ChatCRS: Incorporating External Knowledge and Goal Guidance for LLM-based Conversational Recommender Systems

Anonymous ACL submission

Abstract

This paper aims to efficiently enable large language models (LLMs) to use *external knowledge* and *goal guidance* in conversational recommender system (CRS) tasks. Advanced LLMs (*e.g.*, ChatGPT) are limited in domain-specific CRS tasks for 1) generating grounded responses with recommendation-oriented knowledge, or 2) proactively leading the conversations through different dialogue goals. In this work, we first analyze those limitations through a comprehensive evaluation, showing the necessity of external knowledge and goal guidance which contribute significantly to the recommendation accuracy and language quality. In light of this finding, we propose a novel ChatCRS framework to decompose the complex CRS task into several subtasks through the implementation of 1) a knowledge retrieval agent using a tool-augmented approach to reason over external Knowledge Bases and 2) a goal-planning agent for dialogue goal prediction. Experimental results on two multi-goal CRS datasets reveal that ChatCRS sets new state-of-the-art benchmarks, improving language quality of informativeness by 17% and proactivity by 27%, and achieving a tenfold enhancement in recommendation accuracy[1].

1 Introduction

Conversational recommender system (CRS) integrates conversational and recommendation system (RS) technologies, naturally planning and proactively leading the conversations from non-recommendation goals (e.g., *"chitchat"* or *"question answering"*) to recommendation-related goals (e.g., *"movie recommendation*; Jannach et al., 2021; Liu et al., 2023b). Compared with traditional RS, CRS highlights the multi-round interactions between users and systems using natural language. Besides the **recommendation task** evaluated by the recommendation accuracy as in RS, CRS also focuses on multi-round interactions in **response generation tasks** including asking questions, responding to user utterances or balancing recommendation versus conversation (Li et al., 2023).

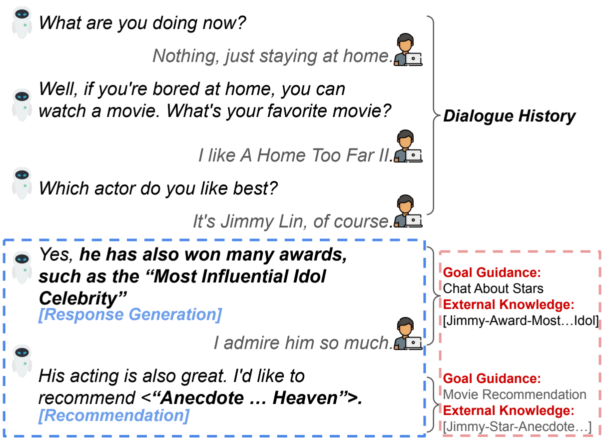


Figure 1: An example of CRS tasks with external knowledge and goal guidance. (Blue: CRS tasks; Red: External Knowledge and Goal Guidance)

Large language models (LLMs; e.g., ChatGPT) that are significantly more proficient in response generation show great potential in CRS applications. However, current research concentrates on evaluating only their recommendation capability (Sanner et al., 2023; Dai et al., 2023). Even though LLMs demonstrate a competitive zero-shot recommendation proficiency, their recommendation performance primarily depends on content-based information (internal knowledge) and exhibits sensitivity towards demographic data (He et al., 2023; Sanner et al., 2023). Specifically, LLMs excel in domains with ample internal knowledge (e.g., English movies). However, in domains with scarce internal knowledge (e.g., Chinese movies[2]), we found through our empirical analysis (§ 3) that their recommendation performance notably dimin-

[1] Our code is publicly available at Anonymous-ChatCRS

[2] The Chinese movie domain encompasses CRS datasets originally sourced from Chinese movie websites, featuring both Chinese and international films.

ishes. Such limitation of LLM-based CRS motivates exploring solutions from prior CRS research to enhance domain coverage and task performance.

Prior work on CRS has employed general language models (LMs; e.g., DialoGPT) as the base architecture, but bridged the gap to domain-specific CRS tasks by incorporating external knowledge and goal guidance (Wang et al., 2021; Liu et al., 2023b). Inspired by this approach, we conduct an empirical analysis on the DuRecDial dataset (Liu et al., 2021) to understand how external inputs[3] can efficiently adapt LLMs in the experimented domain and enhance their performance on both recommendation and response generation tasks.

Our analysis results (§ 3) reveal that despite their strong language abilities, LLMs exhibit notable limitations when directly applied to CRS tasks without external inputs in the Chinese movie domain. For example, lacking domain-specific knowledge (*"Jimmy's Award"*) hinders the generation of pertinent responses, while the absence of explicit goals (*"recommendation"*) leads to unproductive conversational turns (Figure 1). Identifying and mitigating such constraints is crucial for developing effective LLM-based CRS (Li et al., 2023).

Motivated by the empirical evidence that external inputs can significantly boost LLM performance on both CRS tasks, we propose a novel **ChatCRS** framework. It decomposes the overall CRS problem into sub-components handled by specialized agents for knowledge retrieval and goal planning, all managed by a core LLM-based conversational agent. This design enhances the framework's flexibility, allowing it to work with different LLM models without additional fine-tuning while capturing the benefits of external inputs (Figure 2b). Our contributions can be summarised as:

- We present the first comprehensive evaluation of LLMs on both CRS tasks, including response generation and recommendation, and underscore the challenges in LLM-based CRS.
- We propose the ChatCRS framework as the first knowledge-grounded and goal-directed LLM-based CRS using LLMs as conversational agents.
- Experimental findings validate the efficacy and efficiency of ChatCRS in both CRS tasks. Furthermore, our analysis elucidates how external inputs contribute to LLM-based CRS.

[3] In this paper, we limit the scope of external inputs to external knowledge and goal guidance.

2 Related Work

Attribute-based/Conversational approaches in CRS. Existing research in CRS has been categorized into two approaches (Gao et al., 2021; Li et al., 2023): 1) *attribute-based approaches*, where the system and users exchange item attributes without conversation (Zhang et al., 2018; Lei et al., 2020), and 2) *conversational approaches*, where the system interacts users through natural language (Li et al., 2018; Deng et al., 2023; Wang et al., 2023a).

LLM-based CRS. LLMs have shown promise in CRS applications as 1) zero-shot conversational recommenders with item-based (Palma et al., 2023; Dai et al., 2023) or conversational inputs (He et al., 2023; Sanner et al., 2023; Wang et al., 2023b); 2) AI agents controlling pre-trained CRS or LMs for CRS tasks (Feng et al., 2023; Liu et al., 2023a; Huang et al., 2023); and 3) user simulators evaluating interactive CRS systems (Wang et al., 2023c; Zhang and Balog, 2020; Huang et al., 2024). However, there is a lack of prior work integrating external inputs to improve LLM-based CRS models.

Multi-agent and tool-augmented LLMs. LLMs, as conversational agents, can actively pursue specific goals through multi-agent task decomposition and tool augmentation (Wang et al., 2023d). This involves delegating subtasks to specialized agents and invoking external tools like knowledge retrieval, enhancing LLMs' reasoning abilities and knowledge coverage (Yao et al., 2023; Wei et al., 2023; Yang et al., 2023; Jiang et al., 2023).

In our work, we focus on the conversational approach, jointly evaluating CRS on both recommendation and response generation tasks (Wang et al., 2023a; Li et al., 2023; Deng et al., 2023). Unlike existing methods, ChatCRS uniquely combines goal planning and tool-augmented knowledge retrieval agents within a unified framework. This leverages LLMs' innate language and reasoning capabilities without requiring extensive fine-tuning.

3 Preliminary: Empirical Analysis

We consider the CRS scenario where a system $system$ interacts with a user u. Each dialogue contains T conversation turns with user and system utterances, denoted as $C = \{s_j^{system}, s_j^u\}_{j=1}^T$. The target function for CRS is expressed in two parts: given the dialogue history C_j of the past j^{th} turns, it generates 1) the recommendation of item i and 2) the next system response s_{j+1}^{system}. In

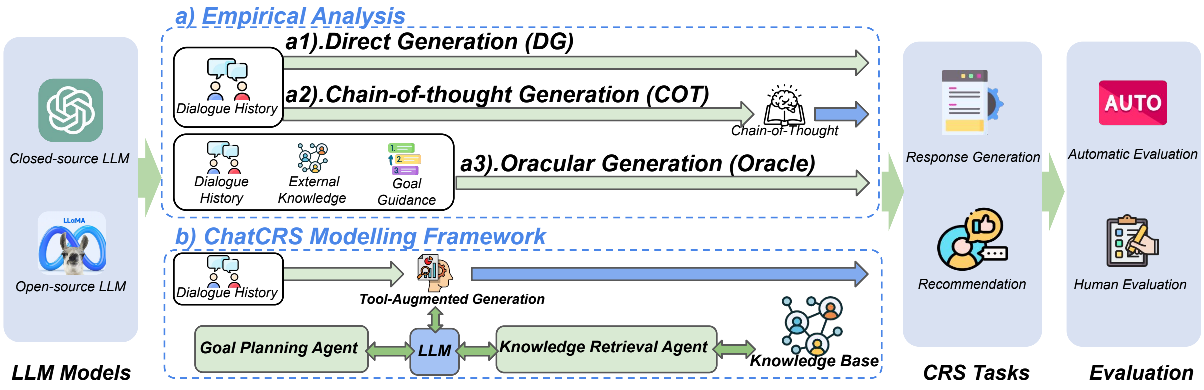


Figure 2: a) Empirical analysis of LLMs in CRS tasks with DG, COT& Oracle; b) System design of ChatCRS framework using LLMs as a conversational agent to control the goal planning and knowledge retrieval agents.

some methods, knowledge K is given as an external input to facilitate both the recommendation and response generation tasks while dialogue goals G only facilitate the response generation task due to the fixed "recommendation" goals in the recommendation task. Given the user's contextual history C_j, $system$ generates recommendation results i and system response s_{j+1}^{system} in Eq. 1.

$$y^{*} = \prod_{j=1}^{T} P_{\theta}\left(i, s_{j+1}^{system} \middle| C_j, K, G\right) \quad (1)$$

3.1 Empirical Analysis Approaches

Building on the advancements of LLMs over general LMs in language generation and reasoning, we explore their inherent response generation and recommendation capabilities, with and without external knowledge or goal guidance. Our analysis comprises three settings, as shown in Figure 2a:

- ***Direct Generation (DG).*** LLMs directly generate system responses and recommendations without any external inputs (Figure 5a).
- ***Chain-of-thought Generation (COT).*** LLMs internally reason their built-in knowledge and goal-planning scheme for both CRS tasks (Figure 5b).
- ***Oracular Generation (Oracle).*** LLMs leverage gold-standard external knowledge and dialogue goals to enhance performance in both CRS tasks, providing an upper bound (Figure 5c).

Additionally, we conduct an ablation study of different knowledge types on both CRS tasks by analyzing 1) factual knowledge, referring to general facts about entities and expressed as single triple (e.g., *[Jiong–Star sign–Taurus]*), and 2) item-based knowledge, related to recommended items and expressed as multiple triples (e.g., *[Cecilia–Star in–<movie 1, movie 2, ..., movie n>]*). Our primary experimental approach utilizes in-context learning (ICL) on the *DuRecDial* dataset (Liu et al., 2021). Figure 5 provides an overview of the ICL prompts, with examples detailed in Appendix A.1 and experiments detailed in § 5. For response generation, we evaluate content preservation ($bleu$-n, $F1$) and diversity ($dist$-n) with knowledge and goal prediction accuracy. For recommendation, we evaluate top-K ranking accuracy ($NDCG@k$, $MRR@k$).

LLM	$Task$	$NDCG@10/50$	$MRR@10/50$
ChatGPT	DG	0.024/0.035	0.018/0.020
	COT-K	0.046/0.063	0.040/0.043
	Oracle-K	**0.617/0.624**	**0.613/0.614**
LLaMA-7b	DG	0.013/0.020	0.010/0.010
	COT-K	0.021/0.029	0.018/0.020
	Oracle-K	**0.386/0.422**	**0.366/0.370**
LLaMA-13b	DG	0.027/0.031	0.024/0.024
	COT-K	0.037/0.040	0.035/0.036
	Oracle-K	**0.724/0.734**	**0.698/0.699**

Table 1: Empirical analysis for recommendation task in DuRecDial dataset (K: Knowledge; **Red**: Best result).

3.2 Empirical Analysis Findings

We summarize our three main findings given the results of the response generation and recommendation tasks shown in Tables 1 and 2.

Finding 1*: The Necessity of External Inputs in LLM-based CRS.* Integrating external inputs significantly enhances performance across all LLM-based CRS tasks (Oracle), underscoring the insufficiency of LLMs alone as effective CRS tools and highlighting the indispensable role of external inputs. Remarkably, the Oracle approach yields over a tenfold improvement in recommendation tasks with only external knowledge compared to DG and COT methods, as the dialogue goal is fixed as "recom-

LLM	*Approach*	*K/G*	*bleu1*	*bleu2*	*bleu*	*dist1*	*dist2*	*F1*	$Acc_{G/K}$
ChatGPT	**DG**		0.448	0.322	0.161	0.330	0.814	0.522	-
	COT	G	0.397	0.294	0.155	0.294	0.779	0.499	**0.587**
		K	0.467	0.323	0.156	0.396	0.836	0.474	**0.095**
	Oracle	G	0.429	0.319	0.172	0.315	0.796	0.519	-
		K	**0.497**	**0.389**	**0.258**	**0.411**	**0.843**	0.488	-
		BOTH	0.428	0.341	0.226	0.307	0.784	**0.525**	-
LLaMA-7b	**DG**		0.417	0.296	0.145	0.389	0.813	0.495	-
	COT	G	0.418	0.293	0.142	0.417	0.827	0.484	0.215
		K	0.333	0.238	0.112	0.320	0.762	0.455	0.026
	Oracle	G	**0.450**	**0.322**	0.164	**0.431**	**0.834**	**0.504**	-
		K	0.359	0.270	0.154	0.328	0.762	0.473	-
		BOTH	0.425	0.320	**0.187**	0.412	0.807	0.492	-
LLaMA-13b	**DG**		0.418	0.303	0.153	0.312	0.786	0.507	-
	COT	G	0.463	0.332	0.172	0.348	0.816	0.528	0.402
		K	0.358	0.260	0.129	0.276	0.755	0.473	0.023
	Oracle	G	**0.494**	**0.361**	0.197	**0.373**	**0.825**	**0.543**	-
		K	0.379	0.296	0.188	0.278	0.754	0.495	-
		BOTH	0.460	0.357	**0.229**	0.350	0.803	0.539	-

Table 2: Empirical analysis for response generation task in DuRecDial dataset (K/G: Knowledge or goal; $Acc_{G/K}$: Accuracy of knowledge or goal predictions; **Red**: Best result for each model; Underline: Best results for all).

mendation" (Table 1). Although utilizing internal knowledge and goal guidance (COT) marginally benefits both tasks, we see in Table 2 for the response generation task that the low accuracy of internal predictions adversely affects performance.

Finding 2*: Improved Internal Knowledge or Goal Planning Capability in Advanced LLMs.* Table 2 reveals that the performance of Chain-of-Thought (COT) by a larger LLM (LLaMA-13b) is comparable to oracular performance of a smaller LLM (LLaMA-7b). This suggests that the intrinsic knowledge and goal-setting capabilities of more sophisticated LLMs can match or exceed the benefits derived from external inputs used by their less advanced counterparts. Nonetheless, such internal knowledge or goal planning schemes are still insufficient for CRS in domain-specific tasks while the integration of more accurate knowledge and goal guidance (Oracle) continues to enhance performance to state-of-the-art (SOTA) outcomes.

Finding 3*: Both factual and item-based knowledge jointly improve LLM performance on domain-specific CRS tasks.* As shown in Table 3, integrating both factual and item-based knowledge yields performance gains for LLMs on both response generation and recommendation tasks. Our analysis suggests that even though a certain type of knowledge may not directly benefit a CRS task (e.g., factual knowledge may not contain the target items for the recommendation task), it can still benefit LLMs by associating unknown entities with their internal knowledge, thereby adapting the universally pre-trained LLMs to task-specific domains more effectively. Consequently, we leverage both types of knowledge jointly in our ChatCRS framework.

Response Generation Task		
Knowledge	*bleu1/2/F1*	*dist1/2*
Both Types	**0.497/0.389/0.488**	**0.411/0.843**
-w/o Factual*	0.407/0.296/0.456	0.273/0.719
-w/o Item-based*	0.427/0.331/0.487	0.277/0.733
Recommendation Task		
Knowledge	*NDCG@10/50*	*MRR@10/50*
Both Types	**0.617/0.624**	**0.613/0.614**
-w/o Factual*	0.272/0.290	0.264/0.267
-w/o Item-based*	0.376/0.389	0.371/0.373

Table 3: Ablation study for ChatGPT with different knowledge types in DuRecDial dataset.

4 ChatCRS

Our ChatCRS modelling framework has three components: 1) a knowledge retrieval agent, 2) a goal planning agent and 3) an LLM-based conversational agent (Figure 2b). Given a complex CRS task, an LLM-based conversational agent first decomposes it into subtasks managed by knowledge retrieval or goal-planning agents. The retrieved knowledge or predicted goal from each agent is incorporated into the ICL prompt to instruct LLMs to generate CRS responses or recommendations.

4.1 Knowledge Retrieval agent

Our analysis reveals that integrating both factual and item-based knowledge can significantly boost the performance of LLM-based CRS. However, knowledge-enhanced approaches for LLM-based CRS present unique challenges that have been relatively unexplored compared to prior *training-based methods* in CRS or *retrieval-augmented (RA) methods* in NLP (Zhang, 2023; Di Palma, 2023).

Training-based methods, which train LMs to memorize or interpret knowledge representations through techniques like graph propagation, have been widely adopted in prior CRS research (Wei et al., 2021; Zhang et al., 2023). However, such approaches are computationally infeasible for LLMs due to their input length constraints and training costs. RA methods, which first collect evidence and then generate responses, face two key limitations in CRS (Manzoor and Jannach, 2021; Gao et al., 2023). First, without a clear query formulation in CRS, RA methods can only approximate results rather than retrieve the exact relevant knowledge (Zhao et al., 2024; Barnett et al., 2024). Especially when multiple similar entries exist in the knowledge base (KB), precisely locating the accurate knowledge for CRS becomes challenging. Second, RA methods retrieve knowledge relevant only to the current dialogue turn, whereas CRS requires planning for potential knowledge needs in future turns, differing from knowledge-based QA systems (Mao et al., 2020; Jiang et al., 2023). For instance, when discussing a celebrity without a clear query (e.g., *"I love Cecilia..."*), the system should anticipate retrieving relevant factual knowledge (e.g., *"birth date"* or *"star sign"*) or item-based knowledge (e.g., *"acting movies"*) for subsequent response generation or recommendations, based on the user's likely interests.

To address this challenge, we employ a relation-based method which allows LLMs to flexibly plan and quickly retrieve relevant "entity–relation–entity" knowledge triples K by traversing along the relations R of mentioned entities E (Moon et al., 2019; Jiang et al., 2023). Firstly, **entities** for each utterance is directly provided by extracting entities in the knowledge bases from the dialogue utterance (Zou et al., 2022). **Relations** that are adjacent to entity E from the KB are then extracted as candidate relations (denoted as $F1$) and LLMs are instructed to plan the knowledge retrieval by selecting the most pertinent relation R^* given the dialogue history C_j. **Knowledge triples** K^* can finally be acquired using entity E and predicted relation R^* (denoted as $F2$). The process is formulated in Figure 3 and demonstrated with an example in Figure 7. Given the dialogue utterance *"I love Cecilia..."* and the extracted entity *[Cecilia]*, the system first extracts all potential relations for *[Cecilia]*, from which the LLM selects the most relevant relation, *[Star in]*. The knowledge retrieval agent then fetches the complete knowledge triple *[Cecilia–Star in–<movie 1, movie 2, ..., movie n>]*.

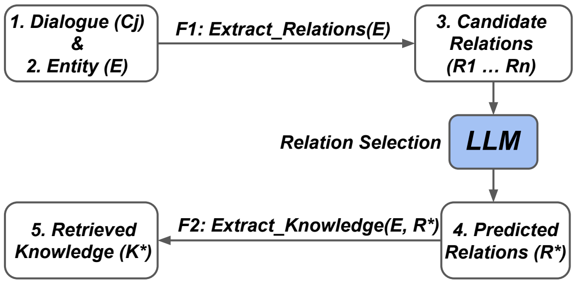


Figure 3: Knowledge retrieval agent in ChatCRS.

When there are multiple entities in one utterance, we perform the knowledge retrieval one by one and in the scenario where there are multiple item-based knowledge triples, we randomly selected a maximum of 50 item-based knowledge due to the limitations of input token length. We implement N-shot ICL to guide LLMs in choosing knowledge relations and we show the detailed ICL prompt and instruction with examples in Table 10 (§ A.2).

4.2 Goal Planning agent

Accurately predicting the dialogue goals is crucial for 1) proactive response generation and 2) balancing recommendations versus conversations in CRS. Utilizing goal annotations for each dialogue utterance from CRS datasets, we leverage an existing language model, adjusting it for goal generation by incorporating a Low-Rank Adapter (LoRA) approach (Hu et al., 2021; Dettmers et al., 2023). This method enables parameter-efficient fine-tuning by adjusting only the rank-decomposition matrices. For each dialogue history C_j^k (j-th turn in dialogue k; $j \in T$, $k \in N$), the LoRA model is trained to generate the dialogue goal G^* for the next utterance using the prompt of dialogue history, optimizing the loss function in Eq 2 with θ representing the trainable parameters of LoRA. The detailed prompt and instructions are shown in Table 11 (§ A.3).

$$L_g = -\sum\nolimits_k^N \sum\nolimits_j^T \log P_\theta \left(G^* | \, C_j^k\right) \quad (2)$$

Model	N-shot	DuRecDial				TG-Redial			
		$bleu1$	$bleu2$	$dist2$	$F1$	$bleu1$	$bleu2$	$dist2$	$F1$
MGCG	*Full*	0.362	0.252	0.081	0.420	NA	NA	NA	NA
MGCG-G	*Full*	0.382	0.274	0.214	0.435	NA	NA	NA	NA
TPNet	*Full*	0.308	0.217	0.093	0.363	NA	NA	NA	NA
UniMIND*	*Full*	0.418	0.328	0.086	0.484	0.291	0.070	0.200	**0.328**
ChatGPT	3	0.448	0.322	**0.814**	0.522	0.262	0.126	**0.987**	0.266
LLaMA	3	0.418	0.303	0.786	0.507	0.205	0.096	0.970	0.247
ChatCRS	3	**0.460**	**0.358**	0.803	**0.540**	**0.300**	**0.180**	**0.987**	0.317

Table 4: Results of response generation task on DuRecDial and TG-Redial datasets. (UniMIND*: Results from the ablation study in the original UniMIND paper.)

Model	N-shot	DuRecDial		TG-Redial	
		$NDCG@10/50$	$MRR@10/50$	$NDCG@10/50$	$MRR@10/50$
SASRec	*Full*	0.369 / 0.413	0.307 / 0.317	0.009 / 0.018	0.005 / 0.007
UniMIND	*Full*	**0.599 / 0.610**	**0.592 / 0.594**	**0.031 / 0.050**	0.024 / 0.028
ChatGPT	3	0.024 / 0.035	0.018 / 0.020	0.001 / 0.003	0.005 / 0.005
LLaMA	3	0.027 / 0.031	0.024 / 0.024	0.001 / 0.006	0.003 / 0.005
ChatCRS	3	**0.549 / 0.553**	**0.543 / 0.543**	**0.031 / 0.033**	**0.082 / 0.083**

Table 5: Results of recommendation task on DuRecDial and TG-Redial datasets.

4.3 LLM-based Conversational Agent

In ChatCRS, the knowledge retrieval and goal-planning agents serve as essential tools for CRS tasks, while LLMs function as tool-augmented conversational agents that utilize these tools to accomplish primary CRS objectives. Upon receiving a new dialogue history C_j, the LLM-based conversational agent employs these tools to determine the dialogue goal G^* and relevant knowledge K^*, which then instruct the generation of either a system response s_{j+1}^{system} or an item recommendation i through prompting scheme, as formulated in Eq 3. The detailed ICL prompt can be found in § A.1.

$$i,\ s_{j+1}^{system} = LLM(\ C_j, K^*, G^*) \tag{3}$$

5 Experiments

5.1 Experimental Setups

Datasets. We conduct the experiments on two multi-goal Chinese CRS benchmark datasets a) DuRecDial (Liu et al., 2021) in English and Chinese, and b) TG-ReDial (Zhou et al., 2020) in Chinese (statistics in Table 12). Both datasets are annotated for goal guidance, while only DuRecDial contains knowledge annotation and an external KB–CNpedia (Zhou et al., 2022) is used for TG-Redial.

Baselines. We compare our model with ChatGPT[4] and LLaMA-7b/13b (Touvron et al., 2023) in few-shot settings. We also compare fully-trained UniMIND (Deng et al., 2023), MGCG-G(Liu et al., 2023b), TPNet(Wang et al., 2023a), MGCG (Liu et al., 2020) and SASRec (Kang and McAuley, 2018), which are previous SOTA CRS and RS models and we summarise each baseline in § A.6.

Automatic Evaluation. For response generation evaluation, we adopt $BLEU$, $F1$ for content preservation and $Dist$ for language diversity. For recommendation evaluation, we adopt $NDCG@k$ and $MRR@K$ to evaluate top K ranking accuracy. For the knowledge retrieval agent, we adopt Accuracy (Acc), Precision (P), Recall (R) and $F1$ to evaluate the accuracy of relation selection (§ A.2).

Human Evaluation. For human evaluation, we randomly sample 100 dialogues from DuRecDial, comparing the responses produced by UniMIND, ChatGPT, LLaMA-13b and ChatCRS. Three annotators are asked to score each generated response with {0: poor, 1: ok, 2: good} in terms of a) general language quality in (Flu)ency and (Coh)erence, and b) CRS-specific language qualities of (Info)rmativeness and (Pro)activity. Details of the process and criterion are discussed in § A.4.

Implementation Details. For both the CRS tasks in Empirical Analysis, we adopt N-shot ICL prompt settings on ChatGPT and LLaMA* (Dong et al., 2022), where N examples from the training data are added to the ICL prompt. In modelling framework, for the goal planning agent, we adopt

[4]OpenAI API: gpt-3.5-turbo-1106

Model	General		CRS-specific		
	Flu	*Coh*	*Info*	*Pro*	*Avg.*
UniMIND	1.87	1.69	1.49	1.32	1.60
ChatGPT	**1.98**	1.80	1.50	1.30	1.65
LLaMA-13b	1.94	1.68	1.21	1.33	1.49
ChatCRS	**1.99**	**1.85**	**1.76**	**1.69**	**1.82**
-w/o K*	**2.00**	**1.87**	1.49 ↓	1.62	1.75
-w/o G*	**1.99**	**1.85**	1.72	1.55 ↓	1.78

Table 6: Human evaluation and ChatCRS ablations for language qualities of (Flu)ency, (Coh)erence, (Info)rmativeness and (Pro)activity on DuRecDial (K^*/G^*: Knowledge retrieval or goal-planning agent).

Model	Knowledge				
	N-shot	Acc	P	R	F1
TPNet	*Full*	NA	NA	NA	0.402
MGCG-G	*Full*	NA	0.460	0.478	0.450
ChatGPT	3	0.095	0.031	0.139	0.015
LLaMA-13b	3	0.023	0.001	0.001	0.001
ChatCRS	3	**0.560**	**0.583**	**0.594**	**0.553**

Table 7: Results for knowledge retrieval on DuRecDial.

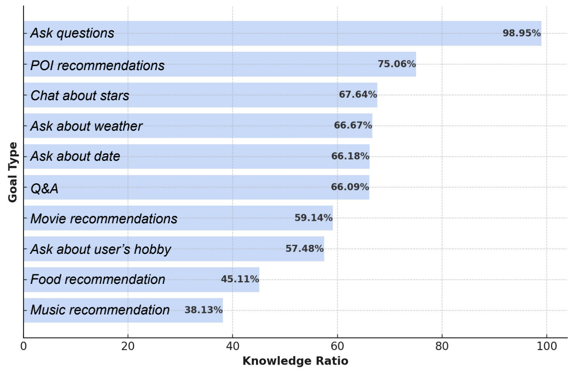


Figure 4: Knowledge ratio for each goal type on DuRecDial. (X-axis: Knowledge Ratio ; Y-axis: Goal type)

QLora as a parameter-efficient way to fine-tune LLaMA-7b (Dettmers et al., 2023). For the knowledge retrieval agent and LLM-based conversational agent, we adopt the same N-shot ICL approach on ChatGPT and LLaMA* (Jiang et al., 2023). Detailed experimental setups are discussed in § A.6.

5.2 Experimental Results

ChatCRS significantly improves LLM-based conversational systems for CRS tasks, outperforming SOTA baselines in response generation in both datasets, enhancing content preservation and language diversity (Table 4). ChatCRS sets new SOTA benchmarks on both datasets using 3-shot ICL prompts incorporating external inputs. In recommendation tasks (Table 5), LLM-based approaches lag behind full-data trained baselines due to insufficient in-domain knowledge. Remarkably, *ChatCRS*, by harnessing external knowledge, achieves a tenfold increase in recommendation accuracy over existing LLM baselines on both datasets with ICL, without full-data fine-tuning.

Human evaluation highlights ChatCRS's enhancement in CRS-specific language quality. Table 6 shows the human evaluation and ablation results. ChatCRS outperforms baseline models in both general and CRS-specific language qualities. While all LLM-based approaches uniformly exceed the general LM baseline (UniMIND) in general language quality, ChatCRS notably enhances coherence through its goal guidance feature, enabling response generation more aligned with the dialogue goal. Significant enhancements in CRS-specific language quality, particularly in informativeness and proactivity, underscore the value of integrating external knowledge and goals. Ablation studies, removing either knowledge retrieval or goal planning agent, demonstrate a decline in scores for informativeness and proactivity respectively, confirming the efficacy of both external inputs for CRS-specific language quality.

5.3 Detailed Discussion

CRS datasets typically contain a huge volume of knowledge. By analyzing dialogues from the DuRecDial datasets, categorized by goal types, we calculated a "Knowledge Ratio" dividing the number of utterances with annotated knowledge $N_{K,G}$ by total number of utterances N_G in each goal type (Eq 4) to measure the necessity of relevant knowledge in CRS task completion. Our analysis, depicted in Figure 4, shows that recommendation tasks rank highly in terms of knowledge necessity, with "POI recommendation" dialogues requiring pertinent knowledge in 75% of cases.

$$\text{Knowledge Ratio (KR)}_G = \frac{N_{K,G}}{N_G} \tag{4}$$

Internal knowledge of LLMs is insufficient for handling domain-specific CRS tasks, which is reaffirmed by the analysis of knowledge retrieval accuracy in Table 7. Contrasting with traditional RS which relies on user data for collaborative recommendations, CRS mainly depends on context/content-based recommendation (He et al., 2023). This shift highlights the limitations of

Case Study 1: Factual Knowledge QA	Case Study 2: Proactive Response Generation
Dialogue history (Dial_ID 39-1): [user]: Do you know Jiong He's zodiac sign ?	**Dialogue history (Dial_ID 89-14):** [user]: I like Cecilia Cheung very much. Her acting is very good. . . . [system]: Yeah, have you seen Cecilia Cheung's One Night in Mongkok? [user]: I've seen it. I don't want to see it again.
System Responses: **[✓] UniMIND**: Of course, Taurus **[✗] ChatGPT**: Jiong He's zodiac sign is Cancer. **[✗] LLaMA**: Jiong He's zodiac sign is Aries. **[✓] ChatCRS**: Yes, I do! Jiong He's zodiac sign is Taurus. ***[Retrieved Knowledge:*** ***('Jiong He', 'zodiac sign ', 'Taurus')]***	**System Responses:** **[✓] UniMIND**: How about <Left Right...Destiny>, a very touching movie. **[✓] ChatGPT**: How about trying her other film, <Lost in Time>, a touching romance that showcases her acting skills? **[✗] LLaMA**: Maybe you can try it again, you may find a different feeling. **[✓] ChatCRS**: If you haven't seen it, you should watch <Left Right...Destiny>, it's Cecilia Cheung's new movie this year. She played the role of a taxi driver, very cool! Have you seen it? ***[Predicted Goal: Movie Recommendation]***

Table 8: Case study for ChatCRS with baselines.

LLMs in harnessing internal knowledge. ChatCRS overcomes these limitations by interfacing LLMs to plan and reason over external KBs through entities and relations. Therefore, it largely improves the recommendation accuracy, outperforming the training-based approach using full data. Given the limitations in LLM-based CRS tasks, (Zhang, 2023; Di Palma, 2023), we anticipate future studies to further explore such approaches in CRS.

Factual knowledge guides the response generation process, mitigating the risks of generating implausible or inconsistent responses. The "Asking questions" goal type which has the highest knowledge ratio, demonstrates the advantage of leveraging external knowledge in answering factual questions like *"the zodiac sign of an Asian celebrity"* (Table 8). Standard LLMs produce responses with fabricated content, but ChatCRS accurately retrieves and integrates external knowledge, ensuring factual and informative responses.

Goal guidance contributes more to the linguistic quality of CRS by managing the dialogue flow. We examine the goal planning proficiency of ChatCRS by showcasing the results of goal predictions of the top 5 goal types in each dataset (Figure 6). DuRecDial dataset shows better balances among recommendation and non-recommendation goals, which exactly aligns with the real-world scenarios (Hayati et al., 2020). However, the TG-Redial dataset contains more recommendation-related goals and multi-goal utterances, making the goal predictions more challenging. The detailed goal planning accuracy is discussed in § A.5.

Dialogue goals guide LLMs towards a proactive conversational recommender. For a clearer understanding, we present a scenario in Table 8 where a CRS seamlessly transitions between "asking questions" and "movie recommendation", illustrating how accurate goal direction boosts interaction relevance and efficacy. Specifically, if a recommendation does not succeed, ChatCRS will adeptly pose further questions to refine subsequent recommendation responses while LLMs may keep outputting wrong recommendations, creating unproductive dialogue turns. This further emphasizes the challenges of conversational approaches in CRS, where the system needs to proactively lead the dialogue from non-recommendation goals to approach the users' interests for certain items or responses (Liu et al., 2023b), and underscores the goal guidance in fostering proactive engagement in CRS.

6 Conclusion

This paper conducts an empirical investigation into the LLM-based CRS for domain-specific applications in the Chinese movie domain, emphasizing the insufficiency of LLMs in domain-specific CRS tasks and the necessity of integrating external knowledge and goal guidance. We introduce ChatCRS, a novel framework that employs a unified agent-based approach to more effectively incorporate these external inputs. Our experimental findings highlight improvements over existing benchmarks, corroborated by both automatic and human evaluation. ChatCRS marks a pivotal advancement in CRS research, fostering a paradigm where complex problems are decomposed into subtasks managed by agents, which maximizes the inherent capabilities of LLMs and their domain-specific adaptability in CRS applications.

Limitations

This research explores the application of few-shot learning and parameter-efficient techniques with large language models (LLMs) for generating responses and making recommendations, circumventing the need for the extensive fine-tuning these models usually require. Due to budget and computational constraints, our study is limited to in-context learning with economically viable, smaller-scale closed-source LLMs like ChatGPT, and open-source models such as LLaMA-7b and -13b.

A significant challenge encountered in this study is the scarcity of datasets with adequate annotations for knowledge and goal-oriented guidance for each dialogue turn. This limitation hampers the development of conversational models capable of effectively understanding and navigating dialogue. It is anticipated that future datasets will overcome this shortfall by providing detailed annotations, thereby greatly improving conversational models' ability to comprehend and steer conversations.

Ethic Concerns

The ethical considerations for our study involving human evaluation (§ 5.1) have been addressed through the attainment of an IRB Exemption for the evaluation components involving human subjects. The datasets utilized in our research are accessible to the public (Liu et al., 2021; Zhou et al., 2020), and the methodology employed for annotation adheres to a double-blind procedure (§ 5.1). Additionally, annotators receive compensation at a rate of $15 per hour, which is reflective of the actual hours worked.

References

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1126–1132, New York, NY, USA. Association for Computing Machinery.

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 Trans. Inf. Syst.*, 41(3).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dario Di Palma. 2023. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1369–1373, New York, NY, USA. Association for Computing Machinery.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

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.

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.

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*.

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

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. *arXiv preprint arXiv:2308.10053*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–an evaluation protocol on conversation recommender systems with system-and user-centric factors. *arXiv preprint arXiv:2404.03304*.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5).

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 304–312, New York, NY, USA. Association for Computing Machinery.

Chuang Li, Hengchang Hu, Yan Zhang, Min-Yen Kan, and Haizhou Li. 2023. A conversation is worth a thousand recommendations: A survey of holistic conversational recommender systems. In *KaRS Workshop at ACM RecSys '23*, Singapore.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Yuanxing Liu, Weinan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li, and Wanxiang Che. 2023a. Conversational recommender system and large language model are made for each other in E-commerce pre-sales dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9587–9605, Singapore. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049. Association for Computational Linguistics.

Zeming Liu, Ding Zhou, Hao Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, Ting Liu, and Hui Xiong. 2023b. Graph-grounded goal planning for conversational recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4923–4939.

Ahtsham Manzoor and Dietmar Jannach. 2021. Generation-based vs retrieval-based conversational recommendation: A user-centric comparison. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 515–520, New York, NY, USA. Association for Computing Machinery.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia, and Eugenio Di Sciascio. 2023. Evaluating chatgpt as a recommender system: A rigorous approach.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language- and item-based preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 890–896, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jian Wang, Dongding Lin, and Wenjie Li. 2023a. A target-driven planning approach for goal-directed dialog systems. *IEEE Transactions on Neural Networks and Learning Systems*.

Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. 2021. Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph. *CoRR*, abs/2110.07477.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023b. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10052–10065, Singapore. Association for Computational Linguistics.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023c. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10052–10065, Singapore. Association for Computational Linguistics.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023d. Recmind: Large language model powered agent for recommendation.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew O. Arnold. 2021. Knowledge enhanced pretrained language models: A compreshensive survey. *CoRR*, abs/2110.08455.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Gangyi Zhang. 2023. User-centric conversational recommendation: Adapting the need of user with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1349–1354, New York, NY, USA. Association for Computing Machinery.

Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1512–1520, New York, NY, USA. Association for Computing Machinery.

Xiaoyu Zhang, Xin Xin, Dongdong Li, Wenxuan Liu, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2023. Variational reasoning over incomplete knowledge graphs for conversational recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 231–239.

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.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey.

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

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1488–1496.

Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving conversational recommender systems via transformer-based sequential modelling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2319–2324. ACM.

A Appendix

A.1 ICL Prompt for Empirical Analysis

In Section § 3, we examine the capabilities of Large Language Models (LLMs) through various empirical analysis methods: Direct Generation (DG), Chain-of-Thought Generation (COT), and Oracular Generation (Oracle). These approaches assess both the intrinsic abilities of LLMs and their performance when augmented with internal or external knowledge or goal directives. The description and testing objective of each empirical analysis methods is shown as follows:

- ***Direct Generation (DG).*** Utilizing dialogue history, DG produces system responses and recommendations to assess the model's inherent capabilities in two CRS tasks (Figure 5a).
- ***Chain-of-thought Generation (COT).*** With dialogue history as input, COT generates knowledge

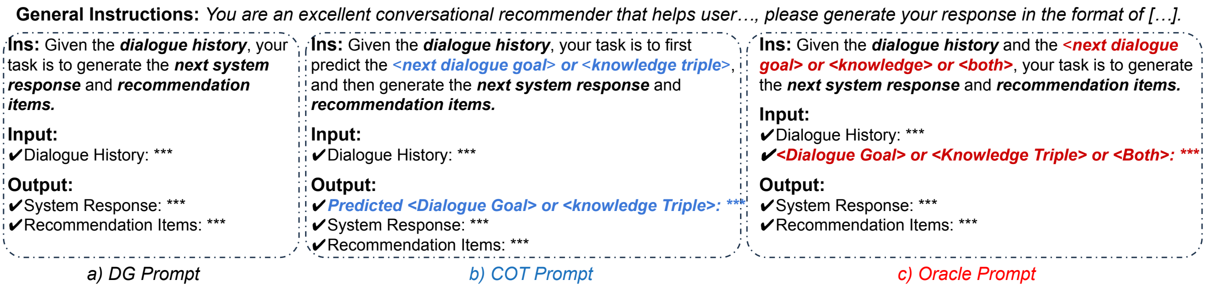
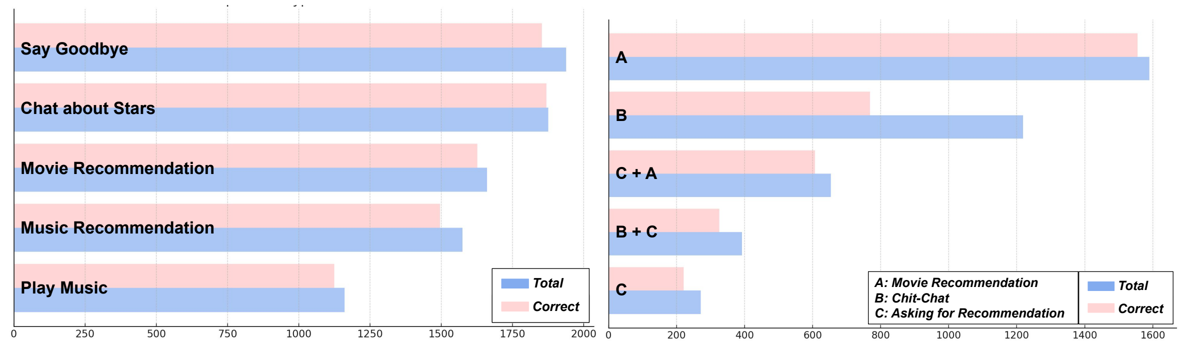


Figure 5: ICL prompt design for empirical analysis, detailed examples are shown in Appendix A.1.



(a) Results of goal predictions for DuRecDial dataset.

(b) Results of goal predictions for TG-Redial datasets.

Figure 6: Results of ChatCRS goal predictions with different goal types.

or goal predictions before generating system responses and recommendations. We evaluate the model's efficacy using only its internal knowledge and goal-setting mechanisms (Figure 5b).

- ***Oracular Generation (Oracle).*** By incorporating dialogue history, and ground truth external knowledge and goal guidance, Oracle generates system responses and recommendations. This yields an upper-bound, potential performance of LLMs in CRS tasks (Figure 5c).

We provide the ICL prompt design in Table 5 and sample instructions within the prompts in Table 13. Furthermore, we detail the actual input-output examples presented in Table 14.

A.2 Detailed Knowledge Retrieval Agent

For the knowledge retrieval agent in ChatCRS, we adopt a 3-shot ICL approach to guide LLMs in planning and selecting the best knowledge for the next utterance by traversing through the relations of the entity, as discussed in § 4.1. For each dialogue history, we first extract the entity in the utterance from the knowledge base and then extract all the candidate relations of the entity from the knowledge base. Given the entity, candidate relations and dialogue history, we use instructions to prompt LLMs in planning and select the relations relevant to the knowledge or topics in the next utterance, as shown in Figure 7. We use 3-shot ICL for our experiment in knowledge retrieval with examples of 3 dialogue histories (C_j) randomly sampled from the training data and each dialogue history may contain up to j-th turn of conversation. The actual examples of the knowledge retrieval prompt are shown in Table 10. Lastly, we retrieve the full knowledge triples using the entity and selected relation. Our knowledge retrieval agent provides a fast way to interface LLMs with external knowledge bases but is limited to one-hop reasoning due to the nature of using a single relation for knowledge retrieval.

For the item-based knowledge, which contains multiple knowledge with the same relation (e.g., *[Cecilia–Star in–<movie 1, movie 2, ..., movie n>]*), we randomly select 50 knowledge triples due to the limit of input token length and only evaluate the correctness of "Entity-Relation" for the item-based knowledge because there is only one ground-truth knowledge for each utterance in DuRecDial dataset (Liu et al., 2021).

A.3 Detailed Goal Planning Agent

Both DuRecDial and TG-Redial datasets have full annotation for the goal types of each utterance. For DuRecDial, each utterance is only related to

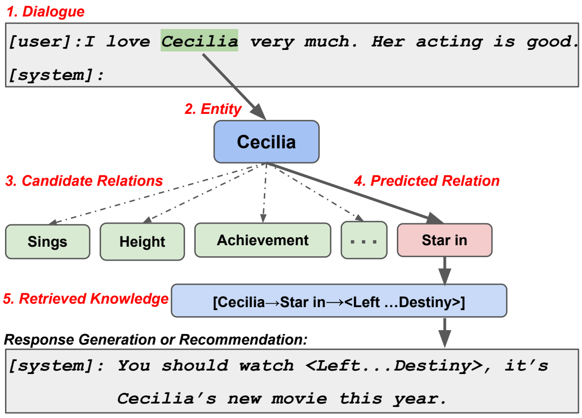


Figure 7: An example of the knowledge retrieval agent.

Model	DuRecDial				TG-RecDial			
	Acc	P	R	F1	Acc	P	R	F1
MGCG	NA	0.76	0.81	0.78	NA	0.75	0.81	0.78
BERT	NA	0.87	0.90	0.89	NA	0.92	0.93	**0.92**
BERT+CNN	NA	0.87	0.92	0.90	NA	**0.93**	**0.94**	**0.92**
UniMIND	NA	0.89	0.94	0.91	NA	0.89	**0.94**	0.91
ChatGPT	0.31	0.05	0.04	0.04	0.38	0.14	0.10	0.10
LLaMA	0.11	0.03	0.02	0.02	0.25	0.06	0.06	0.05
ChatCRS	**0.98**	**0.97**	**0.97**	**0.97**	**0.94**	0.82	0.84	0.81

Table 9: Results of goal planning task.

a single dialogue goal (e.g., 'Asking questions' or 'movie recommendation') while for TG-Redial, one utterance can have multiple dialogue goals (e.g., 'Chit-chat and Asking for recommendation'), which makes it more challenging. Our goal planning uses the dialogue history to prompt the LoRA model in generating the dialogue goals for the next utterance by selecting one or multiple goals from the given goal list. The prompt is "Given the dialogue history C_j, plan the next dialogue goal for the next conversation turns by selecting the dialogue goal G from the <Dialogue Goal List>" and the real example of training samples with a prompt is shown in Table 11. We use the full training data (around 6K and 8K for DuRecDial and TG-ReDial) in each dataset for the fine-tuning LLaMA-7b using LoRA, enhancing parameter efficiency (Dettmers et al., 2023; Deng et al., 2023). The LoRA attention dimension and scaling alpha were set to 16. While the language model was kept frozen, the LoRA layers were optimized using the AdamW. The model was fine-tuned over 5 epochs, with a batch size of 8 and a learning rate of 1 × 10-4. We compare the goal predictions results of ChatCRS with previous LM baselines like BERT (Devlin et al., 2019) and BERT+CNN (Deng et al., 2023), as well as LLM baselines like ChatGPT and LLaMA, as shown in Table 9.

A.4 Human Evaluation

We selected 100 dialogues from the DuRecDial dataset to evaluate the performance of four methodologies: ChatGPT[5], LLaMA-13b[6], UniMIND, and ChatCRS. Each response generated by these methods was assessed by three annotators using a scoring system of 0: bad, 1: ok, 2: good across four metrics: Fluency (F_h), Coherence (C_h), Informativeness (I_h), and Proactivity (P_h). The annotators, fluent in both English and Mandarin, are well-educated research assistants. This human evaluation process received IRB exemption, and the dataset used is publicly accessible. The criteria for evaluation are as follows:

- ***General Language Quality:***
  - ***Fluency:*** It examines whether the responses are articulated in a manner that is both grammatically correct and fluent.
  - ***Coherence:*** This parameter assesses the relevance and logical consistency of the generated responses within the context of the dialogue history.
- ***CRS-specific Language Quality:***
  - ***Informativeness:*** This measure quantifies the depth and breadth of knowledge or information conveyed in the generated responses.
  - ***Proactivity:*** It assesses how effectively the responses anticipate and address the underlying goals or requirements of the conversation.

Human evaluation results and an ablation study, detailed in Table 6, show that ChatCRS delivers state-of-the-art (SOTA) language quality, benefiting significantly from the integration of external knowledge and goal-oriented guidance to enhance informativeness and proactivity.

A.5 Discussion on Goal Predictions

Figure 6 illustrates the five primary goal categories along with their respective predictions across each dataset and Table 9 shows the overall results of goal planning in different models for

[5]OpenAI API: gpt-3.5-turbo
[6]Hugging Face: LLaMA2-13b-hf

♠ **Examples of Single Prompt Design for the Knowledge Retrieval Agent**

General Instruction:
You are an excellent knowledge retriever who helps select the relation of a knowledge triple [entity-relation-entity] from the given candidate relations. Your task is to choose only one relation from the candidate relations mostly related to the conversation and probably to be discussed in the next dialogue turn, given the entity and the dialogue history. Please directly answer the question in the following format: "The relation is XXX.",

Dialogue History:
[user]: Hello, Mr.Chen! How are you doing?
[system]: Hello! Not bad. It's just that there's a lot of pressure from study.
[user]:You should find a way to relax yourself properly, such as jogging, listening to music and so on.
...
[system]:Well, I don't want to watch movies now.
[user]:It's starred by Aaron Kwok, who has won the Hong Kong Film Awards for Best Actor.

Entity: Aaron Kwok

Candidate Relations:
['Intro', 'Achievement', 'Stars', 'Awards', 'Height', 'Star sign', 'Comments', 'Birthplace', 'Sings', 'Birthday']

Output: "The relation is Intro."

♠ **Examples of 3-shot ICL prompt**

Input: (**Words in brackets provide explanations and are omitted in the actual ICL prompt**)
General Instruction: ... *(general instruction for knowledge retrieval agent)*
Dialogue History 1: ... *(dialogue example from* ***training data****)*
Entity 1: ... *(entity in the last utterance of dialogue history 1)*
Candidate Relations 1: ... *(candidate relations of entity 1)*
Output 1: ... *(the ground-truth relation prediction)*

General Instruction: ...*(...)*
Dialogue History 2: ... *(dialogue example from* ***training data****)*
Entity 2: ...*(...)*
Candidate Relations 2: ...*(...)*
Output 2: ...*(...)*

General Instruction: ...*(...)*
Dialogue History 3: ... *(dialogue example from* ***training data****)*
Entity 3: ...*(...)*
Candidate Relations 3: ...*(...)*
Output 3: ...*(...)*

General Instruction: ... *(general instruction for knowledge retrieval agent)*
Dialogue History T: ... *(testing dialogue from* ***testing data****)*
Entity T: ... *(entity in the last utterance of dialogue history T)*
Candidate Relations T: ... *(candidate relations of entity T)*

Output: "The relation is XXX" (the final relation prediction for testing data)

Table 10: Example of prompt design in Knowledge Retrieval Agent.

♠ **Examples of Prompt Design for Goal Planning Agent**

General Instruction: "You are an excellent goal planner and your task is to predict the next goal of the conversation given the dialogue history. For each dialogue, choose one of the goals for the next dialogue utterance from the given goal list: ["Ask about weather", "Food recommendation, ..., "Ask questions"].

Dialogue history
[user]: I like Cecilia Cheung very much. Her acting is very good.
. . .
[system]: Yeah, have you seen Cecilia Cheung's One Night in Mongkok?
[user]: I've seen it. I don't want to see it again.

Output: "The dialogue goal is Movie recommendation".

Table 11: Example of prompt design in Goal Planning Agent.

both datasets. ChatCRS demonstrates high proficiency in predicting overall goals, achieving accuracy rates of 98% and 94% for the DuRecDial and TG-Redial datasets respectively. Within the DuRecDial dataset, ChatCRS shows commendable performance in accurately predicting both non-recommendation goals ("say goodbye" and "chat about stars") and recommendation-specific goals ("movie or music recommendation"), surpassing all comparative baselines. However, in the TG-Redial dataset, characterized by multiple dialogue goals within each utterance, ChatCRS exhibits a slight decline in accuracy for non-recommendation goals (general conversation) compared to recommendation-centric goals, leading to diminished overall accuracy.

A.6 Baselines and Experiment Settings

For the response generation and knowledge retrieval tasks in CRS, we consider the following baselines for comparisons:

- ***MGCG:*** Multi-type GRUs for the encoding of dialogue context, goal or topics and generation of response, focusing only on the response generation task (Liu et al., 2020).
- ***UNIMIND:*** Multi-task training framework for goal and topic predictions, as well as recommendation and response generation, focusing on both CRS tasks (Deng et al., 2023).
- ***MGCG-G:*** GRU-based approach for graph-grounded goal planning and goal-guided response generation, focusing only on the response generation task (Liu et al., 2023b).
- ***TPNet:*** Transformer-based dialogue encoder and graph-based dialogue planner for response generation and goal-planning, focusing only on response generation task (Wang et al., 2023a).

Additionally, we consider the following baselines for recommendation and goal-planning tasks:

- ***SASRec:*** Transformer-based recommendation system for item-based recommendation without conversations (Liu et al., 2020).
- ***BERT:*** BERT-based text-classification task for predicting the goal types given dialogue context (Devlin et al., 2019).
- ***BERT+CNN:*** Deep learning approach that use the representation from MGCG and BERT for next goal predictions (Deng et al., 2023).

Dataset	Statistics		External K&G	
	Dialogues	Items	Knowledge	Goal
DuRecDial	$10k$	$11k$	✓	21
TG-Redial	$10k$	$33k$	✗	8

Table 12: Statistics of datasets

In our Empirical Analysis and Modelling Framework, we implement few-shot learning across various Large Language Models (LLMs) such as ChatGPT[7], LLaMA-7b[8], and LLaMA-13b[9] for tasks related to response generation and recommendation in Conversational Recommender Systems (CRS). This involves employing N-shot In-Context Learning (ICL) prompts, based on Dong et al. (2022), where N training data examples are integrated into the ICL prompts in a consistent format for each task. Specifically, for recommendations, the LLMs are prompted to produce a top-K item ranking list (§ A.1), focusing solely on the knowledge-guided generation because of the fixed dialogue goal of "Recommendations" and we also omit the ablation study of goal type for recommendation task.

For the Modelling Framework's goal planning agent, QLora is utilized to fine-tune LLaMA-7b, enhancing parameter efficiency (Dettmers et al., 2023; Deng et al., 2023). The LoRA attention dimension and scaling alpha were set to 16. While the language model was kept frozen, the LoRA layers were optimized using the AdamW. The model was fine-tuned over 5 epochs, with a batch size of 8 and a learning rate of 1 × 10-4. The knowledge retrieval agent and LLM-based generation unit employ the same N-shot ICL approach as in CRS tasks with ChatGPT and LLaMA-13b (Jiang et al., 2023). Given that TG-Redial (Zhou et al., 2020) comprises only Chinese conversations, a pre-trained Chinese LLaMA model is used for inference[10]. Our experiments, inclusive of LLaMA, UniMIND or ChatGPT, run on a single A100 GPU or via the OpenAI API. The one-time ICL inference duration on DuRecDial (Liu et al., 2021) test data spans 5.5 to 13 hours for LLaMA and ChatGPT, respectively, with the OpenAI API inference cost approximating US$20 for the same dataset. Statistics of two experimented datasets are shown in Table 12.

[7]OpenAI API: gpt-3.5-turbo-1106
[8]Hugging Face: LLaMA2-7b-hf
[9]Hugging Face: LLaMA2-13b-hf
[10]Hugging Face: Chinese-LLaMA2

♠ **Examples of Prompt Design for Empirical Analysis**
General Instruction: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations.
DG Instruction on Response Generation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history, your task is to generate an appropriate system response. Please reply by completing the output template "The system response is []"
DG Instruction on Recommendation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history, your task is to generate appropriate item recommendations. Please reply by completing the output template "The recommendation list is []." Please limit your recommendation to 50 items in a ranking list without any sentences. If you don't know the answer, simply output [] without any explanation.
COT-G Instruction on Response Generation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history, your task is to first plan the next goal of the conversation from the goal list and then generate an appropriate system response. Goal List: ["Ask about weather", "Food recommendation", "POI recommendation", ... , "Say goodbye"]. Please reply by completing the output template "The predicted dialogue goal is [] and the system response is []".
COT-K Instruction on Response Generation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history, your task is to first generate an appropriate knowledge triple and then generate an appropriate system response. If the dialogue doesn't contain knowledge, you can directly output "None". Please reply by completing the output template "The predicted knowledge triples is [] and the system response is []."
COT-K Instruction on Recommendation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history, your task is to first generate an appropriate knowledge triple and then generate appropriate item Recommendations. If the dialogue doesn't contain knowledge, you can directly output "None". Please reply by completing the output template "The predicted knowledge triples is [] and the recommendation list is []". Please limit your recommendation to 50 items in a ranking list without any sentences. If you don't know the answer, simply output [] without any explanation.
Oracle-G Instruction on Response Generation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history and the dialogue goal of the next system response, your task is to first repeat the conversation goal and then generate an appropriate system response. Please reply by completing the output template "The predicted dialogue goal is [] and the system response is []".
Oracle-K Instruction on Response Generation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history and knowledge triple for the next system response, your task is to first repeat the knowledge triple and then generate an appropriate system response. Please reply by completing the output template "The predicted knowledge triples is [] and the system response is []."
Oracle-K Instruction on Recommendation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history and knowledge triple for the next system response, your task is to first repeat the knowledge triple and then generate appropriate item Recommendations. Please reply by completing the output template "The predicted knowledge triples is [] and the recommendation list is []". Please limit your recommendation to 50 items in a ranking list without any sentences. If you don't know the answer, simply output [] without any explanation.
Oracle-BOTH Instruction on Response Generation Task: You are an excellent conversational recommender who helps the user achieve recommendation-related goals through conversations. Given the dialogue history, the conversation goal and knowledge triple for the next system response, your task is to first repeat the conversation goal and knowledge, and then generate appropriate item Recommendations. Please reply by completing the output template "The predicted dialogue goal is [], the predicted knowledge is [] and the system response is []".

Table 13: Example of instruction in prompt design for Empirical Analysis.

♠ **Examples of Input and Output Format in Prompt Design for Empirical Analysis**
Dialogue History: [user]:Hello! Do you know who starred in the movie Flying Dagger? [system]: Yes, of course I know that. It's Jimmy Lin. [user]: OK, thank you. [system]: He is an amazing all-rounder, and he won Chinese Youth Leader in 2014. [user]: He is my favourite star. [system]: **Dialogue Goal:** Movie recommendation **Knowledge:** 'Jimmy Lin', 'Stars', 'To Miss with Love' **Ground-Truth Recommendation:** To Miss with Love **Ground-Truth Response:** Since you like him so much, I wanna recommend to you the movie To Miss with Love, which is starred by him.
DG Input and Output on Response Generation Task: Input: Dialogue History; Output: "The system response is []"
DG Input and Output on Recommendation Task: Input: Dialogue History; Output: "The recommendation list is []."
COT-G Input and Output on Response Generation Task: Input: Dialogue History; Output: "The predicted dialogue goal is [] and the system response is []".
COT-K Input and Output on Response Generation Task: Input: Dialogue History; Output: "The predicted knowledge triple is [] and the system response is []."
COT-K Input and Output on Recommendation Task: Input: Dialogue History; Output: "The predicted knowledge triple is [] and the recommendation list is []".
Oracle-G Input and Output on Response Generation Task: Input: Dialogue History + Dialogue Goal; Output: "The predicted dialogue goal is [] and the system response is []".
Oracle-K Input and Output on Response Generation Task: Input: Dialogue History + Knowledge; Output: "The predicted knowledge triple is [] and the system response is []."
Oracle-K Input and Output on Recommendation Task: Input: Dialogue History + Knowledge; Output: "The predicted knowledge triple is [] and the recommendation list is []".
Oracle-BOTH Input and Output on Response Generation Task: Input: Dialogue History + Dialogue Goal + Knowledge; Output: "The predicted dialogue goal is [], the predicted knowledge is [] and the system response is []".

Table 14: Example of input and output format in prompt design for Empirical Analysis.