# Enhancing Coherence and Interestingness in Knowledge-Grounded Dialogue Generation

**Hiroki Onozeki** **Michimasa Inaba**
The University of Electro-Communications
{h-onozeki, m-inaba}@uec.ac.jp

## Abstract

Open-domain dialogue systems have been increasingly applied in various situations, with a growing need to improve user engagement. One effective approach is to generate responses based on interesting external knowledge using knowledge-grounded response generation models. However, relying solely on interestingness can lead to incoherent responses, potentially diminishing user engagement. This paper proposes a novel method for generating engaging responses while maintaining contextual coherence. Our approach leverages a pre-trained knowledge-grounded response generation model and modifies the knowledge selection process to enhance response coherence and interestingness without requiring additional training. First, knowledge candidates with high contextual relevance are retrieved. These candidates are then reranked based on their interestingness and used to generate the responses. Finally, the method detects dialogue breakdowns and regenerates responses as necessary to ensure coherence. We conducted experiments using the Wizard of Wikipedia dataset and two state-of-the-art response generation models. The results indicate that the proposed method improves both response coherence and interestingness.

## 1 Introduction

In recent years, significant advancements in language models have led to the widespread use of open-domain dialogue systems across various social settings. Consequently, user expectations have risen, particularly regarding interactive experiences and engagement. To meet this demand, research has increasingly focused on generating responses that enhance user engagement, including empathetic, personality-based, and knowledge-grounded responses (Algherairy and Ahmed, 2024).

Owing to this objective, knowledge-grounded response generation models have been attracting considerable attention. These models retrieve relevant knowledge from external knowledge bases and generate responses accordingly. These models can generate diverse and informative responses (Wang et al., 2023a). Typically, these models select the most suitable knowledge as the next response from the externally given knowledge. However, such models are trained to treat only one knowledge as the correct answer, even when multiple suitable knowledges are available. Consequently, they often generate mundane responses that contain well-known information, as shown in Figure 1 (A). Such responses often make the conversation boring for users and make it difficult to capture and maintain user interest, ultimately leading to a decrease in engagement.

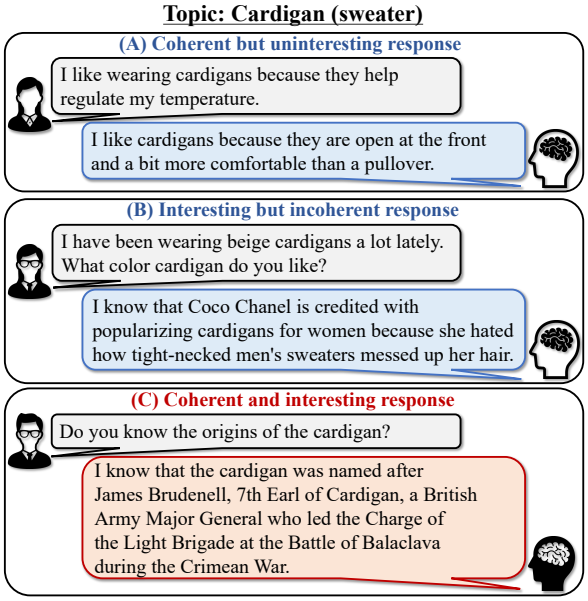


Figure 1: Dialogues between a user and a dialogue system. Generating both coherent and interesting responses is crucial for maintaining user engagement.

This study addresses the challenge of building an engaging dialogue system that captures user interest through captivating responses. Several studies



*Proceedings of the 18th International Natural Language Generation Conference*, pages 1–19
October 29–November 2, 2025. ©2025 Association for Computational Linguistics

have demonstrated that incorporating interesting facts can effectively enhance user engagement in tasks like entity search and conversation. For instance, Tsurel et al. (2017) improved users' overall experience by augmenting entity search results with relevant, interesting information. Leveraging interesting facts to capture user attention is considered an effective approach for building an engaging dialogue system. Several methods have been proposed to incorporate pre-collected interesting facts into open-domain or task-oriented dialogues (Konrád et al., 2021; Vicente et al., 2023). However, user engagement can be significantly diminished even if the responses themselves are interesting, due to incoherence, as illustrated in Figure 1 (B). Therefore, to improve user satisfaction and build an engaging dialogue system, generating responses that balance both coherence and interestingness is crucial, as depicted in Figure 1 (C).

This study proposes a method for selecting knowledge based on both its relevance to the dialogue and its interestingness, allowing us the generation of responses that facilitate coherence and interest. This approach builds on a pre-trained knowledge-grounded response generation model, using it as the base model while modifying the knowledge selection process during inference. First, to maintain response coherence, knowledges are filtered according to their relevance to the dialogue context. Subsequently, the knowledge candidates are reranked by interestingness, assigning a quantitative score using a large language model (LLM). Finally, dialogue breakdown detection is applied to the generated responses using an LLM to ensure response coherence. This method can be employed without the additional training of the base model and can be applied to various knowledge-grounded response generation models. Our contributions can be summarized as follows:

- We propose a novel knowledge-grounded response generation approach that enhances response coherence and interestingness.
- We showed the effectiveness of generating engaging responses by selecting knowledge based on the score predicted by an LLM and the effectiveness of generating coherent responses by detecting dialogue breakdowns using an LLM.
- We demonstrated that applying our proposed approach effectively enhances the consistency and interest of responses through experiments with two strong previous methods.

## 2 Related Work

### 2.1 Knowledge-Grounded Response Generation

Several studies have been conducted on knowledge-grounded response generation models, which generate responses based on external knowledge relevant to the dialogue. Kim et al. (2020) constructed a knowledge selection model with continuous latent variables modeling past knowledge selection. Zhao et al. (2020) also proposed an unsupervised approach to jointly optimize knowledge selection and response generation using a prior learning model. However, most existing methods mainly perform knowledge selection by considering only the dialogue context, leading to responses that contain general information and lack engagement (Xu et al., 2023a). To address this, Xu et al. (2023a) modeled a shift in dialogue topics and developed a model for selecting diverse knowledge while remaining consistent with the dialogue context. Chawla et al. (2024) proposed a method for addressing the trade-off between response consistency and fidelity by planning for desired response features and generating responses. These studies focused on improving diversity. However, to construct a dialogue system that is human-like and engaging, consideration should be given to diversity and interestingness. Even if a system produces diverse responses by drawing from a wide range of knowledge, a lack of interestingness may make the conversation feel dull. In contrast, while novel and unexpected information can enhance the dialogue experience, repeatedly generating similar types of interesting content may still lead to monotonous interactions and reduced user engagement. Our method generates coherent, diverse, and interesting responses by leveraging knowledge that is both relevant and interesting. It also has the advantage of employing an existing response generation model as a foundation, allowing it to retain the ability to generate high-quality responses.

### 2.2 Interestingness of Knowledge

As interest in intriguing facts has grown, research focusing on the automatic extraction of such facts has become increasingly active. Previous studies have defined trivia as any fact about an entity that is interesting due to unusualness, uniqueness, unexpectedness, or weirdness (Prakash et al., 2015).

Prakash et al. (2015) estimated the interestingness of sentences by using SVMrank, which was trained on publicly available data. Tsurel et al. (2017) utilized category information from Wikipedia articles to evaluate the rarity of specific articles within the same category according to similarity, generating trivia sentences using a template-based approach. Kwon et al. (2020) defined a surprise score as the inverse of the similarity to the summary found at the beginning of Wikipedia articles, extracting trivia sentences by leveraging Wikipedia's hierarchical structure. Korn et al. (2019) employed statistical tables from Wikipedia to generate trivia sentences using a template-based approach.

These studies propose methods for automatically retrieving trivia from Wikipedia due to its rich and diverse content. However, they depend heavily on Wikipedia's specific structures, such as categories and tables, which limits their effectiveness for extracting trivia from general text[1]. No existing methods have been proposed for automatically scoring the interestingness of knowledge in a highly accurate and general-purpose manner. Therefore, we assign interestingness scores to each knowledge using an LLM.

### 2.3 Dialogue Systems Using Interesting Knowledge

Several studies have investigated the use of interesting knowledge in dialogues to achieve engaging interactions. Konrád et al. (2021) developed a dialogue system that incorporated interesting knowledge by generating follow-up questions. However, the timing of knowledge insertion was rule-based, and the responses did not consider the broader dialogue context, leading to unnatural conversations. Vicente et al. (2023) proposed a method for incorporating interesting knowledge into a spoken dialogue system to help users perform complex tasks. In this approach, interesting knowledge gathered from web searches is incorporated into the dialogue using templates, which tends to produce monotonous and inconsistent responses. The proposed method addresses these issues by selecting knowledge based on both relevance to the dialogue and interestingness, generating responses without using templates, and ensuring coherence using dialogue breakdown detection.

[1]The knowledge in the Wizard of Wikipedia dataset used in the experiments in this paper is limited to the first paragraph of each Wikipedia page, so these methods cannot be applied.

## 3 Method

We propose a method for selecting knowledge based on both its relevance to the dialogue and its interestingness, generating responses that embody both coherence and interestingness to build an engaging dialogue system. Figure 2 presents an overview of the proposed method. We build upon pre-trained knowledge-grounded response generation models, using them as base models. In these base models, each knowledge is assigned a score that represents its suitability for the next response. The knowledge with the highest score is selected, and a response is generated based on that knowledge and the dialogue context. This score reflects both the appropriateness of the knowledge for the next response and its relevance to the dialogue context, which we refer to this score as the **contextual relevance score**. Our proposed method modifies the knowledge selection process during the inference of the base models to improve response coherence and interestingness. Instead of the base model's knowledge selection process, we introduce a three-step approach encompassing knowledge filtering, knowledge reranking, and dialogue breakdown detection. In the knowledge filtering step, we select knowledge candidates based on contextual relevance to ensure coherence. In the knowledge reranking step, we reorder the candidates by their interestingness to enhance engagement. Finally, in the dialogue breakdown detection step, we assess the generated responses for coherence and regenerate them as needed. Importantly, our method does not require additional model training and can be applied broadly to various knowledge-grounded response generation models.

### 3.1 Task Definition

Suppose we have a case of knowledge-grounded dialogues $(U_t, K_t)$, where $U_t = \{u_1, ..., u_t\}$ denotes a dialogue context up to turn $t$ on a given topic, and $K_t = \{k_{t,1}, ..., k_{t,M}\}$ represents the knowledge items relevant to the dialogue at turn $t$. Here, $u_i$ is the utterance at turn $i$, $k_{t,j}$ is the $j$-th knowledge item at turn $t$, and $M$ is the number of relevant knowledge items. The objective is to generate an engaging and interesting response $u_{t+1}$ by selecting knowledge from $K_t$ that is both contextually relevant and interesting. For simplicity, we omit the subscript turn $t$ in the following explanation.

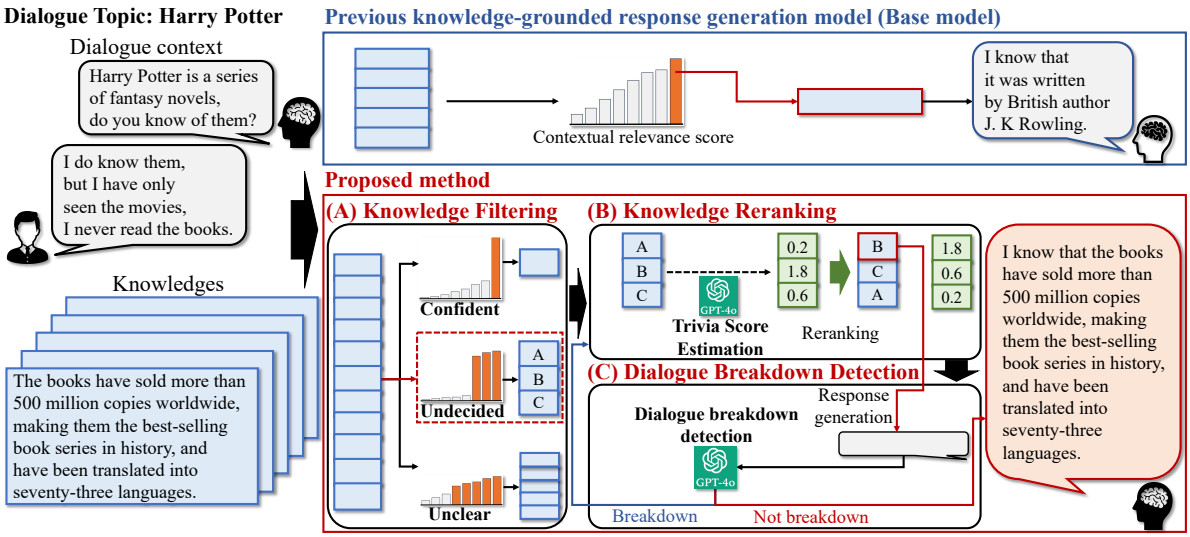


Figure 2: The overview of our approach. The previous knowledge-grounded response generation model assigns a contextual relevance score to each knowledge, selects the knowledge with the highest score, and generates a response. Conversely, our proposed method replaces the inference process of previous methods with a three-step approach encompassing knowledge filtering, knowledge reranking, and dialogue breakdown detection.

### 3.2 Knowledge Filtering

In previous knowledge-grounded response generation models, the knowledge with the highest contextual relevance score is selected to generate a response. Nonetheless, in dialogue, multiple knowledges may be appropriate for a response, depending on the context. For instance, in response to the question "When is Leonardo DiCaprio's date of birth?", using "Leonardo DiCaprio was born on November 11, 1974." would be most appropriate. Conversely, for the question "What were Leonardo DiCaprio's hobbies as a child?", both "His hobbies were collecting baseball cards and comic books." and "He enjoyed visiting museums with his father." could be contextually relevant. When multiple knowledge candidates are contextually appropriate, selecting the more interesting ones among them is expected to enhance the overall interestingness of the response while maintaining their coherence. Thus, as shown in Figure 2 (A), the proposed method initially filters knowledge candidates based on their contextual relevance scores to identify those contextually suitable for response generation.

In this study, we propose to divide into the following three cases based on the contextual relevance scores of the knowledge:

- **Confident**: When one particular knowledge has a significantly higher score than others, only that knowledge is considered suitable for the context.
- **Undecided**: When two or three knowledges have notably high scores, all of them are considered relevant to the context.
- **Unclear**: When all knowledges have roughly similar scores, all knowledges are considered relevant to the context.

We apply a softmax function to the contextual relevance scores of the knowledge to calculate the degree to which each knowledge dominates the overall knowledge. We then divide into the three cases based on the threshold for that ratio and filter the knowledge accordingly. In the Confident case, we filter one knowledge candidate; in the Undecided case, we filter two or three; and in the Unclear case, we filter multiple candidates. All filtered knowledge candidates, denoted as $K'_F$, are considered highly relevant to the dialogue context, ensuring that the selected knowledge contributes to generating coherent and contextually appropriate responses. Appendix A.1 provides the detailed procedure for knowledge filtering.

### 3.3 Knowledge Reranking

Knowledge candidates that are highly relevant to the dialogue context and appropriate for a response have been selected. Hence, we rerank these candidates based on their interestingness to identify the most engaging knowledge, as shown in Figure 2 (B). To effectively rerank knowledge based on interest, a quantitative measure is necessary. Ac-

cordingly, we define the **trivia score** based on the framework established by Prakash et al. (2015); Kwon et al. (2020) as follows:

- **Trivia score 0 (Not Trivia)**: A fact that is common, expected, ordinary, or irrelevant.
- **Trivia score 1 (Trivia)**: A fact that is unusual, unexpected, or unique, but not particularly engaging.
- **Trivia score 2 (Good Trivia)**: An interesting fact that is unusual, unexpected, or unique.

Trivia scores for each knowledge candidate are predicted using GPT-4o[2]. Given the knowledge as input, GPT-4o classifies it into one of the three categories: "Not Trivia," "Trivia," or "Good Trivia." These labels are mapped to scores of 0, 1, and 2. Inspired by the concept of self-consistency (Wang et al., 2023b), the trivia score for each knowledge candidate is predicted five times with the temperature set to $\mathrm{temperature} = 1$. Self-consistency is a method that samples a diverse set of reasoning paths and selects the most consistent answer, enhancing the performance of chain-of-thought prompting. However, unlike self-consistency, this method uses the average of the outputs as the trivia scores rather than a majority vote to provide a more fine-grained representation.

The knowledge candidates $K'_F = \{k'_1, \ldots, k'_F\}$ are reordered according to their trivia scores $T'_F = \{t'_1, \ldots, t'_F\}$, prioritizing those higher scores.

$$K''_F = k''_1, \ldots, k''_F \quad \text{where} \quad t''_1 \geq \ldots \geq t''_F \quad (1)$$

$K''_F$ represents the reranked list of knowledge candidates. If multiple candidates have the same trivia score, they are further ranked based on their contextual relevance scores as determined by the base model. Appendix A.3 provides the prompts used for trivia score prediction and examples of trivia scores.

### 3.4 Dialogue Breakdown Detection

The proposed method selects the top-ranked knowledge $k''_1$ from the reranked knowledge candidates $K''_F$ and generates a response based on that knowledge and the dialogue context. However, even when selecting knowledge relevant to the dialogue and generating a response, there may be instances where the generated response is influenced by the content of the knowledge, resulting in a lack of consistency with the context. Therefore, dialogue breakdown detection is performed to ensure the response coherence, as illustrated in Figure 2 (C).

For dialogue breakdown detection, we utilize GPT-4o. Given the dialogue topic, dialogue context, and generated response as input, it performs binary classification to determine whether the response causes a dialogue breakdown. If no breakdown is detected, then the generated response is retained. If a breakdown is detected, then a new response is generated using the next highest-rank knowledge candidate. This process continues until a coherent response is produced or all knowledge candidates are exhausted. If all knowledge-grounded responses result in breakdowns, the model generates a response using only the dialogue context without relying on external knowledge. Appendix A.4 provides the prompts used for dialogue breakdown detection.

## 4 Experiments

We conduct experiments to compare the response generation performance of previous knowledge-grounded response generation models with and without the proposed method.

### 4.1 Datasets

We use the test set of the Wizard of Wikipedia (WoW)[3] (Dinan et al., 2019) dataset, a large-scale English knowledge-grounded dialogue dataset widely employed in many studies. The WoW comprises dialogues between two participants: the Wizard, who has access to knowledges related to the conversation from Wikipedia articles, and the Apprentice, who does not. The Wizard is given 15 Wikipedia articles: one related to the overall dialogue topic and seven related to each of the two preceding dialogue turns. Each article is divided into sentences from the introductory paragraph, with each sentence treated as a distinct knowledge. From this set, the Wizard selects one knowledge and generates a response based on it. In the WoW, the test data is composed of **Seen**, which includes dialogue topics that appear in the training data, and **Unseen**, which comprises topics not covered in the training data. We use both the Seen and Unseen in our experiments.

[2] https://platform.openai.com/docs/models/gpt-4o

[3] https://github.com/facebookresearch/ParlAI

### 4.2 Models

In the proposed method, the contextual relevance score is employed to identify knowledge that is deemed appropriate for the dialogue context. To ensure high accuracy in determining the appropriateness of a response to the selected knowledge, we use knowledge-grounded response generation models that excel in this task as the base models. Thus, we adopt two state-of-the-art, high-performance knowledge-grounded response generation methods: Sequential Posterior Inference (**SPI**) and Generative Knowledge Selection (**GenKS**).

**SPI** Xu et al. (2023b) proposed a probabilistic model with dual latent variables, a discrete latent variable for knowledge selection, and a continuous latent variable for response generation. This model is characterized by its high knowledge selection accuracy and the high quality of response fidelity and diversity.

**GenKS** Sun et al. (2023) employs BART (Lewis et al., 2020) to perform generative knowledge selection, effectively capturing the interaction between dialogue and knowledge while demonstrating strong performance in both knowledge selection and response generation. This model concatenates knowledge with the dialogue context as input to BART, producing both a knowledge identifier and a response simultaneously. To address BART's input length limitation, a pre-trained DistilBERT (Sanh et al., 2020) is utilized to select the relevant document, and the knowledge from that document is subsequently fed into BART.

Our method can be applied to many pre-trained knowledge-grounded response generation models without additional training or modifying their mechanisms, and the descriptions of SPI and GenKS above are presented in a concise manner.

### 4.3 Ablation Models

To evaluate the impact of each component of the proposed method, experiments were conducted using three ablation models.

- **w/o Knowledge Filtering (- w/o KF)**: The knowledge filtering step was omitted from the proposed method, leading to the selection of the top three knowledge candidates based on their contextual relevance scores.

- **w/o Knowledge Reranking (- w/o KR)**: The knowledge reranking step was eliminated from the proposed method. The knowledge candidates were ordered solely by their contextual relevance scores.

- **w/o Dialogue Breakdown Detection (- w/o DBD)**: The dialogue breakdown detection step was removed from the proposed method. The top-ranked knowledge candidate was always selected, and the generated response was retained.

### 4.4 Evaluation Metrics

**Automatic Evaluation** We use the accuracy (ACC) score, which is the proportion of the number of correct knowledge selections to the total number of knowledge selections, to evaluate the knowledge selection performance. Additionally, we employ perplexity (PPL) of the ground-truth responses, unigram F1 (F1) (Dinan et al., 2019), BLEU-4 (Papineni et al., 2002), ROUGE-2 (Lin, 2004), and distinct score (Dist-2) (Li et al., 2016). These metrics are referred to as **reference-based** metrics.

Moreover, we adopt G-Eval (Liu et al., 2023) and MEEP (Ferron et al., 2023), which are reference-free metrics using LLM. G-Eval was used to evaluate the fluency (Flu.), coherence (Coh.), informativeness (Inf.), and interestingness (Int.) of the responses on a five-point Likert scale. MEEP evaluates response engagement on a scale of 0 to 100, with higher scores indicating higher engagement. Here, interestingness indicates the potential attraction of the information itself to a user, whereas engagement indicates how actively a user is involved in the dialogue. These metrics are referred to as **LLM-based** metrics. The LLM used was GPT-4o.

**Human Evaluation** A/B tests of the proposed method were conducted, comparing it with the base model and each of the three ablation models. For each pair of models, 100 cases were randomly selected from the different responses generated. The crowdsourcing site Amazon Mechanical Turk (AMT)[4] was used to evaluate the responses, with three annotators for each response. The same evaluation metrics were adopted as those used during the evaluation with G-Eval, which was rated on a five-point Likert scale. Appendix A.5 provides a detailed description of the crowdsourcing process.

[4] https://www.mturk.com

| Data | Models | Reference-based | | | | | | LLM-based | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC↑ | PPL↓ | F1↑ | BLEU-4↑ | ROUGE-2↑ | DIST-2↑ | Flu. | Coh. | Inf. | Int. | MEEP |
| **Seen** | SPI | **0.359** | **17.12** | **22.69** | **7.68** | **8.82** | 40.93 | 4.34 | 3.56 | 3.05 | 2.48 | 60.10 |
| | Ours (SPI) | 0.295 | 17.64 | 21.65 | 6.77 | 7.89 | 41.75 | 4.50 | 3.84 | 3.03 | 2.55 | 62.25 |
| | - w/o KF | 0.210 | 18.54 | 20.35 | 5.90 | 6.79 | 40.69 | 4.49 | 3.82 | 3.04 | 2.60 | **62.93** |
| | - w/o KR | 0.318 | 17.46 | 22.14 | 7.05 | 8.29 | **42.12** | **4.51** | **3.87** | 3.01 | 2.48 | 61.57 |
| | - w/o DBD | 0.317 | 17.64 | 21.52 | 7.09 | 8.00 | 39.12 | 4.28 | 3.45 | **3.06** | **2.61** | 60.95 |
| | GenKS | **0.346** | 13.29 | **23.99** | **4.40** | **9.82** | **38.06** | 4.50 | 3.83 | 3.07 | 2.57 | 59.59 |
| | Ours (GenKS) | 0.286 | 13.20 | 23.43 | 4.00 | 9.33 | 37.79 | 4.58 | 3.98 | 3.10 | 2.62 | 61.89 |
| | - w/o KF | 0.210 | 13.48 | 22.13 | 3.46 | 8.40 | 36.88 | 4.57 | 3.98 | **3.12** | **2.68** | **63.71** |
| | - w/o KR | 0.302 | **13.13** | 23.55 | 4.07 | 9.45 | 37.95 | **4.58** | **4.00** | 3.07 | 2.58 | 61.44 |
| | - w/o DBD | 0.312 | 13.37 | 23.47 | 4.26 | 9.52 | 37.55 | 4.48 | 3.78 | 3.10 | 2.64 | 60.93 |
| **Unseen** | SPI | **0.346** | **19.11** | **22.01** | **7.30** | **8.52** | 24.27 | 4.35 | 3.55 | 3.06 | 2.48 | 59.41 |
| | Ours (SPI) | 0.281 | 19.49 | 21.22 | 6.66 | 7.81 | 25.10 | 4.51 | 3.85 | 3.02 | 2.54 | 62.13 |
| | - w/o KF | 0.193 | 20.49 | 19.40 | 5.64 | 6.47 | 24.83 | 4.50 | 3.80 | 3.04 | 2.58 | **62.46** |
| | - w/o KR | 0.299 | 19.30 | 21.54 | 6.89 | 8.11 | **25.15** | **4.52** | **3.89** | 3.01 | 2.47 | 61.44 |
| | - w/o DBD | 0.309 | 19.49 | 21.01 | 6.79 | 7.80 | 22.29 | 4.29 | 3.44 | **3.07** | **2.61** | 60.78 |
| | GenKS | **0.369** | 13.22 | **24.33** | **4.83** | **10.06** | 21.20 | 4.49 | 3.83 | 3.05 | 2.58 | 59.39 |
| | Ours (GenKS) | 0.305 | 13.28 | 23.41 | 4.08 | 9.21 | 21.84 | **4.59** | 4.01 | 3.08 | 2.62 | 61.60 |
| | - w/o KF | 0.220 | 13.64 | 21.89 | 3.36 | 7.97 | 20.95 | 4.59 | 4.00 | **3.16** | **2.70** | **64.37** |
| | - w/o KR | 0.320 | **13.18** | 23.88 | 4.35 | 9.57 | **21.91** | 4.59 | **4.02** | 3.05 | 2.58 | 61.03 |
| | - w/o DBD | 0.340 | 13.40 | 23.44 | 4.33 | 9.35 | 20.55 | 4.46 | 3.77 | 3.08 | 2.65 | 60.72 |

Table 1: Automatic evaluation results on WoW test data. ACC denotes the accuracy of knowledge selection, PPL indicates perplexity, F1 represents token unigram F1, and DIST-2 refers to distinct-2. The best results are highlighted with **bold**, and the second-best results are highlighted with underline.

### 4.5 Implementation Details

SPI and GenKS were trained on the WoW train data using the parameters published in their respective papers (Xu et al., 2023b; Sun et al., 2023).

In SPI, a score is generated for each knowledge during inference, with the knowledge having the highest score selected to formulate the response. Therefore, when using SPI as the base model for the proposed method, we utilize this score as the context relevance score. Conversely, GenKS conducts knowledge selection by generating a knowledge identifier token using BART. Consequently, we use the output probability for the knowledge identifier token from BART as the context relevance score. Furthermore, we control the knowledge selection by regulating BART's output vocabulary.

The GPT-4o parameters were set to $\mathrm{n} = 5$, $\mathrm{temperature} = 1$ when assigning trivia scores to knowledge, and $\mathrm{n} = 1$, $\mathrm{temperature} = 0$ when detecting dialogue breakdowns in responses.

## 5 Results and Analysis

### 5.1 Automatic Evaluation

Table 1 presents the results of the automatic evaluation. We used the published model without additional training and therefore report results from a single run result. Compared to the base models, our method exhibits lower performance on reference-based metrics. This decline stems from its focus on selecting interesting knowledge from a pool of highly relevant candidates, prioritizing user engagement and dialogue coherence over strict contextual appropriateness. Consequently, knowledge selection accuracy decreases. Metrics such as BLEU and ROUGE, which assess the similarity between generated and reference responses, are negatively affected by this decline in accuracy, leading to lower scores. These metrics are insufficient for assessing the appropriateness of a response to a context, as they do not account for situations where multiple responses may be equally appropriate within the dialogue context.

Meanwhile, LLM-based metrics such as G-Eval and MEEP indicated that our method substantially outperformed the base models in overall response quality. Reportedly, these metrics correlate more closely with human judgment than reference-based metrics and exhibit a high reliability (Liu et al., 2023; Ferron et al., 2023). These findings imply that our approach effectively enhances response consistency and interestingness.

Furthermore, ablation studies confirmed that

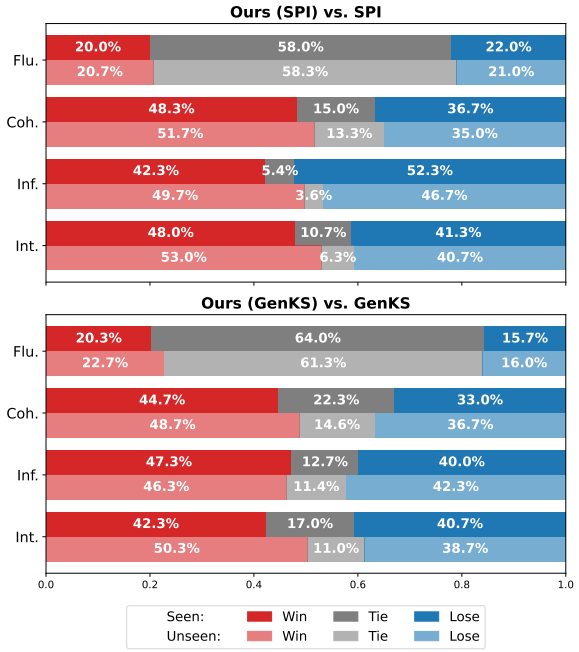


Figure 3: Human evaluation results. The upper and lower bars indicate Seen and Unseen data, respectively.

each component of our method contributed to improving response quality. In particular, knowledge reranking especially enhanced response interest, whereas dialogue breakdown detection improved response consistency. In addition, when knowledge filtering is removed from the proposed method, response quality remains largely unchanged, but knowledge selection accuracy declines significantly. Notably, our method achieves a balanced improvement in both consistency and interestingness.

### 5.2 Human Evaluation

Figure 3 presents the results of the human evaluation. The experimental results confirmed that models incorporating the proposed method maintain fluency but slightly reduce informativeness compared to the base models. This occurs because the model generates a response without using knowledge when all knowledge candidates are judged to cause a dialogue breakdown in our approach. However, notable improvements are observed in the metrics of coherence and interestingness. These results indicate that the proposed method effectively generates consistent and engaging responses by selecting coherent, interesting knowledge and detecting dialogue breakdowns. Appendix A.6 provides human evaluation results comparing our model with each ablation model.

### 5.3 Trivia Score Prediction Performance

To validate the performance of the proposed method in predicting the trivia score, GPT-4o's predictions were compared with human assessments. One hundred dialogues each were extracted from the Seen and Unseen data. Three annotators assigned a trivia score to each knowledge from the dialogues using the crowdsourcing platform AMT. Fleiss' kappa was calculated to assess the agreement among the human annotators. A Fleiss's kappa value of 0.115 indicated that perceptions of the interestingness of knowledge vary based on an individual's knowledge, interests, and experiences, making it challenging to accurately predict trivia scores. The average of the three human trivia scores for each knowledge was calculated and compared with GPT-4o's average predictions using Spearman and Kendall-Tau correlation. A Spearman correlation of 0.459 and a Kendall-Tau correlation of 0.368 suggested a moderate correlation between GPT-4o's predictions and human scores, despite the challenges in predicting trivia scores. Reportedly, the proposed method effectively predicts trivia scores, even when human annotators have differing opinions on knowledge interestingness.

### 5.4 Dialogue Breakdown Detection Performance

To validate the dialogue breakdown detection component of the proposed method, we conduct a binary classification task to determine whether or not a dialogue breakdown has occurred. We used the test data of Dialogue Breakdown Detection Challenge 5 dataset (DBDC5)[5] (Higashinaka et al., 2019). Appendix A.7 provides a detailed description of the dataset and the experimental setup. We compare the accuracy (ACC) and F1 score of GPT-3.5-turbo, GPT-4o-mini, and GPT-4o in the zero-shot setting. Table 2 exhibits the dialogue breakdown detection performance. The results confirm that GPT-4o outperforms the other models in terms of both accuracy and F1 score. Additionally, the overall performance of dialogue breakdown detection is high. These findings suggest that the proposed method is effective in detecting dialogue breakdowns and regenerating responses as necessary to ensure coherence.

[5] https://chateval.org/dbdc5

| Models | ACC | F1 |
|---|---|---|
| GPT-3.5-turbo | 0.677 | 0.802 |
| GPT-4o-mini | 0.771 | 0.849 |
| GPT-4o | **0.816** | **0.864** |

Table 2: Dialogue breakdown detection performance. The best results are highlighted with **bold**.

## 6 Conclusion

In this research, we propose a method that enhances existing knowledge-grounded response generation models by modifying the knowledge selection process during inference. The proposed method selects knowledge that is highly relevant to the dialogue context, reranks it based on its interesting level, and employs dialogue breakdown detection on the generated responses to ensure coherence and engagement. The experiments demonstrate that the implementation of the proposed method results in the generation of consistently informative and engaging responses.

## Limitations

This study has several notable limitations. First, our method requires that a high-performance model with high knowledge selection accuracy be used as the base model. Our method requires accurately estimating the relevance of each knowledge to the dialogue context during the knowledge filtering process. Improved estimations could significantly enhance the proposed method's effectiveness.

Moreover, acknowledging that different individuals perceive interest differently is crucial. As discussed in Section 5.3, humans agree less while judging topics of interest. Even if a model identifies knowledge as interesting and incorporates it into a response, the response may not be engaging for all users. This study focused on knowledge that is unique and less widely known. Future research should deeply explore individual differences in interest and develop methods adaptable to users' personalities and preferences.

## Ethical Considerations

In our experiments, we used the WoW dataset and the models SPI and GenKS. These datasets and models are publicly available, do not contain any personally identifiable information or offensive content, and do not raise any potential ethical concerns. However, we employed an LLM to predict trivia scores and detect dialogue breakdowns, which may introduce ethical considerations. An LLM could mistakenly classify offensive knowledge as interesting, potentially leading to the generation of inappropriate responses. While the proposed method can incorporate any external knowledge, careful consideration is required when selecting both the knowledge source and the LLM to ensure ethical and responsible usage.

## References


Jon Agley, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. 2021. Quality control questions on amazon's mechanical turk (mturk): A randomized trial of impact on the usaudit, phq-9, and gad-7. *Behavior Research Methods*, 54.

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.

Kushal Chawla, Hannah Rashkin, Gaurav Singh Tomar, and David Reitter. 2024. Investigating content planning for navigating trade-offs in knowledge-grounded dialogue. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2316–2335, St. Julian's, Malta. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Marc Dupuis, Karen Renaud, and Rosalind Searle. 2022. Crowdsourcing quality concerns: An examination of amazon's mechanical turk. In *Proceedings of the 23rd Annual Conference on Information Technology Education*, SIGITE '22, page 127–129, New York, NY, USA. Association for Computing Machinery.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2019. *Improving Taxonomy of Errors in Chat-Oriented Dialogue Systems*, pages 331–343.

Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey Bigham. 2018. Striving to earn more: A survey of work strategies and tool use among crowd workers. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 6:70–78.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondrej Kobza, Lenka Hýlová, and Jan Sedivý. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. *CoRR*, abs/2109.07968.

Flip Korn, Xuezhi Wang, You Wu, and Cong Yu. 2019. Automatically generating interesting facts from wikipedia tables. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 349–361, New York, NY, USA. Association for Computing Machinery.

Jingun Kwon, Hidetaka Kamigaito, Young-In Song, and Manabu Okumura. 2020. Hierarchical trivia fact extraction from Wikipedia articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4825–4834, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Catherine C. Marshall, Partha S.R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M. Shipman. 2023. Who broke amazon mechanical turk? an analysis of crowdsourcing data quality over time. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 335–345, New York, NY, USA. Association for Computing Machinery.

Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16.

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

Eyal Péer, Joachim Vosgerau, and Alessandro Acquisti. 2013. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods*, 46:1023 – 1031.

Abhay Prakash, Manoj K. Chinnakotla, Dhaval Patel, and Puneet Garg. 2015. Did you know? mining interesting trivia for entities from wikipedia. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3164–3170. AAAI Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Weiwei Sun, Pengjie Ren, and Zhaochun Ren. 2023. Generative knowledge selection for knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2077–2088. Association for Computational Linguistics.

David Tsurel, Dan Pelleg, Ido Guy, and Dafna Shahaf. 2017. Fun facts: Automatic trivia fact extraction from wikipedia. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 345–354, New York, NY, USA. Association for Computing Machinery.

Frederico Vicente, Rafael Ferreira, David Semedo, and Joao Magalhaes. 2023. The wizard of curiosities: Enriching dialogues with fun facts. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 149–155. Association for Computational Linguistics.

Ming Wang, Bo Ning, and Bin Zhao. 2023a. A review of knowledge-grounded dialogue systems. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)*, pages 819–824.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Lin Xu, Qixian Zhou, Jinlan Fu, and See-Kiong Ng. 2023a. Cet2: Modelling topic transitions for coherent and engaging knowledge-grounded conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3527–3536.

Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, and Ying Nian Wu. 2023b. Diverse and faithful knowledge-grounded dialogue generation via sequential posterior inference. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390. Association for Computational Linguistics.

# A Appendix

## A.1 Knowledge Filtering Procedure

In the knowledge filtering step, the three cases (Confident, Undecided, and Unclear) are categorized using the following procedure. Given a set of knowledge $K = \{k_1, \ldots, k_M\}$, where each knowledge $k_i$ has an associated contextual relevance score $c_i$, the knowledge are sorted in descending order by $c_i$.

$$K' = k'_1, \ldots, k'_M \quad \text{where} \quad c'_1 \geq \ldots \geq c'_M \tag{2}$$

Here, $K'$ represents the sorted list of knowledge. Next, the softmax function is applied to the knowledge scores to obtain the knowledge selection probabilities:

$$s'_i = \text{Softmax}(c'_i) = \frac{e^{c'_i}}{\sum_{j=1}^{M} e^{c'_j}} \quad \forall i = 1, \ldots, M \tag{3}$$

$A$ is defined as the sum of the top three softmax scores:

$$A = \sum_{i=1}^{3} s'_i \tag{4}$$

When $A$ is high, it indicates that the top three knowledges are highly relevant to the dialogue context. $B$ is defined as the ratio of the highest softmax score to the sum of the top three softmax scores:

$$B = \frac{s'_1}{\sum_{i=1}^{3} s'_i} \tag{5}$$

A high value of $B$ suggests that the highest score is considerably greater than the others. $C$ is defined as the ratio of the second-highest softmax score to the highest softmax score:

$$C = \frac{s'_2}{s'_1} \tag{6}$$

A low value of $C$ affirms that the difference between the top two scores is significant. Finally, the cases are classified into three categories based on the values of $A$, $B$, and $C$:

$$\begin{cases} \text{Confident} & \text{if} \quad (A \geq \alpha \wedge B \geq \beta) \vee C \leq \gamma \\ \text{Undecided} & \text{elseif} \quad A \geq \delta \\ \text{Unclear} & \text{otherwise} \end{cases} \tag{7}$$

where $\alpha$, $\beta$, $\gamma$, and $\delta$ are hyperparameters.

A fixed number of knowledge candidates $K'_F = \{k'_1, \ldots, k'_F\}$ are retrieved based on the identified cases. Here, $F$ represents the number of knowledge candidates.

- **Confident**: Only the knowledge with the highest score $k'_1$ is selected.
- **Undecided**: If there is a significant difference between the second and third knowledge scores, then the third knowledge is less suitable for the next response. $D$ is defined as $D = \frac{s'_3}{s'_2}$, the ratio of the third highest softmax score to the second highest. If $D \geq \epsilon$, then the top two knowledges are selected, $k'_1$ and $k'_2$. Otherwise, we select the top three knowledges, $k'_1$, $k'_2$, and $k'_3$ are selected.
- **Unclear**: Knowledge candidates are selected in order of their scores until the cumulative score reaches $\zeta$.

Here, $\epsilon$ and $\zeta$ are hyperparameters. These hyperparameters are determined through preliminary analysis in Appendix A.2.

## A.2 Preliminary Analysis and Hyperparameter Decision

To determine the hyperparameters for the knowledge filtering in the proposed method, a preliminary analysis was conducted focusing on the trends of knowledge contextual relevance scores for SPI and GenKS while using WoW validation data.

Figure 4 illustrates the relationship between knowledge selection accuracy, the sum of the top three knowledge softmax scores (Equation 4), and the ratio of the highest softmax score to the sum of the top three softmax scores (Equation 5) as a

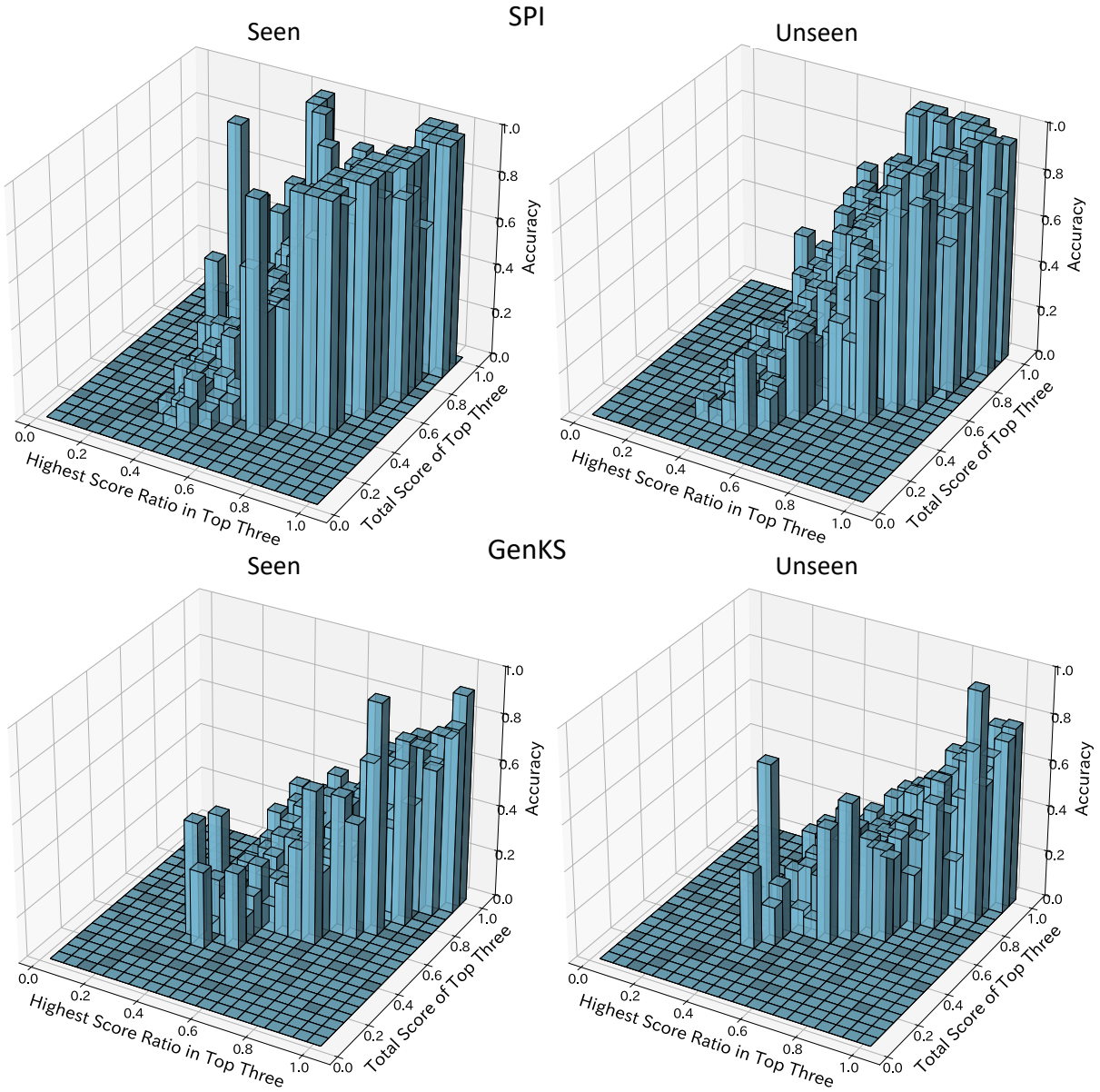


Figure 4: The relationship between knowledge selection accuracy and the sum of the top three knowledge softmax scores and the ratio of the highest softmax score to the sum of the top three softmax scores.

three-dimensional graph. From this graph, it is evident that higher values for both the sum of the top three knowledge softmax scores and the ratio of the highest softmax score to the sum of the top three softmax scores contributed positively to knowledge selection accuracy.

Figure 5 further illustrates the relationship between knowledge selection accuracy and the ratio of the second-highest softmax score to the highest softmax score (Equation 6). The graph shows that lower values for this ratio, combined with a significant difference between the maximum score knowledge and the second-highest score, were associated with an increased knowledge selection accuracy.

Based on this analysis, it was concluded that using these scores for three cases of division in knowledge filtering was effective. Consequently, hyperparameters were established based on these results.

When SPI was utilized as the base model for our approach, the hyperparameters were $\alpha = 0.6$, $\beta = 0.6$, $\gamma = 0.5$, $\delta = 0.4$, $\epsilon = 0.5$, and $\zeta = 0.6$. Similarly, GenKS was utilized, the hyperparameters were $\alpha = 0.8$, $\beta = 0.8$, $\gamma = 0.3$, $\delta = 0.65$, $\epsilon = 0.5$, and $\zeta = 0.6$.

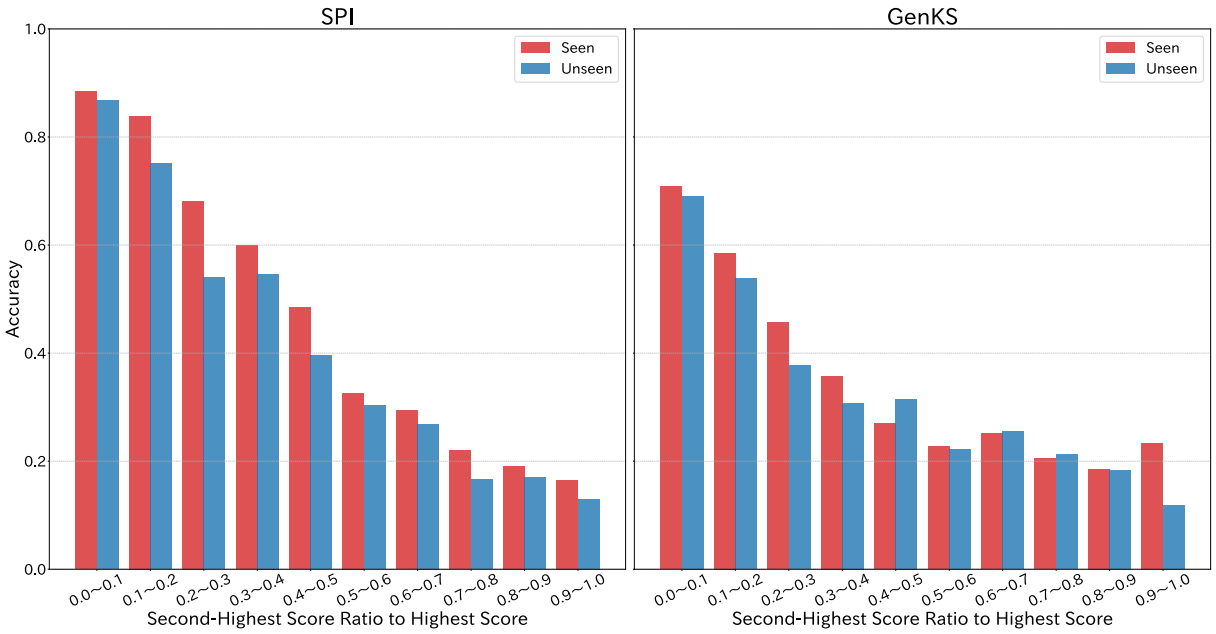


Figure 5: The relationship between knowledge selection accuracy and the ratio of the second-highest softmax score to the highest softmax score.

### A.3 Trivia Score Prediction

During the knowledge reranking process, a trivia score was assigned to each knowledge using GPT-4o. Table 3 presents an example prompt for trivia score prediction. The model was provided with a topic (i.e., the title of the Wikipedia article containing the information) and corresponding knowledge, and it produced a trivia score of "Not Trivia," "Trivia," or "Good Trivia" with reasoning. These labels were mapped to scores of 0, 1, and 2. Five estimates are made for each knowledge, and the average score represents its trivia score.

Table 4 shows an example of trivia scores assigned to knowledge about archery. The knowledge regarding the oldest signs of archery received the highest trivia score, implying that it is considered an interesting fact that is unusual, unexpected, or unique.

### A.4 Dialogue Breakdown Detection

Table 5 presents an example prompt for dialogue breakdown detection. By providing instructions, a dialogue topic, dialogue history, and the generated response as input, the model performed binary classification to determine whether the response caused a dialogue breakdown. The model outputs the reasoning behind the classification and assigns the label "Dialogue Breakdown" or "Not Dialogue Breakdown." Dialogue breakdown refers to a situation in which users are unable to continue the conversation (Martinovski and Traum, 2003).

### A.5 Crowdsourcing

We used Amazon Mechanical Turk (AMT) to collect annotations for trivia score annotation and response evaluation. To ensure high-quality annotations, we required workers to meet the following qualifications: at least 500 approved Human Intelligence Tasks (HITs) and an approval rate of 95% or higher. Additionally, annotators must be located in Australia, Canada, Ireland, New Zealand, the United Kingdom, or the United States. We grouped 20 samples as a HIT and compensated crowdworkers with $2.50 per HIT.

The issue of low worker quality on AMT, where many workers complete tasks inadequately, has been widely referenced (Marshall et al., 2023; Dupuis et al., 2022; Kaplan et al., 2018; Péer et al., 2013; Agley et al., 2021). Given the high volume of submissions, manually reviewing all work is impractical. Moreover, variations in individual judgment can complicate the evaluation of trivia scores and response quality.

To address these challenges, we developed a set of objective tests to assess whether workers are completing their tasks correctly. These tests were designed to minimize subjective variation in evaluations. A HIT was approved only if all included tests were satisfactorily completed. This approach ensured data quality, promoted higher standards

| Prompt |
|---|
| ### Task ###<br>You will be given a topic and one piece of knowledge related to that topic.<br>Your task is to classify this piece of knowledge into one of three labels based on its level of interestingness:<br>Good Trivia, Trivia, or Not Trivia.<br>The definitions for each label are as follows:<br>- Good Trivia: The knowledge is an interesting fact that is unusual, unexpected, or unique.<br>- Trivia: The knowledge is not the interesting fact that is unusual, unexpected, or unique.<br>- Not Trivia: The knowledge is a common, expected, or irrelevant fact.<br>Provide the reasoning for your classification, and then output the label at the end.<br>Start with ### Reason ### and provide a clear explanation for your classification decision,<br>detailing why the knowledge falls into the selected category.<br>End with ### Label ### followed by the chosen label (Good Trivia, Trivia, or Not Trivia).<br><br>### Topic ###<br>Archery<br><br>### Knowledge ###<br>The oldest signs of its use in Europe come from the in the north of Hamburg,<br>Germany and dates from the late Paleolithic, about 10,000–9000 BC.<br><br>### Reason ###<br>The fact that the oldest known evidence of archery usage in Europe dates back to the late Paleolithic era in the north of Hamburg, Germany, is noteworthy. It links the practice of archery to a specific prehistoric time and location, providing insight into the early development and spread of bow and arrow technology. This adds an interesting historical dimension, highlighting the ancient roots of archery, which is not commonly known by the general public.<br>The specificity of the location and era makes it stand out as unusual and unexpected, which qualifies it as a Good Trivia.<br><br>### Label ###<br>Good Trivia |

Table 3: An example prompt for trivia score prediction. The highlights show the passages generated by GPT-4o.

| Knowledge | Trivia Score |
|---|---|
| Archery is the sport, practice or skill of using a bow to propel arrows. | 0.0 |
| The word comes from the Latin “arcus”. | 0.6 |
| Historically, archery has been used for hunting and combat. | 0.0 |
| In modern times, it is mainly a competitive sport and recreational activity. | 0.0 |
| A person who participates in archery is typically called an archer or a “bowman”, and a person who is fond of or an expert at archery is sometimes called a toxophilite. | 1.6 |
| The bow and arrow seems to have been invented in the later Paleolithic or early Mesolithic periods. | 0.8 |
| The oldest signs of its use in Europe come from the in the north of Hamburg, Germany and dates from the late Paleolithic, about 10,000–9000 BC. | 2.0 |
| The arrows were made of pine and consisted of a main shaft and a long fore shaft with a flint point. | 1.0 |
| There are no definite earlier bows; previous pointed shafts are known, but may have been launched by spear-throwers rather than bows. | 1.4 |
| The oldest bows known so far come from the Holmegård swamp in Denmark. | 1.8 |
| Bows eventually replaced the spear-thrower as the predominant means for launching shafted projectiles, on every continent except Australasia, though spear-throwers persisted alongside the bow in parts of the Americas, notably Mexico and among the Inuit. | 1.8 |

Table 4: Example of trivia scores assigned to knowledge about archery.

in the dataset, and enhanced the reliability of the trivia scores and response evaluations collected from AMT workers.

#### A.5.1 Trivia Score Annotation

Figure 6 shows the annotator instruction and an example of the interface for trivia score annotation. Given a keyword, a keyword description, and a knowledge, the annotator is asked to assign a trivia score to the knowledge.

#### A.5.2 Human Evaluation

Figure 7 illustrates the annotator instruction and interface for human evaluation. Given a dialogue topic, a dialogue history, two responses generated by different models, the annotator is asked to select the better response for each metrics: Fluency, Coherence, Informativeness, and Interestingness.

**Prompt**

### Task ###
You will be given a dialogue history between Speaker A and Speaker B, along with a response that follows it.
Your task is to classify the response into one of two labels: Dialogue Breakdown or No Dialogue Breakdown.
Dialogue Breakdown refers to a situation where the flow of conversation becomes unnatural or incoherent.
Provide the reasoning for your classification, and then output the label at the end.
Start with ### Reason ### and provide a clear explanation for your classification decision,
detailing why the response falls into the selected category.
End with ### Label ### followed by the chosen label (Dialogue Breakdown or Not Dialogue Breakdown).

### Dialogue Topic ###
Blue

### Dialogue History ###
Speaker B: Blue is my favorite primary color.
Speaker A: Blue is always nice. I like royal blue.
Speaker B: I once road on The Royal Blue train from New York to D.C
Speaker A: Oh that sounds really nice. I bet there was a lot of scenery and blue skies.

### Response ###
The clear daytime sky and the deep sea appear blue because of an optical effect known as Rayleigh scattering.

### Reason ###
The response from Speaker B introduces a scientific explanation about why the sky and sea appear blue,
which is related to the topic of "blue".
However, it does not directly connect to the previous part of the conversation,
which was about the experience of riding The Royal Blue train and the scenery observed during the trip.
The response shifts the focus from a personal experience to a scientific explanation without a clear transition
or connection to the previous statements. This abrupt change in topic can disrupt the natural flow of the conversation,
making it feel disjointed and causing a dialogue breakdown.

### Label ###
Dialogue Breakdown

Table 5: An example prompt for dialogue breakdown detection. the highlights show the passages generated by GPT-4o.

### A.6 Human Evaluation for Ablation Models

Figure 8 and Figure 9 present the human evaluation results for the ablation models compared to SPI and GenKS with our approach. The experimental results indicate that knowledge reranking significantly enhanced response interest, while dialogue breakdown detection significantly improved response consistency. In addition, when knowledge filtering was removed from the proposed method, response quality remained largely unchanged, but knowledge selection accuracy declined significantly, as shown in Table 1. Notably, our method achieved a balanced improvement in both consistency and interestingness.

### A.7 Dialogue Breakdown Detection Experiments

DBDC5 was created to detect whether a system utterance will lead to a dialogue breakdown within a given dialogue context. This dataset comprises dialogues between the system and humans, with each system utterance labeled by 30 annotators using three dialogue breakdown labels: “breakdown”, “possible breakdown”, and “not a breakdown.” In the current study, dialogue breakdown detection was performed to eliminate contextually inappropriate responses. Thus, “breakdown” and “possible breakdown” were combined into a single category, treating both as breakdowns. Each utterance was assigned a label based on the majority vote between the “breakdown” and “not a breakdown” categories, effectively converting the task into a binary classification problem distinguishing between breakdown and non-breakdown instances.

### A.8 Knowledge Filtering Analysis

The proposed method performed knowledge filtering to ensure that only contextually relevant knowledge was utilized for generating appropriate responses. This filtering divided knowledge selection into three cases, namely, confident, uncertain, and unknown, and filtered knowledge based on the quantity appropriate to each case.

In this section, we analyze the validity of this

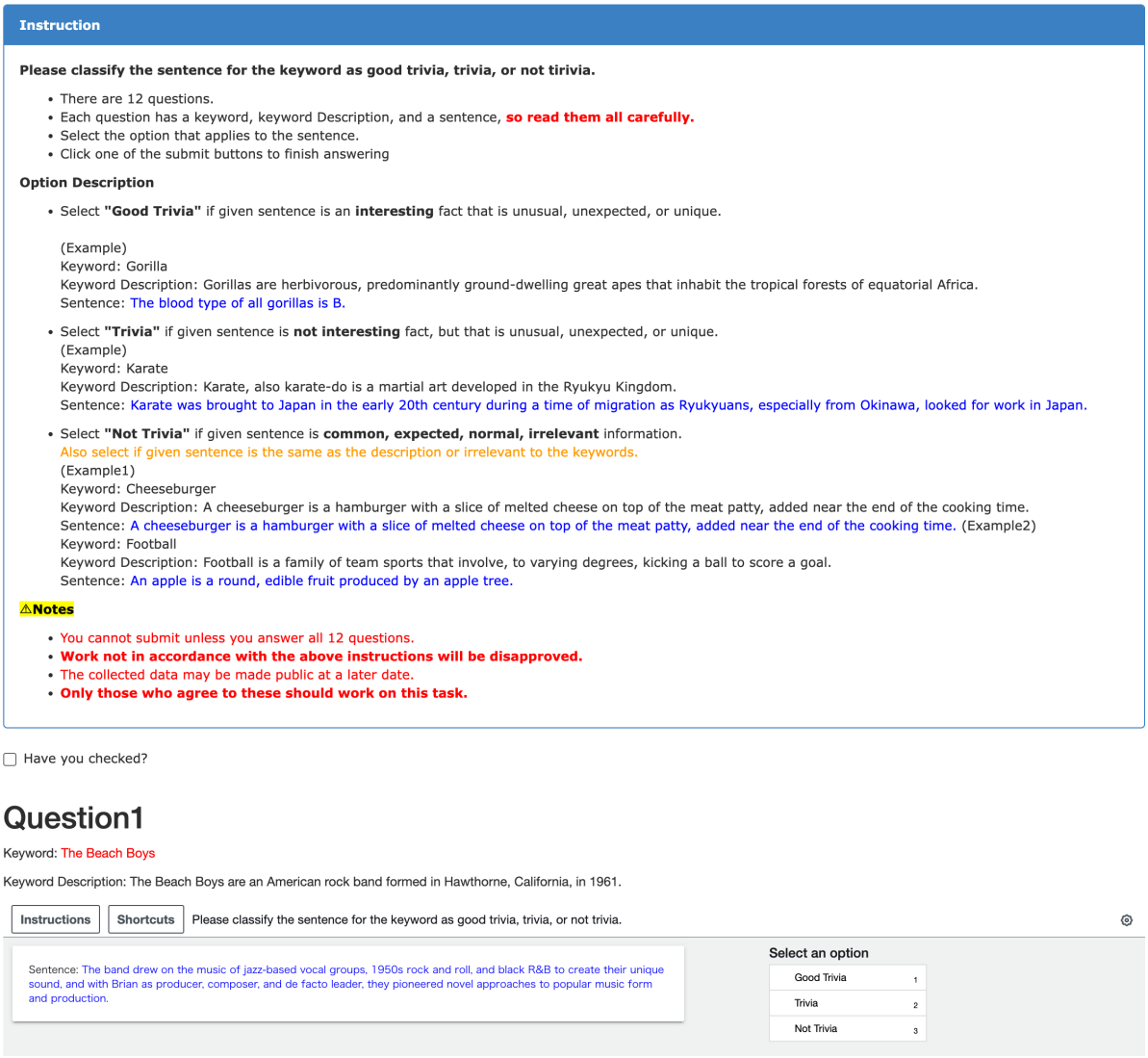


Figure 6: The annotator instruction and an example of the interface for trivia score annotation.

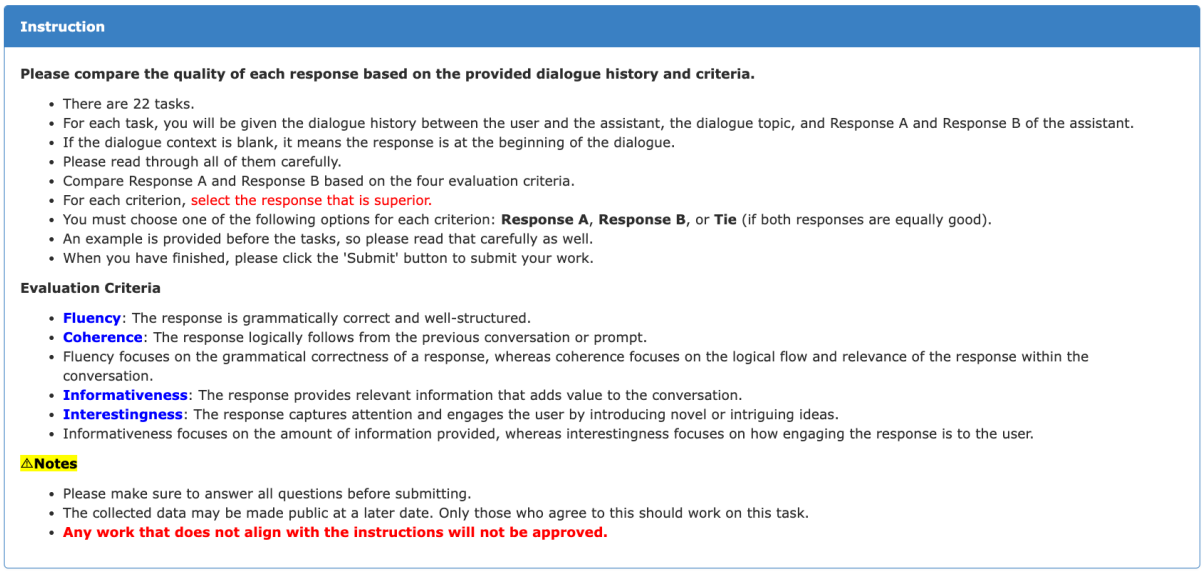


Figure 7: The annotator instruction and interface for human evaluation.

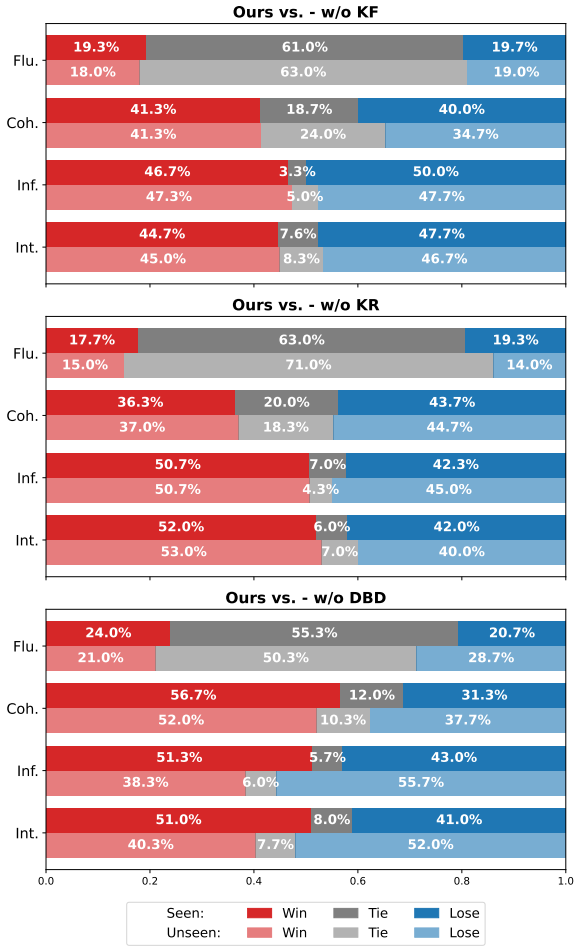


Figure 8: Human evaluation results for the ablation models compared to SPI with our approach. The upper and lower bars indicate Seen and Unseen data, respectively.

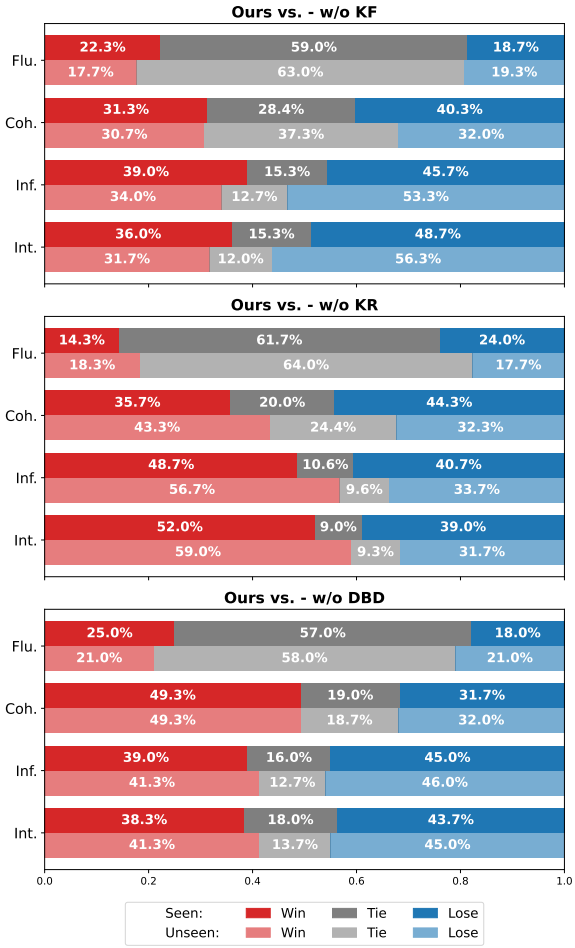


Figure 9: Human evaluation results for the ablation models compared to GenKS with our approach. The upper and lower bars indicate Seen and Unseen data, respectively.

division and its effectiveness in filtering knowledges that are contextually relevant for generating the next response. In the confident case, a single knowledge item was deemed contextually appropriate; in the undecided case, two or three knowledge items were deemed contextually appropriate; and in the unclear case, most knowledge items were deemed contextually appropriate. Therefore, if the filtering process based on these categories is effective, then we would expect to observe significant differences in knowledge selection accuracy across the three cases.

To validate this, we compared the knowledge selection accuracy for each of the three cases using both SPI and GenKS on WoW test data. Figure 10 depicts the knowledge selection accuracy for each case using both SPI and GenKS. The results revealed a clear disparity in accuracy across the three cases for both models. This indicated that the division into three cases and the corresponding filtering of selected knowledge were effective.

To validate the number of knowledge candidates selected for each case, we also compared the recall@k for each. Since GenKS performed document selection first and then selected knowledge from within those documents, it was impossible to calculate recall@k for GenKS. Therefore, we conducted the analysis using only SPI. Figure 11 shows the recall@k for each of the three cases using the SPI. From these results, we observed that the values for Confident's Recall@1, Undecided's Recall@3, and Unclear's Recall@10 were roughly the same. This suggested that the number of knowledge candidates in each case was appropriate and that the filtering process effectively isolated knowledge candidates with high relevance to the conversation.

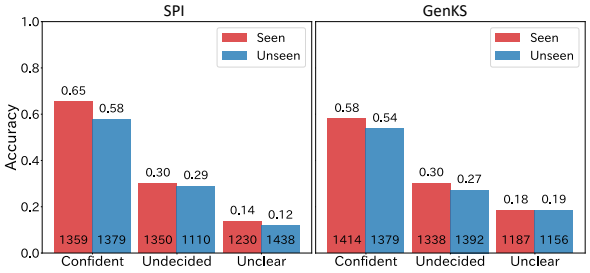


Figure 10: Knowledge selection accuracy for each case using SPI and GenKS on WoW test data. The numbers at the base of the bar graph represent the number of data points corresponding to each case, while the values at the top of the bars indicate accuracy.

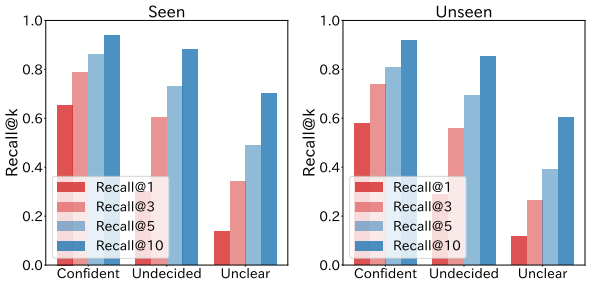


Figure 11: Recall@k for each case using SPI on WoW test data.

### A.9 Case Study

Table 6 and Table 7 display the cases from the test data of WoW, comparing the generated responses of SPI and GenKS with our approach. In the case of Table 6, SPI selected a general fact about France with a low trivia score, whereas Ours (SPI) chose a detailed explanation of the etymology of the word "blonde" with a high trivia score. Consequently, SPI generated an incoherent and less engaging response to the dialogue context, while Ours (SPI) produced a more coherent and engaging response by selecting knowledge that is both contextually relevant and interesting. In the case of Table 7, both GenKS and Ours (GenKS) selected interesting knowledge about Pittsburgh with high trivia scores. Although GenKS chose knowledge with a higher trivia score, its generated response was less engaging and coherent. On the other hand, Ours (GenKS) employed dialogue breakdown detection to filter out out-of-context responses, resulting in responses that were both coherent and engaging.

| | | |
|---|---|---|
| Dialogue History | **Topic** | **Blond** |
| | User | To be honest, I am more attracted to blonde hair than brown or dark. |
| | System | I am actually a blonde. Blonde hair comes from low levels of the pigment eumelanin? |
| | User | I am not really knowledgeable about its scientific aspect, but there is something in it that makes me feel compelled to its beauty. |
| | System | I know a lot about it, strangely. There's a Fischer-Scale that ranges from A to O, light blonde to dark blonde. Mine is an A shade. |
| | User | It is very interesting to know about the evolutionary reasoning behind blond hair. |
| | System | The term blonde was first used in 1481 from old French "blund". |
| | User | I will not be surprised if its history is tied to France as this country along with its people are considered romantic. |
| Selected Knowledge | SPI | France, officially the French Republic, is a country whose territory consists of metropolitan France in western Europe, as well as several overseas regions and territories. [Trivia Score: 0.0] |
| | Ours (SPI) | Some linguists say it comes from Medieval Latin "blundus", meaning "yellow", from Old Frankish "blund" which would relate it to Old English "blonden-feax" meaning "grey-haired", from "blondan/blandan" meaning "to mix" (Cf. [Trivia Score: 1.8] |
| Response | SPI | France is a country whose territory consists of metropolitan France in western Europe. |
| | Ours (SPI) | Some linguists say it comes from Medieval Latin "blundus", meaning "yellow", from Old Frankish "blond" which would relate it to Old English "blondeen-feax" meaning "grey-haired", from "blondsan/blandan" meaning to mix. |

Table 6: One case from test data of WoW, comparing the generated responses of SPI and Ours (SPI).

| | | |
|---|---|---|
| Dialogue History | **Topic** | **Pittsburgh** |
| | User | Do you know about Pittsburgh? |
| | System | Yes! I'm originally from Pittsburgh! Go Steelers! Pittsburgh is located in Allegheny County, Pennsylvania. |
| | User | Yes your right, it has junction of 3 rivers. |
| | System | Yes, it does. The Ohio, Monongahela and the Allegheny! The converge downtown and there's a beautiful park at that spot, Point State Park. |
| | User | Hmm, it is county seat of Allegheny County. |
| | System | Yes it is. Even though Pittsburgh looks like a big city, it's population within the city limits is only 305,704 making Pittsburgh only the 63rd largest in the US. |
| | User | Pittsburgh is kidsburgh! |
| Selected Knowledge | GenKS | For part of the 20th century, Pittsburgh was behind only New York and Chicago in corporate headquarters employment; it had the most U.S. stockholders per capita. [Trivia Score: 2.0] |
| | Ours (GenKS) | Located at the confluence of the Allegheny, Monongahela, and Ohio rivers, Pittsburgh is known as both "the Steel City" for its more than 300 steel-related businesses, and as the "City of Bridges" for its 446 bridges. [Trivia Score: 1.4] |
| Response | GenKS | I'm not sure I'd like to visit there. I know they make a lot of steel there, but I'm sure they also make a ton of other things too. |
| | Ours (GenKS) | Pittsburgh is a great city! It's known as the "Steel City" for its more than 300 steel related businesses and is also known as "the City of Bridges" for having 446 bridges. |

Table 7: One case from test data of WoW, comparing the generated responses of GenKS and Ours (GenKS).