Proactive User Information Acquisition via Chats on User-Favored Topics

Shiki Sato[1] **Jun Baba**[1] **Asahi Hentona**[1] **Shinji Iwata**[1]
Akifumi Yoshimoto[1] **Koichiro Yoshino**[2]
[1]CyberAgent [2]Institute of Science Tokyo
{sato_shiki,baba_jun,hentona_asahi,iwata_shinji}@cyberagent.co.jp
yoshimoto_akifumi_xa@cyberagent.co.jp koichiro@c.titech.ac.jp

Abstract

Chat-oriented dialogue systems that deliver tangible benefits, such as sharing news or frailty prevention for seniors, require proactive acquisition of specific user information via chats on user-favored topics. This study proposes the Proactive Information Acquisition (PIA) task to support the development of these systems. In this task, a system needs to acquire a user's answers to predefined questions without making the user feel abrupt while engaging in a chat on a predefined topic. We created and analyzed a dataset of 650 PIA chats, identifying key challenges and effective strategies for recent LLMs. Our system, designed from these insights, surpassed the performance of LLMs prompted solely with task instructions. Finally, we demonstrate that automatic evaluation of this task is reasonably accurate, suggesting its potential as a framework to efficiently develop techniques for systems dealing with complex dialogue goals, extending beyond the scope of PIA alone. Our dataset is available at https://github.com/CyberAgentAILab/PIA.

TOPIC: Fishing QUESTION: *Do you have high-end earphones?* User information: *I enjoy staring up at the sky. I don't enjoy cold drinks. I have high-end earphones.*	
S:	*Hi! Let's talk about fishing!*
U:	*I went fishing last weekend. Didn't catch much though.*
S:	*That's too bad. So what do you do while you're waiting?*
U:	*Not much, honestly. It can get pretty boring.*
S:	*I heard some people listen to music to kill time when they're not catching anything. Do you ever do that?*
U:	*That's not a bad idea. Maybe I'll give it a shot.*
S:	***Do you use high-end earphones?*** *It might be better not to take them. Dropping them in the ocean would hurt.*
U:	*Plus, salty air could damage them.* ***I'll avoid taking my high-end ones****. Any suggestions for music while fishing?*

Table 1: An example of the PIA chat. S and U respectively represent the system and the user. In this example, the system is engaging in a chat on fishing while also obtaining the user's answer to the QUESTION.

1 Introduction

Developing fluent chat-oriented dialogue systems has become feasible with advances in Large Language Models (LLMs) (Shuster et al., 2022; Yi et al., 2024). Attention is shifting from pure chatting partner systems toward those capable of providing tangible benefits via chats, such as helping senior citizens prevent frailty or sharing news.

To provide beneficial support while maintaining the user's interest, the system must not merely follow the user's requests but also have an objective of collecting user information necessary for providing practical support. This requires striking a balance between information acquisition objectives and responding to the user's immediate requests, as the information necessary for supporting the user often does not directly relate to topics addressed by the user. This concept is closely related to previously studied proactive dialogue (Deng et al., 2023a), and we refer to it as Proactive Information Acquisition (PIA).

PIA arises in practical settings. For instance, research on preventing frailty involves systems collecting user answers to carefully prepared questions (e.g., current season, appetite) during casual chats, subsequently providing the answers to medical professionals for dementia diagnosis (Yoshino et al., 2021; Tanaka et al., 2017; Ujiro et al., 2018). Such diagnostic procedures become more engaging when integrated into chats about user-preferred topics. Another example is chat-based systems that provide the latest news (Yoshino and Kawahara, 2015; Li et al., 2023). When such a system receives news relevant to users with certain traits, such proactive information acquisition enables it to quickly determine whether a user matches the relevant attributes through casual chats, allowing timely and efficient delivery.

Moreover, such proactive information acquisition can be viewed as a domain-independent and general-purpose proactive behavior of asking spe-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2418–2443
November 4-9, 2025 ©2025 Association for Computational Linguistics

cific questions from chat partners under particular constraints. Thus, it could also serve as a framework for developing techniques applicable to diverse dialogue systems tasked with achieving complex goals, such as persuasion and negotiation (Li et al., 2020; Samad et al., 2022).

Given this background, we propose the Proactive Information Acquisition (PIA) task, as exemplified in Table 1. In this task, a system needs to acquire user answers to predefined questions **without** making the user feel abrupt while chatting on a predefined topic. This task's core lies in two key constraints: (1) the system must not stray from the topic, and (2) the predefined questions do not directly relate to that topic. These constraints highlight the system's ability to acquire information on the user's preferred topic.

Section 3 details our proposed PIA task. Section 4 outlines the creation of a dataset consisting of 650 PIA chats between LLM-based systems and human users, serving as a foundational resource for analyzing and training. Section 5 identifies key challenges faced by current LLMs and distinguishing features of successful PIA through dataset analysis. Section 6 demonstrates experimentally that systems designed based on these findings outperform LLMs prompted solely with task instructions. Section 7 explores the feasibility of automatically evaluating PIA chats, suggesting this task's potential as a framework for iterative improvement of techniques applicable to various systems dealing with complex dialogue goals.

2 Related work

Proactive dialogues. Task-oriented dialogue systems assume clearly defined user goals (Chen et al., 2017). However, in some cases, the systems themselves may have goals that conflict with the user's. In such situations, the systems need to chat proactively to achieve their own goals. Such dialogues are called proactive dialogues and have been studied as tasks requiring advanced strategies (Deng et al., 2023a). While proactive dialogues have been mainly studied in specific contexts like negotiation (Li et al., 2020; Samad et al., 2022), the PIA task targets domain-independent proactive dialogues, providing an environment to refine techniques applicable to proactive dialogues across diverse domains.

Target-guided dialogues. One exceptional task dealing with open-domain proactive dialogues is the target-guided dialogue task, in which systems aim to guide topic transitions during chats with users (Tang et al., 2019; Wu et al., 2019; Yang et al., 2022). With the advent of LLMs, achieving natural topic transitions has become feasible (Deng et al., 2023b), likely enabling the systems that collect user information with system-initiated topic transitions. We introduce a more challenging and complementary task, which involves acquiring necessary user information while maintaining chats on user-preferred topics to ensure user engagement.

BDI model. The analysis in Section 5.1 demonstrates that LLMs fail the PIA task by not adequately addressing multifaceted requirements, particularly (1) considering task completion status and (2) avoiding abrupt utterances. This aligns with previous studies highlighting limitations of single-output, prompt-based generation methods like in-context learning in addressing multiple requirements (Madaan et al., 2023). Our study developed a simple BDI-model-based system, which has been widely used in complex dialogue tasks (Ichida and Meneguzzi, 2023; Frering et al., 2025). The BDI model is a framework in artificial intelligence and cognitive science that describes rational agents based on their Beliefs (information about the world), Desires (objectives or goals), and Intentions (current plans or commitments to actions) (Bratman, 1987; Rao and Georgeff, 1991). BDI-model-based systems ensure responses are both status-aware and non-abrupt by decomposing response generation into explicit Belief updating, Desire formulation, and Intention selection stages, allowing the system to (1) explicitly consider the task-completion status in Belief updating and (2) discard abrupt candidate utterances in Intention selection.

3 Our proposed PIA task

In this task, a system must acquire a user's answers to predefined questions (hereafter QUESTIONs) without making the user feel abrupt while chatting on a predefined topic that may not be directly related to the QUESTIONs (hereafter TOPIC).

3.1 Flow of the PIA task

The PIA chat has two participants: a user role and a system role. Before the chat, the user role receives a TOPIC, along with n sentences of user information (hereafter "persona set"). Half of these sentences are affirmative, and the remaining half are negative. Similarly, the system role receives the same TOPIC and m ($\leq n$) QUESTIONs, created by randomly selecting m sentences from the persona set and converting them into Yes-No questions. For simplicity, this study sets $m = 1$.

The chat begins with the system role initiating with the phrase, *Hi! Let's talk about [TOPIC].* The user responds freely, and then participants alternate turns until completing a predefined number of exchanges. The system role must maintain a TOPIC-relevant chat that does not feel abrupt while acquiring sufficient information to objectively infer the user role's answers to the QUESTIONs. The user role engages in the chat about the TOPIC without contradicting the provided persona set.

To ensure diversity, protect participants' privacy, and support reproducibility, a persona set consists of predefined sentences rather than real personal data. The persona set and the TOPIC are independently assigned, and no direct relationship between the TOPIC and QUESTIONs is guaranteed. Each QUESTION is phrased as a Yes-No query to clarify whether the chat provides enough information to infer the user's answer.

3.2 Evaluation metrics

Abruptness. Chats are classified into two categories based on whether the system's utterances feel abrupt when considering the TOPIC. The classification is conducted by three human evaluators not participating in the chat to ensure reproducibility. Each evaluator rates the system's utterances on a 3-point scale: "3 - Most people would not find the utterance as abrupt," "2 - Some people might find the utterance abrupt; it might or might not be considered abrupt, depending on individual interpretation," "1 - Many people would find the utterance abrupt." An utterance is classified as non-abrupt if at least two evaluators rate it as 3. A chat is considered non-abrupt if all utterances within it are classified as non-abrupt. Although abruptness is inherently subjective, by defining scoring criteria based on "how many people would find an utterance as abrupt" and establishing detailed evaluation thresholds (Appendix A), high inter-annotator agreement is achieved, as shown in Section 4.1.

Predictability. The task involves binary classification to determine whether enough user information has been acquired during a chat to objectively infer the user's answer to a QUESTION.[1] This criterion is also assessed by three human evaluators for each chat. Each evaluator rates the chat on a 3-point scale and, for scores of 2 or 3, infers the user's answer (i.e., Yes or No): "3 - The information obtained from the chat allows a clear and accurate inference of the user's answer to the QUESTION," "2 - The information obtained from the chat allows a tentative guess of the user's answer, although it comes with a degree of uncertainty due to ambiguous or incomplete information," "1 - The chat provides insufficient information to make any guess regarding the user's answer." A chat is judged to succeed in information acquisition if at least two evaluators' inferred answers match. Evaluators assigning scores of 2 or 3 also identify the user utterances containing the required information. In our analysis (Section 5), the earliest utterance identified by at least two evaluators is considered the initial information acquisition point.

4 Dataset construction

We constructed a dataset of 650 PIA chats to facilitate analysis and training. The data collection involved two stages: an initial Pilot Collection of 200 chats with human evaluations, followed by a Refined Collection of 450 additional chats.

4.1 Pilot Collection

Data collection for the PIA task can be streamlined by clarifying in advance human and recent LLM performance, as well as observed tendencies in PIA chats. Thus, we conducted a Pilot Collection before the large-scale collection.

Participants. We prepared four types of system role players, including three LLMs known to be particularly high-performance: GPT-4o (OpenAI, 2024), Gemini-1.5-pro (Team, 2024), and Claude-3.5-sonnet,[2] as well as human speakers. To maintain diversity in dialogue behaviors—and because this paper introduces the PIA task, for which no established, widely adopted strategies exist yet—we only provided the LLMs task instructions as prompts, and responses were generated in 0-shot (Prompt 1). Further details are presented in Appendix B. Regarding user roles, we recruited 200 study participants through crowdsourcing.[3]

Topic and user information. We prepared 50 pairs of TOPICs and persona sets. We prepared the TOPICs by randomly selecting 50 noun phrases

[1] For simplicity, the explanation here assumes $m = 1$.

[2] www.anthropic.com/news/claude-3-5-sonnet.

[3] https://www.prolific.com/.

System	N	ACQ	N-ABR	SUC
Pilot Collection				
GPT-4o	50	41 (82%)	11 (22%)	6 (12%)
Claude 3.5 Sonnet	50	46 (92%)	3 (6%)	1 (2%)
Gemini 1.5 Pro	50	42 (84%)	4 (8%)	0 (0%)
Human	50	44 (88%)	10 (20%)	6 (12%)
Refined Collection				
GPT-4o	50	19 (38%)	22 (44%)	5 (10%)
Claude 3.5 Sonnet	100	71 (71%)	35 (35%)	20 (20%)
Claude 3 Opus	100	85 (85%)	33 (33%)	27 (27%)
Gemini 1.5 Pro	50	30 (60%)	15 (30%)	8 (16%)
LLama 3.1 405B	100	77 (77%)	33 (33%)	24 (24%)
Mistral Large 2	50	41 (82%)	9 (18%)	6 (12%)

Table 2: An overview of our dataset. Here, "Human" denotes the Pilot condition in which the system role was performed by human speakers (not LLMs). Note that the evaluation criteria differ between the Pilot and Refined Collection (Sections 4.1 and 4.2).

representing chat topics (e.g., fishing) from the Wizard of Wikipedia dataset (Dinan et al., 2019). We sourced the persona sets from the ConvAI2 dataset (Dinan et al., 2020). In the ConvAI2 dataset, each speaker is assigned a set of 3 to 5 persona sentences (original persona set). For this experiment, we developed 50 persona sets based on 50 randomly sampled original persona sets from this dataset, as detailed in Appendix C.

Number of turns. Following the experimental setup in the research on target-guided dialogue systems (Tang et al., 2019), the system role speaks eight times, excluding the initialization utterance (Section 3.1), and the chat ends when the user role responds to the final system role utterance.

Human evaluation. We hired three dedicated evaluators via crowdsourcing for each of the two metrics. Fleiss' Kappa of the abruptness evaluation for the Pilot Collection was 0.743 for the two-value classification of whether each system role's utterance was abrupt (scores 1 and 2) or not. For the predictability evaluation, Fleiss' Kappa reached 0.764 for the three-value classification, which categorized the predicted user answer to the QUESTION as "Yes," "No," or "Unpredictable."

Pilot Collection results. The results are shown in the upper part of Table 2. Here, N denotes the total number of chats; ACQ denotes chats that acquired enough information to infer the user's answer; N-ABR denotes chats with no abrupt system utterance; and SUC = ACQ ∩ N-ABR. The success rates (SUC) in gathering information without abrupt utterances were comparable between GPT-4o and humans. However, as these rates remain below 20% across all evaluated LLMs, even the recent LLMs face significant challenges in this task. Notably, although all LLMs successfully acquired user information in over 80% of cases, abrupt system utterances occurred in more than 78% of chats. Analysis of these cases is detailed in Section 5.1.

4.2 Refined Collection

In the previous section, we confirmed that recent LLMs achieve low task success rates. To increase successful instances, we prepared an LLM-based response generation framework designed based on the error analysis of the Pilot Collection (Section 5.1). We collected additional PIA chats employing this framework as system roles. We followed the same settings as in the Pilot Collection, except for the following differences.

LLM-based response generation framework. In analyzing the Pilot Collection (Section 5.1), we identified two primary issues with LLM performance on this task: (1) LLMs occasionally generate responses that they themselves recognize as abrupt upon reconsideration, and (2) LLMs sometimes continue discussing the QUESTION even after sufficient information has been collected. To address them, we introduce a simple LLM-based response generation framework. First, each LLM-generated response is evaluated automatically for abruptness, and if necessary, rewritten to be less abrupt before being output. Second, upon receiving each user's utterance, the system automatically checks whether the user's answer to the QUESTION can be inferred from the prior history. If inference is possible, the instructions to gather further information about the QUESTION are omitted from the LLM prompt, preventing unnecessary continuation of the discussion. Further details are provided in Appendix D.

Participants. As the system role players, we used the following six LLMs as the base LLMs that generate responses following the flow described in the previous paragraph: GPT-4o, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, LLama 3.1 405B Instruct (Llama-team, 2024), Mistral Large 2.[4] Compared with the Pilot Collection, we expanded the set of system-role LLMs to broaden behavioral coverage and reduce model-specific bias, thereby increasing dataset diversity and improving

[4] https://mistral.ai/news/mistral-large-2407/.

the generalizability of our analyses. Regarding user roles, we recruited 450 study participants through crowdsourcing.

Topics and persona sets. Since only a limited number of TOPICs and user information sentences were obtained from existing data sets, we prepared 450 additional TOPICs and persona sets using LLMs. Details are given in Appendix G.

Evaluation. Each metric was evaluated by a single evaluator per chat, considering cost. For predictability, the evaluator's results were directly used as final annotations. For abruptness, a more conservative approach was adopted: only utterances judged non-abrupt by both humans and the fine-tuned GPT-4o evaluator (Section 5.1.2) were labeled as non-abrupt; others were labeled abrupt.

Refined Collection results. The collection results are shown at the bottom of Table 2. In the Refined Collection, to maximize the yield of successful chats, we prioritized systems based on higher-success LLMs, which led to an uneven number of collected chats across LLMs.

4.3 Dataset overview

We collected 650 PIA chats, including 103 successful instances. Each chat has 17 utterances between an LLM-based system and a human user. The dataset includes human evaluations of abruptness and predictability for each chat, comprising 5,850 user utterances and 5,200 system utterances, suitable for statistical analysis and training.

5 Data analysis

5.1 Analysis of failure cases

As mentioned in Section 4.1, many chats of the Pilot Collection in which three LLMs performed the system role included abrupt utterances, hindering the task's completion. In this section, we analyze failure instances of the Pilot Collection focusing on the abrupt utterances.

5.1.1 Types of abrupt utterances

We extracted all 132 chats with at least one LLM-generated abrupt utterance from the Pilot Collection. Analyzing each chat's first abrupt utterance, we categorized them into four types (Table 3).[5]

[5] Appendix E shows their examples and distributions.

1	Utterance suddenly starting to talk about the QUESTION without any relevant context
2	Utterance introducing an unnatural relationship to associate the QUESTION with the dialogue context or the TOPIC
3	Utterance focusing too much on the QUESTION after the introduction of a natural relationship to associate the QUESTION with the dialogue context or the TOPIC
4	Utterance trying to continue talking about the QUESTION even though user information has been obtained

Table 3: Types of abrupt utterances.

5.1.2 Analysis of types 1-3

These three utterance types were judged as abrupt because the system attempted to acquire information about the QUESTION when still lacking sufficient information to predict the user's answer.

We investigated whether LLMs generated these utterances despite potentially recognizing their abruptness or were inherently unable to recognize their abruptness. Specifically, we evaluated the LLM's ability to detect abrupt utterances by having them rate system utterances on a 3-point scale, similar to human evaluations (Section 3.2). We divided the 200 chats of the Pilot Collection into equal training and evaluation sets and fine-tuned GPT-4o using the training set. Inputs included task instructions (Prompt 2), the TOPIC, a system utterance to be evaluated, and preceding chat history. The output was a 3-point rating mirroring human evaluations. We validated the fine-tuned evaluator by comparing its binary classifications (rating of 3 or not) with human evaluations. Before fine-tuning, the F1 score for identifying abrupt utterances was 40.1 (recall: 26.5, precision: 82.6). After fine-tuning, it improved significantly to 88.5 (recall: 87.4, precision: 89.5), indicating LLMs effectively detected abruptness with minimal training.[6] Further details are provided in Appendix F.

The above results suggest that LLMs generated utterances whose abruptness they could recognize, provided that they explicitly reconsidered them after careful alignment with human evaluations.

5.1.3 Analysis of type 4

These abrupt utterances occurred when LLMs unnecessarily continued discussing the QUESTION despite already having sufficient information.

We investigated whether these utterances arose

[6] When we later expanded the training set by adding the 450 chats from the Refined Collection and re-tuned the evaluator, the F1 score improved only marginally (from 88.5 to 89.8), suggesting near-saturation at this scale.

SUB-THEME	TOPIC can feature goods, events, or other things related to QUESTION.
PLACE	TOPIC can be the place, organization or event where the event related to QUESTION occurs.
MEANS	TOPIC can be a means to achieve a goal related to QUESTION.
CO-OCCUR	TOPIC can occur or exist at the same time (or before or after) as the event or object related to QUESTION.
CAUSE	TOPIC can be the cause of the event, situation or state related to QUESTION.
PREREQUISITE	TOPIC can be a prerequisite for dealing with something related to QUESTION.
DOER	TOPIC can be done by QUESTION.

Table 4: Categories of relationship types.

because the LLMs unnecessarily continued chat about the QUESTION despite potentially being capable of recognizing that sufficient information had already been obtained, or it was fundamentally difficult for the LLMs to accurately assess the information acquisition state. For each of the Pilot Collection's 200 chats, we had GPT-4o predict the user's answer ("Yes," "No," or "Unpredictable") using the same instructions as the human evaluation in 0-shot (Prompt 3). GPT-4o matched human predictions in 88.0% of cases, indicating that LLMs could accurately judge information acquisition status when explicitly instructed.

The results suggest that while LLMs have the capability to recognize information acquisition status, they sometimes continue discussing QUESTIONs unnecessarily, resulting in abrupt utterances.

5.2 Analysis of successful cases

5.2.1 TOPIC–QUESTION association

An analysis of the 103 successful cases revealed that LLMs primarily selected the most suitable relationship type from the seven listed in Table 4 to connect a TOPIC with a QUESTION. Using this relationship as a starting point, the LLM generated questions related to the QUESTION within the context of the TOPIC. Examples of utterances for each relationship type are provided in Appendix H.

5.2.2 Use of cushion utterance

Here, we define the "key utterance" as the first system role utterance immediately preceding the user's utterance containing sufficient information to infer the user's answer to the QUESTION.

An analysis of 35 randomly selected successful chats[7] showed that in 24 instances, the key utterance was introduced without prior interaction related to the QUESTION. Conversely, in 11 chats, the system role produced at least one preceding "cushion" utterance to smoothly transition to the key utterance. Table 5 shows a cushion utterance. Among the 11 chats, only 3 featured multiple cushion utterances, indicating that generally, a single cushion utterance—or none at all—was adequate. This suggests that it is important to strategically use a single cushion utterance as needed to reduce abruptness of introducing QUESTIONs.

TOPIC: Natural parks QUESTION: *Are you a fan of funk music?*	
S:	*Hi! Let's talk about National parks!*
U:	*Hi, I love National parks and I often meet my friends there for lunch. One of my friends brings her little baby too which is lovely.*
S:	*That's wonderful! National parks are great for gatherings. Do you have a favorite park for your meetups? Some parks have amazing picnic areas.*
U:	*We usually go to the same one. We like the cafe there because they do nice jacket potatoes and coffee. Sometimes we have cake too.*
S:	*Sounds delightful! It's great when parks have good amenities.* ***Do you ever explore the trails or natural areas after your lunch?*** *Some parks offer beautiful hiking opportunities.*
U:	*We will have a walk around the lake but that's all because we have the baby in a buggy. Sometimes, I go cycling at a different park with my husband.*
S:	*A lakeside stroll sounds perfect for a group with a baby. And cycling in parks is fantastic! Do you ever bring music to listen to during your rides or walks?*
U:	*I do like listening to music when I am on my own with headphones. I like funk music. If I am with someone else, I don't listen to music.*

Table 5: An example of a cushion utterance. The seventh system utterance elicits the user's answer. It focuses on the plausible relationship between the TOPIC and the QUESTION: "music to listen to during rides or walks," thereby obtaining an answer without deviating from the TOPIC. Furthermore, "rides or walks" is clearly introduced naturally by the system in the previous turn (the fifth utterance) with the question "Do you ever explore the trails or natural areas after your lunch?" Therefore, the fifth utterance can be regarded as a cushion utterance to reduce the introduction of the seventh utterance.

5.2.3 Inclusion of explanation

Among the dataset's chats with abrupt key utterances, 34 instances lacked explicit explanations of how the QUESTION relates to the TOPIC.[8] [9] After

[7]We randomly selected five chats from each of the seven LLMs used in the dataset construction.

[8]The presence of an explanation was determined by OpenAI o1's (`https://openai.com/o1/`) 0-shot inference (Prompt 11). See Appendix B for the details.

[9]For example, the phrase "Dropping them in the ocean would hurt" in the last system utterance in Table 1 explicitly

adding explicit explanations using GPT-4o (0-shot, Prompt 12), the fine-tuned GPT-4o-based evaluator (Section 5.1.2) reassessed 38% of these utterances as non-abrupt, suggesting the necessity of explicitly stating the QUESTION-TOPIC relationship within key utterances to mitigate abruptness.

6 Validation of strategy-based system

In this section, we propose a simple BDI-model-based system designed based on the Section 5's insights, which we call **strategy-based system**, and confirm that it outperforms LLMs prompted solely with task instructions.

6.1 Design of strategy-based system

Upon receiving a user utterance, it updates its belief by evaluating the information acquisition state. Based on the belief, it determines whether to acquire more information (desire generation). It then selects a response from four types of candidates, which were generated simultaneously with belief update, based on the desire (intention generation).

6.1.1 Belief update and desire generation

When receiving a user utterance, the system employs an LLM to predict the user's answer to the QUESTION in 0-shot, based on the chat history using Section 5.1.3's method (Prompt 3). If the prediction is either "Yes" or "No", the system terminates information gathering and focuses solely on the TOPIC. Otherwise, information gathering continues. This approach leverages our findings (Section 5.1.3) indicating that LLMs can effectively track the information acquisition state, thus preventing abrupt utterances that continue talking about the QUESTION unnecessarily.

6.1.2 Response candidate generation

Key utterance candidates. Many successful chats effectively utilized the seven relationship types (Section 5.2.1) to connect the TOPIC and QUESTION; we explicitly model this strategy. Specifically, an LLM generates seven "key utterance prototypes" prior to the chat by associating the TOPIC with the QUESTION using these relationship types (Prompt 13). During the chat, the LLM rephrases these prototypes to align with the ongoing chat (Prompt 14), using them as key utterance candidates. As emphasized in Section 5.2.3, key utterances must clearly explain how the QUESTION relates to the TOPIC. To achieve this, we instruct the LLM to: (i) identify the specific relationship between the TOPIC and QUESTION based on the given type, (ii) explicitly explain this relationship, and (iii) generate responses based on these explanations. Additional details are provided in Appendix I.1.

explains the relationship between TOPIC and QUESTION.

Cushion utterance candidates. Some successful chats included cushion utterances before key utterances (Section 5.2.2). To emulate this, we have an LLM generate a cushion utterance in 0-shot for each of the seven key utterance prototypes at each turn (Prompt 15) and add them to candidates.

Vanilla candidate. The system also includes a candidate generated by an LLM with solely task instructions (Prompt 1) to retain LLM's flexibility.

Safe candidate. To prepare for cases where all candidates introduced thus far are deemed abrupt in the subsequent process, an LLM instructed to focus exclusively on casual chat on the TOPIC generates an additional response candidate (Prompt 5).

6.1.3 Intention generation

The system only selects the safe candidate after confirming that information acquisition is complete. Otherwise, it selects a suitable candidate for non-abrupt information acquisition. Leveraging the result that LLMs accurately identify abrupt utterances (Section 5.1.2), the selection process employs an LLM with the same input-output format as the GPT-4o-based abruptness evaluator described therein. First, if any "key utterance candidates" are classified as non-abrupt by the evaluator LLM (Section 5.1.2), the system selects the candidate with the highest likelihood of scoring 3, calculated by the evaluator LLM. If none exist, the system sequentially evaluates "cushion utterance candidates," followed by the "vanilla candidate," and finally the "safe candidate," applying the same selection procedure. This approach flexibly integrates cushion utterances and fallback responses while prioritizing key utterances when possible.

6.2 Evaluation settings

Except as described below, we collected chats using the same settings as in Section 4.1.

Compared systems. To evaluate the strategy-based system, we compared it against two alternatives. The first alternative (Standard) was a 0-shot response generation using GPT-4o with only task instructions, similar to the approach described in

System	ACQ	N-ABR	SUC
Standard	74%	38%	16%
Prompt-based	92%	22%	18%
Strategy-based	50%	82%	40%

Table 6: Baseline systems' performance in our task.

System	Recall	Precision	F1
0-shot	35.7	81.2	49.6
3-shot	67.0	67.0	67.0
9-shot	67.7	62.9	65.2
15-shot	71.0	61.4	65.9
Fine-tuned	90.5	85.3	87.8

Table 7: Semi-automatic evaluation performance.

Section 4.1. The second alternative, termed the Prompt-based, was also a 0-shot response generation with GPT-4o but included task instructions and a comprehensive description of insights from Section 5.2's analysis in its prompt (Prompt 16).

Settings of strategy-based system. GPT-4o was utilized for all processes except the response selection. To develop an evaluator LLM for the response selection, GPT-4o was fine-tuned on our dataset's chats, excluding the chats in the experiment's test set of Section 5.1.2. See Appendix I.2 for the training details. When compared to the evaluator fine-tuned in Section 5.1.2, the detection performance (F1 score) for abrupt utterances in Section 5.1.2's test set improved from 88.5 (recall: 87.4, precision: 89.5) to 89.8 (recall: 94.0, precision: 86.0).

Topic and personas. We prepared 50 TOPICs and persona sets like Section 4.1's data collection. We made sure that the persona sentences and the TOPICs did not overlap with those in our dataset.

6.3 Evaluation results

Table 6 summarizes the evaluation results. The strategy-based system had significantly fewer abrupt utterances compared to the two prompt-only systems (Standard and Prompt-based), substantially improving the task success rate. This confirms that insights from the previous section effectively enhanced performance.

However, the success rate of the strategy-based system remains low at 40%, likely due to challenges in generating non-abrupt key utterances. Out of 50 chats totaling 130 turns with key utterance generation, 36% failed to produce viable candidates, as all generated utterances were classified as abrupt by the evaluator LLM.

7 Discussion: evaluation automatability

Based on the results in the previous section, there remains room for improving the generation of effective and non-abrupt key utterance candidates. Continued iterative refinement and evaluation of this point are essential. However, the human evaluation approach previously used is costly and hinders rapid iteration. This section demonstrates that this task can be automatically evaluated with reasonable accuracy, enabling quicker and more cost-effective iterative improvements. Additionally, since the task focuses on domain-independent, general-purpose proactive dialogue actions, it can serve as a benchmark for techniques applicable to various systems dealing with complex dialogue goals.

Evaluating systems for this task currently requires human involvement in two phases: chat collection and assessment. Therefore, we first investigate the feasibility of semi-automatic evaluation, which automatically assesses human-system interactions, thus validating systems based on actual user interactions without annotation costs. We then explore the possibility of fully automating the evaluation workflow, including chat collection. Although this fully automatic method does not allow validation based on actual user interactions, it entirely eliminates human labor costs.

7.1 Semi-automatic evaluation

Section 5.1.3 demonstrated accurate automated predictability assessment (88.0% accuracy) by a 0-shot LLM; this section explores abruptness evaluation.

Previous automated abruptness evaluations (Sections 5.1.2 and 6.2) used identical LLMs for both dataset creation and evaluation. Here, we extend the analysis to cases where dataset creation and evaluation involve different LLMs. For fine-tuning, we randomly selected 50 chats each from three LLMs not in the Pilot Collection (Claude 3 Sonnet, Llama 3.1 405B, Mistral Large 2) from the Refined Collection. The evaluation set consisted of all 200 Pilot Collection samples. Additionally, we conducted inference experiments to evaluate in-context learning (0-shot, 3-shot, 9-shot, and 15-shot prompting), using annotated examples from the fine-tuning dataset. All other settings matched Section 5.1.2. Table 7 summarizes the evaluation results, showing that few-shot learning and fine-tuning both enhance automatic abruptness evaluation relative to 0-shot inference, with fine-tuning

Metric	Pearson r	Spearman ρ
ACQ	0.738 (p=0.094)	0.943 (p=0.005)
N-ABR	0.815 (p=0.048)	0.928 (p=0.008)
SUC	0.377 (p=0.461)	0.435 (p=0.389)

Table 8: Full-automatic evaluation's correlation.

proving especially effective. The high F1 score from fine-tuning indicates accurate evaluation of abruptness even for previously unseen LLMs.

These results confirm the feasibility of highly accurate semi-automatic evaluation for both metrics.

7.2 Full-automatic evaluation

We introduced an LLM-based user simulator to assess the feasibility of fully automated system performance evaluations. The simulator (GPT-4o-mini with Prompt 18) engaged in fifty chats with each of the six systems in the Refined Collection, employing identical settings for TOPICs, persona sets, and QUESTIONs as those used in the Refined Collection. These chats were evaluated automatically: abruptness was assessed using Section 7.1's fine-tuned evaluator, and predictability was measured following the method detailed in Section 5.1.3.[10] Chats were classified as ACQ if predictability evaluation distinctly indicated "Yes" or "No," as N-ABR if no abrupt utterances were detected, and as SUC if both were satisfied. These automated evaluations were then compared with human evaluations from the Refined Collection (Table 6).

Table 8 shows correlation coefficients between automated and manual evaluations. They indicate strong correlations for ACQ and N-ABR, while those for the combined SUC metric were comparatively weaker. These findings suggest that individual automated metrics (ACQ and N-ABR) align closely with human evaluations, highlighting the potential for accurate full-automatic assessments.

8 Conclusion

PIA is crucial for systems offering user benefits, such as health services or tailored news. This study proposes the PIA task as a foundational framework to advance such technologies. We constructed a dataset to facilitate system development, revealing the challenges and effective strategies for LLMs. An insight-based system significantly outperformed LLMs prompted solely by task instructions. Furthermore, we confirmed high accuracy in semi-automatic evaluations and reasonable accuracy in full-automatic evaluations for this task.

This task focuses on a domain-independent, general-purpose action of asking specific questions, which can be reliably evaluated automatically. We anticipate that this task will efficiently support the development of techniques applicable to various dialogue systems addressing complex dialogue goals.

Limitations

Use of artificial persona data. In this study, we conducted experiments by assigning prepared personas to users rather than using real user information from the perspective of protecting the personal information of the crowd-sourcing workers, the tasks' reproducibility, and the diversity of the target user information. Therefore, in addition to the ones mentioned in this paper, different challenges may exist in acquiring actual user information. However, this study focuses not on analyzing user behavior regarding information disclosure but on basic chat strategies for acquiring user information necessary for benefiting users; thus, we recognize that this is not a critical problem in this study.

Question formats. This study exclusively employed Yes-No questions for the PIA task, focusing on simplicity and clarity of responses. However, alternative approaches such as open-ended or ranking questions were not explored. Future research could consider using various questioning formats, including open-ended or ranking-based methods, to enhance the understanding and insights gained from the PIA task.

Setting diversity. While we ensured diversity by sourcing topics from the Wizard of Wikipedia dataset, which includes a wide range of topics derived from Wikipedia article names, and questions from the ConvAI2 dataset, composed of fictional user-generated statements by numerous crowd workers, the random combination method employed may not fully guarantee comprehensive coverage. Thus, although diverse, the dataset might not systematically represent all possible combinations of topics and questions, highlighting room for improvement in coverage and representativeness.

Domain and user dependency. Although we claim the PIA task is domain-independent, our current evaluation does not include detailed analysis

[10]Note that the abruptness evaluator's training data included chats from Claude 3 Sonnet, Llama 3.1 405B, and Mistral Large 2, meaning they were not entirely new to the evaluator.

of how system performance correlates with specific topic-question combinations or user characteristics. The perception of "abruptness" is likely highly sensitive to specific domains and individual user characteristics—certain topic-question pairings may inherently feel more natural than others, and thresholds for abruptness may vary across users. Future work should examine these dependencies through fine-grained analysis to better understand the generalizability of our approach.

Relationship between the system and users. Since this study focuses on basic chat strategies, we did not define the relationship between the user and the system. In actual information acquisition, different behaviors may be displayed depending on the intimacy with the chat partner.

Cushion utterances. Although our analysis did not explicitly identify risks associated with using cushion utterances, it is important to acknowledge potential drawbacks. The strategy-based system prepares cushion utterances based on predefined conditions related to dialogue context. However, if users provide unanticipated responses following a cushion utterance, the introduction of "Key utterance candidates" may become more difficult. Future research is required to explore robust recovery strategies when users' reactions deviate from expected patterns after cushion utterances.

Dependence on prompts. The experimental results may depend on our prepared prompts, although they were carefully created after much trial and error.

Ethical considerations

In this study, topics and persona sentences were prepared from existing datasets and LLM-generated content. The authors manually verified that none of this material contained harmful content prior to use. Although this study involved acquiring information through chat interactions, all information was fictional, and no actual user personal data was collected. For tasks involving human participants, we obtained informed consent after providing comprehensive explanations regarding the risks of participation and data handling procedures. Based on these protocols and safeguards, our institution's ethical review board determined that this study does not require formal ethical review.

It is essential to note that any deployment of PIA must occur only with user consent and in full compliance with applicable regulations, including the EU AI Act,[11] similar to requirements for facial recognition and behavioral profiling technologies. Furthermore, any collection of users' personal information for improving user experience must be conducted with prior user consent to ensure compliance with existing legal frameworks such as the General Data Protection Regulation (GDPR).[12]

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and constructive feedback, which helped improve the quality of this paper. We are also grateful to all the participants who took part in our experiments for their time and valuable contributions to this research.

References

Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, Cambridge.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6583–6591. International Joint Conferences on Artificial Intelligence Organization.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023b. 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*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard

[11]https://artificialintelligenceact.eu/
[12]https://gdpr.eu/

of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Laurent Frering, Gerald Steinbauer-Wagner, and Andreas Holzinger. 2025. Integrating belief-desire-intention agents with large language models for reliable human–robot interaction and explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, 141:109771.

Alexandre Yukio Ichida and Felipe Meneguzzi. 2023. Modeling a conversational agent using bdi framework. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 856–863, New York, NY, USA. Association for Computing Machinery.

Siheng Li, Yichun Yin, Cheng Yang, Wangjie Jiang, Yiwei Li, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. NewsDialogues: Towards proactive news grounded conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3634–3649, Toronto, Canada. Association for Computational Linguistics.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020. End-to-end trainable non-collaborative dialog system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8293–8302.

Llama-team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Anand S. Rao and Michael P. Georgeff. 1991. Modeling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856, Seattle, United States. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *Preprint*, arXiv:2208.03188.

Hiroki Tanaka, Hiroyoshi Adachi, Norimichi Ukita, Manabu Ikeda, Hiroaki Kazui, Takashi Kudo, and Satoshi Nakamura. 2017. Detecting dementia through interactive computer avatars. *IEEE Journal of Translational Engineering in Health and Medicine*, 5:1–11.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Tsuyoki Ujiro, Hiroki Tanaka, Hiroyoshi Adachi, Hiroaki Kazui, Manabu Ikeda, Takashi Kudo, and Satoshi Nakamura. 2018. Detection of dementia from responses to atypical questions asked by embodied conversational agents. In *Interspeech 2018*, pages 1691–1695.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. TopKG: Target-oriented dialog via global planning on knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 745–755, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *Preprint*, arXiv:2402.18013.

Koichiro Yoshino and Tatsuya Kawahara. 2015. Conversational system for information navigation based on pomdp with user focus tracking. *Computer Speech & Language*, 34(1):275–291.

Koichiro Yoshino, Yukitoshi Murase, Nurul Lubis, Kyoshiro Sugiyama, Hiroki Tanaka, Sakti Sakriani, Shinnosuke Takamichi, and Satoshi Nakamura. 2021. *Spoken Dialogue Robot for Watching Daily Life of Elderly People*, pages 141–146. Springer Singapore, Singapore.

A Additional threshold for human evaluation of abruptness

A preliminary survey of human evaluation of abruptness revealed that high annotator agreement was achieved by setting the following two points as thresholds.

- An utterance introducing or discussing a specific instance within the broader context of a given TOPIC should NOT be considered abrupt.
  - For example, "Lakers" is a specific instance within the broader context of "basketball." Therefore, if the TOPIC is "basketball," a response discussing the Lakers should not be classified as abrupt.
- A response introducing a new subtopic should not be considered abrupt, as long as the primary focus of the response clearly remains on the original TOPIC, and the new subtopic works for continuing the chat on TOPIC.
  - However, if the response excessively focuses on the subtopic without clear relevance to the original TOPIC, assign a lower rating (such as 2 or 1).
  - Additionally, if the combination of a new subtopic and the original TOPIC seems unnatural or unusual, please rate it lower (e.g., "Do you enjoy online shopping while painting?").

B Detailed settings of LLMs

The experiments in this study used the versions of LLMs described in Table 9 through the API services listed in the same table. We used the default settings of each API service for all LLMs.

C Detailed settings of persona sets

We sourced the persona sets from the ConvAI2 dataset (Dinan et al., 2020). In the ConvAI2 dataset, each speaker is assigned a set of 3 to 5 persona sentences (hereafter "original persona set"). For this experiment, we developed 50 persona sets based on 50 randomly sampled original persona sets from this dataset. More specifically, we randomly selected half of the persona sentences from each original persona set and automatically converted them into their negated forms.[13] Each modified set was assigned to a user role player. One of the user information sentences in each set was randomly chosen and automatically converted into a Yes-No question,[14] which was presented to the corresponding system role.

D Details of response generation framework for the Refined Collection

This framework generates responses through a three-step process using two distinct LLMs: a base LLM and an evaluator LLM. Firstly, the base LLM generates responses in the same way as the LLMs employed in Section 4.1. Secondly, the evaluator LLM automatically assesses the abruptness of the base LLM's response. When the assessed response is judged abrupt, the base LLM rewrites the response to mitigate its abruptness (Prompt 4). This approach is grounded in the premise that fine-tuned LLMs can identify abrupt utterances with reasonable accuracy (Section 5.1.2). In this paper's data collection, we used the fine-tuned GPT-4o in Section 5.1.2 as the evaluator LLM of this response generation framework, regardless of which LLM was used to generate responses. Thirdly, after outputting the response and then receiving a new user utterance, the base LLM predicts the user's answer to the QUESTION (Prompt 3) based on the chat history up to that point in 0-shot. If the predicted answer is either "Yes" or "No" (as opposed to "Unpredictable"), all auxiliary processes, except for response generation by the base LLM, are stopped for the rest of the chat. Furthermore, the framework rewrites the base LLM's response generation prompt to remove instructions for collecting user information, and the base LLM is made to focus solely on chatting about TOPIC (Prompt 5 and 6). This process is based on our findings that recent LLMs can track the state of user information acquisition, which could prevent the generation of abrupt utterances that continue talking about the QUESTION unnecessarily (Section 5.1.3).

E Examples and distributions of abrupt utterances

Table 10 shows the frequency in the 132 analyzed chats for each of the categories of abrupt utterances found in Section 5.1.1. Tables 11, 12, 13, and 14 show examples of abrupt utterances for the types 1, 2, 3, and 4, respectively.

[13] www.github.com/dmlls/negate.

[14] We used the following library: www.github.com/shiki-sato/nbest-contradiction-analysis.

LLM	Version	API
GPT-4o	2024-05-13	Azure OpenAI Service (https://azure.microsoft.com/products/ai-services/openai-service)
Gemini-1.5-pro	001	Google Vertex AI (https://cloud.google.com/vertex-ai)
Claude-3.5-sonnet	20240620-v1:0	Google Vertex AI
Claude-3-opus	20240229-v1:0	Google Vertex AI
LLama-3.1-405B	-	Google Vertex AI
Mistral-Large-2	2407	Google Vertex AI
OpenAI o1	preview-2024-09-12	OpenAI API (https://openai.com/index/openai-api/)
GPT-4o-mini	2024-07-18	Azure OpenAI Service

Table 9: List of LLM versions and APIs used in the experiments.

	Type of abrupt utterance	Freq. in the 132 chats.
1	Utterance suddenly starting to talk about the QUESTION without any context	30
2	Utterance introducing an unnatural relationship to associate the QUESTION with the dialogue context or the TOPIC	38
3	Utterance focusing too much on the QUESTION after the introduction of a natural relationship to associate the QUESTION with the dialogue context or the TOPIC	50
4	Utterance trying to continue talking about the QUESTION even though user information has been obtained	14

Table 10: Types of abrupt utterances.

F Detailed settings of experiments for automatic abruptness evaluation

The 200 chats obtained in the evaluation experiment were divided into approximately equal parts for the training and evaluation sets. They were divided so that the same questions and topics were not included in both the training and evaluation sets, and the number of chats in the training and test set was finally set at 109 and 91, respectively. Since each chat contains eight system utterances, the split of 109/91 chats corresponds to 872/728 system-utterance training/evaluation instances used for fine-tuning and testing the evaluator. We used the training data to fine-tune GPT-4o-2024-08-06 using the OpenAI API. The OpenAI API automatically set the hyperparameters, which were 3 epochs, 1 batch size, and 2 LR multipliers. We validated the fine-tuned evaluator by comparing its binary classifications (whether each utterance was rated as 3 or not) against the human evaluation results. Specifically, after computing the softmax probabilities for the system's ratings of 1, 2, and 3, an utterance was classified as "non-abrupt" if the probability of receiving a rating of 3 exceeded 50%. Otherwise, it was classified as "abrupt."

G Detailed settings of topic and personas for dataset construction

For the TOPICs, we generated 200 words using each of GPT-4o, Gemini-1.5-pro, and Claude-3.5-opus (Prompt 7) and then had GPT-4o remove duplicate instances (Prompt 8), resulting in a final set of 212. For the persona sentences, we generated 200 in the same way (Prompt 9 and 10) and then added 67 from the ConvAI2 dataset, resulting in a final set of 267. We randomly selected three sentences from these 267 and automatically converted half of them (one or two) into negative sentences to create a persona set. We repeated this process to create 450 persona sets. We finally prepared 450 combinations of these 212 TOPICs and 450 persona sets.

H Details of analysis on association between TOPIC and QUESTION

Table 15 shows the frequency of use of each relationship type in the 103 successful chats of our dataset. This table also shows the frequency of use of each relationship type in 70 randomly sampled unsuccessful chats of our dataset which include abrupt utterances even though the user information was acquired. The table shows that failed chats use COMMONALITY more than successful ones. This suggests that COMMONALITY is likely to be judged as a weak association when talking about QUESTION in a chat on TOPIC. Thus, it is possible that in order to succeed in the task, selecting the best of the seven relationship types (excluding COMMONALITY from the above eight types) for associating QUESTION with TOPIC may lead to the acquisition of user information without a sense

Topic: Fairy tale Question: *Do you like to exercise?*	
S:	*Hi! Let's talk about Fairy tale!*
U:	*I enjoy clothing with logos on it.*
S:	*That's interesting! Fairy tales don't usually feature logos, do they?*
U:	*They do*
S:	*Really? I've never noticed them.* ***Are you physically active?***

Table 11: Example of an abrupt utterance suddenly starting to talk about the QUESTION without any context (bolded for illustration).

Topic: Tattoo Question: *Are you basically addicted to caffeine?*	
S:	*Hi! Let's talk about Tattoo!*
U:	*Okay sure. What is your favorite tattoo style?*
S:	*I love traditional American style tattoos! Bold lines, vibrant colors. Do you have any tattoos yourself? What's your daily routine like when caring for them?*
U:	*Yes. I just make sure to keep them moisturized.*
S:	*That's great! Moisturizing is key. Do you have a specific lotion you prefer?* ***Some people find tattoo care as essential as their morning coffee. Are you a coffee drinker?***

Table 12: Example of an abrupt utterance introducing an unnatural relationship to associate the QUESTION with the dialogue context or the TOPIC.

of abruptness.

Table 16 shows examples of our dataset's system utterances for the relationship types.

I Details of strategy-based system

I.1 Details of key utterance candidates

Preparing response candidates for all seven relationship types in advance and rewriting every candidate at each turn using an LLM would incur a high computational cost. Thus, we automatically evaluated the abruptness of generated key utterance prototypes and selected the top four prototypes that are considered the least abrupt as key utterances before the chat. The selection process is similar to the automatic abruptness evaluation in Section 5.1.2, except for the task instruction (Prompt 17) and the use of chat history; this evaluation process does not use chat history. We fine-tuned GPT-4o for this evaluation process with the training instances extracted from the training set for the automatic abruptness evaluator of Section 6.2.[15] Specifically, we extracted system utterances that elicited the user information and their abruptness scores described in Section 5.1.2 from the training set. We also extracted these system utterances and their abruptness scores from the test set of Section 6.2 as test samples to assess the performance of the key utterance prototype evaluator. Finally 610 and 119 instances were collected for the training and test, respectively. The detection performance (F1) for the non-abrupt utterances in this test was 67 (recall: 60, precision: 75) by the fine-tuned evaluator.

The key utterance prototypes are ranked in descending order of the probability that this evaluator gives them a score of 3 (indicating "non-abrupt").

I.2 Details of training evaluator LLM

The training of the evaluator LLM is the same as the training of the one in Section 5.1.2, except for the following two points. The first is the scale of the training data. In addition to the training samples of Section 5.1.2, 450 chats collected in Section 4.2 were added. The second is the hyperparameters for the fine-tuning. The number of epochs was reduced in line with the increase in the size of the training data. Specifically, we set the number of epochs to 1, the batch size to 2 and the LR multiplier to 2.

[15]We used the OpenAI API for the fine-tuning. The OpenAI API automatically set the hyperparameters, which were 3 epochs, 1 batch size, and 2 LR multipliers.

Topic: Motorcycle Question: *Do you like to spend time in the kitchen?*
S: *Hi! Let's talk about Motorcycle!*
U: *motorcycles? ok, let's talk about them.*
S: *Cool! Do you ride?*
U: *sure! i do*
S: *That's awesome! I bet you've seen some amazing scenery on your rides.*
U: *very often*
S: *Must be a great way to clear your head after a long day, right?*
U: *helps me to cool off*
S: *I bet!* ***What do you do to unwind after a long day?***

Table 13: Example of an abrupt utterance focusing too much on the QUESTION after the introduction of a natural relationship to associate the QUESTION with the dialogue context or the TOPIC.

Topic: Gossip Question: *Do you enjoy swimming?*
S: *Hi! Let's talk about Gossip!*
U: *Okay let's gossip*
S: *Have you heard any interesting gossip lately?*
U: *Yes I traveled to a new place recently*
S: *That sounds exciting! Did you meet any interesting people during your trip?*
U: *Yes I did, I went on a lot of adventures such as swimming, camping*
S: *That sounds amazing!* ***Did you enjoy the swimming part?***

Table 14: Example of an abrupt utterance trying to continue talking about the QUESTION even though user information has been obtained.

Type of abrupt utterance		Freq. in the 103 success chats.	Freq. in the 70 failed chats.
1 SUB-THEME	TOPIC can feature goods, events, or other things related to QUESTION, or vice versa.	31	10
2 PLACE	TOPIC can be the place, organization or event where the event related to QUESTION occurs, or vice versa.	9	1
3 MEANS	TOPIC can be a means to achieve a goal related to QUESTION, or vice versa.	18	10
4 CO-OCCUR	TOPIC can occur or exist at the same time (or before or after) as the event or object related to QUESTION, or vice versa.	13	12
5 CAUSE	TOPIC can be the cause of the event, situation or state related to QUESTION, or vice versa.	10	2
6 PREREQUISITE	TOPIC can be a prerequisite for dealing with something related to QUESTION, or vice versa.	3	2
7 DOER	TOPIC can be done by QUESTION, or vice versa.	5	2
8 COMMONALITY	TOPIC has common points with something related to QUESTION, or vice versa.	5	15
9 NO RELATION	The relationship between TOPIC and QUESTION is not introduced.	9	16

Table 15: The frequency of each relationship type counted in 103 successful chats and 70 randomly sampled failed chats.

1 SUB-THEME	TOPIC: Parkour QUESTION: Are you an avid stamp collector? Utterance: Have you considered incorporating parkour-inspired designs or themes into any collections you might have?
2 PLACE	TOPIC: Luxury yacht QUESTION: Are you american? Utterance: I plan on one (yacht) going on one when I vacation in the USA. Have you ever been there?
3 MEANS	TOPIC: Mountaineering QUESTION: Do you have a fascination with trains? Utterance: Speaking of transportation, some mountaineers use unique ways to reach base camps. Have you ever had a fascination with t-rains in mountainous regions?
4 CO-OCCUR	TOPIC: Indian cuisine QUESTION: Do you like to exercise before work? Utterance: It is said that exercising before eating spicy food might help. Do you like to exercise before working or eating spicy foods?
5 CAUSE	TOPIC: Balloon animals QUESTION: Are you a fan of minimalist living? Utterance: I haven't tried making a goose, but I've seen some impressive balloon animals! Do you prefer simple or complex designs?
6 PREREQUISITE	TOPIC: Butterfly QUESTION: Do you like to raise animals? Utterance: Butterflies undergo 4 different stages - From the egg, to the lava, the next is the pupa stage and finally the Adult stage. Do you like to raise animals?
7 DOER	TOPIC: Zorbing QUESTION: Are you an amateur winemaker? Utterance: That's okay! Do you think aliens would enjoy human activities like art or ballet?
8 COMMONALITY	TOPIC: Parkour QUESTION: Are you an avid stamp collector? Utterance: What are your usual hobbies? Anything adventurous like Zorbing?

Table 16: Examples of system utterances in our dataset based on the eight relationship types.

J Employed prompts

```
The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."
Each chat ends when the line number reaches 18.

In each chat, the chatbot:
   - Naturally and subtly asks the user as many questions listed under "QUESTIONS" as possible and gets the answers by the end of the chat.
      - Ensures that these questions are seamlessly woven into the chat.
      - May rephrase these questions to suit the specified topic and the chat flow, or break a single question into smaller parts and ask them over multiple turns.
      - Keeps trying to ask questions until the chatbot gets clear answers.
      - Gets the user's answers so that non-participants can accurately guess them based on the chat.
   - Does not deviate from the topic specified in "TOPIC."
      - Ensures that as many utterances as possible relate directly to the specified topic.
      - Brings the chat topic to the specified one naturally if it deviates.
   - Avoids utterances that the user might find abrupt.
      - Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the specified topic and the context.
      - Utterances are considered abrupt if they introduce content seemingly unrelated to the specified topic or the context, attempt to delve into the topic from an unnatural angle, or involve unnatural associations.
   - Ensures each response follows the format "Line number CHATBOT: Response."
   - Ensures each response consists of no more than 30 words.


# EXAMPLE-1

## TOPIC
   - [TOPIC]
## QUESTIONS
   - [QUESTION]

## CHAT ABOUT THE SPECIFIED TOPIC "[TOPIC]"
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]
```

Prompt 1: Prompt used to make LLMs generate responses with task instructions only. The parts enclosed by [] are replaced by the actual values.

```
The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."
For each utterance in each chat, rate on a 3-point scale the abruptness of the utterance.
Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the specified topic and the context established by previous exchanges in the chat.
Utterances are considered abrupt if they introduce content seemingly unrelated to the specified topic or the context, attempt to delve into the specified topic from an unnatural angle, or involve unnatural associations.

The 3-point scale is defined as follows:
   3: Most people would not find the utterance as abrupt.
   2: Some people might find the utterance abrupt; it might or might not be considered abrupt, depending on individual interpretation.
   1: Many people would find the utterance abrupt.

The response format is as follows:
1 CHATBOT: 3-point rating score for this utterance
2 USER: 3-point rating score for this utterance
3 CHATBOT: 3-point rating score for this utterance
...


# EXAMPLE-1

## TOPIC
   - [TOPIC]

## CHAT
   1 CHATBOT: [First utterance of the system role]
   2 USER: [First utterance of the user role]
   3 CHATBOT: [Second utterance of the system role]
   4 USER: [Second utterance of the user role, and further utterances, if any.]
```

Prompt 2: Prompt used to make LLMs evaluate the abruptness of system utterances with task instructions only. The parts enclosed by [] are replaced by the actual values.

The examples below show chats between a chatbot (CHATBOT) and its users (USER).
In each chat, the chatbot attempts to extract the user's answers to the questions listed under "QUESTIONS."
For each question in each chat, rate on a 3-point scale whether the chatbot has elicited enough information to infer the user's correct answer.

The 3-point scale is defined as follows:
3: The information obtained from the chat allows a clear and accurate inference of the user's answer to the question.
- For example, for the question "Do you like basketball?" if the user mentions, "I enjoy all ball games," it is logical to infer the user likes basketball and assign a 3 with a "Yes" prediction.
2: The information obtained from the chat allows a tentative guess of the user's answer, although it comes with a degree of uncertainty due to ambiguous or incomplete information.
- For example, if the user says "I like most ball games" in response to liking basketball, infer a tentative "Yes" but note the uncertainty with a 2.
1: The chat provides insufficient information to make any guess regarding the user's answer.
- If the user simply says, "I often play sports," it does not allow for any reasonable inference about their interest in basketball, resulting in a 1 with a "CannotGuess" prediction.
If you rate 2 or 3, infer whether the user's answer to the questions is "Yes" or "No."
If the score is 1, indicate that you cannot guess the user's answer with "CannotGuess."

The response format is as follows:
Q1: {3-point rating score for Q1}/{Predicted user answer to Q1 (Yes/No/CannotGuess)}
Q2: {3-point rating score for Q2}/{Predicted user answer to Q2 (Yes/No/CannotGuess)}
...

EXAMPLE-1

CHAT
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]

QUESTIONS
Q1: [QUESTION]

Prompt 3: Prompt used to make LLMs predict the user information with task instructions only. The parts enclosed by [] are replaced by the actual values.

```
# Background
The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."
Each chat ends when the line number reaches 18.

In each chat, the chatbot:
  - Naturally and subtly asks the user as many questions listed under "QUESTIONS" as possible and gets the answers by the end of the chat.
    - Ensures that these questions are seamlessly woven into the chat.
    - May rephrase these questions to suit the specified topic and the chat flow, or break a single question into smaller parts and ask them over multiple turns.
    - Keeps trying to ask questions until the chatbot gets clear answers.
    - Gets the userś answers so that non-participants can accurately guess them based on the chat.
  - Does not deviate from the topic specified in "TOPIC."
    - Ensures that as many utterances as possible relate directly to the specified topic.
    - Brings the chat topic to the specified one naturally if it deviates.
  - Avoids utterances that the user might find abrupt.
    - Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the specified topic and the context.
    - Utterances are considered abrupt if they introduce content seemingly unrelated to the specified topic or the context, attempt to delve into the topic from an unnatural angle, or involve unnatural associations.
  - Ensures each response follows the format "Line number CHATBOT: Response."
  - Ensures each response consists of no more than 30 words.


# Task
The final utterance of the chatbot in each chat feels abrupt to humans as an utterance in chatting about TOPIC.
Rewrite the utterance so that the main theme of the utterance feels more like the "TOPIC" while considering the Background.


# Tips
  - Utterances that fall into any of the following categories tend to be perceived as abrupt.
    - Utterances that focus on the unnatural or implausible connection between TOPIC and QUESTION.
    - Utterances that focus too much on the QUESTION and have little direct relevance to TOPIC.
  - If the connection between TOPIC and QUESTION that the chatbot is focusing on does not seem plausible at first glance, adding a supplementary explanation of the connection may help to alleviate the abruptness.
  - If asking QUESTION directly is difficult, it may be effective to ask about customs and behaviors related to TOPIC that depend on the answers to QUESTION.


# EXAMPLE-1

## TOPIC
  - [TOPIC]

## QUESTIONS
  - [QUESTION]


## CHAT ABOUT THE SPECIFIED TOPIC "[topic]"
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]
[t] CHATBOT: [The t-th system utterance]
```

Prompt 4: Prompt used to make LLMs rewrite the t-th system utterance to reduce the abruptness of the utterance, with task instructions only. The parts enclosed by [] are replaced by the actual values.

```
The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."
Each chat ends when the line number reaches 18.

In each chat, the chatbot:
  - Does not deviate from the topic specified in "TOPIC."
    - Ensures that as many utterances as possible relate directly to the specified topic.
    - Brings the chat topic to the specified one naturally if it deviates.
  - Avoids utterances that the user might find abrupt.
    - Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the specified topic and the context.
    - Utterances are considered abrupt if they introduce content seemingly unrelated to the specified topic or the context, attempt to delve into the topic from an unnatural angle, or involve unnatural associations.
  - Ensures each response follows the format "Line number CHATBOT: Response."
  - Ensures each response consists of no more than 30 words.


# EXAMPLE-1

## TOPIC
  - [TOPIC]
## CHAT ABOUT THE SPECIFIED TOPIC "[TOPIC]"
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]
```

Prompt 5: Prompt used to make LLMs generate responses without the instruction to acquire user information. The parts enclosed by [] are replaced by the actual values.

```
# Background
The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."
Each chat ends when the line number reaches 18.

In each chat, the chatbot:
  - Does not deviate from the topic specified in "TOPIC."
    - Ensures that as many utterances as possible relate directly to the specified topic.
    - Brings the chat topic to the specified one naturally if it deviates.
  - Avoids utterances that the user might find abrupt.
    - Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the specified topic and the context.
    - Utterances are considered abrupt if they introduce content seemingly unrelated to the specified topic or the context, attempt to delve into the topic from an unnatural angle, or involve unnatural associations.
  - Ensures each response follows the format "Line number CHATBOT: Response."
  - Ensures each response consists of no more than 30 words.


# Task
The final utterance of the chatbot in each chat feels abrupt to humans as an utterance in chatting about TOPIC.
Rewrite the utterance so that the main theme of the utterance feels more like the "TOPIC" while considering the Background.


# EXAMPLE-1

## TOPIC
  - [TOPIC]

## QUESTIONS
  - [QUESTION]


## CHAT ABOUT THE SPECIFIED TOPIC "[topic]"
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]
[t] CHATBOT: [The t-th system utterance]
```

Prompt 6: Prompt used to make LLMs rewrite the t-th system utterance to reduce the abruptness of the utterance, without the instruction to acquire user information. The parts enclosed by [] are replaced by the actual values.

```
The following are examples of topics for casual conversation.
List 200 other topics to augment this list:

[The list of TOPICs used in Section 4.1.]
```

Prompt 7: Prompt used to make LLMs generate TOPIC candidates. The parts enclosed by [] are replaced by the actual values.

```
Please create a list that excludes items that are semantically almost the same from the following topic list:

[The list of generated TOPIC candidates.]
```

Prompt 8: Prompt used to make LLMs remove duplicate TOPIC candidates. The parts enclosed by [] are replaced by the actual values.

```
The following are examples of profile sentences.
List 200 other profile sentences to augment this list:

[The list of user information sentences used in Section 4.1.]
```

Prompt 9: Prompt used to make LLMs generate user information sentence candidates. The parts enclosed by [] are replaced by the actual values.

```
Please create a list that excludes items that are semantically almost the same from the following profile sentence list:

[The list of generated user information sentence candidates.]
```

Prompt 10: Prompt used to make LLMs remove duplicate user information sentence candidates. The parts enclosed by [] are replaced by the actual values.

In the following chat (CHAT) on a predefined topic (TOPIC), a chatbot (CHATBOT) subtly asked questions at the asterisked utterance to get the answer of a user (USER) to the specified QUESTION.
One effective technique for subtly obtaining the answer to a QUESTION in a TOPIC-related chat is to explicitly add the reason for asking the questions to the same utterance.
Your task is to classify whether the reason for asking the questions in the asterisked utterance is explicitly added in the same utterance.
If it is, output "Yes," otherwise output "No."

TOPIC
[TOPIC]

QUESTION
[QUESTION]

CHAT
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]
$*$ [i] CHATBOT: [The i-th system utterance]

Prompt 11: Prompt used to make LLMs determine the presence of explicit explanation on the relationship between TOPIC and QUESTION in the i-th utterance.

In the following chat (CHAT) on a predefined topic (TOPIC), a chatbot (CHATBOT) subtly asked questions at the asterisked utterance to get the answer of a user (USER) to the specified QUESTION.
One effective technique for subtly eliciting the answer to a QUESTION in a TOPIC-related chat is to explicitly add the reason for asking the questions to the same utterance, in a way that mentions its relevance to the TOPIC and previous interactions.
Your task is to rewrite the asterisked utterance by adding a sentence that clearly explains the reason for asking the question in the same utterance in a way that mentions its relevance to the TOPIC and previous interactions.
The only possible change to the utterance is to add a sentence that clearly explains the reasons and you must not change any other part of the utterance.

TOPIC
[TOPIC]

QUESTION
[QUESTION]

CHAT
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]
$*$ [i] CHATBOT: [The i-th system utterance]

Prompt 12: Prompt used to make LLMs explicitly add the explanation to the key utterances.

Background
Given a chat topic (TOPIC) and a question (QUESTION), in a TOPIC-related chat, a chatbot tries to subtly elicit the information from which the user's answer to the specified QUESTION (ANSWER) can be inferred.
One effective way to get ANSWER is to actively introduce the strong and necessary relationship between TOPIC and QUESTION during the chat.

Task
Given TOPIC, QUESTION, and a relationship type (RELATIONSHIP-TYPE), please find a specific relationship between TOPIC and QUESTION in the RELATIONSHIP-TYPE and present an example of the utterance (UTTERANCE) that uses the found relationship to subtly elicit the information from which ANSWER can be inferred.

Output format
SPECIFIC-RELATIONSHIP: A description of the found specific relationship between TOPIC and QUESTION based on the given RELATIONSHIP-TYPE.
EXPLANATION-FOR-RELATIONSHIP-TYPE: Explanation of whether SPECIFIC-RELATIONSHIP is based on the given RELATIONSHIP-TYPE.
EXPLICIT-REASON: Reason for asking the question in a way that mentions its relevance to TOPIC. Note that EXPLICIT-REASON should take into account SPECIFIC-RELATIONSHIP.
UTTERANCE: An example of the utterance that is based on SPECIFIC-RELATIONSHIP and EXPLICIT-REASON to subtly elicit ANSWER. Ensure that the content of the EXPLICIT-REASON is included in the utterance.

Notes on the example utterance
- TOPIC must be the main topic of the utterance.
- EXPLICIT-REASON must be based on the RELATIONSHIP-TYPE.
- Explicitly include EXPLICIT-REASON into UTTERANCE.
- Rephrase QUESTION to better fit RELATIONSHIP-TYPE and TOPIC.
  - Including specific words from QUESTION in UTTERANCE can easily feel abrupt. You can abbreviate or omit such words.
- Avoid making any assumptions about the user's background, interests, or profession.
  - Ensure that the questions remain general and can be relevant to anyone, without implying that the user has specific experiences or roles related to the TOPIC.
  - Use neutral language that does not presume the user's involvement or interest in TOPIC beyond general curiosity.
- Avoid an utterance that the user might find abrupt.
  - Utterances are considered abrupt if they introduce content seemingly unrelated to TOPIC, attempt to delve into TOPIC from an unnatural angle, or involve unnatural associations.
- Ensures the utterance consists of no more than 30 words.

TOPIC
- [TOPIC]
QUESTION
- [QUESTION]
RELATIONSHIP-TYPE
- [One of the seven relationship types in Table 4]

Prompt 13: Prompt used to make LLMs generate key utterance prototypes.

Background
Given a chat topic (TOPIC) and a question (QUESTION), in TOPIC-related chat (CHAT), a chatbot (CHATBOT) tries to subtly elicit the information from which the user's (USER) answer to the specified QUESTION can be inferred.
Specifically, the CHATBOT will elicit the information from the USER by outputting an utterance rewritten from the utterance described in PLANNED UTTERANCE to fit the current CHAT.

Task
Given TOPIC, QUESTION, CHAT, and PLANNED UTTERANCE, please rewrite PLANNED UTTERANCE to make it fit contextually as the next utterance of the CHATBOT following the USER's last utterance in the CHAT.

Notes on the output utterance
- TOPIC must be the main topic of the utterance.
- Avoid an utterance that the user might find abrupt.
  - Utterances are considered abrupt if they introduce content seemingly unrelated to TOPIC, attempt to delve into TOPIC from an unnatural angle, or involve unnatural associations.
- Include reactions to the USER's utterance in the rewritten utterance.
- Ensures the utterance consists of no more than 30 words.
- Ensures the utterance follows the format "Line number CHATBOT: Utterance."

TOPIC
[TOPIC]
QUESTION
[QUESTION]
CHAT
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]

PLANNED UTTERANCE
$[i]$ CHATBOT: [The prepared key utterance]

Prompt 14: Prompt used to make LLMs rephrase prepared key utterance prototypes to fit the ongoing chat.

Background
Given a chat topic (TOPIC) and a question (QUESTION), in TOPIC-related chat (CHAT), a chatbot (CHATBOT) tries to subtly elicit the information from which the user's (USER) answer to the specified QUESTION can be inferred.
Specifically, the CHATBOT will elicit information by introducing the utterance described in PLANNED UTTERANCE below in the next turn.
In order for the CHATBOT to introduce the PLANNED UTTERANCE in its next turn without any abruptness, the CHATBOT must first make an utterance in this turn that will act as a subtle cushion for a non-abrupt introduction of the CHATBOT's PLANNED UTTERANCE.

Task
Given TOPIC, QUESTION, CHAT, and PLANNED UTTERANCE, please present the CHATBOT's next utterance following the USER's last utterance in the CHAT.
The CHATBOT's utterance you present should act as a subtle cushion for a non-abrupt introduction of the CHATBOT's PLANNED UTTERANCE in the next turn.

Notes on the output utterance
- TOPIC must be the main topic of the utterance.
- Avoid an utterance that the user might find abrupt.
  - Utterances are considered abrupt if they introduce content seemingly unrelated to TOPIC, attempt to delve into TOPIC from an unnatural angle, or involve unnatural associations.
- Ensures the utterance consists of no more than 30 words.
- Ensures the utterance follows the format "Line number CHATBOT: Utterance."

TOPIC
[TOPIC]
QUESTION
[QUESTION]
CHAT
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]

PLANNED UTTERANCE
$[i]$ CHATBOT: [The prepared key utterance]

Prompt 15: Prompt used to make LLMs generate cushion utterances.

The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."
Each chat ends when the line number reaches 18.

In each chat, the chatbot:
- Naturally and subtly asks the user as many questions listed under "QUESTIONS" as possible and gets the answers by the end of the chat.
    - Ensures that these questions are seamlessly woven into the chat.
    - May rephrase these questions to suit the specified topic and the chat flow, or break a single question into smaller parts and ask them over multiple turns.
    - Keeps trying to ask questions until the chatbot gets clear answers.
    - Gets the users´ answers so that non-participants can accurately guess them based on the chat.
- Does not deviate from the topic specified in "TOPIC."
    - Ensures that as many utterances as possible relate directly to the specified topic.
    - Brings the chat topic to the specified one naturally if it deviates.
- Avoids utterances that the user might find abrupt.
    - Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the specified topic and the context.
    - Utterances are considered abrupt if they introduce content seemingly unrelated to the specified topic or the context, attempt to delve into the topic from an unnatural angle, or involve unnatural associations.
- Ensures each response follows the format "Line number CHATBOT: Response."
- Ensures each response consists of no more than 30 words.

EFFECTIVE WAYS TO SUBTLY ELICIT ANSWER
- Actively introduce the strong and necessary relationship between TOPIC and QUESTION.
    - The following are examples of the relationship types between TOPIC and QUESTION:
        1. TOPIC can feature goods, events, or other things related to QUESTION, or vice versa.
        2. TOPIC can be the place, organization or event where the event related to QUESTION occurs, or vice versa.
        3. TOPIC can be a means to achieve a goal related to QUESTION, or vice versa.
        4. TOPIC can occur or exist at the same time (or before or after) as the event or object related to QUESTION.
        5. TOPIC can be the cause of the event, situation or state related to QUESTION, or vice versa.
        6. TOPIC can be a prerequisite for dealing with something related to QUESTION, or vice versa.
        7. TOPIC can be done by QUESTION, or vice versa.
- Include the reason for asking the question about QUESTION into the response explicitly in a way that mentions its relevance to TOPIC.
- Refrain from chatting about QUESTION after you have obtained enough information to guess the user's answer to QUESTION.
- Make a response that will act as a subtle cushion for a non-abrupt introduction of the question about QUESTION, when it is difficult to subtly obtain the user's answer to QUESTION with a single turn.

EXAMPLE-1

TOPIC
- [TOPIC]
QUESTIONS
- [QUESTION]

CHAT ABOUT THE SPECIFIED TOPIC "[TOPIC]"
1 CHATBOT: [First utterance of the system role]
2 USER: [First utterance of the user role]
3 CHATBOT: [Second utterance of the system role]
4 USER: [Second utterance of the user role, and further utterances, if any.]

Prompt 16: Prompt used to make LLMs generate responses with task instructions and the insights from Section 5. The parts enclosed by [] are replaced by the actual values.

Given a chat topic (TOPIC), please rate the abruptness of the following utterance (UTTERANCE) as an utterance in a chat about TOPIC on a 3-point scale.
Abruptness here refers to the degree to which an utterance deviates from the expected flow of the chat based on the TOPIC.
Utterances are considered abrupt if they introduce content seemingly unrelated to the TOPIC, attempt to delve into the TOPIC from an unnatural angle, or involve unnatural associations.

The 3-point scale is defined as follows:
3: Most people would not find the utterance as abrupt.
2: Some people might find the utterance abrupt; it might or might not be considered abrupt, depending on individual interpretation.
1: Many people would find the utterance abrupt.

TOPIC
- [TOPIC]
UTTERANCE
- [Key utterance prototype]

Prompt 17: Prompt used to make LLMs evaluate the abruptness of key utterance prototypes. The parts enclosed by [] are replaced by the actual values.

The examples below show chats between a chatbot (CHATBOT) and its users (USER) on the topics specified in "TOPIC."

In each chat, the user:
- Enjoys chatting with the chatbot on the topic specified in "TOPIC" and avoids introducing unrelated subjects or questions.
- Does not make any utterances that naturally lead to the chat's conclusion.
  - For example, the utterance "Thanks for chatting with me today." may be taken as an implicit suggestion to end the chat.

Does not respond inconsistently with the user profiles listed under "PROFILES."
Note that the profiles serve as background information rather than subjects to be forcefully incorporated into the chat.
For example, if a profile states, "My father is a doctor," and the chat turns to medical professionals in the family, it is appropriate to mention this profile.
However, if the chat does not relate to medical topics, such profiles should not be introduced without context.
Ensures each response follows the format "Line number USER: Response."
Ensures each response consists of no more than 30 words.

EXAMPLE-1

TOPIC
- [TOPIC]

PROFILES
- [First persona sentence]
- [Second persona sentence, and further sentences, if any.]

CHAT ABOUT THE SPECIFIED TOPIC "[TOPIC]"

Prompt 18: Prompt used to emulate a user. The parts enclosed by [] are replaced by the actual values.