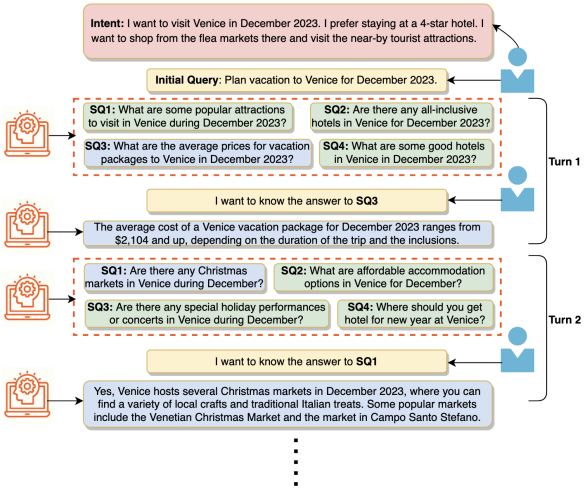

Figure 1: An instantiation of the ProMISe setting: Proactive Multi-turn Information-Seeking Dialogue

To bridge the gap between intent satisfaction, exploration of topics unfamiliar to the user and keeping the information exchange cognitively easy for the users to understand, several popular search engines like Google, Bing, etc. have a "Related Questions/People Also Ask" feature that assists users by providing related queries and web-snippets. However, these are restricted to a single-turn information exchange with the user and fail end-to-end to fully encompass the information-seeking intent of the user. The agent does not have a systematic approach to satisfy the user needs by means of exploring the unfamiliar topic, and continues to generate duplicate questions on aspects of the user intent that have previously been addressed (STAT, 2016). Additionally, in cases when the user intent is complex (spanning diverse facets of a topic), a single all-encompassing response may increase the cognitive load (Sweller, 2011) of the user's understanding of the information exchange.

Previously, Task-Oriented Dialogue (TOD) systems have aimed to help users resolve their intents by means of slot-filling-based frameworks in closed domains eg: MultiWOZ (Eric et al., 2019), STAR (Mosig et al., 2020). However, this restricts their

applications to surrogate real-world scenarios (Lee et al., 2023) and limits their scope for exploration of unfamiliar topics. In contrast, proactive dialogue systems have the capability of leading the conversation direction towards achieving predefined targets or fulfilling certain goals from the system side. While many intelligent systems overlook the property of pro-activity (Deng et al., 2023a), we argue that this is crucial for the domain of satisfying information-seeking intents on unfamiliar topics. A key complexity in this domain is the ever-evolving user intent over the interaction with the agent, as more information on the topic is explored. For example: a user without any prior knowledge on drones might enrich their initial intent of 'Buy a drone under $100' to 'Buy a drone under $100 with a range of 500m and camera resolution of 12MP' as they explore more information on this topic.

To make the interaction with agents more pragmatic and proactive, while keeping the cognitive load of the interaction minimal on the users, we propose a new setting (**ProMISe**: **Pro**active **M**ulti-turn **I**nformation-**Se**eking Dialogue) that involves breaking the user-agent interaction into a conversation of multiple turns where the agent attempts to answer atomic aspects of the user's intent. At each turn of the conversation, the agent generates a set of suggested questions (SQs) and the user selects the most helpful SQ. We empirically observe improved trade-off between satisfaction of user intents, exploration of unfamiliar topics and cognitive load of the interaction on users in the ProMISe setting, when compared to multiple existing interaction settings like single turn QA exchange, single turn SQ exchange or multi-turn free-form conversation with the agent (refer Section 3 for details).

We illustrate a sample conversation under the ProMISe setting in Fig 1 where the user has a predefined intent to fulfill and begins the conversation with an AI-agent by asking a simple question related to the intent. The agent then generates a set of relevant SQs for the user to chose from. At every step/turn, the user can choose one of the relevant SQs from the agent to get the corresponding answer which can help in bridging the gap towards resolving the intent. We curate a dataset for ProMISe by simulating user intents and initial queries from popular Google Trends topics by prompting large language models (LLMs). We simulate the agent to generate SQs using web-retrieval augmented generation. We devise an annotation task to simulate user choices during each turn of the conversation (choosing one of the SQs or indicating that the information need has been satisfied). We analyze the quality of the SQA generation in the dataset on attributes such as well-formedness, relevance, diversity, specificity and web-grounding.

Using the collected dataset, we aim to evaluate how effectively language models can mimic the reasoning of users (humans) in carrying forward an information-seeking exchange with an agent to satisfy an intent. Simulating users effectively is an important paradigm in modern-day NLP research, as this can improve the velocity of collection of dialogue datasets and facilitate privacy-aware evaluations (Zamani et al., 2023). We benchmark the abilities of several popular LLMs such as ChatGPT (OpenAI, 2023), LLaMA (Touvron et al., 2023), MPT (Team, 2023), Vicuna (Zheng et al., 2023), Dolly (Conover et al., 2023) and Falcon (Almazrouei et al., 2023) to replicate user behavior through explanation-guided action generation. Empirically, we observe a significant performance gap between popular LLMs and humans for this task of simulating users with an intent.

We believe that the ProMISe dataset and methodology for collecting it (containing user simulations with information-seeking intents, along with SQAs) can be beneficial to the broader NLP community and researchers working in real-world applications in domains of Question-Answering, Dialogue, Conversational Agents and Language Models. We make the code and the dataset publicly available through our GitHub repository[1]. The key contributions of the paper are summarized below:

- We propose and evaluate a novel interaction setting with intelligent assistive agents termed as **ProMISe** (**Pro**active **M**ulti-turn **I**nformation-**Se**eking) to fulfill information-seeking user requests in an end-to-end manner.
- We create a high quality dataset of 1025 dialogues (containing 4453 turns and 17812 SQAs), created using human feedback for user-simulation aimed at satisfying open-domain real-world user intents using web retrieval-augmented generation with LLMs.
- We benchmark and perform an in-depth analysis of the performance of popular LLMs for the task of simulating user-behavior on the dataset.

[1] https://github.com/amazon-science/promise

2 Related Work

Proactive Conversational Systems Several research studies have explored the topic of clarification question generation (Kumar and Black, 2020; Majumder et al., 2021) and question disambiguation (Gao et al., 2021; Min et al., 2020). Aliannejadi et al. (2021) proposed the ClariQ dataset of open domain dialogue for predicting and generating clarification questions. Guo et al. (2021) and Deng et al. (2022) propose datasets (AbgCoQA and PACIFIC respectively) in this domain for disambiguity prediction, clarification question generation and conversational QA.

Zhang et al. (2018) proposed the proactive 'System Ask User Respond' setting for improving conversational search. (Deng et al., 2021; Zhang et al., 2022; Zhao1 et al., 2023) acquire user preference through multiple turns of interactions using RL-based conversational recommendation systems. These works, however, are constrained to the product domain and only focus on one feature per turn. Zhong et al. (2021) propose a keyword-guided conversational model for reaching a target keyword. Our work extends this by enhancing the complexity of user intent from keywords to open-domain natural language constructs. Gaur et al. (2021) propose a RL-based approach for generating information-seeking questions starting from short initial user queries. However, this approach is restricted to single-turn SQ generation, and does not contain answers to the generated SQs. SeeKeR (Shuster et al., 2022) highlights that search and knowledge augmented dialogue outperforms previous state-of-the-art models in open-domain knowledge-grounded conversations on aspects of consistency, knowledge and per-turn engagement.

LLMs and Dialogue Large Language Models (LLMs) have shown state-of-the-art reasoning abilities, along with zero-shot and few-shot generalization capabilities (Kojima et al., 2023; Wei et al., 2023). Internet-augmented dialogue generation (Komeili et al., 2022) proposes an approach to generate a web search query based on the dialogue and using the search results to condition the LLM's output. Liu et al. (2022) propose multi-stage prompting for knowledgeable dialogue generation that increases knowledge, relevance and engagement without fine-tuning the model. Deng et al. (2023b) propose the Proactive Chain-of-Thought prompting scheme to augment LLMs with goal planning and generating clarification questions. Terragni et al. (2023) use in-context learning to generate diverse questions in task oriented dialogues based on user goals. Wang et al. (2023) use LLMs for planning and reasoning to provide a more personalized and engaging experience for the user query.

3 The ProMISe Setting

We first formally define the Proactive Multi-turn Dialogue for Information-seeking Intent Resolution setting. Consider an interaction between a user **U** and an AI-agent **A**. The user **U** has an information-seeking intent **I**. Based on meta-information that the user has on the topic of **I**, the user formulates an initial question q_0 to ask **A** to initiate the information-seeking dialogue. At each turn i, the agent **A** uses the conversation history with **U** to create a set of L suggested questions (SQs) S^i: $\{s^i_1, s^i_2, \ldots, s^i_L\}$ that may be relevant for the user. The user then chooses SQ s^i_u from the set S^i of SQs created by **A** in turn i, or indicates that none of S^i are relevant to their intent. After making the choice, **A** provides the answer to s^i_u to **U**. At the end of each turn, **U** indicates if their original information-seeking intent **I** has been satisfied or if they still need more information on some aspects of **I**. The conversation continues till the user signals that their information-seeking intent has been satisfied. We illustrate the ProMISe setting in Fig 2. We describe information available to **U** and **A** below:

Agent: At each turn i of the conversation, the agent **A** has access to the conversation history with the user including the initial question q_0, and previously generated SQs and choices made by the user: $\{S^1, s^1_u\}$, $\{S^2, s^2_u\}, \ldots, \{S^{i-1}, s^{i-1}_u\}$. Note that **A** does not have access to the information-seeking intent **I** of the user.

User: At each turn i of the conversation, the user **U** makes a choice s^i_u from the set S^i of SQs created by **A** using the previous conversation history with the agent including: the initial question q_0, previously generated SQs and choices made by the user: $\{S^1, s^1_u\}$, $\{S^2, s^2_u\}, \ldots, \{S^{i-1}, s^{i-1}_u\}$ and the information-seeking intent **I**.

3.1 ProMISe v/s Existing Interaction Settings

ProMISE enables proactive concept exploration, with the agent getting feedback from both the selected and non-selected questions to reach conclusions on what next set of information would be useful for the user. To empirically highlight the benefits of this setting, we conduct user studies to

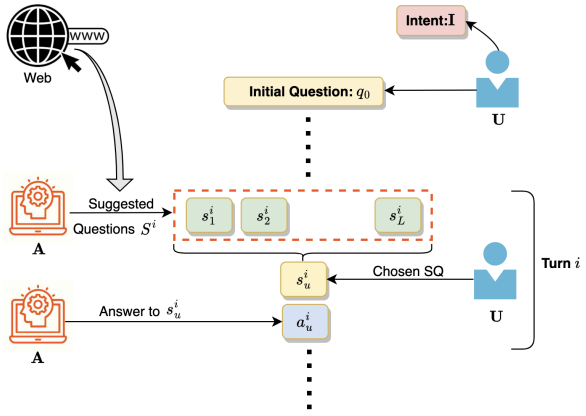

Figure 2: An illustration of choices made by the user and agent at an arbitrary turn i of the ProMISe conversation.

compare ProMISe with three existing information-seeking interaction settings with AI agents:

- **Single Turn QA:** Generating a single answer response to the user's initial question (without offering the user the opportunity to explore beyond their pre-existing information on the topic).
- **Single Turn SQA:** A single turn instantiation of ProMISe, i.e., generating multiple SQs and their answers to the user's initial question. This setting is similar to previously studied methods for generating follow-up questions (Gaur et al., 2021; Zamani et al., 2020; Rosset et al., 2020).
- **Muti Turn QA:** User breaks down the complex intent into multiple atomic questions, and the agent sequentially responds to these atomic questions that the user asks.

The first two settings are based on single-turn information exchange, while the third setting and ProMISe have multiple turns of interaction. We consider user intents from open-domain trending queries on Google Trends and use web-augmented ChatGPT as the AI agent for simulating the different interaction settings (Refer to Appendix A for complete details). We generate the user-agent interactions in each of the four settings and ask annotators to evaluate these interactions (on a 1-5 Likert scale) on five metrics as described below:

1. **Satisfaction:** Does the interaction completely resolve the user intent? We limit the interaction to 8 turns for multi-turn settings.
2. **Naturalness:** Is the interaction natural and instinctive to the user.
3. **Cognitive Load:** Is the information presented by the agent (content, format, etc.) cognitively challenging to understand for the user. A lower score indicates minimal cognitive load.
4. **Ease of Interaction:** For multi-turn settings, how much effort is required on the part of the user to interact with the system.
5. **Exploration:** Does the interaction cover multiple diverse aspects of the user's intent on an unfamiliar topic.

Interaction	Satisfaction	Naturalness	Cognitive Load	Ease of Interaction	Exploration
Single Turn QA	2.2	**4.1**	4.2	-	1.9
Single Turn SQA	2.7	3.9	3.3	-	2.8
Multi Turn QA	4.1	4.0	2.2	2.9	3.1
ProMISe	**4.2**	4.0	**2.1**	**4.5**	**4.1**

Table 1: Empirical evaluation of user-AI agent interaction settings for the task of information-seeking intent-resolution. Best results highlighted in **boldface**.

Table 1 highlights that while users, on average, find all four interaction settings to be similarly instinctive and natural, the multi-turn interactions have a much higher chance of intent resolution and exhibit lower cognitive load in absorbing information on the part of the user. Compared to the naive multi turn QA conversation setting where the user articulates follow-up questions, ProMISe facilitates better exploration of diverse topics, thereby outperforming the former in cases when the user's intent is on unfamiliar topics. Additionally, ProMISe provides an easier mode of interaction for the user who's action is restricted to choosing one of the SQs generated by the agent (compared to formulating a natural language question to ask the agent). The ProMISe setting is an enhancement over (Rosset et al., 2020) which aims to lead conversations and explore topics by providing multiple suggested questions in a single turn. This analysis empirically highlights that the ProMISe setting enables achieving an enhanced trade-off between the satisfaction of user intents, exploration of unfamiliar topics and cognitive load of the interaction on the user.

4 The ProMISe Dataset

To curate the dataset, we implement a two-player setting as shown in Fig 2 where one player acts as agent while the other player acts as user. We use a web-retrieval augmented language model as the agent. We now describe our methodology for simulating the agent and the user below:

4.1 Agent: Web Retrieval-Augmented LLM

The goal of the agent is to generate diverse and useful suggested questions based on the dialogue context that can help the user explore information related to their intent, and get closer to satisfying it. To simulate the agent, we use a popular large language model: ChatGPT (gpt-3.5-turbo-0613)

Algorithm 1 ProMISe Pseudo-code

Query $q \leftarrow q_0$
Dialogue Context $C \leftarrow [\,]$
Action $a \leftarrow$ *None*
for $i \leftarrow 1$ to *Max Turns* **do**
    Passage $\leftarrow$ *BING-API* (q)
    SQ $S^i \leftarrow$ *LLM(Passage, C)*
    $a \leftarrow$ *USER*(S^i, C)
    C.*append*(S^i, a)
    if a is s_u^i **then**
        $q \leftarrow a$
    if a is *'No SQ helps'* **then**
        $q \leftarrow$ Concatenation of all previous q's
    if a is *'Intent Satisfied'* **then**
        Break

available through the OpenAI API [2] in July-2023. Our choice is dictated by complex reasoning capabilities coupled with instruction following and larger context-length of 4k tokens. To improve beyond the parametric memory and to generate SQs over diverse real-world topics, we leverage retrieval augmented generation (Lewis et al., 2021) by extracting relevant web snippets from Bing-API[3].

The suggested questions at a turn i should not only be diverse and exploratory, but also specific to the suggested question s_u^{i-1} chosen by the user in the last turn $(i-1)$. We synthesize a prompt (shown in Table 6) for ChatGPT to generate SQs S^i in turn i of the conversation that are conditioned on the suggested question s_u^{i-1} opted by the user in the last turn $(i-1)$ and the web-snippets from Bing-API. We ensure the intended format of output SQA generation through instructions and in-context examples. Algorithm 1 contains pseudo-code for how the agent generates suggested questions S^i at turn i. As demonstrated in the pseudo-code, we use the last selected query s_u^{i-1} for retrieving the web-snippets. However, in the event that the user chooses 'No Relevant SQs,' we concatenate all preceding selected queries for web-retrieval. This facilitates the exploration and creation of SQs pertaining to topics discussed in the initial turns of dialogue.

4.2 User

At a particular turn, the role of user is to select one of the L SQs generated by the agent which helps towards satisfying the intent, or state that none of the SQs generated in this turn are helpful. If the user gauges that their intent has been satisfied, they can signal the agent to terminate the conversation. To create a high quality dataset, we use qualified crowd-annotators to simulate the user. We also devise an approach to use an LLM to simulate the user, without reliance on annotators through explanation-guided chain-of-thought generation. We first describe how we collect real-world user topics to create user intents for the dataset.

Real-world User Topics For collecting topics from open-domain to be used for creating intents for our dataset, we consider trending and most frequent queries on Google Trends. We scrape ∼30k queries using the PyTrends library [4], and then create 2500 clusters from these web queries using their Word2Vec embedding (Mikolov et al., 2013). From each cluster, we select a single example to serve as the topic for a dialogue.

User Intent and Initial Question We create the intent **I** to verbosely describe the information need of the user. The first user question q_0 represents a brief query that a user asks to initiate the conversation with the agent. Note that q_0 is not the same as **I** due to the complexity of articulating the intent well, and the lack of meta-information on the part of the user for the information-seeking topic. Note that the intent **I** may evolve and expand over the conversation with the agent as the user finds out more information about a particular topic. From the perspective of the dataset, since we want to simulate users, we consider the intent to contain all information that the user would want to know about by *the end of the conversation*, and treat the initial question as a proxy for what the user knows and can articulate properly at *the beginning of the conversation*. We generate the user intent **I** and first user question q_0 by instruction prompting LLMs, specifically LLaMA-13B and MPT-7B: we first create **I** from real-world topics, and then create the q_0 from **I**. Refer to Appendix C for prompts and anecdotal examples.

User Simulation The user action at each turn i can be: (i) choose one of the L generated SQs S^i by the agent which is assists in satisfying the intent **I**, (ii) indicate that none of the L SQs S^i generated by the agent are relevant for satisfying **I**, (iii) indicate end of conversation due to **I** being completely satisfied from the conversation with the agent. For creating a high quality dataset, we select Mechanical Turk[5] workers based on a comprehensive qualification test (refer to Appendix E for annotation guidelines and statistics). At each turn, the annotators are provided the conversation history as context and

[2] OpenAI API model
[3] Bing-Web-Search-API
[4] https://pypi.org/project/pytrends/
[5] https://www.mturk.com/

the intent **I**, and asked to make a choice from L generated SQs S^i provided to them. We take a majority vote from 3 qualified annotators for each turn of each dialogue to make a decision. If the user indicates that none of the L SQs generated by the agent across *two turns* are relevant for satisfying **I**, then the user terminates the conversation with the intent being unsatisfied.

Simulating User through LLM We propose a means to simulate the user through a LLM where the model is provided as context: the user intent **I** and the conversation history, and at each turn it makes a choice from the L generated SQs S^i provided to it. The model can either choose one of the SQs, indicate that none are relevant for the intent, or indicate if the conversation can be marked complete due to the intent being fully satisfied. We prompt LLMs with in-context examples along with the current dialogue history to generate the appropriate responses. (Prompt format in Appendix H) We leverage chain-of-thought prompting (Wei et al., 2023) to make the model generate an intermediate explanation on which suggested questions may be helpful in realizing the intent. Based on the explanation, the model then takes action as whether to select any of the suggested questions or to conclude the conversation. We provide an example of this chain of thought reasoning in Table 12.

4.3 Dataset Evaluation

We set L=4 and create a dataset starting from the real-world user topics. From all the topics we consider, we observe that more than half the dialogues conclude within the first 4 turns of conversation, and thus we set Max Turns to 8 to terminate any conversation if it has not concluded within 8 turns. We preemptively terminate any conversation where 'No SQs help' is chosen twice during the conversation. Our dataset contains 1025 dialogues with user actions taken by human annotators. Employing a high-level intent clustering, we split the 1025 dialogues into a validation and test set such that the intent topics and dialogue outcomes are balanced. The statistics of the validation and test sets are given in Table 2 and Fig 3. The annotated dataset contains 17,812 pairs of SQAs.

4.3.1 User Intent and Initial Question

We want to ensure that the initial question is not excessively verbose, while still capturing essential details relevant to the user intent. To this end, we perform a MTurk evaluation on 500 randomly sampled intents and initial-questions from the dataset. From the study, we observe that: (i) the initial question encompasses important details but leaves out trivial details of the intent in 62.6% of the samples, (ii) the initial question paraphrases the intent in 28.6% of the samples, and (iii) the initial question skips some important details of the intent in 8.2% of the samples. Detailed results are presented in Appendix D.

	Validation	Test	Total
Conversation Outcome (Number of Conversations)			
Intent Satisfied (within 8 turns)	315	315	630
Preemptive Termination (SQs repeatedly not satisfying intent)	118	118	236
Incomplete Conversation (> 8 turns needed to satisfy intent)	79	80	159
Task 1: Intent Satisfaction (Number of Turns)			
Intent not satisfied	1893	1930	3823
Intent satisfied	315	315	630
Task 2: SQ Selection (Number of Turns)			
Choose SQ 1	413	443	856
Choose SQ 2	392	399	791
Choose SQ 3	382	384	766
Choose SQ 4	370	374	744
No SQs help	336	330	666
Aggregate Dataset Statistics			
Total conversations	512	513	1025
Total turns of interaction	2208	2245	4533
Mean turns per conversation	4.31	4.38	4.35

Table 2: The statistics of the dataset collected using human feedback for user-actions.

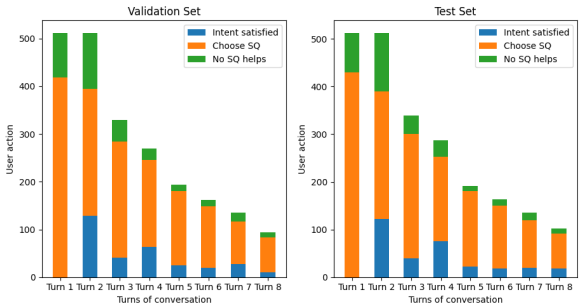

Figure 3: The graphs show the number of instances of action at each turn of dialogue.

4.3.2 Evaluation of Suggested Questions

We evaluate the quality of the suggested questions generated by the agent LLM using both automatic and human metrics as described below. We present consolidated results in Table 3.

Human metrics For each metrics, we get annotations from 3 highly qualified MTurk annotators on 500 turns (2000 SQAs) and take majority voting.

1. **Well-formedness:** We evaluate if the suggested questions are well-formed and sensible. The annotators found 99.8% of the suggested questions to be well-formed.
2. **Specificity:** We ask the annotators if atleast one of the 4 SQs at a turn S^i is relevant to the last selected query s_u^{i-1} to assess the continuity of the conversation. We find that 98.2% of the the times at-least one SQ out of 4 is relevant to the most

	Diversity Amongst SQs			Similarity(Intent I, SQs)	
	Human	Self-BLEU	MS-TTR	BLEU	BERT-Score
Turn 1	3.72	24.20	60.78	11.00	79.61
Turn 2	3.76	26.30	60.63	9.77	79.06
Turn 3	3.74	30.82	59.11	8.28	79.08
Turn 4	3.73	33.15	58.35	7.51	78.82
Turn 5	3.57	36.58	57.75	7.16	79.11
Turn 6	3.57	37.96	57.37	6.47	78.51
Turn 7	3.59	42.17	56.18	6.21	78.32
Turn 8	3.47	45.69	55.03	6.53	78.61
Average	3.69	30.91	59.15	8.72	79.06

Table 3: Evaluating the SQ generation of the agent at a turn-level granularity. The first column is based on MTurk human annotations on the number of unique SQs from 4 at each turn. The second column contains Self-BLEU scores between SQs, corresponding to inverse of diversity. The third column contains lexical diversity - Mean Segmental TTR with segment size of 50 words. The fourth and fifth columns show BLEU-Score and BERT-score of similarity between SQs and the intent.

recent selected question. In the case of 'No Relevant SQ' signalled by the user, the specificity value is 94.64%, while it is 98.65% otherwise. This affirms that once the user indicates that none of the SQs is relevant to the agent, the agent's specificity over the last selected question reduces, facilitating exploration in other directions.

3. **Diversity:** We ask the annotators how many unique SQs (questions that seek different information) are present in each turn among the 4 SQs. A high diversity score is indicative of more exploration. We find that the mean number of diverse questions across all turns is 3.69. The diversity after the 'No Relevant SQ' signal by the user is 3.77, and otherwise is 3.66. As shown in the table 3, we see that diversity decreases as the turns of the conversation increase.
4. **Relevance:** We ask the annotators to label whether the answer to each of the SQs is relevant. Annotators label that 99.4% of the times answer is relevant to the question, indicating a high QA relevance quality in the dataset.
5. **Groundedness:** We ask the annotators to label if the question or answer contains external information not present in the web-retrieved passage. For specialized real-world open-domain topics, any external domain-specific information should only be derived from the passage. This ensures that: (i) SQAs are grounded in the web-snippets with less agent LLM hallucination, and (ii) SQA generation can be conditioned through the web-snippets provided to the agent. Human evaluation showed that the questions are grounded in the web-retrieved passage 97.6% of the times, and answers are grounded in the web-retrieved passage 94.8% of the times.

Automatic metrics

1. **Diversity amongst SQs:** We use Self-BLEU (Zhu et al., 2018) as an approximation of the inverse of diversity. We also evaluate the lexical diversity - Mean Segmental TTR. Table 3 shows that the diversity of SQs decreases according to both human evaluation and automatic metrics across turns of conversation. This can be attributed to the contents of suggested questions converging towards the user intent as the conversation progresses.
2. **Similarity of SQs with the intent:** We evaluate the similarity using two popular metrics BLEU score (Papineni et al., 2002) and BERT-Score (Zhang et al., 2020). For calculating the BLEU score, we consider the intent as the candidate and the 4 SQs as the reference. For BERT-Score, we find the mean of the similarity between the intent and each of the 4 SQs. The table 3 shows that while BLEU-score decreases across the turns of conversation, BERT-Score remains the same. This can be attributed to the observation than across turns of dialogue, the entities contained in the SQs change compared to the first user question which is based directly on the user intent. However, semantic similarity between intents and SQs remains roughly the same.

Failure analysis of Agent: Based on human evaluation, some plausible reasons for the user selecting 'No SQ helps' can be mapped to factors such as the first user-question being non-representative of the intent, the user-intent being personalized, etc. We provide some anecdotal examples of these failure cases in Appendix I.

5 Simulating Human Users using LLMs

Using the collected dataset, we want to study how effectively can language models mimic the reasoning of users (humans) in carrying forward an information-seeking exchange with an agent to satisfy an intent. Simulating users effectively can improve the velocity of collection of dialogue datasets and facilitate privacy-aware evaluations. The problem of simulating the user can be split into two tasks (statistics in Table 2):

- **Task 1: Intent Satisfaction Prediction** Given the user intent and conversation history as the context, decide whether the intent has been satisfied by all the SQs chosen in the dialogue context or not. Specifically, this task is detection of satisfactory dialogue termination.

Model	F1- Intent Satisfaction Prediction				F1-SQ Selection	
	Micro	Macro	Not satisfied	Satisfied	Micro	Macro
Few-shot						
Dolly-v2-7b	79.73	48.71	88.60	8.82	22.23	14.65
LLaMA-7b	22.27	22.10	18.50	25.71	21.40	19.10
Vicuna-7b	40.91	37.85	51.64	24.05	24.82	**21.49**
Falcon-7b	42.00	36.35	55.32	17.39	20.78	15.64
Falcon-7b-instruct	60.94	**52.93**	72.34	33.51	21.66	14.44
MPT-7b	66.90	49.25	79.18	19.33	21.97	14.02
MPT-7b-instruct	28.15	27.78	32.99	22.56	24.56	15.28
MPT-7b-chat	69.62	44.83	81.81	7.84	26.11	20.68
MPT-7b-story	84.90	49.70	91.78	7.63	21.71	17.71
LLaMA-13b	43.96	41.52	53.48	29.56	22.75	19.10
Vicuna-13b	81.20	**58.59**	89.19	27.99	25.65	**23.58**
ChatGPT (turbo-3.5)	72.03	**55.92**	82.57	29.28	32.44	**31.87**
Fine-tuned						
BERT	74.57	58.00	84.38	31.62	23.63	22.00
RoBERTa	76.66	59.51	85.86	33.16	25.96	**24.47**
DeBERTa	78.08	**60.02**	86.89	33.15	25.44	24.26
LLaMA-7b (LoRA)	44.77	42.65	53.66	31.64	39.02	39.15
Vicuna-7b (LoRA)	55.63	**49.74**	66.95	32.52	43.11	**43.33**

Table 4: Benchmarking performance of popular language models (discriminative and generative) on the two user tasks in the ProMISe dataset. We use Macro-F1 for evaluation and highlight the best models of each category of models (discriminative, generative models of different sizes) for both the tasks in bold.

- **Task 2: SQ selection** Given the user intent, conversation history as the context and the list of L SQs generated by the agent at the turn i, select the most appropriate SQ that helps to satisfy the intent. If none of the SQs are relevant to satisfy the intent, select 'No SQ helps'.

Models: We benchmark the following models on the two tasks defined above: (i) *Discriminative Encoder LMs:* fine-tuned BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERT-v2-xlarge (He et al., 2021) by providing the intent and the dialogue context separated by appropriate tokens, (ii) *Generative LLMs:* few-shot instruction prompting ChatGPT, LLaMA, MPT, etc. Additionally, we select two LLMs: LLaMA-7B and Vicuna-7B and fine-tune them using (Dettmers et al., 2023) with LoRA. For details refer Appendix G.

Results: Table 4 contains the benchmarking results of the models over the two tasks. We use the Macro-F1 score to compare the different models. We observe that fine-tuned encoder LMs (BERT, RoBERTa, DeBERTa) are able to beat the performance for almost all few-shot prompted LLMs for Task-1 : Intent Satisfaction Prediction (some LLMs like Falcon-7b-instruct,Vicuna-13b and ChatGPT are able to achieve performance in the same range).

We observe that some models like Dolly-v2-7b and MPT-7b-story are unable to effectively follow instructions and end up generating 'Intent Not Satisfied' for a majority of samples (thereby obtaining imbalanced F1 scores for the two classes). The QLoRA fine-tuned LLaMA-7b and Vicuna-7B perform significantly better than their few-shot counterparts that use in-context learning and explanation-guided prompting. Among the 7 billion parameter sized LLMs, Falcon-7b-instruct and Vicuna-7b perform the best in Task 1 and 2 respectively. Task-2 (SQ Selection) is a significantly harder problem than Task-1 (as indicated by the lower F1 scores on the former). For Task 2, we observe that most of the LLMs show recency bias and tend to generate actions similar to the one present in the last in-context example.

Model	Task1 Macro-F1		Task2 Macro-F1	
	With CoT	W/o CoT	With CoT	W/o CoT
Falcon-7b-instruct	**52.93**	51.25	14.44	**15.38**
Vicuna-7b	37.85	**44.79**	**21.49**	13.38
Vicuna-13b	**58.59**	49.35	23.58	**25.79**
ChatGPT	**55.92**	49.34	31.87	**38.30**

Table 5: We examine the best-performing models from Table 4 to assess how their performance is influenced by explanation-guided chain-of-thought (CoT) prompting.

We notice that none of the models are able to achieve very high Macro-F1 scores for either of the two tasks (Task 2 having significantly lower Macro-F1 scores than Task 1). This highlights a big performance gap in the performance of state-of-the-art LLMs with humans for this task of resolving information-seeking user intents. Given how fundamental this task is for virtual assistants and search engines, we believe that our ProMISe dataset will help encourage research on this problem and improve performance of LLMs on this task.

Ablation 1: Explanation-guided Prompting We study the effect of removing the explanation-guided prompting from the best performing in-context baselines in each category of Table 4, and present the results in Table 5. We provide the same instructions and in-context examples to all the models, but remove the explanation from the prompt. We observe that for Task 1, the explanation-guided prompting helps the model achieve improved performance. Surprisingly, adding explanation-guided prompting deteriorates model performance for Task 2. We conjecture that this may be due to the following two reasons. First, we observe that some LLMs struggle to generate explanations and actions in the intended format compared to solely generating the action, which may lead to a reduction in performance. Second, instruction-prompted models expect SQs to precisely have the missing attributes of the intent rather than allowing a lenient selection which leads to over-prediction of the 'No SQs help' choice. In the case of explanation-guided generation, LLMs seem to amplify this behavior leading to a reduced F1-score performance.

Ablation 2: Turnwise performance We analyze the performance of a subset of models at a turn-level granularity. We present results for Task 1 in Fig 4, and for Task 2 in Fig 5. We observe that for Task 1, the performance of discriminative encoder LMs either remains the same or increases as the number of turns of dialogue increase. With the exception of Vicuna-13b, the performance of in-context learning based LLMs decreases as the dialogue context get larger. Additionally, for Task 1 we observe that the in-context learning based LLMs have an implicit bias to state 'Intent Satisfied' as the dialogue context gets longer.

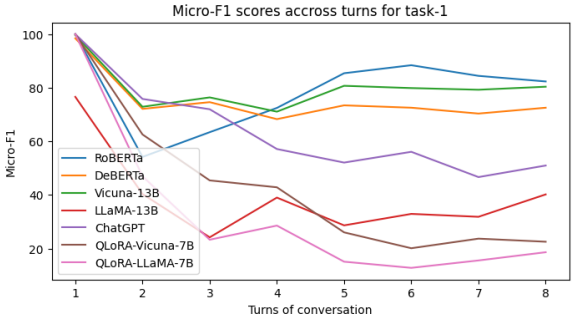

Figure 4: Turn-level performance of some models for Task 1.

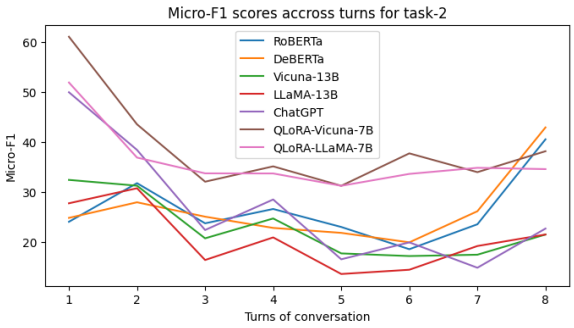

Figure 5: Turn-level performance of some selected baselines on Task 2.

6 Conclusion

We introduce a new setting: ProMISE aimed at improving AI-based virtual assistants and search systems to resolve information-seeking user intents in an end-to-end manner. We create and release a dataset of high-quality conversational data collected using human annotations and LLMs. We analyze the quality of the dataset and benchmark the performance of popular LLMs as user-simulators. The ProMISe framework and dataset will be beneficial in enhancing intelligent systems' user experience by making it interactive and proactive.

7 Limitations:

The generated SQs in our dataset are dependent of search results from Bing API. However, whenever the retrieved web-snippets for a question are similar to those for the previous question, there is a possibility of the generated SQs being similar or less diverse than the previous turn. We utilize ChatGPT (gpt-3.5-turbo-0613) with a maximum sequence length of 4000 tokens for simulating the agent which limits the previous dialogue context that can be fed to the model. In cases where we can't fit the entire conversation history in terms of generated SQs and user-actions, we keep the maximum possible number of recent turns that fit in the prompt. Our dataset collection and benchmarking experiments require access to large GPU resources. Finally, we only consider the English language for dataset and experiments in this paper, however we conjecture that our techniques should work similarly for other languages with limited morphology.

8 Ethics Statement:

For aggregating topics for our dataset, we use the open source implementation of Google Trends, which to the best of our knowledge contains anonymized user queries with no personally identifiable information. The dataset may have a linguistic bias, since we restrict the trending queries only to the English language, and filter out other languages. We use a LLM: ChatGPT for simulating the agent and generating suggested questions, which does not disclose the data sources it has been pre-trained on. Based on quality checking (both through human annotations and automatic evaluations), we believe that our dataset does not contain any personally identifiable information that crept in from the usage of the LLM. We acknowledge the fact that the usage of LLMs in the collection of the dataset may have introduced some unaccounted for biases (like racial stereotypes, gender bias, etc.). Building secure and fair LLMs remains an open challenging question, and we look forward to actively incorporating improvements made in this domain in the future to refine the biases that may have crept in the dataset. We use Mechanical Turk for obtaining annotations for the dataset, and present details of all the choices made with annotations in Appendix E including qualification task, choice of turkers, payment given to the turkers, etc.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: Problems, methods, and prospects.

Yang Deng, Wenqiang Lei, Lizi Liao, and Tat-Seng Chua. 2023b. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 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*, pages 6970–6984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Yifan Gao, Henghui Zhu, Patrick Ng, Cicero Nogueira dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Answering ambiguous questions through generative evidence fusion and round-trip prediction. 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)*, pages 3263–3276, Online. Association for Computational Linguistics.

Manas Gaur, Kalpa Gunaratna, Vijay Srinivasan, and Hongxia Jin. 2021. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coQA: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Vaibhav Kumar and Alan W Black. 2020. ClarQ: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.

Sang-Woo Lee, Sungdong Kim, Donghyeon Ko, Donghoon Ham, Youngki Hong, Shin Ah Oh, Hyunhoon Jung, Wangkyo Jung, Kyunghyun Cho, Donghyun Kwak, Hyungsuk Noh, and Woomyoung Park. 2023. Can current task-oriented dialogue models automate real-world scenarios in the wild?

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 1317–1337, Dublin, Ireland. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *EMNLP*.

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning.

OpenAI. 2023. Introducing chatgpt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Corby Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *The Web Conference 2020 (formerly WWW conference)*.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 373–393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

STAT. 2016. What's the deal with people also ask boxes?

John Sweller. 2011. Chapter two - cognitive load theory. volume 55 of *Psychology of Learning and Motivation*, pages 37–76. Academic Press.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Manso, and Roland Mathis. 2023. In-context learning user simulators for task-oriented dialog systems.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Hongru Wang, Rui Wang, Fei Mi, Zezhong Wang, Ruifeng Xu, and Kam-Fai Wong. 2023. Chain-of-thought prompting for responding to in-depth dialogue questions with llm.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 418–428, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational information seeking.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. 2022. Multiple choice questions based multi-interest policy learning for conversational recommendation.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond.

Sen Zhao1, Wei Wei, Yifan Liu, Ziyang Wang, Wendi Li, Xian-Ling Mao, Shuai Zhu, Minghui Yang, and Zujie Wen. 2023. Towards hierarchical policy learning for conversational recommendation with hypergraph-based reinforcement learning.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

Appendix

A ProMISe v/s Existing Settings

To compare ProMISE with alternate existing interaction settings between users and AI agents, we collect user studies for three additional settings for 100 intents sampled from our dataset. We use web-augmented ChatGPT as the AI agent for all the settings to have a fair comparison. The settings we used are:

1. **Single Turn Question Answering**: We extract web snippets based on the entire user intent, and prompt ChatGPT to generate an answer response that resolves the intent.
2. **Single Turn SQA:** We extract web snippets based on the entire user intent and prompt ChtGPT to generate as many suggested question-answers (SQAs) as possible. We instruct ChatGPT to make them diverse and provide in-context examples following what we do in the ProMISe setting.
3. **Multi Turn QA:** The user is tasked with providing a question at each turn. Based on the user query, we utilize Bing-API to retrieve the web snippets which are used to generate the answer. We limit the interaction to 8 turns of conversation similar to ProMISe.
4. **ProMISe:** This is the multi-turn iterative multi-SQA framework that we propose. We limit the interaction to 8 turns of conversation.

Based on the conversation, the users are asked to rate five different metrics on a Likert scale of 1 to 5, as described below: 1 Strongly Disagree, 2 Disagree, 3 Neither Agree nor Disagree, 4 Agree and 5 Strongly Agree. We use the mean of ratings across the intents to get the final scores. For the "ease of interaction" aspect, we only measure the score for the multi-turn settings where the user has to take an action at each turn.

B Prompts for Agent

The table 6 shows the prompt provided to the agent for generating SQAs.

C Prompts: User Intent + Initial Question

Table 7 shows generated user intent and first user question for two examples of initial topics. Table 8 and 9 show the format for prompting LLMs to obtain intents and first user-query respectively.

Prompt for Agent (LLM)	
Instruction	Generate 4 diverse suggested questions and generate their answers for the given query. Use the Passage for reference. Refer to the sample query and sample question -answers for format. Suggested questions should be different from any of the queries or sample questions.
Passage	Passage: {Web Retrieved Snippets}
Dialogue context	Sample query: {$sample_query} Sample question 1: {$sample_question_1} Sample answer 1: {$sample_answer_1} ... Sample question L: {$sample_question_L} Sample answer L: {$sample_answer_L}
Target query	Sample query: {$sample_query}

Table 6: Prompt format for agent LLM: the LLM is instructed to generate SQAs conditioned on the target query and the passage. 'Passage' contains web-snippets retrieved from Bing-API. Previous conversation turns are provided to also serve as in-context examples.

Topic	iPhone11 case
Intent	I want to buy a case for my iphone11. I want a case that is waterproof and has a kickstand. The case should be under $20.
Initial Question	What are some iPhone cases under $20?
Topic	New York advertising
Intent	I want to find an advertising agency that can help me with my business. The agency should have a good reputation and is located in New York city. I want to know what is the average time and price charged by them
Initial Question	What are reputed business advertising agencies in New York?

Table 7: Examples of generated intent and first user question starting from an open-domain user topic.

Instruction	Convert the topics in into an intention question. Cover all the keywords in topics and add user preferences such as price, availability, location, quantity, use-case, etc. Refer to the examples given.
In-context Examples	Topic: $topic Intent: $intent
Target Example	Topic: $topic

Table 8: The intents are expanded into intents by instructing appropriately.

Instruction	Convert the topic and intent into a very short user query. The user query may not have broader information mentioned in the intent but must have specifics. Refer to the examples given
In-context Examples	Topic: $topic Intent: $intent Query: $query
Target Example	Topic: $topic Intent: $intent

Table 9: We use the intent and topic to generate a concise initial user question that the user asks the agent to start the conversation.

D Evaluation: Intent + Initial Question

We prompt the LLMs to generate the first user-question from the user-intent. The first user-question corresponds to a short query that a real-world user may ask to the intelligent agent. Ideally, the first user-question should have important details of the intent, but may skip trivial or ambiguous aspects of the intent. To analyze the generated user questions, we conduct a human evaluation of 500 randomly sampled intents from the dataset through MTurk. We ask the annotators to select from the 5

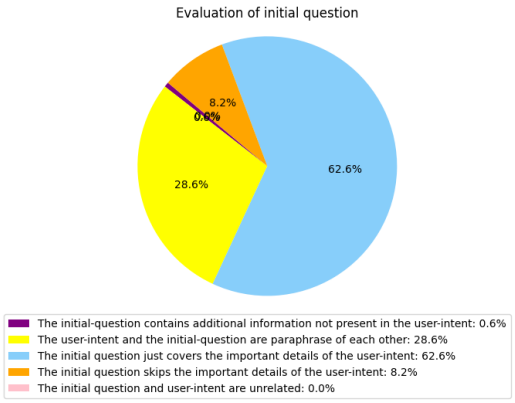

Figure 6: MTurk evaluation for initial questions generated by the user

options shown in Fig 6. We take majority vote from the 3 votes collected for each sample and break ties at random. The figure shows that while 62.8% of the user-queries cover important aspects of the intent, 28.2% of the user-queries are paraphrases of the intent, and only 8.2% of the user-queries miss the important details in the intent.

E Details of MTurk Annotations

Qualification Task: We created a comprehensive qualification test covering all edge-cases to shortlist 60 MTurk annotators. We only allowed highly qualified turkers having 'HIT approval rate' greater than 95% and 'Number of HITs approved' greater than 500 to take the qualification task. The instructions are shown in Fig 7. The annotators were informed about the task being a qualification task set-up for getting user-data for academic research. The annotators had to get a full score in the qualification task to qualify. We did not set any demographic filters for the turkers. We paid the turkers $0.5 for the 10 minute test. We shortlisted 60 workers for performing the actual annotations for the dataset. Having more number of annotators who are qualified for the task helped to reduce the bias in the data.

Annotations for User Actions: We pay shortlisted MTurk workers $0.08 for completing each task of annotation. For collecting the annotations of user-simulators for the dataset, each annotator is presented with the predefined intent, dialogue context and 6 choices as listed below:

1. Intent already satisfied by previous question.
2. SQ1
3. SQ2
4. SQ3
5. SQ4
6. None of the above questions help.

We combine the two tasks of user-simulation to ease with the annotation process. Annotators could choose choice 1 or choice 6 or one or more from choices 2 to 5. When an annotator made a decision to select a SQ, on average they selected 2.15 SQs. It implies that an annotator found 2.15 out of 4 SQs relevant to satisfy the intent of the user. We take 3 annotations for each sample and user majority voting to decide the user action. When there is a tie, it is resolved randomly. After getting all the annotations, we re-order the SQs to maintain a balance of all the selected index for Task 2. Though they are asked to annotate from six choices, we observed that the MTurk workers had a clear majority 66.56% of the times when at least 2 out of 3 annotators voted for the same choice. For 8.64% of the samples, all three annotators unanimously pointed to the same choice.

F Anecdotes: User Intent and Topics

Table 10 contains different categories of intents generated from the trending topics in the dataset.

G Details of LLM User Simulation

- **Discriminative Encoder LMs:** We fine-tune BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERT-v2-xlarge (He et al., 2021) by providing the intent and the dialogue context separated by appropriate tokens. Task 1 is a binary classification problem while Task 2 is 5-way classification problem. We fine-tune all the three models on the validation set using 4-cross validation for 3 epochs each.
- **Generative LLMs:** We prompt various LLMs: ChatGPT, LLaMA, MPT, etc. in a few-shot manner with instructions and in-context examples containing reasoning and action. We ensure that the prompt length is within the 'maximum sequence length' of all the models, and feed the same prompt to all the models. We parse the generation to extract reasoning and action. Additionally, we select two LLMs: LLaMA-7B and Vicuna-7B and fine-tune them using (Dettmers et al., 2023) with LoRA rank 64 and scaling factor of 16 for 300 steps on the validation set, and then evaluate them on the test set.

All experiments are performed using Transformers (Wolf et al., 2020) on NVIDIA Tesla V100 GPUs.

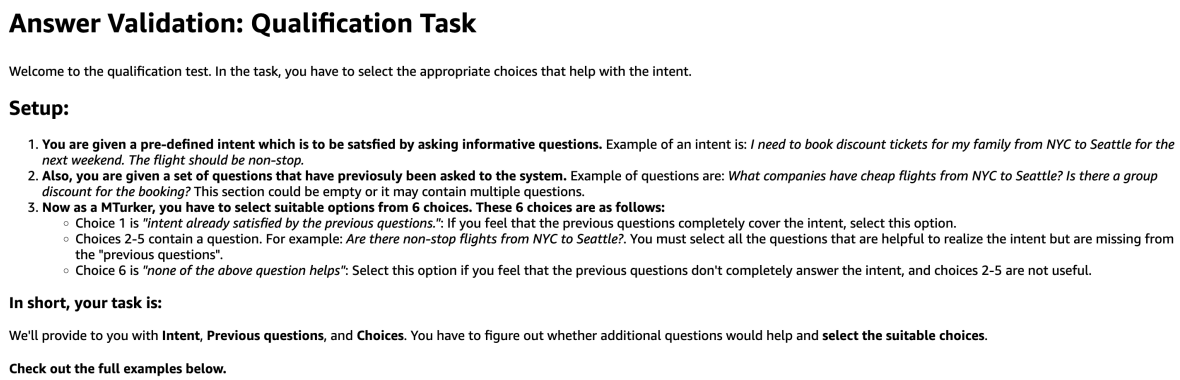

Figure 7: Instructions for MTurk qualification test. We define the task and provide sample examples.

Category	Example
Technical Support	How can I change my privacy settings on Facebook? Can I deactivate my account temporarily if I want to take a break from social media?
Entertainment	I want to watch a romantic comedy movie on Netflix. I want to watch it with my girlfriend. I want to watch it in English. I want to watch it in HD. I want to watch it on my laptop. I want to watch it in the next 2 days.
News	How can I get live scores and updates for the upcoming IPL match between Mumbai Indians and Royal Challengers Bangalore? Is there an app or website that provides live commentary as well?
Event planning	I want to attend the 2023 super bowl in Miami. I want to buy a ticket for the game. I want to buy a ticket for the game.
Curiosity	Can you explain to me what an economic recession is and how it affects individuals and businesses? Additionally, what are some strategies that can be used to mitigate the negative impacts of a recession?
Product purchase	I want to buy anker soundcore liberty air 2 pro. I want to buy it from amazon.com. I want to buy it for $100. I want to buy it in black color. I want to buy it with prime shipping.
Metrics conversion	Can you tell me how many 16 oz water bottles I need to buy to fill a gallon? Also, where can I find these water bottles in bulk and at a reasonable price?
Cooking recipe	I am a beginner in cooking. Can you tell me the steps to boil an egg perfectly? Should I use cold or hot water? How long should I boil it for in order to get a soft yolk?

Table 10: We list some of the different intents that were generated using trending topics fed to the LLMs. Although, we label a single general open-domain category, intents can belong to multiple categories.

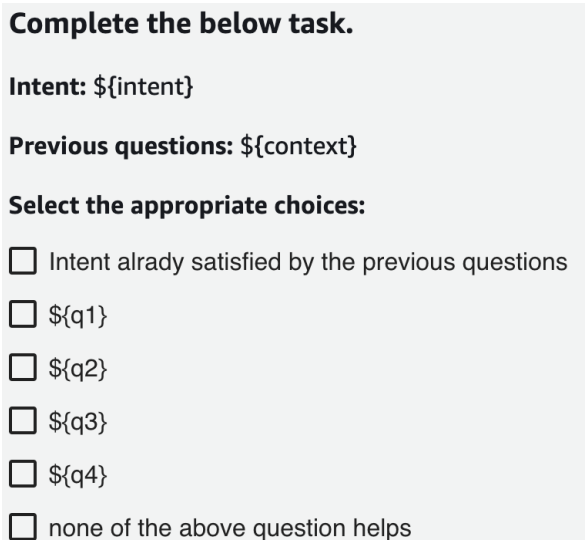

Figure 8: The shortlisted annotator is shown an intent, corresponding context and new questions. Annotator has to select suitable choices.

H Prompt: Simulating User with LLM

Table 11 shows the prompts used to simulate the user end-to-end with an LLM. We provide in-context examples to help model reason and generate actions in the intended format. The 'explanation' helps the model to reason about the ideal user action to take. User action can be one of the following:

- **Done:** Signal that the intent has been satisfied by the questions in the context.
- **Choose x:** Select SQ s_x that helps with the intent.
- **None:** Signal the agent none of the SQs help and another set of SQs is required.

An example of explanation-guided response generation is given in Table 12. Using these prompts we we further generate another 1200 examples using ChatGPT as the LLM to simulate the user.

I Failure Cases for Agent LLM

In this section, we present some anecdotes for cases where none of the generated SQs from the agent LLM are helpful for resolving the user intent.

- **Low similarity between the first user-question and user intent** In the following example, while the first user-question is relevant, it has a low similarity with the user intent.
  Intent: I want to buy a gift for my mom for Christmas.
  First user question: How many days are left for Christmas?
- **The user-intent being personalized.** In a few cases, the LLM-generated intents are personalized, making it difficult for the agent to help

User Simulator Prompt	
Task	Given intent, context and questions, give an explanation response. Here are the rules: 1. 4 questions will be given. Select the single most suitable question out of the 4 that helps with intent. State the response as "choice x" where x is a number representing the question. 2. If it seems that the intent is satisfied by questions in the context and no additional question is required, response is "done". 3. If none of the current questions increases the coverage of intent, response is "none".
In-context examples	Intent: {$intent} Context: {$context} Questions: {$SQ1} {$SQ2} {$SQ3} {$SQ4} Explanation: {$Reasoning} Choice: {$Action}
Target example	Intent: {$intent} Context: {$context} Questions: {$SQ1} {$SQ2} {$SQ3} {$SQ4}

Table 11: The format of the prompt that can simulate the user end-to-end. We provide in-context examples of all possible actions. The model then generates explanation and action for the target query

Explanation-guided action generation	
Intent	I want to watch the football match between England and Germany. I want to watch it on my mobile phone. I want to watch it in English. I want to watch it in HD. I want to watch it live.
Context	Can I watch the England vs Germany football match live on my mobile or tablet? Which mobile or tablet app is providing the live streaming for the England vs Germany football match? Can I use the BBC iPlayer app to watch the England vs Germany football match live for free on my mobile or tablet in the UK?
Suggested questions:	1) Is it possible to watch the England vs Germany football match live on the BBC iPlayer app on a smart TV or game console in the UK? 2) Will the England vs Germany football match be available on demand on the BBC iPlayer app after the live broadcast has ended in the UK? 3) What is the minimum internet speed required to watch the England vs Germany football match live on BBC iPlayer app on mobile or tablet in the UK? 4) Can I download the England vs Germany football match on the BBC iPlayer app and watch it offline on my mobile or tablet in the UK?
Explanation	Question 1 is irrelevant since it asks about watching match on smart TV or game console but the intent is to watch it on mobile phone. Question 2 and 4 are irrelavant since they inquire about watching match offline instead of live. Question 3 is helpful to know while watching football on BBC iPlayer live.
Response	Choice 3

Table 12: Chain of thought reasoning in 'Explanation' guides model to take action that helps with the intent. We prompt multiple in-context examples as shown in Table 11

with SQ generation since the agent lacks meta-information of the user.
Intent: I want to know if it is a holiday today.
Intent: What restaurants will be open in the evening?

- **Agent over-fits on certain aspects of the last selected query.** Sometimes agents generates question on certain aspects of the last selected query that are not crucial to the user. In such cases, updating the web-retrieved passage and last selected query according to the code 1 helps. For example, in the turn below, the agent starts generating questions related to price of Starbucks coffee:
Intent: I want to buy a cup of coffee. I want to buy it from a coffee shop. I want to buy it from a coffee shop that is close to my home.
Last selected Query: What is the price range for a cup of Starbucks coffee?
SQ1: Are Starbucks coffee prices the same worldwide?
SQ2: Do Starbucks prices differ between their company-owned stores and licensed locations?
SQ3: Are there any promotions or discounts available for Starbucks coffee?
SQ4: Are there any additional charges for customizations or add-ons to Starbucks coffee?