# Proactive Guidance of Multi-Turn Conversation in Industrial Search

**Xiaoyu Li, Xiao Li, Li Gao*, Yiding Liu, Xiaoyang Wang**
**Shuaiqiang Wang, Junfeng Wang, Dawei Yin**
Baidu Inc., Beijing, China
demo.xyli@icloud.com, {emilyxiao0512, gaoli.sinh, liuyidingyd}@gmail.com,
{wangxiaoyang06, wangshuaiqiang, wangjunfeng}@baidu.com, yindawei@acm.org

## Abstract

The evolution of Large Language Models (LLMs) has significantly advanced multi-turn conversation systems, emphasizing the need for proactive guidance to enhance users' interactions. However, these systems face challenges in dynamically adapting to shifts in users' goals and maintaining low latency for real-time interactions. In the Baidu Search AI assistant, an industrial-scale multi-turn search system, we propose a novel two-phase framework to provide proactive guidance. The first phase, Goal-adaptive Supervised Fine-Tuning (G-SFT), employs a goal adaptation agent that dynamically adapts to user goal shifts and provides goal-relevant contextual information. G-SFT also incorporates scalable knowledge transfer to distill insights from LLMs into a lightweight model for real-time interaction. The second phase, Click-oriented Reinforcement Learning (C-RL), adopts a generate-rank paradigm, systematically constructs preference pairs from user click signals, and proactively improves click-through rates through more engaging guidance. This dual-phase architecture achieves complementary objectives: G-SFT ensures accurate goal tracking, while C-RL optimizes interaction quality through click signal-driven reinforcement learning. Extensive experiments demonstrate that our framework achieves 86.10% accuracy in offline evaluation (+23.95% over baseline) and 25.28% CTR in online deployment (149.06% relative improvement), while reducing inference latency by 69.55% through scalable knowledge distillation.

## 1 Introduction

The remarkable progress in Large Language Models (LLMs) (Achiam et al., 2023; Yang et al., 2024; Grattafiori et al., 2024; Guo et al., 2025) has propelled conversational AI systems into a new era, where they are increasingly capable of understanding users' queries and providing precise answers. This advancement has spurred the development of multi-turn conversation systems (Aliannejadi et al., 2020; Vadhavana et al., 2024; Yi et al., 2024; Zhang et al., 2025).

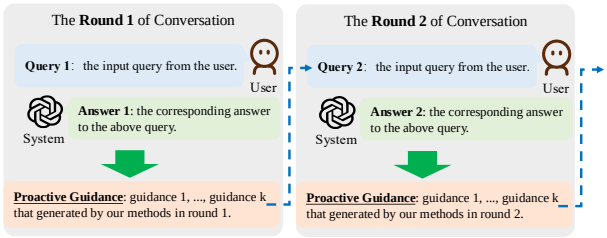


Figure 1: Illustration of the Proactive Guidance task in the multi-turn conversation system scenario. In each turn, given the user's query and the corresponding answer, our method generates $k$ proactive guidance to guide the user to click for the next turn of the conversation.

Contemporary systems are increasingly valued for their ability to anticipate and guide conversational turns (Zhang et al., 2018; Gao et al., 2021; Fang et al., 2024). Instead of requiring users to precisely formulate their next query or even fully understand their own needs, systems can provide proactive guidance as follow-up questions that align with users' conversational goals and significantly enhance the convenience of interactions by minimizing the cognitive load on users. Despite their importance, crafting proactive guidance still remains challenging, particularly in multi-turn conversation systems where users' goals may undergo multiple shifts during interactions (Deng et al., 2023; Bordes et al., 2016).

Traditional methods that utilize LLMs with historical conversation as contextual information have shown impressive results in guidance quality (Li et al., 2024; Duan et al., 2025; Feng et al., 2023). However, they face several challenges when de-

*Corresponding author.



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 706–717

ployed in real-world scenarios. Firstly, these methods often struggle to dynamically adapt to changes in user conversational goals (Li et al., 2024), as incorporating the entire conversation history can inadvertently introduce irrelevant information, which may result in misaligned guidance (i.e., query shifts from food allergy to the stock market may cause LLMs to persistently recommend food safety, losing track of the user's new conversational goal). Secondly, redundant historical context, especially lengthy answers, introduces computational overhead and increased latency, severely affecting real-time interactions (Lapov et al., 2024). Lastly, the high computational demands of LLMs further amplify these issues, hindering their practicality in generating rapid responses.

To address these challenges, we propose an innovative framework that combines Goal-adaptive Supervised Fine-Tuning (G-SFT) with Click-oriented Reinforcement Learning (C-RL) to solve the proactive guidance task, as illustrated in Figure 1.

In the G-SFT phase, our Goal Adaptation Agent (GAA) dynamically identifies and adapts to user goal shifts through three core outputs: explicit goal analysis, shift detection signals, and concise goal-relevant summary. By replacing redundant historical context with these signals in the generation of guidance, we achieve 65.5% faster processing in later turns and 10.18% higher click-through rates. Alongside this, scalable knowledge transfer distills LLMs' vast world knowledge into a more compact model, the G-SFT model, maintaining guidance quality while further reducing inference latency.

The C-RL phase further optimizes the G-SFT model, leveraging user click signals to construct preference pairs for alignment. Various forms of reinforcement learning (Kaelbling et al., 1996; Schulman et al., 2017; Rafailov et al., 2023; Amini et al., 2024; Ethayarajh et al., 2024) have been proposed and implemented in conversation systems due to their ability to adapt responses to better align with user preferences. The key challenge lies in generating meaningful training samples of $k$ guidance from single-clicked guidance, as the model must provide $k$ guidance options per turn. We address this using a generate-rank paradigm: (1) training an augmentation model on 1-pair click data, (2) generating diverse candidate guidance groups using Diverse Beam Search (DBS) (Vijayakumar et al., 2016), and (3) ranking and sampling $k$-pair data using a click estimator and a novel diversity-aware group sampling strategy. Experimental results demonstrate significant improvements, with accuracy increasing by 3.47% and click-through rates increasing from 20.81% to 25.28% in industrial deployment environments.

Our contributions can be summarized as follows:

- We introduce a goal adaptation agent that dynamically identifies and adapts to shifts in user goals, generating concise, goal-aligned summaries that streamline context for guidance generation without additional latency.
- We develop a generate-rank paradigm that leverages the DBS-based generation method, coupled with a group sampling strategy, to address the gap between single-preference data and multi-output requirements, thereby further enhancing the guidance quality.
- Comprehensive experiments demonstrate significant improvements in accuracy, task-related gains ($\Delta$GSB), and click-through rate, validating the effectiveness of our framework in real-world conversational search scenarios.

## 2 Methodology

In this section, we first provide a formal definition of the proactive guidance task in the multi-turn conversation system, then present our innovative two-phase framework, as illustrated in Figure 2.

### 2.1 Proactive Guidance

The task aims to generate a set of guidance phrases, $G_i = \{G_{i1}, G_{i2}, \ldots, G_{ik}\}$, during the $i$-th round of the conversation, where $k$ is a predefined constant. Specifically, in each round $i$, given user's query $Q_i$, the corresponding answer $A_i$ and contextual information $C_i$, our objective is to determine the optimal function $f_i^*$ to generate $G_i$ that maximizes the well-designed evaluation function $\mathbb{Y}$:

$$f_i^*(Q_i, A_i, C_i) = \arg\max_{G_i} \mathbb{Y}(G_i \mid Q_i, A_i, C_i), \tag{1}$$

where $\mathbb{Y}$ comprises two components: the offline and online evaluations. Offline evaluation, $\mathbb{Y}_{\text{offline}}$, assesses 1) Relevance: this evaluates the relevance of $G_i$ in the context of the conversation; 2) Applicability: this dimension measures the practical utility of $G_i$; 3) Diversity: this criterion evaluates the variety and breadth of $G_i$, ensuring a relatively comprehensive range of perspectives. The $\mathbb{Y}_{\text{offline}}$ is conducted through manual scoring by trained annotators, with full evaluation criteria provided in

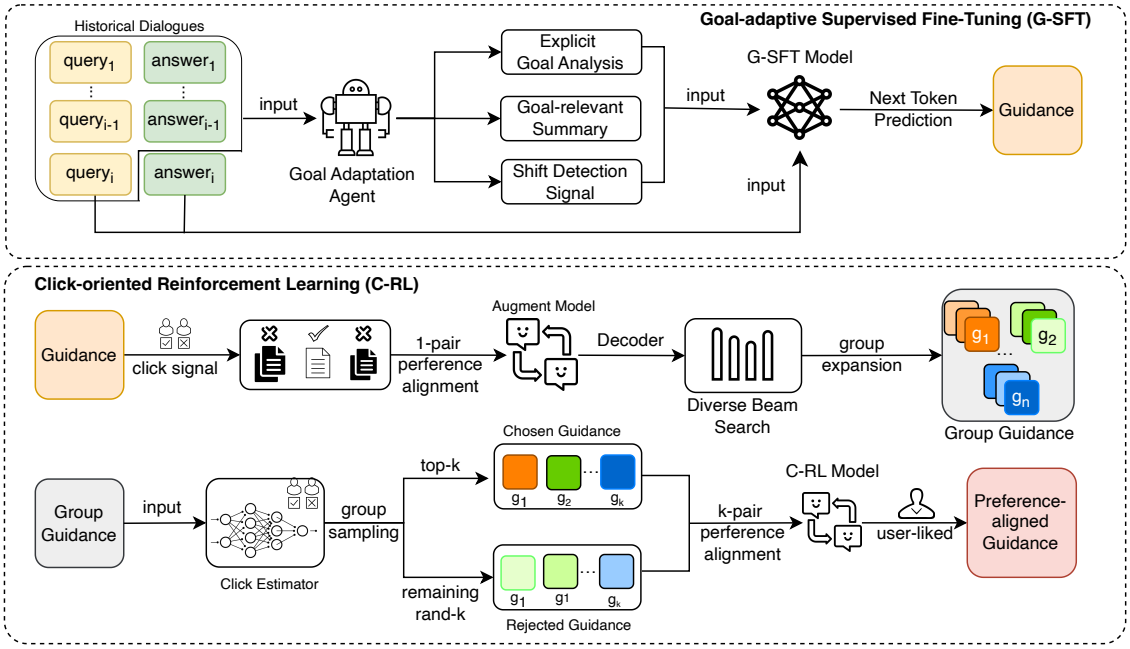


Figure 2: Architecture of the proposed framework.

Appendix D. Online evaluation, $\mathbb{Y}_{\text{online}}$, evaluates the effectiveness of the guidance $G_i$ in stimulating user engagement and promoting users' further interactions, which is quantified using the Click-Through Rate (CTR) metric.

### 2.2 Goal-adaptive Supervised Fine-Tuning

This phase is meticulously designed to produce a model capable of dynamically adapting to shifts in users' goals, providing high-quality guidance, and meeting the stringent latency requirements of industrial applications.

#### 2.2.1 Goal Adaptation Agent

Users' goals are defined as their explicit or implicit query intentions, which may undergo multiple shifts during interaction. By providing **E**xplicit goal analysis $E_i$, goal-relevant **S**ummary $S_i$ and shift **D**etection signal $D_i$, all together as contextual information $C_i$, the Goal Adaptation Agent (GAA) effectively assists the guidance model in dynamically adapting to these shifts.

The process is described in the following. In the initial round ($i = 1$), the GAA is not activated. During the second round ($i = 2$), it analyzes the current query $Q_2$ with the previous dialogue $(Q_1, A_1)$ to generate $\{E_i, S_i, D_i\}$. In subsequent rounds ($i > 2$), besides previous QA pair, the GAA additionally incorporates $S_{i-1}$ to seamlessly maintain context. This process is facilitated through the use of prompts, as described in Appendix A, which details the specific prompts employed by GAA.

**Explicit Goal Analysis.** GAA initially performs a detailed goal analysis by examining the correlation between the current query and the previous dialogue, identifying shifts and evolutions in the user's goals; then it provides explicit textual descriptions of the current intentions and infers potential underlying needs.

**Goal-relevant Summary.** GAA generates concise, goal-aligned contextual information based on $\mathrm{E}_i$ by (1) filtering goal-relevant segments from $A_{i-1}$ and $S_{i-1}$, and (2) inheriting pertinent information from $S_{i-1}$ while summarizing key points from $A_{i-1}$, omitting irrelevant details, to produce $S_i$, which focuses on the most relevant information, enabling the guidance agent to maintain coherence during dynamic goal shifts.

**Shift Detection Signal.** The detection signal $D_i$ serves as an indicator of whether a goal shift has occurred. When a goal shift is detected, $D_i$ prompts the system to reset $S_i$, thereby eliminating outdated information.

Two critical aspects of the GAA should be highlighted: First, the current answer $A_i$ is not used in GAA since it does not reflect the user's intent, allowing GAA to function simultaneously with answer generation and avoiding extra latency. Second, the contextual information $C_i$ provided by GAA

is more concise than the raw chat history, significantly reduces the computational load for guidance generation, and ultimately decreases response latency.

#### 2.2.2 Scalable Knowledge Transfer

Although LLMs deliver impressive results, their latency can be prohibitive. Conversely, smaller models often lack the world knowledge needed to handle the diverse scenarios in reality. To address this, we propose a scalable knowledge transfer method.

Initially, we utilize LLMs to process various conversations, denoted as $Q_i$, $A_i$ and $C_i$, where $C_i$ is provided by GAA. Then LLMs are prompted to produce a chain of thought, $CoT_i$, paired with a list of $n$ guidance candidates, denoted as:

$$\{CoT_i, G_{i1}, \ldots, G_{in}\} = \text{LLM}(Q_i, A_i, C_i). \quad (2)$$

Subsequently, these $n$ candidates undergo a manual selection process based on $\mathbb{Y}_{\text{offline}}$, and $CoT_i$ is strategically discarded for efficiency, resulting in a refined subset of $k$ guidance, where $k < n$. We then fine-tune a significantly smaller model on this refined dataset through a loss function defined as follows:

$$L = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x), \quad (3)$$

where $T$ is the length of the target sequence; $y_t$ is the target word at time step $t$; $y_{<t}$ denotes the sequence of words generated before time step $t$; $x$ is the input context.

Through scalable knowledge transfer, we have effectively equipped a more compact model, referred to as the G-SFT model, with the capability to offer insightful guidance whose quality rivals that of its larger counterparts.

### 2.3 Click-oriented Reinforcement Learning

During the deployment of the G-SFT model, we collected substantial data on users' interactions that inherently reflect user preferences. To fully exploit these valuable data, we introduced an innovative generate-rank paradigm, which effectively bridges the gap between the actual single-clicked guidance and the practical need for $k$ instances.

#### 2.3.1 Generate

In this section, we demonstrate the process of generating multiple guidance phrases as candidates.

**Preference-Aligned Augmentation Model.** We leverage user interaction data to create training samples consisting of preference pairs. Each instance is composed of a question, an answer, and contextual information, collectively referred to as input $x$. The guidance clicked by a user is considered as the preferred response $y_w$, while the others are treated as dispreferred $y_l$, forming preference pairs $(x, y_w, y_l)$. Then, we apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) to the G-SFT model. The goal of the DPO loss function is to optimize the model's response probability, increasing the relative probability of the preferred response. The formula is as follows:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[ \log \sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Big)\Big]. \quad (4)$$

Through this process, we produce a preference-aligned model that has the ability to generate guidance that users are more likely to click on.

**DBS-based Decoding.** To generate multiple guidance outputs using the aligned model trained with single guidance, we incorporate the Diverse Beam Search (DBS) (Vijayakumar et al., 2016) decoding strategy. DBS is an enhanced version of the beam search algorithm. It employs a grouping strategy that divides beams into multiple groups $\mathbf{Y}$ to explore different sequences independently. Additionally, DBS imposes a similarity penalty, discouraging the selection of tokens similar to those in other sequences.

For a sequence $\mathbf{y}_{[t]}$, its dissimilarity against the group $g$ at time step $t$, $\mathbf{Y}^g_{[t]}$ , is measured as:

$$\Delta(\mathbf{y}_{[t]}, \mathbf{Y}^g_{[t]}) = \sum_{b=1}^{B'} \delta(\mathbf{y}_{[t]}, \mathbf{y}^g_{b,[t]}), \quad (5)$$

where $\delta(\cdot, \cdot)$ quantifies sequence dissimilarity, e.g., a negative cost for each co-occurring n-gram in two sentences, distance between distributed sentence representations.

DBS decoding allows the aligned model to produce multiple responses in a single inference, providing guidance with significant differences in semantics, styles, or structures as candidates.

#### 2.3.2 Rank

This section describes how to construct preference pairs with $k$ guidance phrases.

**Click Estimator.** The Click Estimator is developed to predict the clicking likelihood of the guidance. It employs a sophisticated 12-layer ERNIE encoder (Sun et al., 2020) that processes user interactions through a triplet format $(Q_i, G_{ij}, y)$, $j = 1, \ldots, k$ and distinguishes between clicked $(y = 1)$ and unclicked $(y = 0)$ guidance. The training objective is:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N} \sum_{m=1}^{N} \Big[ y_m \cdot \log(\hat{y}_m) + (1 - y_m) \cdot \log(1 - \hat{y}_m) \Big], \quad (6)$$

where $\hat{y}$ denotes the predicted probability. This approach enables the click estimator to effectively predict the probability that a guidance $G_{ij}$ is clicked.

**Diversity-Aware Group Sample Strategy.** The sampling strategy that relies solely on click probability suffers from semantic redundancy, since the click estimator tends to assign similar scores to semantically equivalent guidance.

Based on the traits of DBS, we propose a diversity-aware group sampling strategy that ensures semantic richness. It works as follows: (1) Organize candidates into $n$ groups where each group $P_i$ contains the $i$-th candidate from each beam, then select the highest-CTR candidate per group to yield $n$ diverse choices as a candidate pool $P$; (2) Apply Maximum Marginal Relevance (MMR) (Guo and Sanner, 2010) with

$$\arg\max_{g_i \in P} \Big[ \lambda \cdot \mathrm{CE}(g_i) - (1 - \lambda) \cdot \max_{g_j \in S} \mathrm{sim}(g_i, g_j) \Big], \quad (7)$$

where $P$ denotes the candidate pool and $S$ denotes the selected set, $\mathrm{CE}(\cdot)$ is the click probability predicted by the click estimator. $\lambda$ is a trade-off parameter that balances click probability and semantic diversity, which is set to 0.5 in our implementation. The selecting procedure starts with the guidance clicked by real users as the initial point, then selects $k - 1$ guidance from $P$. These $k$ guidance are combined and seen as the preferred response. Then we randomly sampled $k$ guidance from the unselected ones as dispreferred, ensuring that the maximum $\mathrm{CE}(\cdot)$ score of the dispreferred guidance is less than the minimum score of the preferred guidance. The formats of training data are detailed in Appendix B.

Through this meticulous process, we create the $k$-pair preference-aligned dataset. Subsequently, we employed DPO to optimize the G-SFT model, resulting in the development of our final model being perceptible to user click preferences, referred to as the C-LR model. This model has significantly improved CTR in real-world application scenarios.

## 3 Experiments

To validate the effectiveness of our proposed method, we conducted comprehensive offline evaluations and online experiments within the Baidu Search AI assistant.

### 3.1 Experimental Setup

**Datasets.** We evaluate our models using QA pairs collected from the Baidu Search AI assistant, an industrial-scale multi-round conversation system, to ensure authenticity and diversity. For the G-SFT model, we constructed a training set of 6,072 QA pairs following Section 2.2.2. The C-RL model utilizes 12,000 preference pairs constructed using the generate-rank paradigm described in Section 2.3.

**Evaluation Metrics.** We evaluate the model's performance using three metrics: 1) Accuracy (ACC): The proportion of guidance that meets the $\mathbb{Y}_{\text{offline}}$ as introduced in Section 2.1; 2) Good vs. Same vs. Bad ($\Delta$ GSB): Comparatively evaluates the performance of two models (details in Appendix E); 3) Click-Through Rate (CTR): The ratio of turns with click behavior to total turns.

**Baselines.** We adopt ERNIE Speed (21B) (Sun et al., 2020, 2021), a publicly accessible foundation model[1], as our baseline model. The predefined number of guidance phrases $k$ is set to 3.

### 3.2 Implementation Details

**G-SFT Phase.** We use ERNIE Speed as the base model, where the learning rate is 3e-6, the max sequence length is 4,096, the batch size is 16, and the model training epoch is 3. For scalable knowledge transfer, GPT-4o is chosen as the teacher model (Hurst et al., 2024).

**C-RL Phase.** Parameters are initialized with the best checkpoint of the G-SFT model. During the DPO process, the learning rate is set to 1e-6 with a batch size of 16, and the validation steps are set to 8. The training is conducted for 2 epochs. For DBS decoding parameters, the batch size is set to 16, the number of beam groups is 4, and the beam size within each group is 4.

[1] https://cloud.baidu.com/product-s/qianfan_home

Table 1: Performance comparison of different models.

| Model | Offline | | Online |
|---|---|---|---|
| | **ACC** | **Δ GSB** | **CTR** |
| BaseLine | 62.15% | — | 10.15% |
| SKD model | 71.82% | +2.43% | 14.62% |
| G-SFT model | 82.63% | +4.24% | 20.81% |
| C-RL model | 86.10% | +5.60% | 25.28% |

*Note:* **SKD model** refers to the model after Scalable Knowledge Transfer without the use of GAA. The **G-SFT model** is the model produced after the G-SFT stage of our proposed method, which incorporates both SKD and GAA. The **C-RL model** is the G-SFT model fine-tuned with DPO on the dataset constructed using our proposed generate-rank method.

Table 2: Ablation Studies of Goal Adaptation Agent (**GAA**).

| Model | Offline | | Online |
|---|---|---|---|
| | **ACC** | **Δ GSB** | **CTR** |
| SKD model | 71.82% | — | 14.62% |
| + S | 75.62% | +2.97% | 16.43% |
| + SD | 78.21% | +3.11% | 17.81% |
| + DE | 81.16% | +3.67% | 19.72% |
| + GAA | 82.63% | +4.24% | 20.81% |

*Note:* **SKD model** refers to the model after Scalable Knowledge Transfer without the use of GAA. The table illustrates the impact of different components on model performance. **S** represents the goal-relevant summary, **D** denotes the detection signal of goal shift, and **E** stands for explicit goal analysis.

### 3.3 Results and Analysis

**Overall Results.** Experiments demonstrate significant improvements across offline and online metrics. As shown in Table 1, the baseline model achieves 62.15% ACC and 10.15% CTR, while the C-RL model achieves improved performance with 86.10% ACC, +5.60% ΔGSB and 25.28% CTR. In particular, compared to the SKD model, the G-SFT model increases ACC by 10.81% and CTR by 6.19%, validating the superior goal management capabilities of GAA. Meanwhile, the C-RL phase further enhances CTR by 4.47% with ACC gains (+3.47%), demonstrating the ability of the C-RL model to capture implicit user preferences through click data. These results confirm the effectiveness of our two-phase framework, which excellently performs the task of proactive guidance. Appendix C provides a real sample.

**Consistency Analysis.** There is a strong correlation between offline and online metrics (Spearman's $\rho = 0.986$, $p < 0.01$), indicating that our proposed strategy not only improves objective accuracy but also effectively enhances user experience. The scalable knowledge transfer model shows improvements in ACC and CTR of +9.67% and +4.47% respectively, GAA with improvements of +10.18%/+6.19%, and C-RL with improvements of +3.47%/+4.47%. In particular, the excess gain in CTR of the reinforcement learning phase highlights its ability to capture implicit features of user goals through click behavior.

**Latency Analysis.** Our system achieves industrial-grade efficiency through two techniques: (1) Scalable knowledge transfer, transferring LLMs' world knowledge to a more compact model and further removing the CoT, significantly reduces inference latency by 69.55% (from 2.89s to 0.88s). (2) by replacing raw chat history with GAA-generated concise contextual information, latency decreases by 65.5% (3.25s → 1.12s). The combined optimizations enable real-time responsiveness with end-to-end latency around 1s, meeting industrial deployment requirements.

### 3.4 Ablation Studies

**Goal Adaptation Agent.** The ablation studies of the GAA in Table 2 highlight the critical roles of its components: (1) goal-relevant **S**ummary, (2) **D**etection signal of goal shift, and (3) **E**xplicit goal analysis. The complete GAA achieves optimal performance with 82.63% ACC and 20.81% CTR, underscoring the importance of component synergy for effective multi-turn guidance.

Retaining only **S** results in a notable performance decrease (ACC -7.01% , CTR -4.38%), emphasizing the necessity of comprehensive goal management to maintain conversational coherence. Adding **D** helps recover some performance (ACC 78.21%, CTR 17.81%) by detecting goal shifts and prompting adjustments. However, **E** has a greater impact, achieving 81.16% ACC and 19.72% CTR, by providing a deeper understanding of user intentions. The results indicate that **D** and **E** are essential for maintaining coherent and context-aware guidance in multi-turn conversation.

**DBS Decoding Strategies.** This study examines the impact of the BEAM_GROUP_NUM **B** on generation quality using the DBS decoding strategy. As shown in Table 3, setting **B** to 4 achieves the op-

Table 3: Ablation studies of DBS decoding parameters.

| Model | Offline | | Online |
|---|---|---|---|
| | **ACC** | **Δ GSB** | **CTR** |
| G-SFT model | 82.60% | — | 20.81% |
| $B = 1$ | 82.14% | +2.88% | 22.54% |
| $B = 2$ | 84.87% | +3.23% | 24.78% |
| $B = 4$ | 86.10% | +3.60% | 25.28% |
| $B = 8$ | 84.31% | +3.11% | 24.16% |

*Note:* **B** represents the BEAM_GROUP_NUM used in the diverse beam search decoding strategy.

timal balance with an ACC of 86.10% and a CTR of 25.28%. A group count of 1 limits the diversity, reducing CTR to 22.54%, while 8 groups introduce noise, lowering CTR to 24.16%. Notably, setting **B** to 2 maintains a high CTR of 24.78% and improves decoding efficiency, offering a practical strategy for real-world deployment.

## 4 Conclusion

In this paper, we propose a novel framework for proactive guidance in multi-turn conversation systems, integrating G-SFT with C-RL to address challenges in dynamic goal adaptation and real-time responsiveness. Our approach demonstrates significant improvements in both guidance quality and system efficiency. Experimental results demonstrate that the framework effectively encourages user interaction and significantly increases click-through rates, highlighting its practical value in industrial scenarios.

## 5 Future Work

Despite the progress made in the proactive guidance for multi-turn conversation systems, there remain several areas for improvement and further investigation:

- **Refinement of summary reset mechanisms:** our current methodology resets $S_i$ when goal shifts are detected, failing to accommodate temporary shifts in user goals, resulting in loss of information when users return to previous intentions. Future enhancements could utilize a more sophisticated state-tracking system, allowing for a more flexible and coherent interaction experience.
- **Exploring more diverse baseline models:** The comparison with baseline models in the current study has provided a foundational understanding of our framework's capabilities. However, the rapid advancement in neural network architectures and language models suggests that integrating and comparing newer models could yield further insights.
- **Expansion of evaluation metrics:** the offline evaluation metrics used in this study, while comprehensive, could be expanded to include more diverse criteria that capture other aspects of user experience, such as user satisfaction or the system's ability to handle unexpected queries. Future studies could explore additional metrics that provide a deeper understanding of the qualitative aspects of conversation.

By addressing these future directions, we aim to enhance the functionality and applicability of proactive guidance, paving the way for more intelligent, adaptable, and user-centric conversational agents. This continued research could have a profound impact on the development of AI-driven communication tools across various domains.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing evolution of multi-turn conversations for effective answer retrieval. In *Proceedings of the 2020 conference on human information interaction and retrieval*, pages 33–42.

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A unified multi-task learning framework for multi-goal conversational recommender systems. *ACM Transactions on Information Systems*, 41(3):1–25.

Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, et al.

2025. Guidellm: Exploring llm-guided conversation with applications in autobiography interviewing. *arXiv preprint arXiv:2502.06494*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135*.

Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. A large language model enhanced conversational recommender system. *arXiv preprint arXiv:2308.06212*.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten De Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open*, 2:100–126.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Shengbo Guo and Scott Sanner. 2010. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 833–834.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Viktor Lapov, Nicholas Laurent, Lawrence Araya, Gabriel Ortiz, and Samuel Albrecht. 2024. Dynamic context integration in large language models using a novel progressive layering framework.

Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. 2024. Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. *arXiv preprint arXiv:2405.01868*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Vaishali Vadhavana, Krishna Patel, Brinda Patel, Bansari Patel, Naina Parmar, and Vaibhavi Patel. 2024. Conversational question answering systems: A comprehensive literature review. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 1088–1095. IEEE.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint arXiv:2501.09959*.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

## A Prompt for the Goal Adaptation Agent

This appendix presents the structured prompt for the goal adaptation agent.

**Prompt:** You are a **Goal-Tracking Model** specifically designed for multi-turn dialogue scenarios. Your task is to understand and track the user's evolving goals throughout the dialogue and produce coherent summaries that capture the history and progression of the conversation. This process involves preserving contextual continuity and relevance to the user's current objectives. To accomplish this, you will utilize the following inputs:

- **[** $Q_i$ **]**: The current user question in the dialogue, which may indicate a continuation of previous goals or the introduction of new goals.
- **[** $(Q_{i-1}, A_{i-1})$ **]**: The immediate previous question and answer pair, providing context for $Q_i$ and potentially containing clues about changes in the user's intent since the last turn.
- **[** $S_{i-1}$ **]**: A comprehensive summary of the dialogue history up to the interaction immediately preceding $Q_i$, encapsulating key points and actions taken that are relevant to the evolving goals of the user.

**Task:**

(1) **Explicit Goal Analysis:**

- Perform a detailed analysis of **[** $Q_i$ **]** in the context of **[** $(Q_{i-1}, A_{i-1})$ **]**, to detect nuanced changes in the user's goals. Provide a clear and explicit textual explanation that articulates the current user's intent, and infer any underlying or potential needs that may be driving this intent.

(2) **Goal-relevant Summary:**

- Based on the results of the explicit goal analysis, selectively extract content from **[** $S_{i-1}$ **]** and **[** $(Q_{i-1}, A_{i-1})$ **]**, that is directly related to the user's current goals. Integrate these key points into a new, updated summary **[** $S_i$ **]**, ensuring that it is concise yet comprehensive. Prune any elements that are no longer relevant to the current context or the user's goals to maintain focus and clarity in the evolving conversation.

(3) **Detection Signal:**

- Provide a detection signal **[** $D_i$ **]** that indicates whether a goal transition has occurred between the previous turn and the current turn. If such a transition is detected, trigger a reset of **[** $S_i$ **]** to ensure that the summary remains relevant and does not retain outdated information that could interfere with the user's current goal orientation.

**Expected Output Format:**

The expected output should be a structured JSON object, as follows:

```
{
  "explicitGoalAnalysis": "Description of the user's current intent, and inferred
  potential needs of the user",
  "goalRelevantSummary": "Coherent summary incorporating key points relevant
  to the user's current goals",
  "detectionSignal": "Boolean indicating whether a goal transition has been
  detected"
}
```

## B Data format of G-SFT and C-RL

### B.1 Prompt format

Here is the detailed prompt used for G-SFT and C-RL.

**Background:** As a Proactive Guidance Model, you are tasked with enhancing user experience in a multi-turn dialogue system by predicting potential future inquiries. Through careful analysis of the current and past interactions, you will help drive the conversation towards fulfilling the user's objectives.
**Input Explanation:** The following elements are provided for your analysis:

- Current round's user query (**[** $Q$ **]**).
- The corresponding system's answer (**[** $A$ **]**).
- Contextual information from previous rounds, which includes:
  - A summary of the dialogue thus far (**[** $S$ **]**).
  - Explicit goal analysis, detailing the objectives and needs of the user (**[** $E$ **]**).

**Thought Process:** In predicting the user's next questions, you should:

1. Assess if the current round's answer (**[** $A_n$ **]**) has adequately addressed the user's query (**[** $Q_n$ **]**).
2. Utilize the contextual information, particularly the summary and explicit goal analysis, to comprehend the user's continuous journey and objectives within the dialogue.
3. Anticipate the user's potential next steps by considering the dialogue's progression and any identified goals or needs.
4. Generate $k$ relevant and contextually appropriate questions as guidance that the user might ask next.

**Output Format Requirements:** Present your predictions structured as follows:

```
Guidan_1\n...\nGuidance_k
```

### B.2 Response format:

Here shows the response format of different tasks.

**For G-SFT:**

```
response: Guidan_1\nGuidance_2\nGuidance_3
```

**For 1-pair DPO(Augmentation model as in section 2.3.1):**

```
Chosen: Guidance(clicked)
Rejected: Guidance(unclicked)
```

**For k-pair DPO(C-RL model as in section 2.3):**

```
Chosen: Guidance_pos1\nGuidance_pos2\nGuidance_pos3
Rejected: Guidance_neg1\nGuidance_neg2\nGuidance_neg3
```

note: $Guidance\_pos*$ stands for the chosen guidance sampled through the method in section 2.3.2, while $Guidance\_neg*$ stands for rejected guidance.

## C Showcase

Figure 3 demonstrates proactive guidance in the Baidu Search AI assistant, an industrial-scale multi-turn conversation system.

On the left side of the image, the user poses the question "How to manage emotions?" The guidance is organized into three key areas: cultivating long-term emotional management habits, recommending books on emotional management, and identifying actions for immediate mood improvement. Cultivating long-term habits focuses on sustainable practices, building resilience over time. Book recommendations offer resources for deeper learning, while immediate mood improvement actions provide practical strategies for real-time relief. This structured approach effectively refines the inquiry into specific, actionable advice, enhancing user satisfaction.

On the right side of the image, the user inquires, "Which Taylor Swift song is suitable for a marriage proposal?" The guidance here is thoughtfully structured into three suggestions: Are there any more song recommendations for a proposal? What are the lyrics to "Love Story"? What other classic songs does Taylor Swift have? Each recommendation serves a distinct purpose, ensuring comprehensive support for the user's inquiry. The first expands song options, enhancing satisfaction by offering a wider array of choices. The second caters to users interested in song lyrics, allowing a deeper connection with the thematic elements. The third broadens the user's musical horizon with classic Taylor Swift songs, aiding in discovering songs that resonate with their proposal vision.

Overall, the guidance in both scenarios is diverse and non-overlapping, addressing potential user goals and enhancing engagement through structured, actionable advice.


Figure 3: Proactive guidance in Baidu Search AI assistant. The left query is "How to manage emotions?" and the right query is "Which Taylor Swift song is suitable for a marriage proposal?"

## D Evaluation Criteria

This appendix outlines the evaluation criteria used for assessing the effectiveness of the guidance phrases generated during the conversation rounds. Our evaluation framework consists of three main components: relevance, applicability, and diversity. Each component is crucial for ensuring the quality and utility of the guidance provided. The evaluation is conducted by trained annotators based on the following detailed criteria:

### D.1 Relevance

- **Contextual Relevance**: The guidance phrases must be directly related to the user's query and the ongoing conversation. They should address the user's needs without introducing unrelated topics.

- **Coherence**: The phrases should maintain logical consistency with the conversation history, avoiding contradictions and repetition.

### D.2 Applicability

- **Intent Clarification**: When the user's intent is unclear or comprises multiple potential directions, the guidance should help the user to clarify their intent.

- **Identifying Hidden Demands**: If the current query is only part of the user's fundamental needs, the guidance should aim to uncover underlying requirements, offering comprehensive or extended guidance.

- **Personalized Information Supplementation**: When the user's intent is clear but requires personalized information, the guidance should prompt the user to provide necessary context for a tailored response.

### D.3 Diversity

- **Comprehensiveness**: The guidance should cover a wide range of dimensions or options. It should be supported by expert knowledge or strong a posteriori information justifying the necessity of each guidance element.
- **Mutual Exclusivity**: The guidance should not repeat or overlap with the user's original query or with content already adequately addressed in previous answers. Different guidance options should be distinct from one another, avoiding intersections or inclusions.

### D.4 Redline Criteria

- **Legal and Ethical Compliance**: Guidance must not violate national laws, involve sensitive political or adult content, or touch on sensitive topics.
- **Accuracy and Truthfulness**: The information provided must be factual and free from rumors or misinformation.
- **Emotional Impact**: Guidance should avoid content that is excessively violent, discomforting, or sensationalist, such as exaggerated or eye-catching lowbrow titles.

## E Good vs. Same vs. Bad (GSB) Calculation Details

Good vs. Same vs. Bad (GSB) is a metric judged by professionally trained annotators. For each user query, annotators are presented with the answer, historical conversations, and the guidance generated from both model A and model B. Based on the quality of the guidance, annotators independently assign one of the following labels:

- **Good**: Results from model A are better than model B.
- **Bad**: Results from model B are better than model A.
- **Same**: Results from model A and model B are of equal quality (either good or bad).

To quantify the human evaluation, we use a unified metric $\Delta$GSB, defined as:

$$\Delta\text{GSB} = \frac{\#\text{Good} - \#\text{Bad}}{\#\text{Good} + \#\text{Same} + \#\text{Bad}}.$$