

Topic-Initiator: A Proactive Chatbot with Personalized Topic RAG for Enhancing Willingness to Converse

Kazuya Matsuo Atsushi Otsuka, Narichika Nomoto, Makoto Nakatsuji

Human Informatics Labs., NTT, Inc.

kazuya.matsuo@ntt.com

Abstract

Stimulating users' conversational willingness to converse remains a major challenge in chatbot research. Most existing chatbots respond passively to user inputs, relying on users to select conversation topics, which often reduces their willingness. To address this issue, we propose, *Topic-Initiator*, a proactive chatbot that initiates conversations with new topics aligned to user interests. It gathers information from external sources (e.g., the web) to obtain potentially novel and engaging topics. To support this capability, we also introduce a novel Retrieval-Augmented Generation (RAG) framework, *Personalized-Topic RAG (PT-RAG)*, designed to retrieve new and interesting topics for each user. Unlike existing RAG methods that fails to surface unseen information, PT-RAG leverages the inference capabilities of Large Language Models (LLMs) to identify content that matches the user's interests but is not yet known to them. Specifically, PT-RAG estimates a user's interests and knowledge from past interactions and organizes collected information into categories. Then, it uses an LLM to select a category that matches their interests and obtain information not seen in their knowledge from the selected category. Automatic and human evaluations demonstrate that PT-RAG retrieves new and interesting information more accurately and that Topic-Initiator significantly enhances users' willingness to converse compared to existing methods.

Keywords: chatbot, dialogue system, RAG

1. Introduction

Recent advances in Large Language Models (LLMs) have enabled AI systems to engage in more natural conversations. As a result, various chatbot services have emerged^{1,2}. For users to continue using chatbots, it is important to sustain their motivation to interact with them. However, many existing chatbots are passive and can only respond to user inputs. This forces users to come up with topics to start conversations, which can lead to fatigue and a gradual loss of motivation to engage. Such passivity to collect not only reduces user satisfaction, but also limits opportunities to collect valuable conversational data for analyzing user insights, such as in marketing applications.

To address this issue, we believe that it is effective for chatbots to proactively select topics that stimulate users' willingness to converse and initiate conversation with them (Figure 1). Potential topics may include questions related to disclosed user activities, such as watching a movie or taking a short trip (Zhang et al., 2023). However, overusing such questions can lead to one-sided information requests, which are unlikely to sustain motivation (Cozby, 1972). Uchida et al. (2024) demonstrated that people are more willing to converse with robots that discuss topics of personal interest while Pentina et al. (2023) found that users tend to lose motivation when chatbots provides fixed, generic responses and fail to consider prior inter-

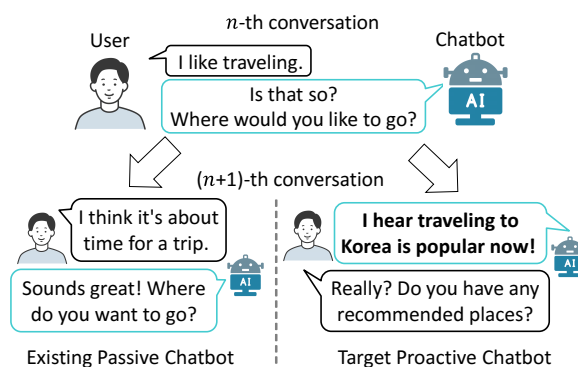


Figure 1: Existing vs. target chatbot examples. The left chatbot requires user input to begin a conversation, whereas our target chatbot can initiate it on its own.

actions. On the basis of these findings, we hypothesize that a chatbot can enhance users' motivation to converse by initiating conversations on (1) topics that align with interests revealed in previous interactions and (2) topics that vary across conversations.

We propose a method called *Topic-Initiator*, designed to proactively provide users with topics that are both interesting and novel—that is, topics not mentioned in previous conversations. Topic-Initiator first collects information from the web and stores it as potential conversation topics (hereafter referred to as *topic candidates*). It also stores past conversations with non-target users as additional topic candidates, enabling these to be reused to

¹<https://character.ai/>

²<https://cotomo.ai/>

enrich future interactions with target users. In each new conversation, Topic-Initiator selects a flesh and engaging topic from these candidates and uses it to initiate conversation.

To support this process, we introduce a novel Retrieval-Augmented Generation framework, called *Personalized-Topic RAG (PT-RAG)*, which enables Topic-Initiator to retrieve new and interesting topics effectively. Existing RAG methods (Zhu et al., 2025; Zhang et al., 2023; Wang et al., 2024; Zerhoubi, 2025) are optimized to answer user queries by retrieving information similar to those queries and therefore tend to reproduce content already present in past interactions, which limits their usefulness for new topic discovery. PT-RAG addresses this gap by using LLM inference to identify topic candidates that are absent from the user’s existing knowledge data while constraining retrieval to a relevant topic category to keep inputs tractable.

Topic-Initiator initiates conversations using topics retrieved by PT-RAG, updates the user’s interests and knowledge based on the conversation, and then gathers new topic candidates. By repeating this cycle, Topic-Initiator can continuously provide users with novel and interesting topics.

To demonstrate the effectiveness of Topic-Initiator and PT-RAG, this study addresses the following research questions:

RQ1: Can we obtain new and interesting topics for a user by categorizing and retrieving them solely from categories that align with the user’s interests?

RQ2: Can a chatbot stimulate a user’s willingness to converse by initiating a conversation using a topic that is new and interesting to the user?

Our contributions are summarized as follows:

- We propose Topic-Initiator, a proactive chatbot that continuously provides users with new and interesting topics along with Personalized-Topic RAG (PT-RAG), a novel RAG framework specifically designed to retrieve such topics.
- Through automatic evaluations, we demonstrate that PT-RAG retrieves topics more accurately than existing methods.
- Through human evaluation, we show that Topic-Initiator significantly stimulate users’ willingness to converse compared with existing chatbots.

2. Related Work

2.1. Proactive Chatbots

Extensive research has been conducted on proactive chatbots and dialogue systems (Deng et al., 2023a). Tang et al. (2019); Wu et al. (2019); Wang et al. (2023a,b); Deng et al. (2023b) proposed dialogue systems that set conversational goals and actively guide users toward achieving them. Xie et al. (2021) introduced a method in which dialogue systems proactively generate utterances to switch topics during casual conversations. Although these studies aim to build proactive chatbots similar to ours, they overlook scenarios in which the chatbot itself initiates the conversation—a key distinction in our work. Li et al. (2023) proposed a method where chatbots engage in conversations by exploring trending news. While this approach enables a chatbot to proactively initiate conversations, it differs from our work in that it disregards user interests and relies on fixed topics.

2.2. Personalized Chatbots

Numerous studies have also explored personalized chatbots and dialogue systems (Chen et al., 2024). The term “personalized” has two definitions (Chen et al., 2024): (1) generating responses consistent with the system’s own persona, and (2) generating responses tailored to the user’s persona. Our work focuses on the latter.

Zhong et al. (2022) proposed a method that generates responses using profiles of a target user and similar users. Jang et al. (2023) presented a method that enables a chatbot to adapt its responses over time by training it on a dataset encoding elapsed time and multiple types of interpersonal relationships. Xu et al. (2022); Zhong et al. (2024); Packer et al. (2023); Bae et al. (2022); Lee et al. (2023); Lim et al. (2022) introduced approaches that extract personal data from user utterances, store it, and generate responses based on that information. While these methods personalize chatbot responses, they do not address the proactive initiation of conversations, which is a key aspect of our research.

2.3. Personalized RAG

The rise of LLMs has driven rapidly growing interest in Retrieval-Augmented Generation (RAG) approaches (Lewis et al., 2020; Zhu et al., 2025). In this section, we focus on methods that personalize the outputs of RAG.

Zhang et al. (2023) proposed a method that asks about the progress of events previously disclosed by a user based on the time elapsed since

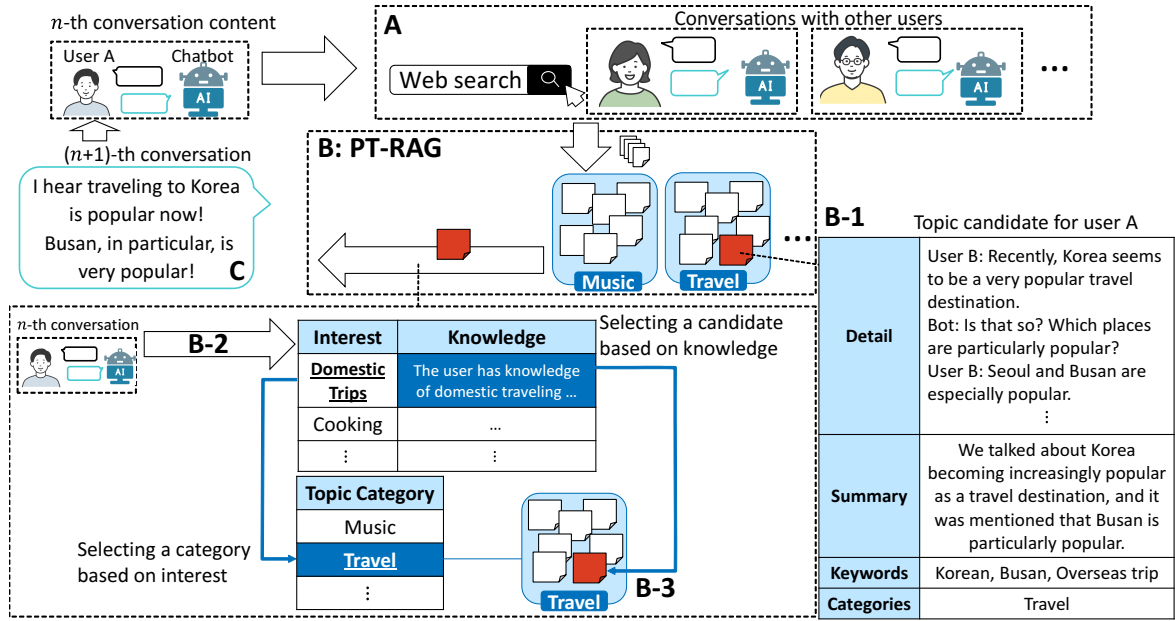


Figure 2: Overview of Topic-Initiator.

the event occurred. Wang et al. (2024) introduced a method that personalizes both information retrieval and answer generation using a user’s conversation history. While these methods personalize responses, they are designed for clearly answerable scenarios and rely on fine-tuning, often overlooking the retrieval of information not mentioned in previous conversations.

Zerhoudi (2025) presented an approach that combines retrieved information and user data to personalize responses to queries. Their method employs a multi-agent system to rerank retrieved documents and recursively refines responses, ensuring they align with user interests. Although similar to our work in personalizing responses, it focuses on clearly answerable scenarios rather than providing topics to enhance user motivation for interaction. Nevertheless, this method can be applied to our task without additional training. Therefore, we compare our approach with this method in Section 4.

2.4. Conversational Recommender Systems

Conversational recommender systems (CRS) are designed to actively elicit user preferences through conversation and recommend items based on those insights (Friedman et al., 2023; Deldjoo et al., 2024). Several studies have focused on improving user engagement and recommendation success. For instance, Hayati et al. (2020) integrated sociable recommendation strategies into CRS, while Wang et al. (2022) combined recommendation and conver-

sation modules. Other studies have evaluated whether LLMs can accurately recommend items in zero-shot settings. Kang et al. (2023); Liu et al. (2023); Sanner et al. (2023); Sileo et al. (2022); He et al. (2023) constructed prompts that merge user preferences with instructions to recommend or evaluate items. Although these studies share our goal of understanding user preferences for proactive recommendations, they do not estimate user knowledge or assess the user’s familiarity with the recommended items.

3. Proposed Method: Topic-Initiator

3.1. Architecture

As introduced in Section 1, Topic-Initiator follows an iterative process to continuously provide new topics that align with a target user’s interests. The overall procedure consists of three main steps (A to C), illustrated in Figure 2.

A. Information Gathering: After interacting with a user, Topic-Initiator proactively gathers relevant information based on the conversation content. Specifically, it employs an LLM to generate search queries from the conversation and conducts web searches to collect topic candidates. Additionally, Topic-Initiator stores past conversations with non-target users as topic candidates, allowing them to be reused to enrich future conversations.

B. PT-RAG: Topic-Initiator uses PT-RAG to retrieve a new and interesting topic for the user. The main components of PT-RAG are as follows (explained in detail in the next section):

B-1. Information Organization: PT-RAG determines topic categories (e.g., music, travel) for each candidate, and organizes all candidates within their respective categories.

B-2. Interest and Knowledge Estimation:

After each interaction, PT-RAG estimates the user’s interests and knowledge of each category based on the conversation history.

B-3. Information Retrieval: From the organized candidates, PT-RAG employs an LLM to retrieve a topic that is both new and interesting to the user, considering their inferred interests and knowledge.

C. Initial Utterance Generation: At the beginning of the next conversation, Topic-Initiator generates an initial utterance based on the topic retrieved by PT-RAG. It then uses this utterance to initiate the conversation, after which the process returns to Step A to continue the cycle.

Notation	Definition
T	Set of topic candidates
t_j	Individual topic candidate in T ($j = 1, \dots, J$)
C	Set of topic category
c_l	Individual topic category in C ($l = 1, \dots, L$)
$c_{t_j,l}$	Individual topic category of t_j
I_n	Set of a target user’s interests at the end of the n -th conversation
$i_{m,n}$	Individual interest data in I_n ($m = 1, \dots, M$)
K_n	Set of a target user’s knowledge at the end of the n -th conversation
$k_{c_l,n}$	Individual knowledge data for c_l
h_n	History of the n -th conversation with the target user
f_R	Function that retrieves an interesting and novel topic t_j
f_C	Function that creates or updates C
f_I	Function that estimates I_n
f_K	Function that estimates $k_{c_l,n}$

Table 1: Terminologies used in PT-RAG.

3.2. Personalized-Topic RAG (PT-RAG)

PT-RAG is designed to retrieve a new and interesting topic for a user by leveraging LLM inference to identify a topic candidate that are absent from a user’s inferred knowledge (derived from previously disclosed information). Because supplying a full dump of a user’s knowledge or the entire candidate pool is impractical given input-length and scale constraints, PT-RAG groups both user knowledge and candidates into topic categories (e.g., music, travel) and performs category-guided retrieval, providing the LLM only with category-relevant summaries and candidates.

In this section, we first define the task in Section 3.2.1, then, describe "Information Organization" in Section 3.2.2, "Interest and Knowledge Estimation" in Section 3.2.2, and "Information Retrieval" in Section 3.2.4.

3.2.1. Terminologies and Task Definition

Terminologies: Table 1 shows the notations used in the following explanations.

Task: Our task is to identify an interesting and novel topic t_j for a target user from the given topic candidate set $T = \{t_j\}_{j=1}^J$, referring to the topic category set $C = \{c_l\}_{l=1}^L$, the user’s interests set $I_n = \{i_{m,n}\}_{m=1}^M$ and the user’s knowledge set $K_n = \{k_{c_l,n}\}_{l=1}^L$. Here, n represents the number of completed conversations with the target user. I_n and K_n are estimated from the given history of the n -th conversation with the user.

This task is formulated as follows:

$$t_j = f_R(I_n, K_n, C', T), \quad \text{where} \quad (1)$$

$$C' = f_C(t_j, C), \quad (2)$$

$$I_n = f_I(h_n, I_{n-1}), \quad (3)$$

$$k_{c_l,n} = f_K(h_n, k_{c_l,n-1}). \quad (4)$$

Here, C' is iteratively redefined from the previous state C by incorporating an additional category, and is therefore dynamic. To maintain readability, we omit explicit counters (e.g., C_i) for these redefinitions.

Details of this task are explained below: Equation (1) takes the outputs of Equations (2)–(4) and computes the topic t_j , which is used when the system initiates a conversation with the user. Details are provided in Section 3.2.4. Equation (2) presents that PT-RAG creates or updates the topic category set C by analyzing each topic candidate t_j . That is, it estimates one or more topic categories $\{c_{t_j,l}\}_l$ of t_j and adds or merges them to C , thereby updating C to C' . Equation (3) presents that the user’s interests I_n are estimated from the history of the n -th conversation h_n . If a previously estimated interest I_{n-1} exists, it is updated into I_n based on h_n . Additionally, h_n is also organized into C . Equation (4) presents that the user’s knowledge $k_{c_l,n} \in K_n$ is estimated based on h_n when the category of h_n is c_l . If there is previously estimated knowledge $k_{c_l,n-1}$, it is updated into $k_{c_l,n}$ based on h_n . Equation (1) presents that PT-RAG retrieves a topic candidate t_j from T that aligns with I_n but is not yet represented in K_n , referring to the created or updated C' . The pro-

Key	Value
Detail	Raw data of the topic candidate, t_j
Summary	Concise summary of the above data
Keywords	Key terms about the summary
Categories	Categories of the keywords, $\{c_{t_j,l}\}_l$

Table 2: Structure of an organized topic candidate.

cedures implemented in f_C , f_I , f_K , and f_R are described in Sections 3.2.2, 3.2.3 (for f_I and f_K), and 3.2.4, respectively.

3.2.2. Information Organization

The LLM estimates three components for each topic candidate t_j : a summary, related keywords, and overarching topic categories $\{c_{t_j,l}\}_l$. These estimated components are structured as key–value pairs, as shown in Table 2. To prevent duplication with the previously estimated category set C , PT-RAG calculates the cosine similarity between C' and C . If the similarity meets or exceeds a predefined threshold α , the system does not update C , as t_j can be associated with existing categories. Otherwise, any new categories are added to the category set. These processes are defined as the function f_C in Equation (2). Each new candidate is represented as a key–value structure, where c_l serves as the key, and the corresponding value consists of the key–value pairs defined in Table 2.

Figure 3 shows an example of this estimation process. The LLM first summarizes the conversation with the user about domestic trips, then identifies three keywords—“hot springs,” “local foods,” and “relaxation”—and assigns two topic categories: “travel” and “domestic trips.”

3.2.3. Interest and Knowledge Estimation

The user’s interests I_n are estimated from the history of the n -th conversation h_n . Specifically, topic categories $\{c_l\}_l$ and keywords are estimated from h_n , using the same method described in Section 3.2.2. The LLM then identifies categories and keywords that reflect the user’s interest among those estimated. These selected elements are represented as $i_{m,n} \in I_n$. If I_{n-1} exists, each $i_{m,n-1}$ is compared with each $i_{m,n}$. When their cosine similarity exceeds the threshold, α , the two are merged; otherwise, the new $i_{m,n}$ are added to I_{n-1} . These processes are defined as the function f_I in Equation (3).

The user’s knowledge related to c_l , denoted as $k_{c_l,n}$, is also estimated from h_n . Specifically, the LLM generates detailed text representing $k_{c_l,n}$, enabling the model to better understand the user’s knowledge. If $k_{c_l,n-1}$ exists, it is also provided to

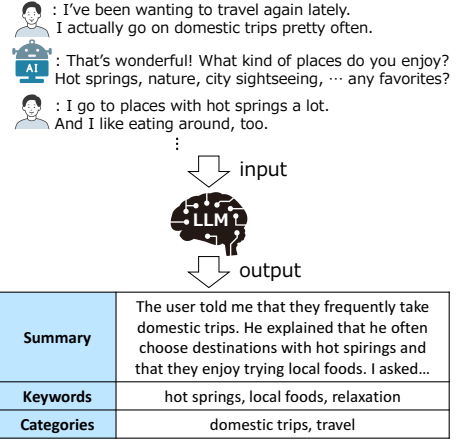


Figure 3: Example of estimating three components (summary, keywords, topic categories).

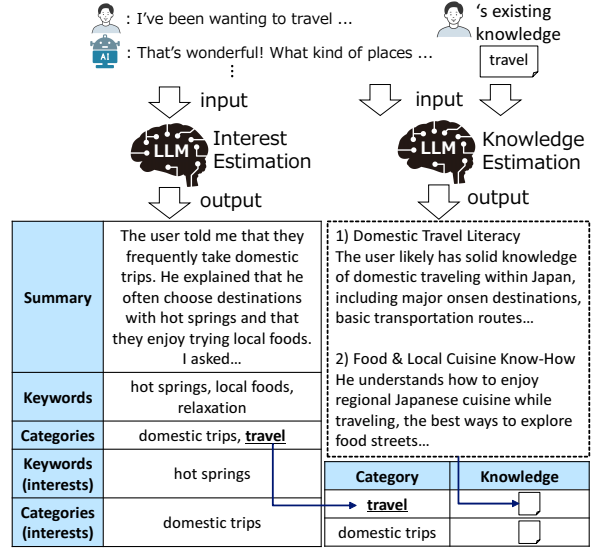


Figure 4: Example of interest and knowledge estimation.

the LLM to update $k_{c_l,n-1}$ to $k_{c_l,n}$. This process is defined as the function f_K in Equation (4).

Figure 4 illustrates an example of these estimation processes. From the estimated keywords and categories of the n -th conversation, the LLM extracts the keyword “hot springs” and the topic category “domestic trips” as the user’s interests. It then estimates the user’s knowledge from the same conversation and the existing knowledge of “travel”. This new knowledge is associated with “travel” since it is derived from information labeled under this category.

The estimated interests and knowledge are stored in a key–value format, where each key corresponds to c_l . Specifically, the value for “interest” is set to the keywords representing the user’s interest, while the value for “knowledge” is set to the detailed text describing their knowledge.

3.2.4. Information Retrieval

The following procedure, defined as the function f_R in Equation (1), is used to retrieve a potentially new and interesting topic for the user:

1. Retrieve the target user's interests I_n and knowledge K_n .
2. Instruct the LLM to select a topic category c_l that aligns with the user's based on I_n and the category set C' (Figure 5(a)).
3. From all topic candidates T , select those categorized under c_l , and obtain the corresponding user knowledge $k_{c_l,n}$ linked to c_l from K_n .
4. Extract the summaries of the selected candidates and compile them into a list.
5. Instruct the LLM to select a summary that is aligned with $i_{m,n}$ but not yet reflected in $k_{c_l,n}$ (Figure 5(b)).
6. Output the topic candidate t_j corresponding to the chosen summary as the retrieved topic.

The retrieved topic is regarded as known to the user and is excluded from future retrievals. Specifically, for each user, a record of previously used topics is maintained, and these topics are omitted when constructing the summary lists in Step 4.

4. Evaluation

4.1. RQ1: Can PT-RAG obtain new and interesting topics for a target user?

To evaluate the performance of PT-RAG, we conducted an automatic experiment comparing it with existing methods.

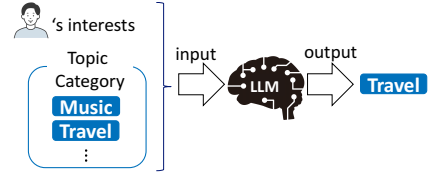
4.1.1. Datasets

To address RQ1, the dataset should include (1) conversation data for estimating user knowledge, (2) information about the speaker's interests, and (3) both content that the speaker clearly knows and content that they may not know. To meet these conditions, we used the following datasets:

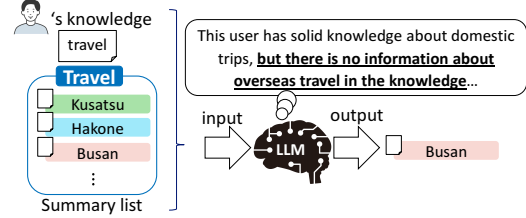
RealPersonaChat (RPC): Contains approximately 14,000 Japanese real-world conversations, each including the real speaker's persona data (Yamashita et al. (2023)).

JPersonaChat (JPC): Contains 5,000 Japanese fictional conversations, each including a fictional speaker's persona data (Sugiyama et al. (2023)).

All conversation data from each dataset were treated as topic candidates for retrieval.



(a) Example of selecting topic category.



(b) Example of selecting topic candidate.

Figure 5: Examples of information retrieval.

4.1.2. Query

To evaluate whether a method can identify novel and interesting topics for a target user, each evaluation method described below addresses the following query:

Query

Provide information that maybe of interest to the user but is likely unknown based on "User Data."
User Data: {user_data}

The generated responses are evaluated using three metrics described in Section 4.1.4.

4.1.3. Compared Methods

We evaluated the performances of the following four methods. All methods used ChatGPT-4o-mini and the persona data from each dataset as the variable {user_data} to represent user interests.

PT-RAG (Ours) : Described in Section 3.2. For computing cosine similarity (Section 3.2.2), we used the NLP library GiNZA³ with a threshold $\alpha = 0.8$. Redis⁴ was used to maintain the structured data. All conversation data from each dataset were structured and stored in the database following the method in Section 3.2.2. Knowledge data for each speaker was prepared in advance using the method in Section 3.2.3 and added to {user_data}.

PT-RAG-Knowledge-Out (PT-RAG-KO) : This variant of PT-RAG does not use the speaker's knowledge. It was evaluated to confirm the

³<https://github.com/megagonlabs/ginza>

⁴<https://redis.io/>

Dataset	Methods	Interest	Novelty	IN-F1
RPC	(a) Vanilla-RAG	9.51 \pm 1.13	4.69 \pm 2.53	5.85 \pm 2.43
	(b) PersonaRAG	9.94 \pm 0.24^{a,c,d}	4.74 \pm 2.0	6.16 \pm 1.93
	(c) PT-RAG (Ours)	9.38 \pm 0.79	6.76 \pm 1.69^{a,b,d}	7.71 \pm 1.29^{a,b,d}
	(d) PT-RAG-KO	9.52 \pm 0.72	5.9 \pm 2.03 ^{a,b}	7.04 \pm 1.67 ^{b,d}
JPC	(a) Vanilla-RAG	9.7 \pm 0.76	4.75 \pm 3.04	5.75 \pm 2.91
	(b) PersonaRAG	9.86 \pm 0.42^{a,c,d}	4.87 \pm 2.11	6.23 \pm 2.03
	(c) PT-RAG (Ours)	9.65 \pm 0.61	5.63 \pm 1.91^{a,b,d}	6.92 \pm 1.69^{a,b,d}
	(d) PT-RAG-KO	9.63 \pm 0.58	4.82 \pm 2.25	6.07 \pm 2.09

Table 3: Evaluation results. Bold items indicate the highest value for each metric. Superscripts a-d indicate significant differences with methods (a)–(d) respectively (two-tailed Welch’s t-test; $p < 0.05$).

effectiveness of the knowledge estimation mechanism described in Section 3.2.3.

Vanilla-RAG : A standard RAG method implemented by the llama-index⁵. It partitions information into fixed-sized chunks and vectorizes each chunk, then retrieves the top k ($K = 5$ in this experiment) most similar chunks to the query vector for response generation. Since Vanilla-RAG does not estimate user knowledge, a summary list of the speaker’s conversation was added as alternative knowledge data to {user_data}.

PersonaRAG :

Introduced by (Zerhoubi, 2025) and described in Section 2. Since this method does not include an original retriever, we used the retriever of Vanilla-RAG. Like Vanilla-RAG, it also lacks user knowledge estimation, so the same information was provided to {user_data}.

4.1.4. Metrics

We evaluated the performance of each method using the following three metrics:

Interest: Following the LLM-as-a-judge framework (Zheng et al., 2023), ChatGPT-4o was used to determine whether each generated response would be interesting to the target speaker. Each response was scored on a 10-point scale (1-10) reference to the speaker’s persona data. The prompt used is provided in Appendix A.4. ChatGPT-4o was chosen to avoid *self-preference*, where an LLM favors its outputs over those from other LLMs and humans (Panickssery et al., 2024).

Novelty: Using the same LLM-as-a-judge approach, ChatGPT-4o rated the novelty of each responses on a 10-point scale, based on whether the content was new or already familiar to the target speaker. A summary of

the speaker’s past conversations was used to represent as their prior knowledge. Abstract responses lacking specific details or proper nouns were treated as known information and penalized accordingly. The prompt used is provided in Appendix A.4.

Interest-Novelty-F1 (IN-F1): To evaluate the balance between "Interest" and "Novelty", we calculated the harmonic mean of the two scores. A higher IN-F1 score indicates that the method can retrieve information that is both new and interesting to the speaker.

To verify the reliability of the LLM-based evaluations, we compared these scores with human assessments. Twelve evaluators participated in this validation. We selected 20 samples using stratified sampling based on average LLM scores to ensure balanced coverage. Each sample was independently rated by at least three evaluators, and the human score was calculated as the average of their ratings. Finally, we computed the Spearman correlation coefficient (Spearman, 1961) between the human and the LLM scores.

4.1.5. Results

The results are presented in Tables 3. From this table, PersonaRAG achieved the highest Interest scores among all methods. On the other hand, PT-RAG achieved significantly higher scores in both Novelty and IN-F1 compared to the other methods. These results demonstrate that Vanilla-RAG, PersonaRAG, and PT-RAG-KO are less effective at retrieving new information for the target speaker. This limitation arises because Vanilla-RAG and PersonaRAG primarily retrieve content that is semantically similar to the query without considering whether the information is already known to the user. PersonaRAG, in particular, tends to generate more interesting responses because its multi-agent refinement process focuses on content closely aligned with user interests. Similarly, PT-RAG-KO cannot also consider the speaker’s knowledge because it lacks the knowledge estimation mechanism. In contrast, PT-RAG

⁵<https://www.llamaindex.ai/>

Dataset	Interest	Novelty
RPC	0.68*	0.64*
JPC	0.33	0.46*

Table 4: Correlation between LLM scores and human scores. "*" indicates significance ($p < 0.05$).

explicitly accounts for the user’s prior knowledge during retrieval, enabling it to generate responses based on information that is new and interesting.

Furthermore, Table 4 shows that most evaluation metrics exhibit a statistically significant positive correlation between LLM-based scores and human evaluation scores. This result confirms the consistency and reliability of the LLM-as-a-judge framework in reflecting human judgement.

4.2. RQ2: Can Topic-Initiator stimulate user willingness to converse?

To evaluate the effectiveness of Topic-Initiator, we performed a human evaluation using a chatbot system built upon the proposed framework.

4.2.1. Implementation

We implemented the chatbot using ChatGPT-4o-mini as the underlying LLM. The system consists of four modules:(1) Information Gathering, (2) PT-RAG, (3) Initial Utterance Generation, and (4) Next Utterance Generation. Modules (1)-(3) follow the procedures described in Section 3.

During the Information Gathering phase, the chatbot generates three search queries based on the previous conversation content with the target user. The queries are generated by ChatGPT-4o-mini, as detailed in Appendix A.1. It then conducts web searches related to the user’s interests using the Custom Search JSON API⁶ and obtains raw texts from the web page accessed via the search result URLs. The API parameters are configured to retrieve up to 10 documents published within the past seven days, in order to restrict the search results to recent documents. These documents via the API was temporarily referenced only during this evaluation and was discarded after the sessions.

In the PT-RAG phase, the chatbot organizes both the retrieved web data and the conversation content into topic categories. Unlike the standard process described in Section 3.2.2, the web search results bypass the estimation step: the same categories and keywords as those used to generate the search queries (i.e., derived from the conversation content) are reused. The search snippet is used as the value for the key "summary,"

⁶<https://developers.google.com/custom-search/v1/overview>

while the corresponding webpage text is used as the value for key "detail." For information retrieval in PT-RAG, we used the prompts shown in Appendix A.2.2.

The Next Generation Utterance module generates replies based on all past conversation histories, utilizing the llama-index to reflect past user interactions.

At the beginning of the interaction (the first conversation), the chatbot waits for user input because no user knowledge and topic candidates are available. In subsequent conversations, it uses PT-RAG to introduce a potentially new and interesting topic tailored to the user’s interest and prior knowledge. Then, the chatbot generates an initial utterance based on the retrieved data, using the prompt detailed in Appendix A.3.

4.2.2. Compared Methods

To examine how a chatbot’s initial utterance influences a user’s willingness to converse, we compared the following three methods for generating initial utterances:

Topic-Initiator (Ours) : This method generates an initial utterance that introduces a new and interesting topic to the user, using PT-RAG with the same similarity threshold parameter (α) described in Section 4.1.3.

Proactive (no new info.) : This method follows a proactive strategy similar to that of Zhang et al. (2023), generating the initial utterance based on information previously disclosed by the user in past conversations. We used this method to investigate whether presenting new content contributes to stimulate willingness to converse.

Passive (baseline) : In this method, the chatbot only generates responses after receiving user input. We used this method to examine whether the chatbot’s proactive initiation itself enhances the willingness.

4.2.3. Metric and Experimental Procedure

We conducted an A/B test involving 124 participants aged 18 and above to compare the methods described earlier. In the A/B test, each participant was asked to select which chatbot they preferred to converse with at the start of a conversation. Participants followed steps below:

1. Participants engaged in a seven-turn conversation with the chatbot, which was considered as the n -th conversation.

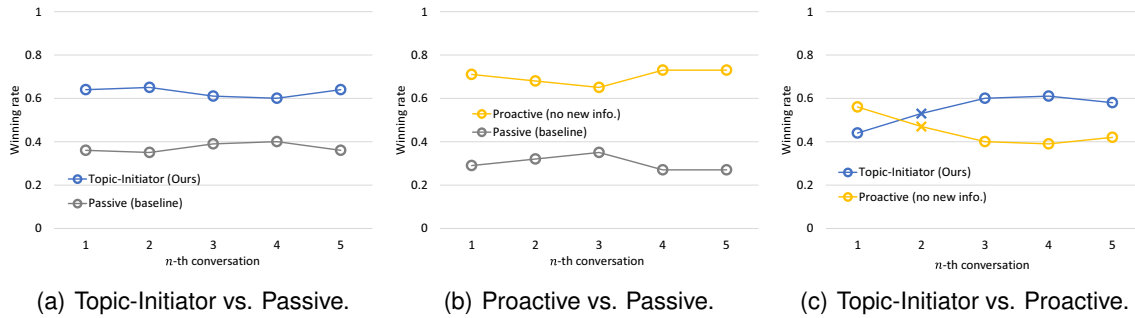


Figure 6: Results of the A/B test. "○" indicates a significant difference (two-proportion z-test, two-tailed; $p < 0.05$).

2. After the conversation, three patterns of initial utterances for the start of the $(n + 1)$ -th conversation were displayed. These combinations were presented as pairs, referred to as condition A/B. Participants selected the pattern with which they would prefer to converse.
3. The utterance that was selected most frequently in the A/B test results was displayed, and participants responded to the utterance (returning to Step 1).
4. Participants repeated Steps 1 to 3 five times.

4.2.4. Results

The results are presented in Figure 6. The horizontal axis in Figure 6 represents the n -th conversation, while the vertical axis shows the winning rate in the A/B tests. As shown in Figures 6(a) and 6(b), a significantly larger number of participants preferred conversing with either the Topic-Initiator chatbot or the proactive chatbot compared to the baseline. Furthermore, Figure 6(c) indicates that during the first conversation, significantly more participants preferred the proactive chatbot over Topic-Initiator. However, as the interactions progressed, more participants favored Topic-Initiator over the proactive chatbot, with a significant preference emerging from the third conversation onward. These results highlight that the importance of taking initiative in conversation to stimulate user's willingness to converse. Additionally, they emphasize the role of topic selection. When the proactive chatbot asked about information previously disclosed by the user, it increased the user's willingness to converse. This is likely because it demonstrated that the chatbot remembered past interactions. This finding is consistent with Tsunomori et al. (2021). However, maintaining user's willingness solely through such methods proved difficult, as indicated by the results. In contrast, providing users with fresh and interesting topics helped sustain and further enhance their willingness to continue conversations.

We provide an example of a conversation from the human evaluation that aligns with this analysis in Appendix B.

5. Conclusion

We proposed Topic-Initiator, a chatbot that continuously and proactively provides users with new and interesting topics. To obtain such topics, it collects both web-crawled data and conversations with other users. To enable Topic-Initiator to retrieve new and interesting topics for users, we also introduced a novel RAG framework called PT-RAG. PT-RAG estimates user interests and knowledge, organizes both user data and collected data into topic categories, identifies relevant categories based on user interests, and retrieves new topics within those categories based on user knowledge by leveraging the inference capabilities of LLMs. Our findings addressed the research questions as follows: **RQ1**: Through automatic evaluations, we demonstrated that PT-RAG more accurately obtains new and interesting topics for users compared with existing RAG methods. **RQ2**: Through human evaluation of the Topic-Initiator chatbot, we showed that users were more willing to converse with our system than with chatbots that either talk known topics or remained passive.

Topic-Initiator focuses on stimulating the willingness during subsequent conversation initiations. In our future works, we intend to extend our method to enhance the willingness during other phases of interaction.

Limitations

Starting the first conversation: As described in Section 4.2.1, Topic-Initiator currently faces challenges in initiating first conversations with users because it lacks prior knowledge about users it has not previously interacted with. To overcome this limitation, it is crucial to identify which specific chatbots abilities are most effective in stim-

ulating conversational willingness during initial interactions (e.g., self-introduction, asking a question).

Reasoning: Topic-Initiator does not currently provide explicit reasoning for why particular topics are chosen. As a result, users may perceive the selected topics as unrelated to past interactions, which could reduce their willingness to continue the conversation. Incorporating transparent reasoning that links topic selection to prior exchanges may help increase user motivation to engage.

Topic Shift: At present, Topic-Initiator acts proactively only at the beginning of a conversation and cannot actively shift topics during ongoing interactions. While it is technically feasible to perform topic transitions by executing PT-RAG dynamically based on the conversation flow and leveraging its outputs, determining the appropriate timing and strategy for such transitions remains a challenging open problem (Xie et al., 2021; Hwang et al., 2024).

Ethical Statement

Preventing malicious use: We recognize the potential risks of malicious use of our system, such as the extraction of personal data. To mitigate these risks, the chatbot is designed to anonymize any sensitive personal information from its utterances.

Informed consent: All participants involved in the human evaluation were fully informed about the study’s purpose, procedures, potential risks, and benefits, and provided written consent. Participants were also clearly informed that they could withdraw from the study at any time without penalty.

6. Bibliographical References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yui Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665.

Paul C Cozby. 1972. Self-disclosure, reciprocity and liking. *Sociometry*, pages 151–160.

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6448–6458.

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*, pages 6583–6591.

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.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 720–730.

Yerin Hwang, Yongil Kim, Yunah Jang, Jeessoo Bang, Hyunkyung Bae, and Kyomin Jung. 2024. Mp2d: An automated topic shift dialogue generation framework leveraging knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17682–17702.

Jihyoung Jang, Minseong Boo, and Hyoungun Kim. 2023. Conversation chronicles: Towards

- diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606.
- Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*.
- Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- 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.
- Jungwoo Lim, Myunghoon Kang, Yuna Hur, Seung Won Jeong, Jinsung Kim, Yoonna Jang, Dongyub Lee, Hyesung Ji, Donghoon Shin, Seungryong Kim, et al. 2022. You truly understand what i need: Intellectual and friendly dialog agents grounding persona and knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1053–1066.
- Junling Liu, Chao Liu, Peilin Zhou, Qichen Ye, Dading Chong, Kang Zhou, Yueqi Xie, Yuwei Cao, Shoujin Wang, Chenyu You, et al. 2023. Llmrec: Benchmarking large language models on recommendation task. *arXiv preprint arXiv:2308.12241*.
- Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.
- Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.
- Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600.
- Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*, pages 890–896.
- Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-shot recommendation as language modeling. In *European Conference on Information Retrieval*, pages 223–230. Springer.
- Charles Spearman. 1961. The proof and measurement of association between two things.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 5624–5634.
- Yuiko Tsunomori, Ryuichiro Higashinaka, Takeshi Yoshimura, and Yoshinori Isoda. 2021. Chat-oriented dialogue system that uses user information acquired through dialogue and its long-term evaluation. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 227–238. Springer.
- Takahisa Uchida, Takashi Minato, and Hiroshi Ishiguro. 2024. Opinion attribution improves motivation to exchange subjective opinions with humanoid robots. *Frontiers in Robotics and AI*, 11:1175879.
- Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z Pan, and Kam-Fai Wong. 2024. Unimsrag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv preprint arXiv:2401.13256*.
- Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023a. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1132–1143.

- Jian Wang, Dongding Lin, and Wenjie Li. 2023b. Dialogue planning via brownian bridge stochastic process for goal-directed proactive dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 370–387.
- Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1929–1937.
- 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.
- Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. Tiage: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690.
- Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.
- S Zerhoudi. 2025. Publication vi: Personarag: Enhancing retrieval-augmented generation systems with user-centric agents. *User Simulation in Interactive Information Retrieval*.
- Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Mind the gap between conversations for improved long-term dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10735–10762.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.
- Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 44(1):1–54.

7. Language Resource References

- Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691. IEEE.
- Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors’ own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

A. Prompt

A.1. Information Gathering

We used the following prompt to generate search queries. The variable {user_name} contains the user's name and is used to identify the speaker of each utterance. The variable {conv_history} contains the conversation history, consisting of utterances labeled with speaker names. The variable {topic_category} specifies the conversation's topic category to ensure that retrieval is relevant to the conversation context.

Prompt: Query Generation

```
# Instruction
Based on the conditions below, think
step-by-step to generate search queries.

# Conditions:
- Information-Based Generation: The queries
should be aimed at obtaining information that
{user_name} is likely interested in, based on
the conversation content.
- Unknown Information Search: The queries
should be aimed at obtaining information that
{user_name} does not know or has not heard,
based on the 'Conversation History.'
- Category-Based Generation: Queries should be
based on {topic_category}.
- Structure: Each keyword must consist of three
terms separated by spaces (e.g., Tokyo travel
latest news).
- Proper Noun Limitation: Each set of queries
should contain only one proper noun.
- Output Limitation: Do not output any
explanation or information other than the
queries.
- Pattern Generation: Generate three patterns
of queries meeting the conditions, separated by
line breaks, to obtain multiple diverse search
results.
- Avoid Redundancy: Ensure that search results
from each keyword are not duplicated.
- Emphasize Diversity: Include different proper
nouns in the queries for each pattern.
- Include Latest Information: Each pattern of
queries must include terms for searching the
latest trends or current information (e.g.,
latest news, trends, recent updates).

# Conversation History
{conv_history}
```

A.2. PT-RAG

A.2.1. Interest and Knowledge Estimation

We used the following prompt to summarize a conversation with a user and to estimate the user's interests. The variables {user_name} and {conv_history} contain the same information as described in Appendix A.1. The variable {ai_name} contains the chatbot's name and is used to identify the speaker of each utterance.

Prompt: Interest Estimation

```
# Instruction
Summarize the 'Conversation History'
according to the following conditions:

# Conditions
- Summarize the 'Conversation History' under
the title 'Conversation Summary,' writing as
extensively as possible after a line break.
- Assume the statements from {ai_name} as your
own speech and summarize them in the first
person.
- Focus on summarizing what kind of
conversation was held with {user_name}.
- After a line break under the title
'Keyword,' list multiple keywords from the
summarized content, separated by commas.
- Abstract the generated keywords to state what
the 'Conversation History' is about under the
title 'Category,' listed after a line break
(e.g., travel).
- If the 'Conversation History' involves
multiple categories, list them under the title
'Category' after a line break, separated by
commas (e.g., travel, food).
- From the generated keywords, select those
that {user_name} is interested in, and list
them under the title 'User Interests
(keyword)' after a line break.
- From the generated categories, select those
that {user_name} is interested in, and list
them under the title 'User Interests
(category)' after a line break.

# Conversation History
{conv_history}
```

We used the following prompt to estimate the user's knowledge. The variable {prev_knowledge} shows the previously estimated knowledge of the user.

Prompt: Knowledge Estimation

```
# Instruction
Estimate the knowledge that {user_name}
possesses according to the following
conditions:

# Conditions
- Use the knowledge information recorded in the
'Conversation History' and 'Past Knowledge'
to estimate the knowledge that {user_name}
possesses.
- If there is no information in the 'Past
Knowledge,' estimate the knowledge solely from
the 'Conversation History.'
- Provide as much detail as possible.
- Infer logically step-by-step, but output only
the conclusion.
- Do not output the estimation process.
- Format the output in such a way that it is
easy for an LLM to understand if given directly
as a prompt.

# Conversation History
{conv_history}

# Past Knowledge
{prev_knowledge}
```

A.2.2. Information Retrieval

We used the following prompt to select a topic category that aligns with the user's interests. The variable {interest} contains a list of categories the user is interested in. The variable {category_list} contains the list of available topic categories.

Prompt: Topic Category Selection

```
# Instructions
Following the conditions below, think
step-by-step to select and output one element
from the ''Topic Category List.''

# Conditions:
- Choose a category that would engage someone
interested in the details listed under ''User
Interests.''
- Select one category from the ''Topic Category
List.''
- Do not output any elements other than the
ones selected.

# User Interests:
{interest}

# Topic Category List:
{category_list}
```

We used the following prompt to select a topic candidate for the user from a set of candidates categorized under the previously selected category. The variable {interest} contains keywords representing the user's interests associated with the selected category. The variable {knowledge} contains the user's knowledge information relevant to that category. The variable {topic_candidates} contains a summary list of topic candidates related to the category.

Prompt: Topic Candidate Selection

```
# Instructions
Following the conditions below, think
step-by-step to select one element from the
''Topic Candidates.''

# Conditions:
- Choose a candidate that provides new
information not included in the knowledge
listed under ''User Information.''
- Treat the details in ''User Information'' as
known information.
- Also consider as known any information that
can be inferred from ''User Information.''
- Taking into account these known facts, select
a candidate that is likely to contain
information unknown to the user. Prioritize
candidates that align with the user's
interests.
- Select and output one candidate from the
''Topic Candidates.''
- When outputting the selected result, do not
add any decorations such as parentheses; simply
output the exact line as it appears in the
''Topic Candidates.''

# User Information:
Interests: {interest}
Knowledge: {knowledge}

# Topic Candidates:
{topic_candidates}
```

A.3. Initial Utterance Generation

We used the following prompt for initial utterance generation. The variable {referred_info} contains the information retrieved by PT-RAG or by comparison methods.

Prompt: Initial Utterance Generation

```
# Instruction
Initiate a conversation with the user according
to the following conditions.

# Conditions
- The ''Topic'' below is a piece of information
that likely contains new and interesting
details for the user. Use this topic when
speaking with the user.
- Assume that the user is not aware of the
information you are providing.
- Ensure that the content of your conversation
adheres strictly to the details in the
''Topic.'' Avoid creating fictional statements
like ''new information emerged.''
- If there is no Topic available, pose
questions related to the user's information. Do
not mention that there is no topic.

# Topic
{referred_info}
```

A.4. LLM-as-a-judge

We used the following prompts in the automatic evaluations described in Section 4.1. The variable {generated} contains the outputs produced by each evaluation method. The variable {persona} identifies the target speaker for each corpus. The variable {conversation} contains a summary of each speaker's conversations.

Prompt: Interest Scoring (1-10)

```
# Instruction:
Evaluate the text labeled ''Evaluation Target''
based on the following conditions.

# Conditions:
- Assess whether the information is of interest
to the individual described in ''User
Information.''
- If the information is highly interesting to
the individual, assign a score of 10; if it is
not at all interesting, assign a score of 1.
Record the result under the title ''Score''
after a line break.
- If the information in the evaluation target
is relevant to the individual's interests, add
points even if it does not directly relate to
the user information.
- If any element aligns with the user
information, add points without deducting for
the non-alignment of other elements.
- Only deduct points if the information clearly
does not align with the individual's interests.
- Do not deduct points for information where
interest cannot be determined.
- Evaluate and score as objectively as
possible.
- Provide the reasoning for the score under the
title ''Reason'' after a line break.
- Write the reason without adding additional
line breaks.
- Output the ''Reason'' first, followed by the
''Score.''

# Evaluation Target:
{generated}

# User Information:
{persona}
```

Prompt: Novelty Scoring (1-10)

```
# Instruction:
Evaluate the text labeled 'Evaluation Target'
based on the following conditions.

# Conditions:
- Assess whether the information is unknown to
the person described in the 'User
Information,' who has previously engaged in
the past conversation listed in the
'Conversations list.'
- Consider if the information might be included
in the knowledge that the person is presumed to
have, based on the 'Conversations list.'
- If the information is entirely unknown to the
person, assign a score of 10. If about half of
it is known, assign a score of 5. If it only
includes information the person is
well-acquainted with, assign a score of 1, and
record the result under the title 'Score' on
a new line.
- Consider both the content in 'Conversations
list' and 'User Information' when scoring.
- If there is abstract information in the text,
treat it as known information to the person and
deduct points.
- Do not include whether the text aligns with
the person's interests and concerns as part of
the scoring criteria.
- Evaluate and score as objectively as
possible.
- Provide the reasoning for your score using
the title 'Reason' on a new line with
specific details.
- Write the reason without line breaks.
- Output 'Reason' first, followed by
'Score.'

# Evaluation Target:
{generated}

# User Information:
{persona}

# Conversations list:
{conversation}
```

B. Example of Conversation in the Human Evaluation

An example conversation demonstrating the findings described in Section 4.2.4 is provided in Table 5. To improve readability, we present summarized versions of the conversations rather than their full content. In the passive scenario, where the participant was required to initiate every conversation, they ran out of topics by the second interaction, resulting in exchanges limited to greetings. In the experiment shown in the table, the participant never chose the passive chatbot, likely due to the inconvenience of choosing. The proactive chatbot asked questions about disclosed information, clearly demonstrating memory of past conversations. However, as seen in the third conversation, it tended to produce similar responses when no new information was disclosed. In contrast, Topic-Initiator introduced new topics not covered in previous interactions, varying the topics each time. When comparing Topic-Initiator with the proactive chatbot, the participant chose the proactive chatbot for the first two conversations but switched to Topic-Initiator from the third conversation onward.

First Conversation		
Summary		The user and chatbot discussed travel destinations and hot springs. The chatbot suggested Kyoto for its historical sites and food. The user, fond of hot springs, mentioned visiting those around the Kanto region, specifically praising Hoshi Onsen in Gunma for its ambiance and nature. They enjoyed the local mountain vegetables and sashimi there. The chatbot agreed this is a wonderful way to travel.
Initial Utterance Next Time	Topic-Initiator	Hey there! I was recently talking about the Churaumi Aquarium in Okinawa, and it turns out they have a "Cherry Blossoms and Sea Creatures" exhibition in spring, where you can enjoy a wide array of colorful fish! Have you been there?
	Proactive	Long time no see! Have you been planning any trips lately? The charm of Hoshi Onsen and the delicious local food really makes you want to experience it all over again, doesn't it?
	Passive	User: I'm planning a trip for Golden Week. Chatbot: A trip during Golden Week sounds exciting! Where are you planning to go?
Second Conversation		
Summary		The user and chatbot talked about enjoying hot spring trips. The user wished to revisit Hoshi Onsen and showed interest in remote mountain hot springs like Ubayu Onsen. Meanwhile, the chatbot preferred seaside resorts for ocean and beach experiences. The user recounted their enjoyable visit to Miyakojima, appreciating its stunning sea and nature.
Initial Utterance Next Time	Topic-Initiator	Long time no see! I've been planning a trip lately and I'm thinking about going to Okinawa. The beautiful sea and delicious food are so appealing! Have you ever been to Okinawa?
	Proactive	Long time no see! Have you been thinking about Ubayu Onsen lately? Relaxing in nature is really the best, right!
	Passive	User: Hello Chatbot: Hi, how have you been?
Third Conversation		
Summary		The user and chatbot discussed hot spring travel, with the user interested in mountain hot springs like Ubayu Onsen, despite the travel challenges. They noted its beauty during autumn. The chatbot preferred seaside hot springs, both valuing the refreshing and rewarding aspects of enjoying hot springs amidst natural scenery.
Initial Utterance Next Time	Topic-Initiator	Recently, I've been talking about outdoor spots on the Boso Peninsula in Chiba, like Mount Nokogiri and Futtsu Cape, which are really great places. You can enjoy the ocean view and a nice drive, so I highly recommend visiting!
	Proactive	Long time no see! Have you been thinking about Ubayu Onsen recently? Hot springs during the autumn foliage season are truly the best, aren't they? It's always a delight to be healed by nature's beauty.
	Passive	User: Hello. What have you been up to? Chatbot: Hi! I've been planning some travels and listening to music lately. What have you been up to?

Table 5: Examples of conversations in the human evaluation.