Capture Your Interest Rapidly: Emotional Intensity-based Topic Preference Learning in Large Language Models

Anonymous ACL submission

Abstract

In recent years, incorporating speaker persona for topic selection has become mainstream in Topic Grounded Dialogue (TGD). In real-world scenarios, however, such models rigidly select topics relevant to the user profile, while disregarding the user's actual willingness to engage, leading to dialogue breakdowns. Such failures stem from two root causes: (1) existing evaluations (e.g., Hit@k) focus solely on top-k matching and ignore the overall quality of topic preference ranking; and (2) current methods lack explicit modeling of user satisfaction, failing to utilize dynamic emotional feedback during conversations. To address evaluation limitations, we formalize the topic preference task, which shifts the focus from single-topic hitting to ranking the preference order of persona-derived topics. To facilitate this, we augment the PANDA dataset with fine-grained preference annotations. To tackle the modeling limitations, we propose the **E**motional **I**ntensity-based **T**opic **P**reference (EITP) framework. Motivated by the correlation between emotional intensity and user engagement, EITP introduces a Bradley-Terry based preference ranking algorithm that leverages turn-by-turn emotional signals to dynamically re-rank persona topics. Furthermore, to ensure generalizability to scenarios lacking preference annotations, we employ a pseudo-labeled convergence predictor to evaluate historical ranking behaviors. Extensive experiments demonstrate the effectiveness of EITP, which improves the proposed preference metric $NDCG_{tp}$ by 0.292. Furthermore, it increases the standard Hit@k by 0.147, validating its underlying topic retrieval capability.

1 Introduction

Dialogue systems are evolving to meet the demands of complex real-world scenarios. Compared to earlier work, modern dialogue tasks emphasize not only response quality (e.g., coherence, diversity) (Zhang et al., 2018; Zhao et al., 2023), but also proactivity and personalization (Li et al., 2023b; Jang et al., 2022). Topic-Grounded Dialogue (TGD) focuses on generating responses centered around specific topics, improving both information relevance and user experience (Wen et al., 2022), and is widely used in personalized dialogue (Hong et al., 2024) and recommendation systems (Ren et al., 2022).

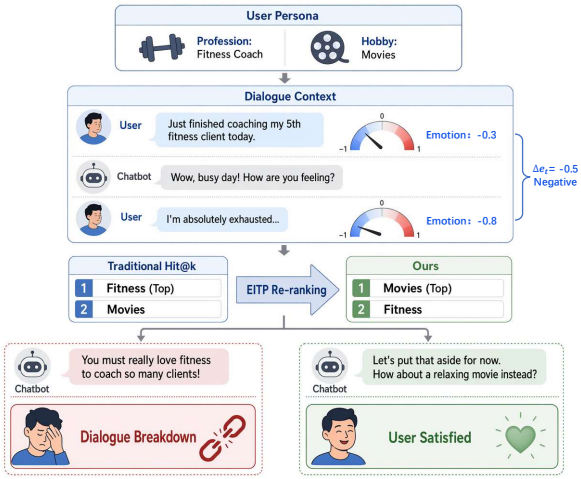


Figure 1: An example of a model engaging in dialogue with a user demonstrates that, although the model responds with content relevant to both the user's persona and the dialogue context, it ultimately fails due to its disregard for the user's willingness to continue the topic.

Recent studies on TGD have attempted to improve topic relevance by incorporating persona information (Fan et al., 2024). However, in real-world applications, these methods reveal a notable shortcoming, as illustrated in Figure 1: although the model can predict topic direction based on context and persona, it often selects topics the user is unwilling to discuss, leading to dialogue failure. We attribute this to two main limitations: 1)The hit@k metric (Zhou et al., 2020) is limited, as it only checks if relevant topics appear in top-k results, ignoring the quality of the overall preference

ranking. 2) Existing methods neglect dynamic user emotional cues, failing to capture whether the user actually wants to discuss a persona-derived topic at a given moment.

To address these issues, this paper proposes three core research questions: **RQ1:** *How can we define and evaluate users' real topic preferences, overcoming the limitations of relevance-focused metrics like Hit@k?* **RQ2:** *How can we dynamically model user topic preferences during a conversation?* **RQ3:** *How can we ensure the model's effectiveness in scenarios lacking explicit preference annotations?*

To address **RQ1**, we incorporate persona information into the dialogue setting and propose the concept of topic preference, defined as a ranking of the user's willingness to engage with persona-derived topics. Unlike traditional settings that rely on relevance-focused metrics like Hit@k, our paradigm shifts the focus to modeling the complete preference order. To support this, we augment the PANDA dataset with fine-grained preference annotations, constructing a reliable testbed for evaluating real user preferences.

To address **RQ2**, we propose the Emotional Intensity-based Topic Preference (EITP) framework. Motivated by the correlation between emotional intensity and user engagement, EITP introduces a Bradley-Terry based preference ranking algorithm. This algorithm leverages turn-by-turn emotional signals to dynamically and continuously re-rank persona topics, enabling the agent to make more informed and acceptable topic selections.

To address **RQ3**, we tackle the generalizability challenge. Since fine-grained preference annotations are expensive and often unavailable in real-world applications, we design a pseudo-labeled convergence predictor. This predictor evaluates historical ranking behaviors, allowing the system to perform effectively even on data without annotated topic preference rankings.

Extensive experiments demonstrated the effectiveness of our proposed approach. EITP significantly outperforms state-of-the-art baselines, improving the Hit@k score by 0.138 and our proposed preference metrics $NDCG_{tp}$ by 0.292, while also achieving a 0.72 improvement in user satisfaction.

In summary, our contributions are threefold: 1. We introduce the concept of topic preference, shifting the TGD paradigm from coarse-grained topic matching to preference ranking. To support this, we augment the PANDA dataset with fine-grained preference annotations. 2. We propose the EITP framework, which introduces a Bradley-Terry based preference ranking algorithm incorporating emotional intensity signals, enabling dynamic modeling of topic preferences. 3. We design a pseudo-labeled convergence predictor to ensure generalizability to data without preference labels.

2 Related Work

2.1 Topic-Grounded Dialogue

TGD task aims to guide generative models to interact around specific topics. Systems typically predict upcoming topics or generate topic-relevant responses based on dialogue history, external knowledge, or contextual information. Existing approaches can be broadly divided into three categories: Knowledge-Based Methods, Global Topic-Based Methods, and User-Based Methods. Knowledge-Based Methods incorporate external resources such as knowledge graphs to support topic selection (Zhong et al., 2021; Li et al., 2023a). Global Topic-Based Methods rely on global conversational context for topic reasoning (Zou et al., 2021; Wen et al., 2022). User-Based Methods encode user-specific traits into the model to guide topic prediction (Ren et al., 2022; Zhou et al., 2020). However, these approaches usually treat persona information as static auxiliary features and struggle to maintain deep dialogue consistency for long-term engagement. Although (Fan et al., 2024) improves consistency by incorporating relevant personas based on dialogue history and topic trajectory, it still faces challenges in maximizing user satisfaction, often selecting topics users are less willing to pursue. This difficulty stems from the fact that people may continue conversations on uninteresting topics to preserve the interlocutor's "face" (Brown and Levinson, 1987), making true preferences harder to infer. Moreover, experiments in Section 5.3 show that large language models still have limitations in modeling such nuanced topic preferences.

2.2 Persona-enhanced Dialogue Modeling

Beyond general topic modeling, recent advancements integrate persona information to enhance stylistic and semantic consistency. Early research focused on direct persona integration, such as leveraging dialogue history for character sensitivity (Li et al., 2023b) or using data augmentation to improve generation stability (Zheng et al., 2020; Hong

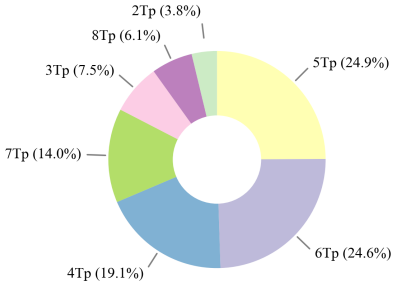


Figure 2: Distribution of the number of topics involved in annotated dialogues, where "n-Tp" indicates that the dialogue involves n topics.

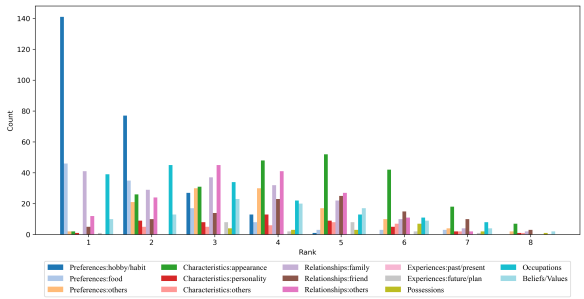


Figure 3: Rank distribution of 14 topics over the entire dataset. where the X-axis and Y-axis represent the rank and the frequency, respectively.

et al., 2024). These frameworks have also been extended to specialized domains like emotional support (Wu et al., 2025; Cheng et al., 2022; Zheng et al., 2022). Increasingly, research emphasizes the synergy between persona and topic modeling: Fan et al. (2024) treats "Persona × Topic" as a joint signal for selection; Kim et al. (2023) explores persona expansion to broaden conversational breadth; and Wu et al. (2021) aligns persona attributes with topics via minimal edits. Building on these insights, our work addresses the limitations of current TGD by focusing on personalized topic preferences to achieve more natural and consistent dialogue.

3 Topic Preference Task

3.1 The TGD task with Persona

In the TGD task, we let $C = \{u_1, u_2, \ldots, u_t\}$ as the dialogue context, which records the dialogue history up turn t. The topic path of the dialogue is denoted as $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_t\}$, indicating the topic discussed at each turn. Additionally, the task provides a persona set $P = \{p_1, p_2, \ldots, p_n\}$, where each p_i is a sentence describing the user's persona. Given the dialogue context C, the topic path $\mathcal{T}$, and the persona set P, the goal of the topic-grounded dialogue task is to predict the next topic $\mathcal{T}_{t+1}$ and generate a corresponding response R.

3.2 Topic Preference Modeling and Evaluation

To address **RQ1**, we formally define the concept of *topic preference* and propose a task framework to model it. Furthermore, we introduce a novel metric, $NDCG_{tp}$, designed to overcome the limitations of traditional relevance-focused metrics (e.g., Hit@k) in evaluating ranking performance. Finally, we augment the PANDA dataset with fine-grained preference annotations to establish a robust testbed for our proposed task.

3.2.1 Topic Preference Task

We define $\mathcal{T}$ as the set of topics derived from the persona descriptions P. The objective of the topic preference task is to rank the elements in $\mathcal{T}$ based on the user's underlying preferences, which are inferred from the dialogue context C. We evaluate the model's ability to capture these preferences from two primary perspectives:

- **Learning Efficiency:** This reflects the number of dialogue turns required for the model to successfully acquire the user's stable preference ranking.
- **Ranking Accuracy:** This measures the alignment between the predicted preference ranking and the ground-truth ranking derived from human annotations.

3.2.2 Evaluation Metric: $NDCG_{tp}$

To assess the performance of topic preference ranking in TGD, we adapt the standard Normalized Discounted Cumulative Gain (*NDCG*) (Järvelin and Kekäläinen, 2002) into a specialized version, $NDCG_{tp}$, defined as follows:

$$NDCG_{tp} = \frac{DCG}{IDCG} = \frac{\sum_{i=\text{predict}(1)}^{\text{predict}(k)} \frac{pref_i}{\log_2(i+1)} - offset}{\sum_{i=\text{ground}(1)}^{\text{ground}(k)} \frac{pref_i}{\log_2(i+1)} - offset} \quad (1)$$

Where $pref_i = |\mathcal{T}| - rank(\mathcal{T}^{\rangle}) + 1$, and it decreases as the ranking increases. Specifically, when $NDCG_{tp} > \delta$, we consider the rankings to

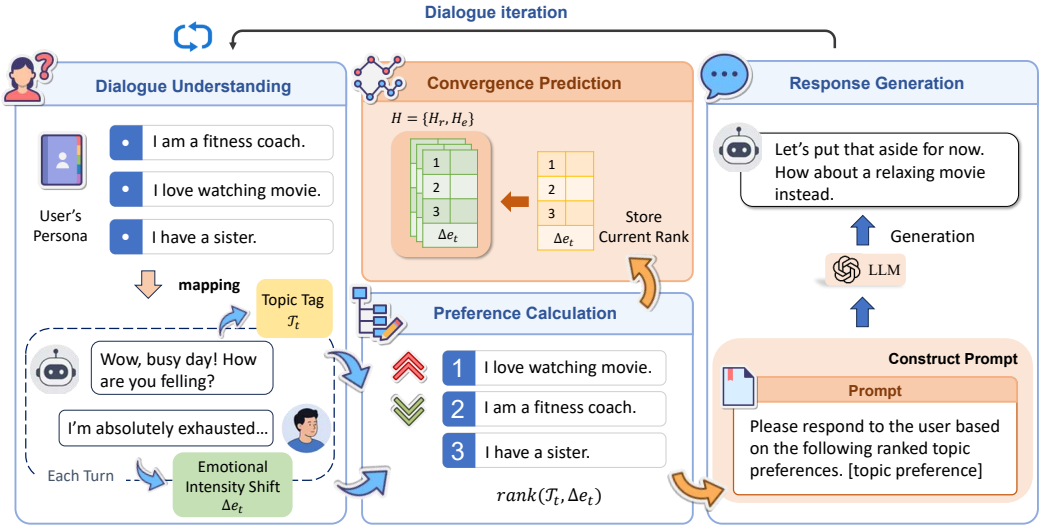


Figure 4: Overview of the designed emotional intensity-based topic preference learning (EITP) framework. The gray part represents the main method process, while the blue part is the convergence detector for semi-supervised training, which is used in real-world scenarios without labeled data.

be similar. δ [1] denotes the tolerance, granting the system a permissible margin of ranking error. $offset = \frac{\sum_{i=1}^{k} pref_i}{log_2(k+1)}$ is a scaling factor used to magnify the differences between two ranks [2].

3.2.3 Dataset Annotation

Building upon the PANDA dataset (Kim et al., 2024), we perform fine-grained annotation to identify user preference levels for topics mentioned throughout the dialogue. Annotators were tasked with scoring each topic according to our predefined guidelines (Appendix C.1). The final ground-truth preference ranking for each dialogue segment is determined by calculating the average scores across annotators. As shown in Figures 2 and 3, we provide the statistical distribution of the annotated dataset. Figure 2 illustrates the distribution of topics per dialogue, where "n-Tp" denotes dialogues involving n topics. Notably, approximately 80% of dialogues contain between 4 and 7 topics, while instances with extreme topic counts are statistically rare. Figure 3 visualizes the preference ranking distribution across 14 topic categories. Observations suggest that users generally exhibit higher preferences for topics such as *hobby*, *food*, and *family*, whereas topics related to *appearance* tend to receive lower preference scores.

[1]The selection of this parameter is described in detail in Appendix C.3.

[2]The details of this formula are provided in Appendix A.4.

4 Emotional Intensity-based Topic Preference Learning

Addressing **RQ2**, we design the EITP framework, which is illustrated in Figure 4. The framework consists of three components: Dialogue Understanding, Preference Calculation, and Response Generation. Additionally, the method learns the user's topic preferences through continuous interaction, accumulating experience during the conversation to avoid topics the user is not interested in.

4.1 Dialogue Understanding

In each turn of the dialogue (assuming the agent initiates by default), we extract the topics mentioned and infer the corresponding emotional shifts experienced by the user. These features are subsequently leveraged as inputs for user preference calculation. Drawing upon the approach outlined in (Kim et al., 2024), we identify the set of topics $\mathcal{T}$ from the input sentence x through a two-step process: (1) first extracting the persona attributes present in x, and then (2) mapping these attributes to their corresponding dialogue topics. [3]

Emotional shift $\Delta e_t \in [-1, 1]$ is defined as the difference in emotional intensity between the current turn and the previous turn, which reflects the impact of the agent's current response on the user's emotions:

$$\Delta e_t = f_{ES}(x_t) - f_{ES}(x_{t-1}) \quad (2)$$

[3]The specific implementation process is shown in Appendix A.1.

where $f_{ES}(.)$ measures the positive emotion level of an utterance using the emotion classification model developed by (Zhou et al., 2023), the specific calculation of $f_{ES}(.)$ is defined in Appendix A.2.

4.2 Preference Calculation

This part dynamically sorts the topic preference of user through an iterative algorithm. For a given set of topics $\mathcal{T}$, we design a preference algorithm inspired by the Bradley-Terry model (Bradley and Terry, 1952). First, we define a user preference value α_i for each topic $\mathcal{T}^i \in \mathcal{T}$ (The superscript i denotes the i-th topic in the set $\mathcal{T}$.), it represents the user's willingness to talk about this topic. Next, the probability $P(i > j)$, which denotes the preference of topic i over topic j, is formally defined as:

$$P(i > j) = \frac{\pi_i}{\pi_i + \pi_j} \tag{3}$$

Where $\pi_i = exp(\alpha_i)$ is used to ensure the non-negativity of the probability values. The maximum likelihood formulation of P is as follows:

$$lnL = \sum_{ij} n_{ij} lnP(i > j) \tag{4}$$

n_{ij} represents the number of times topic i is preferred over topic j. Notably, we dynamically count topics preferences during the dialogue process: In the initial turn of conversation, we set the α value of topic $\mathcal{T}_1$ to Δe_1. For each subsequent time step t, if $\Delta e_t > 0$, we consider that topic i is preferred over topic j once; otherwise, the reverse holds.

Specifically, when $\mathcal{T}_t$ is the same as topic $\mathcal{T}_{t-1}$, no update is performed. This is because the concept of habituation, as proposed in (Rankin et al., 2009), suggests that users' responses are more informative when a topic is first introduced, whereas continued discussion of the same topic tends to elicit weaker reactions—indicating that the user has habituated to the topic. Finally, by maximizing lnL [4] in equation (4), we obtain the preference values for each topic and derive the final ranking $rank\{\mathcal{T}\}$.

4.3 Response Generation

We feed the current computed preference ranking along with the dialogue history into the LLM, and LLM uses the prompt to revise its generation and respond to the user. Meanwhile, the model adjusts its dialogue strategy based on the result of the convergence prediction: when convergence has not been reached, the model is encouraged to explore previously unmentioned topics; On the contrary, once convergence is achieved, the model selects topics according to the established preference ranking. The details of the Prompt are described in Appendix B.

[4]The specific formula derivation is in Appendix A.3

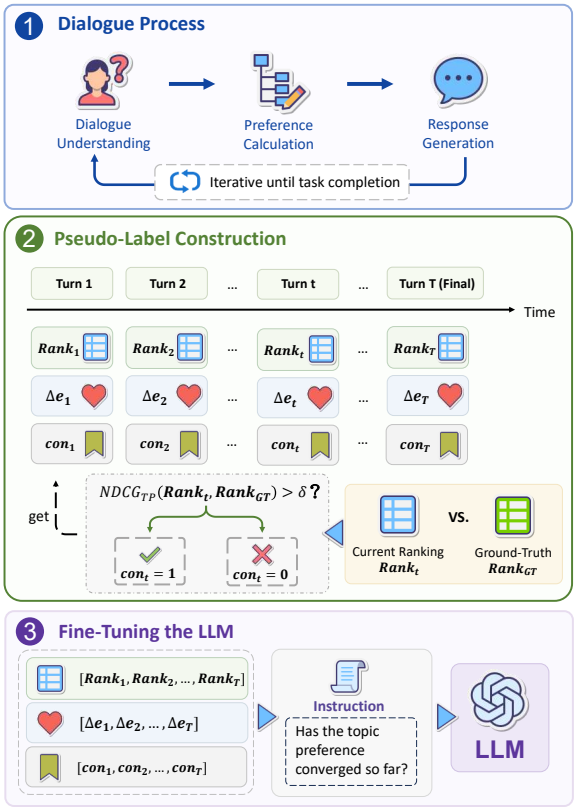


Figure 5: The training pipeline of the convergence prediction module.

4.4 Convergence prediction

To address **RQ3**, we design a convergence predictor to determine when the algorithm has sufficiently learned user preferences and could stop iterating. It uses an LLM classifier over turn-level metadata $H = \{H_r, H_e\}$, Where $H_r = \{Rank_1(\mathcal{T}), Rank_2(\mathcal{T}), ..., Rank_t(\mathcal{T})\}$ and $He = \{\Delta e_1, \Delta e_2, ..., \Delta e_t\}$, tracking preference ranking and emotion histories up to turn t, respectively. The binary classification is defined as:

$$LLM(H) = \begin{cases} 1, & \text{if } NDCG_{tp}(Rank_t, Rank_{GT}) > \delta \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Where δ is the tolerance threshold mentioned in section 3.2.2

During training, we automatically generate convergence labels for each turn based on Equation

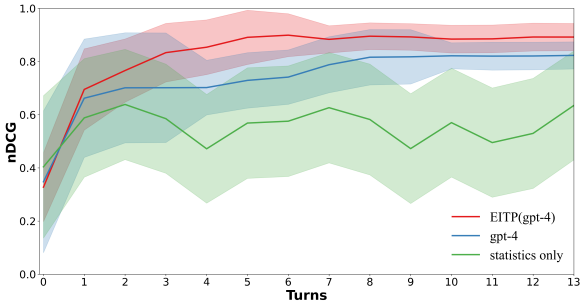


Figure 6: Ablation experiment of preference calculation.

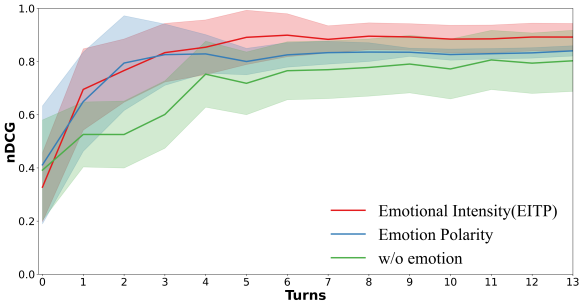


Figure 7: Ablation experiment of emotion intensity.

5 by comparing the current ranking with the final ground-truth ranking. The LLM is instruction-tuned to learn the mapping from H to these labels. During inference, where ground-truth rankings are unavailable, the trained predictor relies solely on H to predict 1 or 0, guiding the system to stop iterating once convergence is achieved. As shown in Figure 5, we collect the metadata $rank_t$ and Δe_t for each round of conversation, and then obtain the convergent pseudo-label con_t for the t-th round of conversation according to Equation 5. This forms a complete set of metadata $\{rank_t, \Delta e_t, con_t\}$, where $t \in [1, T]$ constitutes a complete dialogue. Finally, we input the metadata of multiple dialogues into the LLM for instruction fine-tuning.

5 Experiment and Evaluation

5.1 Models & Implementation Details

We evaluate various large language models on our topic preference and convergence prediction tasks. The models are categorized by usage into two types: 1) Closed-source models accessed via API: GPT-4 (OpenAI, 2023), deepseekR1 (Guo et al., 2025); 2) Open-source large models: LLaMA3 8B (Meta, 2024), Mistral-7B-Instruct (Chaplot, 2023), and Gemma2-9B (Team et al., 2024). Finally, we also included comparisons with several state-of-the-art lightweight models: TG-CRS (Zhou et al., 2020), UPCR (Ren et al., 2022), and PETD (Fan et al., 2024). We performed inference and fine-tuning of the LLMs on an RTX 5090. We fine-tuned the convergence predictor. For models with less than 10B parameters, we used LoRA. For GPT-4, we called the API [5] to perform instruction fine-tuning.

[5] https://platform.openai.com/

5.2 Evaluation Metric

Automatic Evaluation. We evaluate the tasks described in Sections 3.2. For the TGD task with Persona, we adopt the following metrics: 1) For dialogue selection accuracy, we employ $Hit@k$ ($k = 1, 3, 5$) following previous works (Wen et al., 2022; Fan et al., 2024). 2) To evaluate the word overlap of generated responses, we utilize *BLEU-n* (Papineni et al., 2002). 3) To measure the diversity and fluency of generated utterances, we adopt *Distinct-n* (Li et al., 2015) and Rouge-L (Lin, 2004). 4) Additionally, for the Topic Preference Task, we adopt the following metrics: *NDCG* and the average emotional shift, denoted as $\mathbb{E}(\Delta e)$.

Human Evaluation. Following previous works (Wen et al., 2022; Zhou et al., 2020), we adopt Relevance, Fluency and Informativeness of the generated utterances with the rating range of [0, 2]. The evaluation details are shown in the Appendix C.2.

5.3 Experimental on EITP

5.3.1 Primary Result

Table 1 presents the results of selected methods on both the TGD task and the topic preference task with persona. It can be observed that, compared to traditional models, our approach achieves at least a 0.147 improvement in Hit@1 (EITP (LLaMA3-8B) vs. PETD), and nearly reaches the upper bound on Hit@3 and Hit@5. For the topic preference task, the performance gap between EITP and other methods becomes even more significant: compared to traditional models, EITP achieves improvements of at least 0.72 and 0.292 on the $\mathbb{E}(\Delta e)$ and $NDCG_{tp}$ metrics, respectively (EITP (LLaMA3-8B) vs. PETD); compared to baseline LLMs that do not adopt the EITP framework, the gains are still at least 0.07 and 0.058 (EITP (DeepSeekR1) vs. DeepSeekR1). Moreover, further experimental results show that the NDCG scores of all baseline LLMs fall below the toler-

Method	Topic Selection					DialogueGeneration						
	Hit@1	**Hit@3**	**Hit@5**	$\mathbb{E}(\Delta e)$	$NDCG_{tp}$	**BLEU-2**	**Distinct-2**	**Rouge-L**	**Relevance**	**Fluency**	**Informativeness**	**Interest Level**
TG-CRS	0.657	0.846	0.863	-0.75	0.425	0.181	0.237	0.042	1.32	1.26	1.26	0.65
UPCR	0.685	0.855	0.865	-0.65	0.432	0.216	0.192	0.035	1.37	1.35	1.25	0.88
PETD	0.727	0.904	0.951	-0.57	0.562	0.252	0.254	0.041	1.54	1.49	1.36	1.02
LLaMA3-8B	0.854	0.956	0.979	-0.03	0.762	0.387	0.372	0.065	1.58	1.69	1.43	1.33
Mistral-7B	0.869	0.962	0.981	0.07	0.743	0.385	0.354	0.063	1.60	1.72	1.39	1.37
Gemma2-9B	0.831	0.959	0.982	0.06	0.744	0.381	0.346	0.057	1.57	1.75	1.45	1.35
DeepseekR1	0.852	0.964	0.984	0.11	0.815	0.404	0.398	0.078	1.65	1.75	1.54	1.45
GPT-4	0.875	0.970	0.985	0.12	0.823	0.423	0.408	0.089	1.68	1.77	1.56	1.48
EITP(LLaMA3-8B)	0.874	0.975	0.981	0.15	0.854	0.444	0.411	0.084	1.66	1.70	1.65	1.59
EITP(Mistral-7B)	0.872	0.980	0.987	0.18	0.861	0.442	0.417	0.095	1.62	1.71	1.62	1.65
EITP(Gemma2-9B)	0.865	0.978	0.988	0.16	0.858	0.446	0.409	0.087	1.68	1.74	1.58	1.62
EITP(DeepseekR1)	0.954	0.982	0.995	0.18	0.873	**0.456**	0.435	**0.114**	1.74	1.78	**1.71**	**1.74**
EITP(GPT-4)	**0.965**	**0.985**	**0.997**	**0.21**	**0.892**	0.451	**0.442**	0.109	**1.77**	**1.81**	1.69	**1.74**

Table 1: The performance of EITP and all baselines ($p < 0.05$).The results of the best performance in each column are in boldface. Table is divided into three sections: the upper part lists traditional models, the middle part shows prompt-based baseline LLMs, and the lower part presents the results of our proposed approach.

ance threshold δ. This indicates that prompt-based approaches alone are insufficient for handling the topic preference task effectively, highlighting the necessity of EITP in optimizing baseline models. Additionally, we conducted experiments on the TG-ReDial dataset, achieving at least a 0.063 improvement in the Hit@1 metric. Detailed results are provided in Appendix D.1, further demonstrating the strong generalization ability of our method even without annotated preference orderings.

5.3.2 Topic Preference Learning Efficiency

In this section, we provide a detailed discussion of EITP's performance on the topic preference task. As mentioned in Section 3.2.1, we evaluate the models from two perspectives: learning efficiency and accuracy.

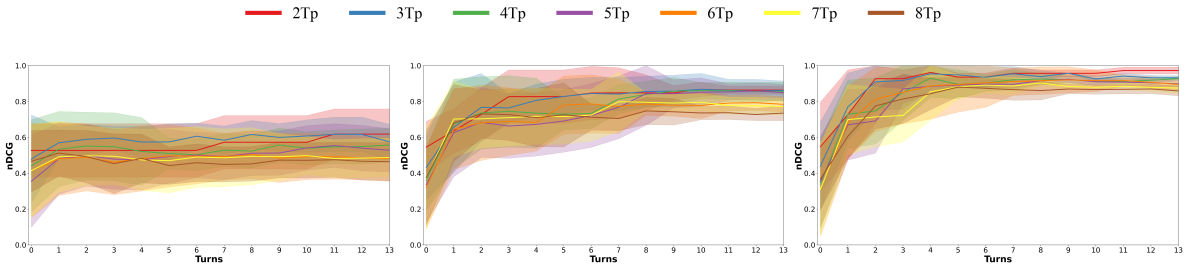
Figure 8 illustrates the performance curves of three approaches in dialogue flow: traditional models, a baseline model (GPT-4), and our EITP method. Since the number of topics covered in each dialogue varies in the dataset, we use different colors to indicate different topic counts. This helps analyze model performance more precisely. It can be observed that compared to traditional models, large language models can dynamically update their understanding of users' pereformance as the dialogue progresses: as the number of dialogue turns increases, the model's NDCG score consistently improves, and the variance gradually decreases. When comparing EITP with the baseline model, we see that under low topic conditions (e.g., 2Tp, 3Tp), the baseline model also converges quickly. However, when the number of topics increases, our method converges significantly earlier—approximately two turns ahead—and with lower variance, indicating higher efficiency in real-time preference learning. Moreover, in terms of ranking accuracy, EITP also outperforms the baseline model.

5.3.3 Effectiveness of Preference Computation

To verify the advantages of preference computation, we design a comparison among three methods: preference ranking using EITP, prompting the baseline LLM to perform preference ranking, and ranking preferences solely based on topic occurrence frequency (a statistical method). As shown in Figure 6, the EITP approach converges faster with smaller variance. In contrast, the statistical method struggles to differentiate preferences due to limited data in the early stages of the dialogue and the similar frequency of topic mentions, resulting in significantly higher variance.

5.3.4 Effectiveness of Emotion Intensity

For the emotion intensity experiment, we compare two settings: one that only distinguishes emotion polarity, and another without considering emotion at all. The "emotion polarity only" setting refers to a three-class classification of emotions into positive, negative, and neutral (i.e., $f_{ES}(.) = \{1, -1, 0\}$). The comparison results are shown in Figure 7. It can be observed that the method without emotional modeling performs poorly, even slightly worse than the baseline LLM. The method that only distinguishes emotion polarity shows improved performance, but still falls short compared to our approach. This is because the former can effectively separate user preferences into liked and disliked categories, but it is insensitive to finer-grained ranking within each category.



(a) PETD (b) GPT-4 (c) EITP(GPT-4)

Figure 8: The performance of three methods on topic preference ranking — the traditional method (PETD), the baseline large language model (GPT-4), and EITP. The x-axis represents turns as the dialogue progresses, and the y-axis indicates the ranking accuracy. Lines in different colors represent different numbers of topics. The experimental results of the other models are presented in Appendix D.2.

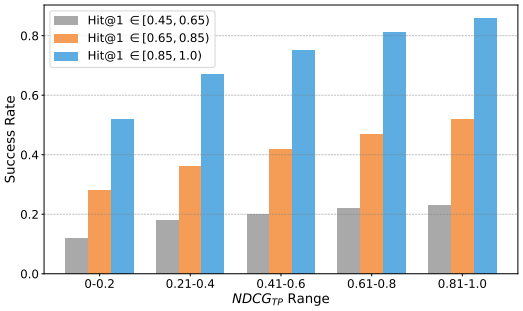


Figure 9: The relationship between NDCG, Hit@1 and the success rate of the conversation. We stipulated that if $\Delta e > 0$ at the end of the dialogue, we considered it a successful dialogue.

5.4 The effectiveness of $NDCG_{tp}$

To verify the effectiveness of the $NDCG_{tp}$ metric, we collected the generated data of the models LLaMa, Mistral, Gemma, DeepSeek, and GPT-4 from Section 4.3, and calculated the success rate of conversations for different $Hit@1$ values within different NDCG score ranges. The results are shown in Figure 9. It can be seen that a higher $Hit@k$ does not necessarily mean a high success rate of the dialogue, and as the $NDCG_{tp}$ increases, the success rate of the dialogue also gradually increases.

5.5 Performance on Convergence Prediction

This section presents the performance of the convergence prediction module. We split the pseudo-labeled data into training and test sets in a 4:1 ratio and perform instruction tuning on the models. We visualize the convergence prediction outcomes: Figure 10 illustrates the results using GPT-4 as an example, showing good performance across different numbers of topics. The accuracy rates of all models are presented in Appendix D.3.

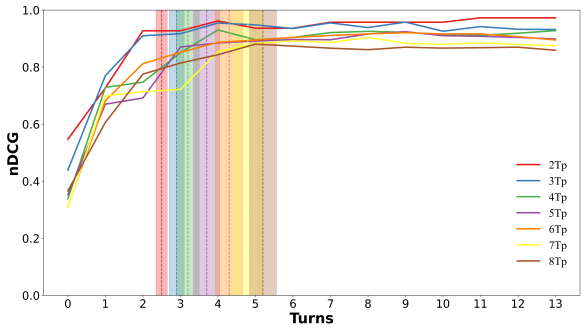


Figure 10: The performance of the convergence prediction model across different numbers of topics. Curves and vertical lines in the same color indicate the ranking accuracy and the predicted convergence point under the same topic number. Solid lines represent the mean values, and the shaded areas denote the standard deviations.

6 Conclusion

In this work, we incorporate persona information into the Topic-Grounded Dialogue task and introduce the concept of topic preference, shifting the focus from relevance matching to preference ranking. To support this paradigm, we augment the PANDA dataset with fine-grained preference annotations. Building on this, we propose the EITP framework, which leverages emotional intensity signals and a Bradley-Terry based ranking algorithm to dynamically model user preferences. Finally, we design a pseudo-labeled convergence predictor to enhance generalizability on data without explicit preference annotations. Experimental results demonstrate that EITP achieves substantial improvements in topic selection quality and user satisfaction over existing methods.

Limitations

Although the proposed EITP framework enables large language models to rapidly and dynamically track users' topic preferences, while maintaining dialogue coherence and improving user satisfaction in TGD tasks, several limitations remain. First, the framework can only perform ranking and selection within the given information space, namely the predefined persona information and existing dialogue history. It is unable to proactively introduce novel topics to expand the candidate topic set. This limitation is also related to the inherent constraint of current TGD tasks, where agent proactivity remains underexplored. Second, topic prediction is intrinsically subjective. Our approach is designed to align with the preference patterns observed in most common conversational scenarios; however, it may not generalize well to special cases, such as conversations with frequent abrupt topic switching or scenarios requiring strictly fixed topics.

7 Ethics Statement

The data used in this study are based on annotations conducted on the publicly available PANDA dataset (Kim et al., 2024). For both the annotation process and the subsequent human evaluation experiments, we recruited experienced crowd workers and provided comprehensive instructions before the tasks to ensure the quality and consistency of annotations and evaluations. In addition, none of the annotation or evaluation tasks involved the collection of annotators' personal privacy information. Finally, all participants received fair and appropriate compensation for their annotation work.

References

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2022. Pal: Persona-augmented emotional support conversation generation. *arXiv preprint arXiv:2212.09235*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Shixuan Fan, Wei Wei, Xiaofei Wen, Xianling Mao, Jixiong Chen, and Dangyang Chen. 2024. Personalized topic selection model for topic-grounded dialogue. *arXiv preprint arXiv:2406.01988*.

James Paul Gee. 2014. *An introduction to discourse analysis: Theory and method*. routledge.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Mengze Hong, Chen Jason Zhang, Chaotao Chen, Rongzhong Lian, and Di Jiang. 2024. Dialogue language model with large-scale persona data engineering. *arXiv preprint arXiv:2412.09034*.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10803–10812.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Donghyun Kim, Youbin Ahn, Chanhee Lee, Wongyu Kim, Kyong-Ho Lee, Donghoon Shin, and Yeonsoo Lee. 2023. Concept-based persona expansion for improving diversity of persona-grounded dialogue. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3471–3481.

Jinsung Kim, Seonmin Koo, and Heui-Seok Lim. 2024. Panda: Persona attributes navigation for detecting and alleviating overuse problem in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12005–12026.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Siheng Li, Wangjie Jiang, Pengda Si, Cheng Yang, Qiu Yao, Jinchao Zhang, Jie Zhou, and Yujiu Yang. 2023a. Enhancing dialogue generation with conversational concept flows. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1514–1525.

Yunpeng Li, Yue Hu, Yajing Sun, Luxi Xing, Ping Guo, Yuqiang Xie, and Wei Peng. 2023b. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13157–13165.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2024. Personalitychat: Conversation distillation for personalized dialog modeling with facts and traits. *arXiv preprint arXiv:2401.07363*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Mark EJ Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

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.

Catharine H Rankin, Thomas Abrams, Robert J Barry, Seema Bhatnagar, David F Clayton, John Colombo, Gianluca Coppola, Mark A Geyer, David L Glanzman, Stephen Marsland, and 1 others. 2009. Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of learning and memory*, 92(2):135–138.

Zhaochun Ren, Zhi Tian, Dongdong Li, Pengjie Ren, Liu Yang, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Variational reasoning about user preferences for conversational recommendation. In *proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 165–175.

Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5):447–470.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Xiaofei Wen, Wei Wei, and Xian-Ling Mao. 2022. Sequential topic selection model with latent variable for topic-grounded dialogue. *arXiv preprint arXiv:2210.08801*.

Chen Henry Wu, Yinhe Zheng, Xiaoxi Mao, and Minlie Huang. 2021. Transferable persona-grounded dialogues via grounded minimal edits. *arXiv preprint arXiv:2109.07713*.

Shenghan Wu, Yang Deng, Yimo Zhu, Wynne Hsu, and Mong Li Lee. 2025. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. *arXiv preprint arXiv:2502.11451*.

Ernst Zermelo. 1929. Die berechnung der turnierergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Kang Zhao, Wei Liu, Jian Luan, Minglei Gao, Li Qian, Hanlin Teng, and Bin Wang. 2023. Unimc: A unified framework for long-term memory conversation via relevance representation learning. *arXiv preprint arXiv:2306.10543*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2022. Augesc: Dialogue augmentation with large language models for emotional support conversation. *arXiv preprint arXiv:2202.13047*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14568–14576.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994*.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. *arXiv preprint arXiv:2109.04084*.

A Details of Preliminary Studies

A.1 Topic Extraction

We adopt the method proposed by (Kim et al., 2024) to obtain the topic set $\mathcal{T}$.

Dialogue Labeling Given a text x, we identify its contained persona attributes $P^x \subseteq P$ using the function $TAG_p(.)$:

$$TAG_P(x) = P^x = \{p_i \in P \mid p_i \in x\} \quad (6)$$

where p_i denotes the i-th attribute in the persona set.

Persona–Topic Mapping We map the identified attributes P^x to a topic set $\mathcal{T} \subseteq T$ (T represents the complete set of all known topics) using the function $TAG_T(.)$. First, each attribute $P_i^x \in P^x$is mapped to its associated topics $\mathcal{T}_i$:

$$TAG_T(p_i^x) = \mathcal{T}_i = \{\mathcal{T}_j \in T \mid \mathcal{T}_j \in p_i^x\} \quad (7)$$

where $\mathcal{T}_j$ denotes the j-th topic. The final topic set for x is the union of these mappings:

$$\mathcal{T} = TAG_T(P^x) = \bigcup_{p_i^x \in P^x} \mathcal{T}_i \quad (8)$$

This framework allows us to get the topic sets $\mathcal{T}$ from the dialogue.

A.2 Emotion Intensity

$f_{ES}(.)$ is specifically implemented based on the work of (Hartmann, 2022), It performs classification prediction over Ekman's 6 basic emotions (Ekman, 1992) plus an additional neutral class, and outputs an extra score $b \in (0, 1]$ to represent the intensity of the emotion, e.g., *{'label': 'anger', 'score': 0.745}*. In our work, we do not require fine-grained emotion categories to compute Δe, so we define **anger**, **disgust**, **fear**, and **sadness** as negative emotions, with their $f_{ES}(.)$ assigned as $\breve{}b$. The fes of **neutral** is defined as 0. The specific mapping is as follows:

$$f_{ES}(x) = \begin{cases} b, & \text{if label} \in \{\text{anger, disgust,} \\ & \quad \text{fear, sadness}\} \\ 0, & \text{if label} = \text{neutral} \\ -b, & \text{if label} \in \{\text{joy, surprise}\} \end{cases} \quad (9)$$

A.3 Bradley-Terry Model Optimization

We follow the efficient method proposed in (Newman, 2023) to solve the Bradley-Terry formulation. Although Equation (3) in section 4.2 involves two variables, π_i and π_j, due to symmetry, we only need to focus on one of them (Zermelo, 1929), namely π_i. Therefore, we take the derivative with respect to π_i, set the resulting expression to zero, and perform the following transformations:

$$\begin{aligned} &\frac{\Delta}{\Delta \pi_i} \sum_{ij} \ln P(i > j) \\ &= \frac{\Delta}{\Delta \pi_i} \sum_{ij} n_{ij} \ln \frac{\pi_i}{\pi_i + \pi_j} \\ &= \frac{\Delta}{\Delta \pi_i} \left(\sum_j n_{ij} \ln \frac{\pi_i}{\pi_i + \pi_j} + \sum_j n_{ji} \ln \frac{\pi_j}{\pi_j + \pi_i} \right) \\ &= \frac{\Delta}{\Delta \pi_i} \sum_j \left\{ n_{ij} \left[\ln \pi_i - \ln (\pi_i + \pi_j) \right] + n_{ji} \ln \frac{\pi_j}{\pi_j + \pi_i} \right\} \\ &= \frac{1}{\pi_i} \sum_j n_{ij} - \sum_j \frac{n_{ij} + n_{ji}}{\pi_i + \pi_j} \\ &= 0 \end{aligned} \quad (10)$$

Thus, the iterative update formula for π_i can be derived as follows:

$$\pi_i = \left(\sum_j \frac{n_{ij} \pi_j}{\pi_i + \pi_j} \right) / \left(\sum_j \frac{n_{ji}}{\pi_i + \pi_j} \right) \quad (11)$$

We initialize the value of π_i corresponding to the topics mentioned in the first round as $exp\{\Delta e\}$, and set all other values to 1. By performing iterative updates, we compute the value of each π. In our study, the number of iterations is set to $k = 10$. Finally, we sort the topics according to their respective π values to obtain the ranking $rank(\mathcal{T})$.

A.4 Scaling Factor

The NDCG evaluation metric has certain limitations, as it may assign high scores to low-quality results (Sakai and Kando, 2008). This can have a significant impact on our task, making the evaluation insensitive to poor outcomes. To address this, we apply a cut-off to the original formula in order to amplify the effective scoring range. Our choice of cut-off is based on the rescaling of the formula under the worst-case scenario, as follows:

$$\sum_{i=1}^{k} \frac{pref_i}{\log_2(i+1)} > \frac{\sum_{i=1}^{k} pref_i}{\log_2(k+1)} \quad (12)$$

The order of the $pref$ set from i to k does not affect the outcome of the rescaling.

A.5 Why PANDA

A widely used dataset for the dialougue system task is PersonaChat (Zhang et al., 2018), which contains approximately 10,000 multi-turn dialogues that reflect specific persona characteristics. However, as research has progressed, PersonaChat has gradually fallen short of meeting various research needs. As a result, many studies have expanded upon it (Lotfi et al., 2024; Wu et al., 2025), (Kim et al., 2024) explicitly indicated whether participants' utterances in the conversation are personalized or not, and annotated which persona attributes are grounded in the utterances. Therefore, in this work, we conducted the annotation based on PANDA to complete our task.

B Prompt Templates

Table 2 shows the prompt template for EITP Response Generation. LLMs aim to enhance user engagement in dialogue by generating appropriate responses based on computed rankings and dialogue history. Table 4 presents the prompt template used for fine-tuning the convergence prediction model. Table 3 illustrates the template used in the prompt-only setting for evaluating the effectiveness of preference computation.

Prompt
You are a conversational AI that selects the most appropriate topic for the next turn based on the user's topic preferences and the dialogue history. Your task is to: 1. Predict the top 5 most likely topics the user would be interested in discussing next. 2. Choose the most appropriate topic among them based on the current context. 3. Generate a natural and coherent reply to the user that reflects that topic. Output format: [Predicted Next Topics]: [topic_1, topic_2, topic_3, topic_4, topic_5] [Reply]: <your generated sentence here> Here is the information you are given: - # User's topic preference ranking (from most to least interested): {{topic ranking}}

Table 2: Prompt template example for EITP Response Generation.

C Annotation Guidelines

C.1 Topic Preference Annotation

When annotating preference rankings in dialogues, we followed the guidelines of the theoretical framework proposed in (Gee, 2014), specifically drawing on the concepts outlined in Chapter 2, 'Building Tasks'. The annotation was informed by three key dimensions: Significance, Activities, and Identities, which were used to guide annotators in identifying and ranking preferences. The detailed annotation guidelines are illustrated in Table 5. We recruited five annotators to label 300 dialogue segments, following a Fleiss Kappa score of 0.62, indicating substantial inter-annotator agreement.

C.2 Human Evaluation Metrics

Relevance is used to evaluate the relevance of selected topics and generated sentences to historical conversations and the user's personality. *Fluency* is used to measure the fluency of generated utterances. *Informativeness* is used to evaluate whether the generated utterance revolves around the topics and user personas. In addition, we introduced a new metric called *Interest Level*, which measures the overall nature of the speaker's engagement throughout the dialogue. The detailed scoring criteria are shown in Table 6. We recruited three individuals with a high proficiency in English to perform the evaluation, and the results are statistically significant ($p < 0.05$).

C.3 Tolerance Parameter

Regarding the selection of the tolerance parameter δ, we invited five crowdworkers to participate in the evaluation. We divided the range of $NDCG_{tp}$ scores from 1.0 to 0.7 into six intervals, each spanning 0.05. Then, we randomly sampled 10 examples from the dataset. For each example, we generated six different ranked lists corresponding to the six $NDCG_{tp}$ intervals.

For instance, given the ground-truth ranking:

["Preferences:hobby/habit", "Preferences:food", "Possessions", "Occupations"]

a predicted ranking such as:

["Preferences:hobby/habit", "Possessions", "Occupations", "Preferences:food"]

has an $NDCG_{tp}$ score of 0.91 and therefore falls

Prompt

Instruction:

You are given a multi-turn dialogue between a user and an assistant.

Your task is to infer the user's topic preferences, and output a ranking of topics from most to least preferred, based only on the topics that appeared in the topic set.

Do not introduce any topics that were not mentioned in the dialogue.

Topic Set: {{topic set}}

- -

Input:

Dialogue History: {{dialogue history}}

- -

Output:

Topic Rank: {{rank}}

Table 3: Prompt template for ranking based solely on instruction tuning.

into the $[0.95, 0.90]$ interval.

After preparing the dataset in this manner, we asked the crowdworkers to examine the dialogue, the ground-truth ranking, and the predicted rankings across intervals. They were instructed to determine the lowest $NDCG_{tp}$ score beyond which the predicted rankings are no longer acceptable—that is, cannot be considered valid preference orderings for the dialogue. The endpoint of that interval was then designated as δ.

Finally, we averaged the judgments from all participants and obtained $\delta = 0.85$.

D More Experiment

D.1 Topic Selection Task on TG-ReDial Dataset

Due to limited resources, we did not perform preference ranking annotation on the TG-ReDia dataset. Therefore, we did not conduct $NDCG_{tp}$ evaluations

Prompt

Instruction:

You are given the ranking history of topics over multiple time steps. For each step, you also receive:

- Emotion intensity changes for the topics, and

- Whether the ranking was converged at that specific step (True or False).

Your task is to determine whether the ranking has finally converged at the last time step.

Please answer only with "Yes" or "No".

- -

Input:

{ Turn 1: {Rank1:[], e1, con1},
Turn 2: {Rank2:[], e2, con2},
...
Turn n: {Rankn:[], en, conn}, }

- -

Output:

Yes or No

Table 4: Instruction-tuning prompt for the convergence prediction module.

on this dataset. Although we did not perform annotation on the TG-ReDial dataset, the training of the convergence prediction module enables the model to learn topic ranking even in the absence of ground-truth rankings. The experimental results are shown in Table 7.

D.2 Topic Preference Experiment

The performance curve of the topic preference task for the remaining model is shown in Figure 11.

D.3 Convergence Prediction Experiment

The accuracy of each model on the convergence prediction task after fine-tuning are shown in Table 8. For Deepseek, we fine-tuned the open-source

Topic Preference Annotation

You will be presented with a segment of daily conversation along with a set of topics mentioned within the dialogue. Please assign an interest score (0–9) for the respondent for each topic, where 0 indicates no interest and 9 indicates the highest level of interest. You may refer to the following criteria to guide your judgment:

Significance: Does the speaker attribute particular importance to the topic?

Activities: Does the speaker elaborate on or engage actively with the topic?

Identities: Does the speaker express personal identity, values, or roles through this topic?

[Topic Set]:
1. Preferences:hobby/habit
2. Relationships:friend
(...)

[Dialogue]:
(...)
A: That is cool, I work at a grocery store when I am not in school
B: I love going to grocery stores. does your grocery store ever play rock music ?
(...)

[Score annotation]:
1. Preferences:hobby/habit : 7
2. Relationships:friend : 4
(...)

Table 5: Guidelines for topic preference annotation.

Human Evaluation

Relevance
2: Relevant to the current dialogue and fully consistent with the given persona information.
1: Relevant to the current dialogue and/or partially consistent with the given persona information.
0: Completely inconsistent with both the persona information and the dialogue context.

Fluency
2: Fluent and easy to read
1: Grammatically formed
0: Not a complete sentence or hard to read

Informativeness
2: Have clear and specific meaning
1: Contain a few informative words
0: Meaningless sentence

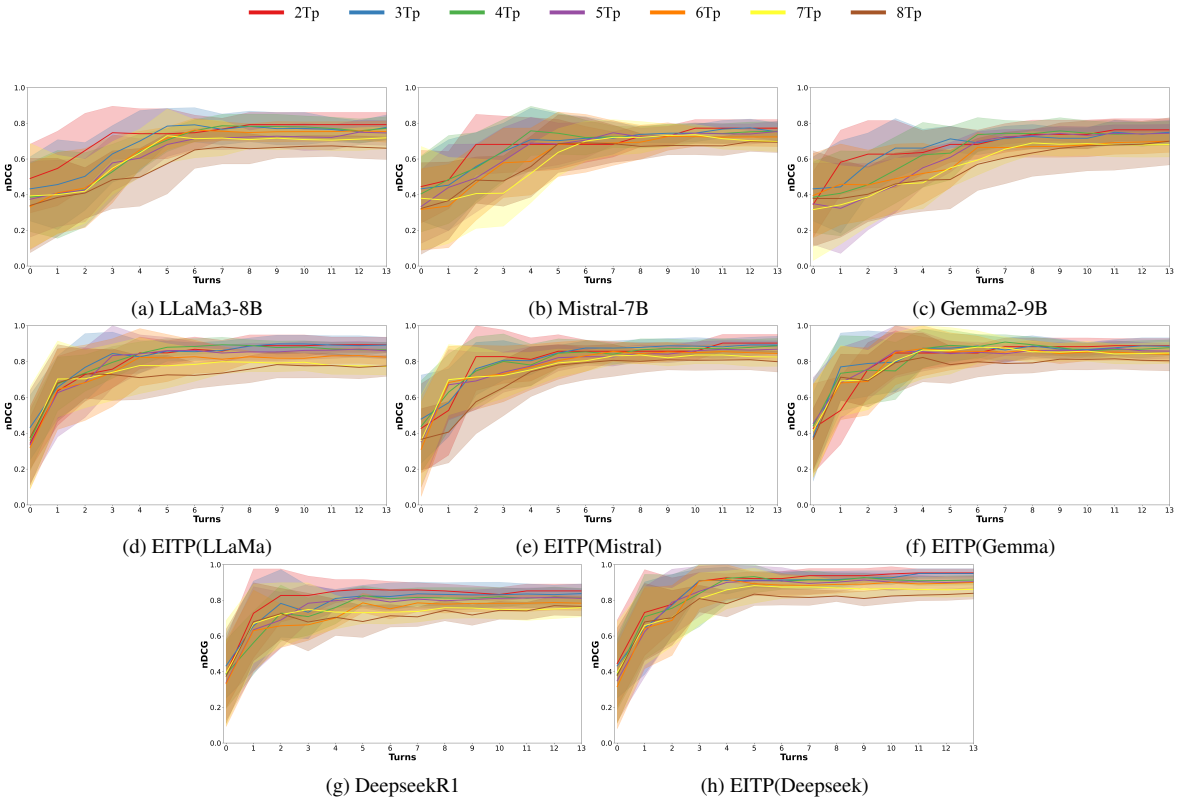
Interest Level
2: The partner interlocutor is highly enthusiastic and shows a strong desire to continue the conversation.
1: The partner interlocutor is moderately interested and continues the conversation with a calm attitude.
0: The partner interlocutor shows impatience with the current topic or explicitly expresses unwillingness to continue the conversation.

Table 6: Criteria of human evaluation.

Method	Topic Selection				DialogueGeneration						
	Hit@1	Hit@3	Hit@5	$\mathbb{E}(\Delta e)$	BLEU-2	Distinct-2	Rouge-L	Relevance	Fluency	Informativeness	Interest Level
TG-CRS	0.613	0.816	0.830	-0.82	0.173	0.094	0.038	1.29	1.24	1.22	0.74
UPCR	0.808	0.883	0.907	-0.74	0.200	0.132	0.030	1.32	1.31	1.24	0.92
PETD	0.837	0.899	0.920	-0.60	0.224	0.176	0.037	1.50	1.44	1.37	1.09
LLaMA3-8B	0.842	0.942	0.949	0.05	0.353	0.293	0.061	1.57	1.68	1.45	1.43
Mistral-7B	0.845	0.956	0.959	0.02	0.332	0.302	0.062	1.62	1.71	1.42	1.40
Gemma2-9B	0.844	0.957	0.963	0.01	0.399	0.285	0.059	1.59	1.67	1.42	1.41
DeepseekR1	0.878	0.960	0.962	0.12	0.371	0.314	0.074	1.67	1.72	1.56	1.52
GPT-4	0.880	0.965	0.970	0.15	0.398	0.339	0.082	1.65	1.75	1.53	1.54
EITP(LLaMA3)	0.905	0.972	0.975	0.17	0.416	0.356	0.087	1.64	1.70	1.61	1.62
EITP(Mistral)	0.900	0.975	0.977	0.15	0.408	0.351	0.080	1.67	1.73	1.63	1.66
EITP(Gemma2)	0.904	0.971	0.974	0.16	0.413	0.350	0.085	1.65	1.71	1.60	1.68
EITP(Deepseek)	**0.921**	0.979	0.981	0.18	0.427	0.395	0.102	**1.76**	1.76	**1.77**	1.78
EITP(GPT-4)	0.919	**0.980**	**0.983**	**0.20**	**0.435**	**0.409**	**0.110**	1.74	**1.78**	1.76	**1.80**

Table 7: The performance of EITP and all baselines on TG-ReDial ($p < 0.05$). The results of the best performance in each column are in boldface. Table is divided into three sections: the upper part lists traditional models, the middle part shows prompt-based baseline LLMs, and the lower part presents the results of our proposed approach.

7B version.



(a) LLaMa3-8B (b) Mistral-7B (c) Gemma2-9B

(d) EITP(LLaMa) (e) EITP(Mistral) (f) EITP(Gemma)

(g) DeepseekR1 (h) EITP(Deepseek)

Figure 11: The topic preference experiment result curve of the remaining models. The x-axis represents turns as the dialogue progresses, and the y-axis indicates the ranking accuracy. Lines in different colors represent different numbers of topics.

	Large Language Model				
	LLaMA3-8B	Mistral-7B	Gemma2-9B	Deepseek-7B	GPT-4
Accuracy	86.7	84.6	85.2	88.6	90.4

Table 8: The results of the five large models after convergence and prediction fine-tuning.