Agentic Conversational Search with Contextualized Reasoning via Reinforcement Learning

Fengran Mo[1*], **Yifan Gao**[2], **Sha Li**[2], **Hansi Zeng**[3*], **Xin Liu**[2], **Zhaoxuan Tan**[4*]
Xian Li[2], **Jianshu Chen**[2], **Dakuo Wang**[5], **Meng Jiang**[4]
[1]University of Montreal; [2]Amazon.com; [3]University of Massachusetts Amherst
[4]University of Notre Dame; [5]Northeastern University
`fengran.mo@umontreal.ca, yifangao@amazon.com`

Abstract

Large Language Models (LLMs) have become a popular interface for human–AI interaction, supporting information seeking and task assistance through natural, multi-turn dialogue. To respond to users within multi-turn dialogues, the context-dependent user intent evolves across interactions, requiring contextual interpretation, query reformulation, and dynamic coordination between retrieval and generation. Existing studies usually follow static "rewrite, retrieve, and generate" pipelines, which optimize different procedures separately and overlook the mixed-initiative action optimization simultaneously. Although the recent developments in deep search agents demonstrate the effectiveness in jointly optimizing retrieval and generation via reasoning, these approaches focus on single-turn scenarios, which might lack the ability to handle multi-turn interactions. We introduce a conversational agent that interleaves search and reasoning across turns, enabling exploratory and adaptive behaviors learned through reinforcement learning (RL) training with tailored rewards towards evolving user goals. The experimental results across four widely used conversational benchmarks demonstrate the effectiveness of our methods by surpassing several existing strong baselines.

1 Introduction

Large Language Models (LLMs) have become a popular interface for human–AI interaction, supporting information seeking and task completion assistance through natural, multi-turn dialogue (Gao et al., 2022; Zamani et al., 2023; Mo et al., 2024b). To respond to users within multi-turn dialogues, the context-dependent user intent usually evolves across historical interactions, requiring contextual interpretation (Jin et al., 2023), mixed-initiative actions (Aliannejadi et al., 2021), and dynamic coordination between information retrieval and response generation (Su et al., 2024; Qian and Liu, 2025; Lai et al., 2025; Qi et al., 2026).

To facilitate the multi-turn conversational scenarios, existing studies (Roy et al., 2024; Ye et al., 2024; Mo et al., 2026a) usually follow static "rewrite, retrieve, and generate" workflows with separated models. Such a pipeline optimizes various procedures separately and requires different task-specific supervision signals, including rewritten queries, relevance judgments, and ground-truth answers for each turn. Besides, it might overlook the mixed-initiative action optimization, e.g., asking a clarification question at a suitable moment, simultaneously. Aiming for a dynamic procedure, recent developments in deep search agentic models (Jin et al., 2025; Zheng et al., 2025; Zeng et al., 2026) achieve remarkable performance on complex information-seeking tasks by interleaving multi-round reasoning and external search. The flexible architecture allows it to dynamically explore external knowledge sources and iteratively refine its reasoning trajectory for open-domain question answering (Zhang et al., 2024b; Zhu et al., 2025b; Zhang et al., 2025a; Su et al., 2025). However, these advanced deep search agentic models are designed for single-turn information accessing tasks, which might lack the ability to handle multi-turn conversational scenarios (Zhang et al., 2024a; Zhu et al., 2025a; Mo et al., 2025c).

Under user-system multiple interactions, the user's queries are context-dependent and involve topic-switch (Mao et al., 2022a; Mo et al., 2023b; Yoon et al., 2025). Thus, it is crucial for the conversational agent to learn to leverage useful historical context to reply to the current turn via reasoning ability. One important aspect is to generate a stand-alone search query condition on comprehending the historical context (Jang et al., 2023; Mo et al., 2024a), which is different from the existing single-turn deep search agent that decomposes

---

[*] Work done during the internship in Amazon.

Findings of the Association for Computational Linguistics: ACL 2026, pages 9125–9138
July 2-7, 2026

a self-contained query. Another crucial factor lies in optimizing the model to effectively leverage information obtained from search results through de-contextualization, enabling accurate answer generation that integrates retrieved knowledge with the model's inherent parametric knowledge (Wang et al., 2024a; Su et al., 2025; Mo et al., 2026b), while the existing studies on deep search lack search results optimization (Jin et al., 2025). Besides, the flexible interaction interface enables the system to respond in various ways, which make it possible to equip the conversational agents with mixed-initiative actions capabilities to response in suitable forms, such as asking clarification questions (Zou et al., 2023), providing short or long form answers with dynamic decisions depending on the query type (Jeong et al., 2024; Mo et al., 2025b), and rejecting to answer when without enough evidence (Wang et al., 2024b). The mixed-initiative actions are useful for the system to engage with the user (Aliannejadi et al., 2021; Mo et al., 2024d; Wang et al., 2024c).

To address the aforementioned issues, we aim to optimize the agentic conversational search task from multiple aspects within a single agentic framework with contextualized reasoning, including history-conditional search query generation, the utilization of search results for accurate answer generation, and interactive action decision via mixed-initiative capacities. To achieve it, we propose *ConvAgent* to decompose the overall reward into three complementary components under a reinforcement learning (RL) training framework (Shao et al., 2024) to interleave search and reasoning across turns, enabling exploratory and adaptive behaviors learned through tailored design. Specifically, to help the agentic model optimize search queries through de-contextualization reasoning in conversations and learn to extract useful information from retrieved results for answer generation, we introduce a search optimization reward. This reward measures the information overlap between the retrieved content and the ground-truth answers, providing additional supervision beyond the ground-truth answer signal alone. Besides, a mixed-initiative action reward is used to enhance the alignment with evolving user goals and facilitate the mixed-initiative capacity of the models. The experimental results across four widely used conversational search benchmarks demonstrate the effectiveness of our methods by surpassing several existing strong baselines with different principles under various settings. Detailed ablation analyses show the contributions of different components and insightful observations to improve multi-turn capacity for conversational agents.

Our contributions can be summarized as follows:

(1) We introduce an agentic framework to optimize the conversational search task from multiple aspects via contextualized reasoning, including history-conditional search query generation, the utilization of search results, and interactive action decision via mixed-initiative capacities.

(2) We propose to decompose the RL training rewards with intermediate components, including a search optimization and a mixed-initiative action reward in addition to the outcome, which enables the model with exploratory and adaptive behaviors to interleave search and reasoning across turns.

(3) We conduct experiments on four widely used conversational search benchmarks and take analyses to understand how agentic models with search and mix-initiative actions impact generation performance in conversational search.

2 Related Work

Conversational Question Answering is the task of responding to user queries within multi-turn dialogues, with dependence on previous turns and often grounding from external knowledge (Qu et al., 2020). The retrieval-augmented generation (RAG) technique (Lewis et al., 2020) is a common practice to obtain external factual grounding in answers. Previous studies in conversational search (Wu et al., 2022; Mo et al., 2023a; Abbasiantaeb et al., 2024; Lupart et al., 2025b) aim to enhance the understanding of user intent and more complex user interactions. A series of approaches are proposed to improve the search results performance in conversations, including query rewriting (Mao et al., 2023; Ye et al., 2023; Lai et al., 2024), conversational dense retrieval fine-tuning (Lin et al., 2021; Kim and Kim, 2022; Mao et al., 2024; Mo et al., 2024c), and data augmentation (Mao et al., 2022b; Mo et al., 2025d). Besides, several studies (Meng et al., 2023; Owoicho et al., 2023) have explored improving mixed-initiative action detection in conversational settings to provide more effective support during interactions. However, these approaches do not directly explore how to leverage the retrieved information to improve answer generation in the era of LLMs (Gao et al., 2023; Mo et al., 2024b; Bhaskar et al., 2025; Zhang et al., 2025b,c). Our research aims to develop an agentic conversational search model that can understand context-dependent user intent, optimize search queries, enhance mixed-initiative interaction, and reason over search results to generate accurate answers.

Agentic Search. Recent studies (Zhang et al., 2024b) focus on developing agentic LLMs for information accessing tasks via RL training. The principle is to decompose the complex questions via integrating in-context reasoning with dynamic search tool invocation when needed. Representative studies such as Search-r1 (Jin et al., 2025), R1-Searcher (Song et al., 2025), and DeepResearch (Zheng et al., 2025) optimize RL policies for generating multiple queries through multi-turn search interactions and integrating retrieved evidence, thereby achieving strong performance on knowledge-intensive QA tasks. However, they are limited to handling self-contained questions as one-shot querying and do not require understanding historical context. Different from them, we aim to define an intermediate reward tailored to answer questions in conversations that directly optimizes the agentic model to utilize search results when generating rewritten queries for search in capturing user intent at each dialogue turn. Concurrent work ChatR1 (Lupart et al., 2025a) and ConvRecR1 (Zhu et al., 2025b) attempt to extend this line of work, while their designed reward for conversational scenarios relies on the ground-truth of rewritten queries as training signals and does not consider the mixed-initiative action capability.

3 Methodology

3.1 Task Formulation

A conversational agent should be capable of performing various actions through reasoning in conversation to arrive at the final answer. Formally, given a dataset of user–system conversations, each composed of multiple turns, and a collection of passages $\mathcal{C}$. At each turn n, the agent system receives the conversation history $\mathcal{H}_n = \{q_i, a_i\}_{i=1}^{n-1}$ and the current user query q_n. The conversational agent requires generating an answer a_n to query q_n, leveraging context from $\mathcal{H}_n$ and grounding in external information $\mathcal{C}$. In some cases, the agent should also be capable of handling unanswerable or ambiguous queries through rejection and clarification, resulting in requiring mixed-initiative action capabilities. We aim to develop such an agentic model via contextualized reasoning in this study.

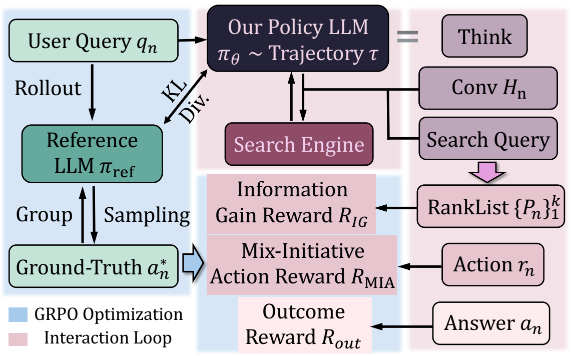

Figure 1: Overview of our designed agentic conversational search framework. The GRPO Optimization is guided by full trajectories sampled from the policy LLM, which consists of reasoning, search, mix-initiative action, and answer steps, with rewards assigned to each signal. The policy model interleaves the same action set while interacting with a search environment within the Interaction Loop for both training and inference.

3.2 Rewards Modeling

Reinforcement learning (RL) plays a crucial role in training agentic models by enabling self-exploration of diverse reasoning trajectories and rewarding those that are effective. This aligns model behavior with the objectives of information seeking for users' context-dependent intents and efficient decision-making for actions in conversations. However, since these trajectories are self-generated and do not always yield correct final answers due to noisy conversational context, relying solely on outcome-based rewards yields sparse supervision signals, making the training process harder to optimize (Chen et al., 2025; Zhang et al., 2026). To this end, we decompose the overall reward into three complementary components: outcome reward, search optimization reward, and mixed-initiative action reward. These rewards guide the model to reason more strategically with the conversational context towards the final answer, effectively capture user intent for search optimization, and respond precisely to various types of queries with mixed-initiative needs in conversations. We describe the overview of our framework in Figure 1.

3.2.1 Outcome Generation Reward

The outcome generation reward directly reflects the final task success. At the end of each rollout, the predicted answer a_n is extracted to evaluate by comparing with the ground-truth answer a_n^* using a task-specific metric $\mathcal{S}_{\text{ans}}$ as $\mathcal{S}_{\text{ans}}(a_n, a_n^*)$, e.g., Exact Match or F1 score. The formulation of the outcome generation reward is

$$\mathcal{R}_{\text{outcome}} = \mathcal{S}_{\text{ans}}(a_n, a_n^*) \tag{1}$$

3.2.2 Information Gain Reward

Solely relying on the supervision signal from the ground-truth answer is not enough to enable the conversational agent to interpret user intent and establish high-quality intermediate results for external search. Different from previous studies (Jin et al., 2025; Song et al., 2025), which only generate search queries for calling the search interface, the agentic model requires optimizing the search query via de-contextualization in conversations and learn to utilize the gained information via external retrieval, i.e., the useful parts of the search results in terms of approaching the final answer. To this end, the information gain reward is designed to measures the information overlap between the retrieved information and the ground-truth a_n^* via applying a specific metric $\mathcal{S}_{\text{Info}}$ on them as $\mathcal{S}_{\text{Info}}(\{P_n\}_1^k, a_n^*)$. The $\{P_n\}_1^k$ is the top-k relevant passages obtained by an ad-hoc retriever $\mathcal{R}\text{et}$ via the generated rewritten-query q_n' for the n-th turn as $\{P_n\}_1^k = \mathcal{R}\text{et}(q_n')$.

Such an optimization is expected to improve the quality of the agent-rewritten search query by aligning its search results with the associated final answer without requiring the relevance judgments between query and gold passages as supervision signals. The information gain reward $\mathcal{R}_{\text{IG}}$ is formulated as the cumulative coverage of the search results in each retrieval calling iteration. According to the length of the ground-truth answer for each query (Bolotova et al., 2022), we use either F1-score or Accuracy as an evaluation metric to validate the information gain reward from long or short types of answers, which is defined as

$$\begin{aligned}\mathcal{R}_{\text{IG}} &= \mathcal{S}_{\text{Info}}(\mathcal{R}\text{et}(q_n'), a_n^*) = \mathcal{S}_{\text{Info}}(\{P_n\}_1^k, a_n^*)\\ &= \begin{cases}\max_{q_n'} \text{F1}(\{P_n\}_1^k, a_n^*) & \text{type}(a_n^*) = \text{long}\\ \max_{q_n'} \text{Acc}(\{P_n\}_1^k, a_n^*) & \text{type}(a_n^*) = \text{short}\end{cases}\end{aligned}$$

where the accuracy is measured by whether the ground-truth a_n^* is a sub-string contained in the retrieved information $\{P_n\}_1^k$ as $\mathbb{I}\big[a_n^* \subseteq \{P_n\}_1^k\big]$.

3.2.3 Mixed-Initiative Action Reward

The mixed-initiative plays an important role in conversational scenarios, where the agentic model is expected to clarify ambiguous query turns or refuse to answer when their knowledge is insufficient. To facilitate the mixed-initiative capacity of our model, we formulate the mixed-initiative as a classification task to validate whether the agentic model can decide the necessary supportive reactions. Specifically, we include four types of mixed-initiative action following (Wu et al., 2023) in the reaction set as $\mathcal{A}\text{ct} = \{\text{answer}, \text{clarify}, \text{no answer}\}$. In addition to generating a final answer as a common case, the agentic model should generate clarification questions, return only relevant search results with the final answer, or refuse to answer to accommodate the unanswerable cases. Then, the mixed-initiative action (MIA) reward $\mathcal{R}_{\text{MIA}}$ is designed similarly to the format reward to measure whether the specific label tokens, i.e., <clarify> and <noanswer>, are included in the generated sequence r_n among the turns that are annotated with requiring specific mixed-initiative actions in the ground-truth reaction set $\mathcal{A}\text{ct}_n^*$ as

$$\mathcal{R}_{\text{MIA}} = \begin{cases}1, & \text{if } r_n \subseteq \mathcal{A}\text{ct}_n^*,\\ -0.5, & \text{otherwise.}\end{cases} \tag{2}$$

The negative reward assigned to incorrect mixed-initiative actions serves as a penalty that discourages the model from aggressively generating unnecessary clarifications or refusals.

Finally, the reward assigned to a rollout trajectory τ aggregates the three components as

$$\mathcal{R}(\tau) = \mathcal{R}_{\text{outcome}} + 0.5 \times (\mathcal{R}_{\text{IG}} + \mathcal{R}_{\text{MIA}}) \tag{3}$$

3.3 Contextualized Reasoning in Conversations with RL

With the designed rewards, the optimization objective of our model with parameters θ is to maximize the aggregated reward $\mathcal{R}(\tau)$ given the history $\mathcal{H}$, the user query q at the last turn n, and the search engine $\mathcal{R}\text{et}$, while minimizing the distance between the optimized and original policies as Eq. 4.

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Big[\sum_i \min\Big(\phi_i(\theta)\mathcal{A}_i,\ \text{clip}\Big(\phi_i(\theta),\\ 1-\epsilon, 1+\epsilon\Big)\mathcal{A}_i\Big) - \gamma D_{KL}(\pi_\theta || \pi_{ref})\Big]\end{aligned} \tag{4}$$

where $\phi_i(\theta) = \frac{\pi_\theta(\tau_i|q,\mathcal{H},\tau_{<i},\mathcal{R}\text{et})}{\pi_{\theta_{old}}(\tau_i|q,\mathcal{H},\tau_{<i},\mathcal{R}\text{et})}$ denotes the probability between the new and old policies and $\mathcal{A}_i = \frac{\mathcal{R}(\tau_i) - \text{mean}(\mathcal{R}(\tau))}{\text{std}(\mathcal{R}(\tau))}$ represents the estimated advantage of the i-th rollout trajectory within the

current group calculated by our designed reward function $\mathcal{R}(\tau)$ in Eq. 3. Besides, the parameter ϵ controls the clipping threshold, while γ regulates the KL divergence penalty.

We adopt the efficient Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024), which eliminates the need for explicit reward and value models. As an alternative optimization approach, we also implement the Proximal Policy Optimization (PPO) to optimize our approach. The corresponding analysis compared to GRPO is shown in the experimental Section 4.3.

3.4 Usage of Clarification Results

The goal of generating necessary clarified questions in mixed-initiative actions for some ambiguous turns is to support the user in obtaining their desired answers. Thus, the evaluation of clarification actions should consider not only whether they are taken at appropriate moments but also how effective they are in improving downstream task performance. To this end, we design the usage of clarification results in downstream tasks. Specifically, in the retrieval procedure interacting with the search engine, the clarified question q_n^c generated by the model is used as an expansion, concatenated with the previous rewritten query q'_n as $\mathcal{R}\text{et}(q'_n \circ q_n^c)$ to obtain search results. For the answer generation procedure, we replace the original query q_n with a clarified one q_n^c, which is used as a simulated question to obtain the final answer a_n.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct the main evaluation on four widely-used conversational search datasets, including TopiOCQA (Adlakha et al., 2022), INSCIT (Wu et al., 2023) for in-domain evaluation, and QReCC (Anantha et al., 2021), and CORAL (Cheng et al., 2025a) for out-of-domain evaluation. More statistical details are provided in Appendix A.

Evaluation Metric. We investigate three different capabilities of the conversational agent: search optimization with query generation, answer generation with retrieval augmentation, and action determination in conversations. For search evaluation, we adopt the traditional retrieval metric NDCG@3 and the token-/string–based overlap InfoGain metric to measure the overlap between retrieved passages and ground-truth answers, thereby assessing information gain (Qian and Liu, 2025), which aligns with our search optimization reward. For answer generation, we use F1-score following previous studies (Lupart et al., 2025a; Mo et al., 2025a) for fair comparison, Exact Match (EM) for the short answer dataset TopiOCQA, and LLM-as-a-Judge (Gu et al., 2024) based on Qwen-3-4B model for the remaining towards semantic measurement. For mixed-initiative actions, the evaluation is conducted as a classification task and then further evaluates the improvement with supportive actions.

Baselines. We compare our model with various baselines, which is categorized into two types: i) Separated retriever-generator pipeline systems, including ChatQA (Liu et al., 2024), UniConv (Mo et al., 2025a), REFRAG (Lin et al., 2025), and EvoRAG (Cheng et al., 2025b), and ii) agentic systems, including Search-R1 (Jin et al., 2025), ConvSearch-R1 (Zhu et al., 2025a), ChatR1 (Lupart et al., 2025a), and the AgenticLM (Yang et al., 2025) using latest Qwen-3 models as backbone with agentic capacity. Among them, ConvSearch-R1 and ChatR1 require the rewritten queries from original datasets or LLM-generated pseudo data as additional supervised signal during the model training. In addition, we include a supervised fine-tuning (SFT) baseline trained with the standard next-token prediction loss for answer generation, as well as direct inference results from proprietary large language models, including ChatGPT (OpenAI, 2022) and Claude (Anthropic, 2023), for reference. The detailed information about each baseline method is provided in Appendix B.

Implementation Details. We implement our conversational agent model based on Qwen-2.5-3b/7b (Yang et al., 2025) and launch a retriever model with E5 (Wang et al., 2022) based on four NVIDIA A100 80GB GPUs. The top-3 retrieved passages are used as search results for each turn among the compared baselines and our method to ensure a fair comparison. The TopiOCQA and INSCIT are used as the training set, and evaluated on all datasets with in-domain and out-of-domain settings. The mixed-initiative action reward is applied on the INSCIT dataset, which is the only one containing corresponding supervision signals. The batch size of the policy model is 256, a max prompt length of 8192 tokens, and a learning rate of 1e-6. The total training steps are 150.

Method	LLM	TopiOCQA		INSCIT		QReCC		CORAL	
		F1	EM	F1	LLM-Jud.	F1	LLM-Jud.	F1	LLM-Jud.
SFT	Qwen-2.5-3b	18.2	14.7	23.7	52.6	17.0	50.4	15.2	42.3
Search-R1-3b	Qwen-2.5-3b	26.1	13.1	5.8	50.1	5.9	48.7	3.9	41.2
ConvSearch-R1	Qwen-2.5-3b	7.6	3.5	15.7	48.5	-	-	-	-
AgenticLM-4b	Qwen-3-4b	29.4	11.2	21.6	50.8	7.9	46.3	22.1	42.6
ChatR1-3b	Qwen2.5-3b	29.4	-	28.2	-	28.0	-	-	-
ConvAgent-3b (ours)	Qwen2.5-3b	25.2	14.1	23.5	51.8	24.1	56.3	22.4	43.2
ChatQA	LLaMA-3-8b	18.1	-	25.7	-	23.7	-	20.3	-
UniConv	Mistral-2-7b	29.6	-	**33.2**	-	26.2	-	24.3	-
EvoRAG	Qwen2.5-7b	26.8	-	26.8	-	24.0	-	25.1	-
REFRAG	LLaMA-2-7b	28.2	-	-	-	17.4	-	-	-
SFT	Qwen-2.5-7b	23.6	17.8	24.5	53.9	19.1	54.3	18.8	43.0
Search-R1-7b	Qwen2.5-7b	37.0	20.7	9.1	57.7	8.6	60.3	3.8	43.0
AgenticLM-8b	Qwen-3-8b	34.6	15.3	23.2	52.1	21.1	49.6	24.1	44.7
ChatR1-7b	Qwen2.5-7b	30.6	-	27.8	-	**31.0**	-	-	-
ConvAgent-7b (ours)	Qwen2.5-7b	**40.3**†	**23.3**†	30.1†	**58.5**†	30.4†	**64.6**†	**28.9**†	**46.8**†
For Reference - Direct Inference with Proprietary LLMs									
ChatGPT	GPT-3.5	25.5	-	22.6	-	22.8	-	26.8	-
Claude	Sonnet-3.5	27.2	-	25.0	-	27.0	-	27.4	-

Table 1: Performance of different systems for the answer generation task. **Bold** and underline indicates the best and the second-best results. Superscript † denotes significant improvements with paired t-test at $p < 0.05$ over two comparable main competitors, agentic systems Search-R1 and AgenticLM.

Method	TopiOCQA	INSCIT	QReCC	CORAL
Search-R1-3b	22.8	21.4	35.9	32.9
AgenticLM-4b	18.7	24.1	33.2	28.6
ChatR1-3b	24.1	-	36.4	-
ConvAgent-3b	22.7	24.7	37.2	31.6
Search-R1-7b	26.5	22.3	36.7	33.3
AgenticLM-8b	20.4	23.8	36.0	30.9
ChatR1-7b	26.7	-	37.0	-
ConvAgent-7b	**29.4**	**30.8**	**39.6**	**34.9**

Table 2: Performance of the retrieval task.

4.2 Main Results

In this section, we evaluate the performance of answer generation, search results, and the correlation analysis between search results quality and answer generation performance.

4.2.1 Answer Generation Performance

The result of the answer generation is shown in Table 1, where our approach consistently achieves the best or second-best performance across nearly all datasets and metrics. Specifically, our agent model with 7B variant achieves the highest F1-scores on TopiOCQA, QReCC, and CORAL datasets by surpassing the comparable agentic systems such as AgenticLM, Search-R1, and ChatR1, and even exceeds the performance of proprietary LLMs with direct inference (ChatGPT and Claude). Such observations demonstrate the importance of optimizing the model to learn to interact with the search engine and utilize the useful parts of search results in conversations via our designed reward signals.

Besides, the agentic systems (Search-R1, AgenticLM, ChatR1, and ours) perform better than the traditional "retrieval-generation" pipeline approaches (ChatQA, UniConv, and EvoRAG) with separated fine-tuning using different supervision signals, which indicates the implicit reasoning capacity is important for conversational engagement and interacting with the search engines. Among the agentic systems, the Search-R1 does not perform well on the datasets with long-form answers due to the lack of intermediate rewards. The ChatR1 should rely on rewritten queries as additional supervision signals to capture user intent, and this might be the potential reason for its better performance on 3b variants. Nevertheless, by scaling to 7b size, our approach outperforms ChatR1, which indicates the scaling law (Chung et al., 2022) should still be helpful in such a scenario.

In addition, our approach also achieves better results on LLM-as-a-Judge scores, although with less significance. The scores judged by LLM might

Method	TopiOCQA			INSCIT			QReCC			CORAL		
	N@3	F1	EM	N@3	F1	LLM-Jud.	N@3	F1	LLM-Jud.	N@3	F1	LLM-Jud.
Our Full Model	29.4	40.3	23.3	30.8	30.1	58.5	39.3	27.4	63.6	34.9	28.9	46.8
- IG Reward	26.7	37.8	20.8	23.5	17.6	54.8	37.0	11.2	60.7	32.6	14.8	44.7
- MIA Reward	28.1	39.5	21.2	29.5	24.3	60.3	39.6	30.4	64.6	33.6	24.1	46.4
- both	26.5	37.0	20.7	22.3	9.1	50.7	36.7	8.6	60.3	33.3	3.8	43.0
Replace by PPO	26.4	37.4	21.6	25.7	29.4	57.5	37.2	24.7	63.0	33.8	20.6	46.5

Table 3: Ablation studies in terms of various reward signals based on the 7b model across four datasets.

not always align with the F1 scores, which is still an open question in evaluating long-form answers.

4.2.2 Search Results Performance

Table 2 presents the retrieval performance across four benchmark datasets, comparing our approach with strong agentic baselines. Overall, our model consistently achieves the best NDCG@3 scores on all datasets for both 3B and 7B model sizes, demonstrating its effectiveness in identifying relevant information among conversational contexts. This is attributed to the search optimization rewards for generating rewritten queries with supervision signals via contextualized reasoning. Specifically, the improvement compared to existing systems is more notable on TopiOCQA and INSCIT datasets, where the topic-switch and mixed-initiative phenomenon make the conversational retrieval more challenging but are optimized directly by our designed rewards. These gains indicate that our conversational agent implicitly identifies the contextual relevance better via contextualized reasoning and emphasize the necessity of optimization toward adaptive user behaviors and search quality.

4.2.3 Correlation between Search Quality and Answer Generation Performance

The search results evaluated by traditional IR metrics (e.g., NDCG@3) might not always reflect the effectiveness for answer generation. To this end, we investigate the correlation between search quality and answer generation performance, where the search quality is measured by the information gain during the retrieval procedure in terms of approaching the final answer as described in Section 3.2.2.

We depict the results in Figure 2. We observe that even with higher information gain via search results, the Search-R1 still cannot perform better in answer generation than AgenticLM and our approach with optimized capacity to utilize the retrieved knowledge from external resources. In other words, without specific optimization for leveraging search results, the agent models would still tend to answer the question using parametric knowledge. In contrast, our method can better integrate externally retrieved information with original knowledge via our designed information gain reward.

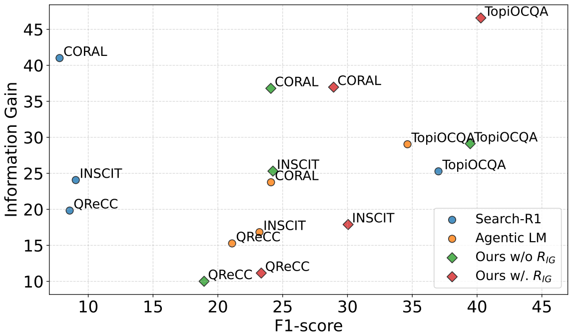

Figure 2: Correlation between search quality and answer generation performance on two main competitors and our variants of w/. or w/o information gain reward.

4.3 Ablation Studies

The results of ablation studies are shown in Table 3, where we aim to investigate the effectiveness of various reward signals during RL training. We find that the full model integrating both Information Gain (IG) reward and mixed-initiative action (MIA) reward achieves the best overall performance. Such results confirm these rewards as complementary benefits for reasoning when using search results for response generation in conversational scenarios, and our conjecture that solely relying on the supervision signal from ground-truth answers is not enough. Besides, removing either reward leads to clear performance degradation. The IG reward contributes more than the MIA reward, while the MIA reward might not always be effective in the datasets (i.e., QReCC) that do not contain mixed-initiative phenomena. In addition, excluding both rewards results in a sharp decline across all datasets, and replacing the RL optimization algorithm with standard PPO yields consistently lower results, which

Model	No Answer Acc.	Clarify Acc.	NDCG@3 Improve	F1 Improve
ConvAgent w/o $\mathcal{R}_{\text{action}}$ (Base-7b)	1.42	3.67	-	-
ConvAgent w/. $\mathcal{R}_{\text{action}}$ & sep. label	11.51	83.19	3.9	5.8
ConvAgent w/. $\mathcal{R}_{\text{action}}$ & comb. label	6.99	78.35	0.2	0.8
Qwen-3-4b	24.86	80.21	-0.4	-2.3
Qwen-3-8b	20.99	82.00	0.8	-0.2
Qwen-3-14b	16.57	83.70	4.2	4.6
Qwen-3-32b	2.21	81.50	4.8	4.4

Table 4: Accuracy of mixed-initiative action and the effectiveness of clarifications on downstream tasks.

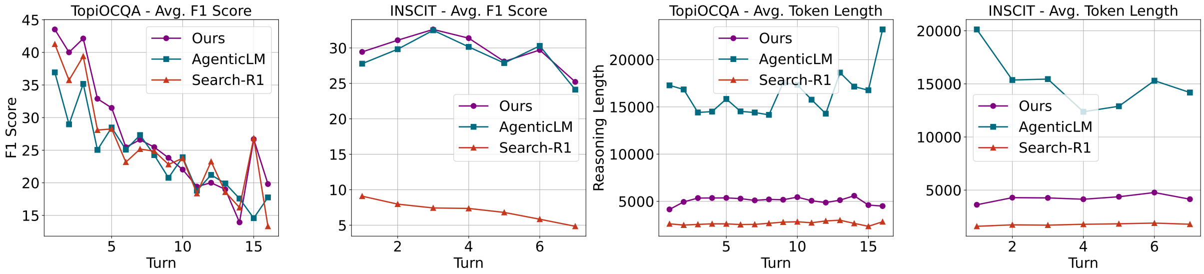

Figure 3: Impact of conversational context on the reasoning effectiveness and efficiency with different models.

demonstrate that the GRPO algorithm performs much better and provides more stable training in this scenario.

4.4 Investigation of Mixed-Initiative Actions

In this section, we investigate the performance of mixed-initiative actions. We evaluate the accuracy of taking the associated mixed-initiative actions, whether to generate clarification (Clarify Acc.) or refuse to answer when no answers are available, at specific turns across different model capacities, including variants of our models based on "Qwen-2.5-7b" with various training strategies and the latest agentic models with multiple sizes.

The results are shown in Table 4. We observe that the non-agentic backbone model training without mixed-initiative action reward $\mathcal{R}_{\text{action}}$ will almost not take mixed-initiative actions. After training with our designed reward, the accuracy of determined actions improved, as well as the F1 scores of generated answers. Among them, the format of the special labels in the input template affects the final performance, where separating the labels into <clarify>, <noanswer>, and <answer> with later parsing (sep. label) can obtain better performance than combining all mixed-initiative result content in a single span (comb. label). Moreover, the agentic models in the Qwen-3 series exhibit a clear scaling effect: larger models achieve better performance in both answer generation and clarification tasks, although they tend to respond more aggressively, even in the absence of ground-truth evidence. Nevertheless, incorporating accurate mixed-initiative actions leads to notable improvements in both retrieval performance (NDCG@3) and answer generation (F1) compared to models without mixed-initiative optimization.

4.5 Impact of Conversational Context

In this analysis, we study the impact of the context (multi-turn conversations) on the reasoning capacity to leverage the historical context learned by different models. We use their per-turn answer generation performance and reasoning token length for effectiveness and efficiency evaluation.

As shown in Figure 3, all models drop as the conversation goes on in TopiOCQA datasets, while our model consistently performs slightly better than the others. In the INSCIT dataset, the agentic models (ours and Agentic LM) exhibit their capacity in leveraging abundant historical context in mix-initiative scenarios compared to the non-agentic one (Search-R1) and thus actually increase in longer conversations. However, the agentic LM cannot perform well in terms of efficiency compared to Search-R1. With the intermediate rewards, our agentic model achieves a better effectiveness-efficiency trade-off, with better performance than Search-R1 and one-third of the reasoning token cost of Agentic LM.

5 Conclusion

In this paper, we present a contextualized reasoning RL framework to improve multi-turn capacity for conversational agents, which aims to address the scenarios of user intent evolving across turns and must be inferred from context. We interleave search and reasoning across turns in conversations, enabling exploratory and adaptive behaviors learned through decomposed rewards by guiding the model to reason conversational context towards the final answer, effectively capture user intent for search optimization, and respond precisely to various types of queries. The experimental results across four widely used conversational benchmarks demonstrate the effectiveness of our methods by surpassing several existing strong baselines.

Limitations

Despite our comprehensive studies, some potential limitations can be addressed in the future:

Efficiency. While the RL training facilitates the multi-turn capacity of the agentic conversational search model via contextualized reasoning, the additional computational cost is added for both the training and inference phases. Thus, efficiency optimization is required.

Broader Experimental Configuration. Although our approach could generalize to other backbone models, we did not test on more types and sizes of LLMs. Besides, we only leverage the fixed hyperparameters for reward modeling, where the exploration within broader experimental configurations could lead to better performance.

In addition, how to design better intermediate reward to facilitate the utilization of search results for better response generation quality is still an open question for further exploration.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing mixed initiatives and search strategies during conversational search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 16–26.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.

Anthropic. 2023. Introducing claude.

Adithya Bhaskar, Xi Ye, and Danqi Chen. 2025. Language models that think, chat better. *arXiv preprint arXiv:2509.20357*.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1196–1207.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025a. Coral: Benchmarking multi-turn conversational retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1308–1330.

Yiruo Cheng, Hongjin Qian, Fengran Mo, Yongkang Wu, Zhonghua Li, Qi Ye, Ji-Rong Wen, and Zhicheng Dou. 2025b. Evolving graph-based context modeling for multi-turn conversational retrieval-augmented generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen

Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Seungpil Won, Hwanhee Lee, and Kyomin Jung. 2023. Itercqr: Iterative conversational query reformulation without human supervision. *arXiv preprint arXiv:2311.09820*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Instructor: Instructing unsupervised conversational dense retrieval with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6649–6675.

Sungdong Kim and Gangwoo Kim. 2022. Saving dense retriever from shortcut dependency in conversational search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10278–10287. Association for Computational Linguistics.

Yilong Lai, Jialong Wu, Congzhi Zhang, Haowen Sun, and Deyu Zhou. 2024. Adacqr: Enhancing query reformulation for conversational search via sparse and dense retrieval alignment. In *COLING*.

Yilong Lai, Yipin Yang, Jialong Wu, Fengran Mo, Zhenglin Wang, Ting Liang, Jianguo Lin, and Keping Yang. 2025. Crmweaver: Building powerful business agent via agentic rl and shared memories. *arXiv preprint arXiv:2510.25333*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015.

Xiaoqiang Lin, Aritra Ghosh, Bryan Kian Hsiang Low, Anshumali Shrivastava, and Vijai Mohan. 2025. Refrag: Rethinking rag based decoding. *arXiv preprint arXiv:2509.01092*.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. *Advances in Neural Information Processing Systems*, 37:15416–15459.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025a. Chatr1: Reinforcement learning for conversational reasoning and retrieval augmented question answering. *arXiv preprint arXiv:2510.13312*.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025b. Disco meets llms: A unified approach for sparse retrieval and contextual distillation in conversational search. In *SIGIR*.

Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, and Zhicheng Dou. 2024. Chatretriever: Adapting large language models for generalized and robust conversational dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1227–1240.

Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022a. Curriculum contrastive context denoising for few-shot conversational dense retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 176–186.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022b. Convtrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946.

Chuan Meng, Mohammad Aliannejadi, and Maarten de Rijke. 2023. System initiative prediction for multi-turn conversational information seeking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1807–1817.

Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, et al. 2025a. Uniconv: Unifying retrieval and response generation for large language models in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6936–6949.

Fengran Mo, Yifan Gao, Zhuofeng Wu, Xin Liu, Pei Chen, Zheng Li, Zhengyang Wang, Xian Li, Meng Jiang, and Jian-Yun Nie. 2026a. Leveraging historical information to boost retrieval-augmented generation in conversations. *Information Processing & Management*, 63(2):104449.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024a. Chiq: Contextual history enhancement for improving query rewriting in conversational search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2268.

Fengran Mo, Yuchen Hui, Yuxing Tian, Zhaoxuan Tan, Chuan Meng, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2025b. Towards adaptive personalized conversational information retrieval. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2137–2147.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024b. A survey of conversational search. *arXiv preprint arXiv:2410.15576*.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023a. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012.

Fengran Mo, Chuan Meng, Mohammad Aliannejadi, and Jian-Yun Nie. 2025c. Conversational search: From fundamentals to frontiers in the llm era. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023b. Learning to relate to previous turns in conversational search. In *29th ACM SIGKDD Conference On Knowledge Discover and Data Mining (SIGKDD)*.

Fengran Mo, Chen Qu, Kelong Mao, Yihong Wu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024c. Aligning query representation with rewritten query and relevance judgments in conversational search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1700–1710.

Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024d. History-aware conversational dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13366–13378.

Fengran Mo, Zhan Su, Yuchen Hui, Jinghan Zhang, Jia Ao Sun, Zheyuan Liu, Chao Zhang, Tetsuya Sakai, and Jian-Yun Nie. 2026b. Opendecoder: Open large language model decoding to incorporate document quality in rag. *arXiv preprint arXiv:2601.09028*.

Fengran Mo, Jinghan Zhang, Yuchen Hui, Jia Ao Sun, Zhichao Xu, Zhan Su, and Jian-Yun Nie. 2025d. Convmix: A mixed-criteria data augmentation framework for conversational dense retrieval. *arXiv preprint arXiv:2508.04001*.

OpenAI. 2022. Openai: Introducing chatgpt.

Paul Owoicho, Ivan Sekulić, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. *arXiv preprint arXiv:2304.13874*.

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Jinan Xu, Meng Jiang, Jian-Yun Nie, and Kaiyu Huang. 2026. Language-coupled reinforcement learning for multilingual retrieval-augmented generation. *arXiv preprint arXiv:2601.14896*.

Hongjin Qian and Zheng Liu. 2025. Scent of knowledge: Optimizing search-enhanced reasoning with information foraging. *arXiv preprint arXiv:2505.09316*.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Nirmal Roy, Leonardo Ribeiro, Rexhina Blloshmi, and Kevin Small. 2024. Learning when to retrieve, what to rewrite, and how to respond in conversational qa. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10604–10625.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013.

Zhan Su, Fengran Mo, Jinghan Zhang, Yuchen Hui, Jiaao Sun, and Jian-yun Nie. 2025. Parametric retrieval-augmented generation using latent routing of lora adapters. *arXiv preprint arXiv:2511.17044*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024a. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024b. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911.

Zhenduo Wang, Zhichao Xu, Vivek Srikumar, and Qingyao Ai. 2024c. An in-depth investigation of user response simulation for conversational search. In *Proceedings of the ACM Web Conference 2024*, pages 1407–1418.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, and Gaurav Singh Tomar. 2022. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Inscit: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006.

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2301–2305.

Chanwoong Yoon, Gangwoo Kim, Byeongguk Jeon, Sungdong Kim, Yohan Jo, and Jaewoo Kang. 2025. Ask optimal questions: Aligning large language models with retriever's preference in conversation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5899–5921.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456.

Hansi Zeng, Zoey Li, Yifan Gao, Chenwei Zhang, Xiaoman Pan, Tao Yang, Fengran Mo, Jiacheng Lin, Xian Li, and Jingbo Shang. 2026. Synplanresearch-r1: Encouraging tool exploration for deep research with synthetic plans. *arXiv preprint arXiv:2603.07853*.

Jinghan Zhang, Fengran Mo, Tharindu Cyril Weerasooriya, Ruimin Dai, Xiaoyan Han, Yanjie Fu, Dakuo Wang, and Kunpeng Liu. 2026. Starpo: Stability-augmented reinforcement policy optimization. *arXiv preprint arXiv:2604.08905*.

Jinghan Zhang, Fengran Mo, Tharindu Cyril Weerasooriya, Xinyue Ye, Dongjie Wang, Yanjie Fu, and Kunpeng Liu. 2025a. Blind spot navigation in large language model reasoning with thought space explorer. *Preprint*, arXiv:2410.24155.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025b. Entropy-based exploration conduction for multi-step reasoning. *arXiv preprint arXiv:2503.15848*.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025c. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Jintian Zhang, Cheng Peng, Mengshu Sun, Xiang Chen, Lei Liang, Zhiqiang Zhang, Jun Zhou, Huajun Chen, and Ningyu Zhang. 2024a. Onegen: Efficient one-pass unified generation and retrieval for llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4088–4119.

Yongfeng Zhang, Zhiwei Liu, Qingsong Wen, Linsey Pang, Wei Liu, and Philip S Yu. 2024b. Ai agent for information retrieval: Generating and ranking. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5605–5607.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

Changtai Zhu, Siyin Wang, Ruijun Feng, Kai Song, and Xipeng Qiu. 2025a. Convsearch-r1: Enhancing query reformulation for conversational search with reasoning via reinforcement learning. *arXiv preprint arXiv:2505.15776*.

Yaochen Zhu, Harald Steck, Dawen Liang, Yinhan He, Vito Ostuni, Jundong Li, and Nathan Kallus. 2025b. Rank-grpo: Training llm-based conversational recommender systems with reinforcement learning. *arXiv preprint arXiv:2510.20150*.

Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM Transactions on Information Systems*, 41(1):1–25.

Appendix

A Datasets Details

	TopiOCQA	QReCC	CORAL	INSCIT
#Conv.	205	2,775	200	469
#Turns(Qry)	2,514	16,451	2,153	2,767
#Collection	25M	54M	0.2M	49M
#Avg. Qry	12.9	5.3	10.8	5.9
#Min Qry	5	2	8	2
#Max Qry	25	12	19	7
#Avg. Psg	9.0	1.6	2.4	1.6

Table 5: Statistics of the four used datasets.

The statistics of each dataset are presented in Table 5. For the CORAL dataset, we only use the most challenging subset (level-d) with the longest conversations for the evaluation in our experimental scenario. Additionally, we utilize the INSCIT dataset for evaluating mixed-initiative actions, as it is the only dataset annotated with the associated ground-truth information.

B Baseline Details

We provide a more detailed introduction to the following baselines used for comparison:

Supervised fine-tuning (SFT): A simplest training method with next-token prediction loss to generate the answer on the backbone language models.

ChatQA (Liu et al., 2024): A two-stage instruction fine-tuned model for conversational QA and RAG, with synthetic data generation and human-annotated high-quality data.

UniConv (Mo et al., 2025a): A unified framework to improve the seamless consistency between conversational retrieval and its augmented response generation within a single LLM.

REFRAG (Lin et al., 2025): An optimized architecture to compress only a small subset of retrieved documents that are directly related to the query for effective and efficient RAG.

EvoRAG (Cheng et al., 2025b): A novel framework that dynamically maintains an evolving knowledge graph to model the logical relations among user queries, system responses, and the relevant passages across conversational turns.

Search-R1 (Jin et al., 2025): A new and effective paradigm for enhancing input quality of the LLMs by integrating in-context reasoning with dynamic search tool invocation when needed.

ConvSearch-R1 (Zhu et al., 2025a): A conversational query rewriting approach that eliminates dependency on external rewrite supervision based on the paradigm of Search-R1.

ChatR1 (Lupart et al., 2025a): An RL-based reasoning model for Conversational QA, which optimizes multi-turn retrieval and generation end-to-end with dynamic behavior learning.

AgenticLM (Yang et al., 2025): Leveraging the capacity of agentic LLMs to perform explicit reasoning in conversational scenarios.

C Instruction Template and AI Usage

The full instruction template to guide our agentic conversational model based on different backbone LLMs is provided in Table 6. Besides, we only use AI assistants (e.g., ChatGPT and Claude) to test in our experimental scenarios, without the use of coding and writing.

```
You are a helpful assistant. Answer the
question in multi-turn conversations. The
historical context is provided inside
<context> and </context>.
- You must conduct reasoning inside <think>
and </think>first every time you get new
information. After reasoning, if you find you
lack some knowledge, you should first rewrite
the context-dependent question into a
rewritten query by refering the historical
context.
- Then, you can call a search engine by
<search>rewritten query + your answer
</search>, and it will return the top
searched results between <information>and
</information>. You can search as many times
as you want.
- If you find no available answer could be
provided, you can say you cannot find any
information inside <answer>and </answer>,
without detailed illustrations. For example,
<answer>Sorry, I did not find any useful
information. </answer>.
- If you find the question is ambiguous to
answer or want to engage the conversation,
you can generate a clarification question
inside <clarify>and </clarify>, without
detailed illustrations. For example, <clarify>
There are commercial and homemade substitutes
available, which ones would you like to know
about? </clarify>.
- If you find no further external knowledge
needed, you can directly provide the answer
inside <answer>and </answer>, without
detailed illustrations. For example, <answer>
Beijing </answer>.

Context Begin: <context> {Historical Context}
</context>.
Question: {Current User Query}
```

Table 6: Instruction Template for the policy LLM.