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

Chuang Li[12], **Yang Deng**[13], **Hengchang Hu**[1], **Min-Yen Kan**[1], **Haizhou Li**[14]
[1]National University of Singapore
[2]NUS Graduate School for Integrative Sciences and Engineering
[3]Singapore Management University, Singapore
[4]Chinese University of Hong Kong, Shenzhen
{lichuang, hengchanghu}@u.nus.edu
{ydeng, kanmy, haizhou.li}@nus.edu.sg

Abstract

We enable large language models (LLMs) to efficiently use *external knowledge* and *goal guidance* in conversational recommender system (CRS) tasks. LLMs currently achieve limited effectiveness in domain-specific CRS tasks for 1) generating grounded responses with recommendation-oriented knowledge, or 2) proactively leading the conversations through different dialogue goals. We analyze these limitations through a comprehensive evaluation, showing the necessity of external knowledge and goal guidance which contribute significantly to the recommendation accuracy and language quality. This finding leads us to propose the ChatCRS framework, which decomposes the complex task of CRS into sub-tasks 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. By incorporating these inputs, LLMs proactively plan interactions and generate outputs with rich information. Experiments on two multi-goal CRS datasets reveal that ChatCRS sets new state-of-the-art performance, improving language quality of informativeness by 17% and proactivity by 27%, with a tenfold recommendation accuracy enhancement[1].

1 Introduction

Conversational recommender system (CRS) integrates conversational and recommendation system (RS) technologies, naturally planning and proactively leading conversations between non-recommendation (e.g., *"chitchat"* or *"question answering"*) and recommendation goals (e.g., *"movie recommendation"*; Jannach et al., 2021; Liu et al., 2023b). Compared with traditional RS that focus on the **a) recommendation task** , CRS highlights the multi-round interactions between users and systems using natural language as part of the **b) response generation task**. This second task includes

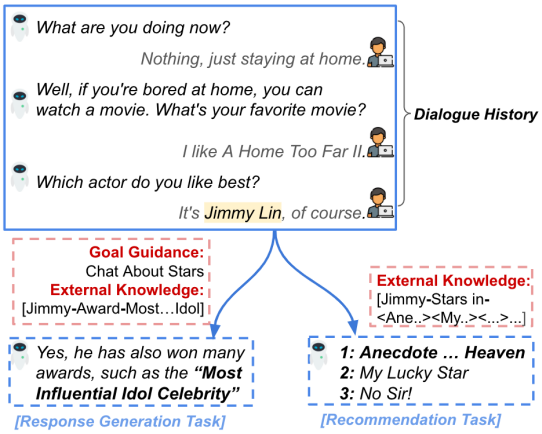

Figure 1: An example of CRS tasks with external knowledge and goal guidance. (Blue: CRS tasks; Red: External Knowledge and Goal Guidance)

asking questions, responding to user utterances, or balancing item recommendation versus conversation. It is evaluated by language quality in natural language generation (Li et al., 2023; Jannach et al., 2021; Deng et al., 2023c; Liu et al., 2023b).

Large language models (LLMs; e.g., ChatGPT) which have demonstrated significant proficiency in response generation show great potential in CRS. However, current research concentrates on evaluating only their recommendation capability (Sanner et al., 2023; Dai et al., 2023). While LLMs demonstrate competitive recommendation proficiency, the performance on the recommendation task primarily depends on its access to 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 movie domain). However, in sparse data domains (e.g., Chinese movies), our empirical analysis (§ 3) validates that the scarce internal knowledge notably diminishes their recommendation performance.

Inspired by prior CRS works using general language models (LMs; e.g., DialoGPT), which incorporates external knowledge and goal guidance for domain-specific CRS tasks (Wang et al., 2021; Liu

[1]Our code is available at github/lichuangnus/ChatCRS.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 295–312

et al., 2023b), we conduct an empirical analysis on the DuRecDial dataset (Liu et al., 2021) to understand how external inputs (specifically, *external knowledge* and *goal guidance*) can efficiently adapt LLMs to a target domain, enhancing performance on both tasks. Despite their strong language abilities, our analysis reveal that LLMs exhibit notable limitations when directly applied to CRS tasks without external inputs. In Figure 1, 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. Identifying and mitigating such constraints is crucial for developing effective LLM-based CRS (Li et al., 2023).

Empirical evidence suggests that external inputs significantly enhance LLMs in both CRS tasks. This prompts us to incorporate external inputs, to plug the knowledge gap in the target domain. However, knowledge-enhanced approaches for LLM-based CRS presents unique challenges that have less studied compared to prior *training-based methods* (Li et al., 2022; Liu et al., 2022; Zhou et al., 2020; Wang et al., 2022) or *retrieval-augmented methods* (Zhang, 2023; Di Palma, 2023; Dao et al., 2024). *Training-based methods*, which train LMs to memorize knowledge representations, have been widely adopted in prior CRS research (Wei et al., 2021; Zhang et al., 2023; Deng et al., 2023c; Liu et al., 2023b). However, such approaches are computationally infeasible for LLMs due to input length constraints and training costs. In contrast, *Retrieval-augmented methods*, which collect evidence and generate responses, face two key limitations in CRS (Manzoor and Jannach, 2021; Gao et al., 2023). First, without a clear query formulation in CRS, retrieval-augmented 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 retrieval of accurate knowledge is challenging. Second, retrieval-augmented methods retrieve knowledge relevant only to the current dialogue turn. In contrast, CRS requires planning for potential knowledge needs in future turns. 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"*) to enrich the conversations, or item-based knowledge (e.g., *"acting movies", "singing songs"*) for subsequent response generation or recommendation, based on the user's likely interests. These challenges highlight the need for more efficient methods of knowledge retrieval without extensive fine-tuning of LLMs.

We thus propose **ChatCRS**, a framework to tackle these challenges. 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. The knowledge retrieval agent uses tool-augmented methods to actively interface and reason over the external knowledge base through entity–relation–entity knowledge paths. The goal planning agent generates the dialogue goals that proactively lead the conversation. This two-part design applies atop any LLM backbone, capturing the benefits of external inputs, without needing costly fine-tuning (Figure 3). Our contributions can be summarised as:

- We comprehensively evaluate LLMs on both CRS tasks and underscore the challenges in LLM-based CRS;
- We propose the ChatCRS framework as the first knowledge-grounded and goal-directed LLM-based CRS with multi-agents;
- Experiments validate the efficacy and efficiency of ChatCRS and our analysis elucidates how external inputs contribute to the model.

2 Related Work

Attribute-based or 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 in entity space without conversations (Zhang et al., 2018; Lei et al., 2020; Deng et al., 2021), and 2) *conversational approaches*, where the system interacts users through natural language generation (Li et al., 2018b; Deng et al., 2023c; Wang et al., 2023a). The majority of CRS work featuring conversational approaches use language models as the backbone for language generation (Li et al., 2018a; Hayati et al., 2020; Liu et al., 2021). However, due to the limitation of universally pre-trained LM in domain-specific CRS applications, they leverage external knowledge or other external guidance like goals or topics to improve performance (Li et al., 2018a; Wang et al., 2022, 2021).

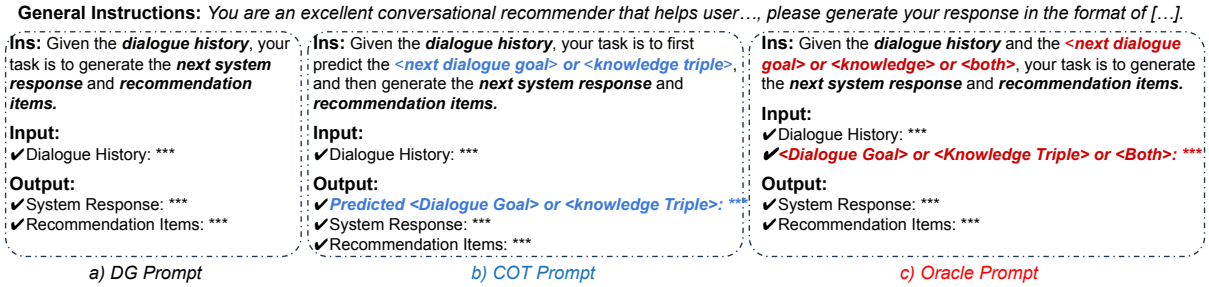

Figure 2: ICL prompt design for empirical analysis.

LLM-based CRS. LLMs have shown promise in CRS applications as 1) zero-shot or few-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; Qin et al., 2024) to generate recommendation results; and as 2) AI agents controlling pre-trained CRS or LMs that distribute CRS subtasks and optimize the ensemble system to synthesize final outputs (Feng et al., 2023; Liu et al., 2023a; Huang et al., 2023); and as 3) user simulators to generate CRS datasets or evaluate 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. As conversational agents, LLMs 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 or function calling, enhancing LLMs' reasoning abilities and knowledge coverage (Yao et al., 2023; Wei et al., 2023; Yang et al., 2023; Jiang et al., 2023; Zhang et al., 2024).

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., 2023c). 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 sys interacts with a user u. Each dialogue contains T conversation turns with user and system utterances, denoted as $Conv = \{s_j^{sys}, s_j^u\}_{j=1}^T$. The target function for CRS is expressed in two parts: given the dialogue history Cov, it generates 1) the recommendation of item i and 2) a next-turn system response s_{j+1}^{sys}. In some methods, knowledge K is given as an external input to facilitate both tasks, while dialogue goals G serve only the response generation task due to the static recommendation goal. The CRS process is formulated as Eq. 1:

$$(i, s_{j+1}^{sys}) = CRS\ (Cov, K, G) \quad (1)$$

where we consider two types of knowledge in our study including: ***1) Factual knowledge***, general facts about entities in a single triple (e.g., *[Jiong–Star sign–Taurus]*) and ***2) Item-based knowledge***, items/entities that are expressed as multiple triples (e.g., *[Cecilia–Star in–*$<i_1><i_2>...<i_n>$*]*).

LLM	*Task*	N@10	N@50	M@10	M@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**
LlaMA7B	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**
LlaMA13B	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 (K: Knowledge; N: NDCG; M: MRR.)

3.1 Empirical Analysis Approaches

Building on the advancements of LLMs, we explore their inherent response generation and recommendation capabilities, with and without external knowledge or goal guidance in three settings using prompt examples in Figure 2:

- ***Direct Generation (DG).*** LLMs directly take the input of dialogue history and prompt to generate system responses or recommendations without any external inputs (Fig. 2a);
- ***Chain-of-thought Generation (COT).*** With the input of dialogue history, LLMs internally reason their built-in knowledge and goal-planning

Approach	G	K	$bleu1$	$bleu2$	$bleu$	$dist1$	$dist2$	$F1$
ChatGPT (DG)			0.448	0.322	0.161	0.330	0.814	0.522
ChatGPT (COT)	✓		0.397	0.294	0.155	0.294	0.779	0.499
		✓	0.467	0.323	0.156	0.396	0.836	0.474
ChatGPT (Oracle)	✓		0.429	0.319	0.172	0.315	0.796	0.519
		✓	**0.497**	**0.389**	**0.258**	**0.411**	**0.843**	0.488
	✓	✓	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
LLaMA-7b (COT)	✓		0.418	0.293	0.142	0.417	0.827	0.484
		✓	0.333	0.238	0.112	0.320	0.762	0.455
LLaMA-7b (Oracle)	✓		**0.450**	**0.322**	0.164	**0.431**	**0.834**	**0.504**
		✓	0.359	0.270	0.154	0.328	0.762	0.473
	✓	✓	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
LLaMA-13b (COT)	✓		0.463	0.332	0.172	0.348	0.816	0.528
		✓	0.358	0.260	0.129	0.276	0.755	0.473
LLaMA-13b (Oracle)	✓		**0.494**	**0.361**	0.197	**0.373**	**0.825**	**0.543**
		✓	0.379	0.296	0.188	0.278	0.754	0.495
	✓	✓	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; **Bold**: Best result for each model; **Bolded Underline**: Best results across all models).

Response Generation Task					
Knowledge	$bleu1$	$bleu2$	$F1$	$dist1$	$dist2$
Both Knowledge	*0.497*	*0.389*	*0.488*	*0.411*	*0.843*
w/o Factual Know.	0.407	0.296	0.456	0.273	0.719
w/o Item Know.	0.427	0.331	0.487	0.277	0.733

Recommendation Task				
Knowledge	$N@10$	$N@50$	$M@10$	$M@50$
Both Knowledge	*0.617*	*0.624*	*0.613*	*0.614*
w/o Factual Know.	0.272	0.290	0.264	0.267
w/o Item Know.	0.376	0.389	0.371	0.373

Table 3: Ablation over knowledge types utilising ChatGPT as the LLM backbone (N/M: NDCG/MRR).

scheme and then generate recommendations or responses for both CRS tasks (Fig. 2b);

- ***Oracular Generation (Oracle).*** With the dialogue history input, LLMs leverage gold-standard external knowledge and dialogue goals to enhance final performance in both CRS tasks, providing an upper-bound performance (Fig. 2c).

Additionally, we conduct an ablation study of different knowledge types on both CRS tasks by analyzing 1) factual knowledge and 2) item-based knowledge. Our primary experimental approach utilizes in-context learning (ICL) on the *DuRecDial* dataset (Liu et al., 2021). The prompt design and experiment settings are detailed in Tables 11 & 12 as well as in § 5. For response generation, we evaluate content preservation ($bleu$-n, $F1$) and diversity ($dist$-n) for language quality. For recommendation, we evaluate top-K ranking accuracy ($NDCG@k, MRR@k$).

3.2 Empirical Analysis Findings

We summarize our three main findings given the results of the empirical analysis and ablation study shown in Tables 1, 2 and 3.

Finding 1*: The Necessity of External Inputs in LLM-based CRS.* For both CRS tasks (Tables 1 & 2), integrating external inputs (Oracle) significantly enhances performance across all LLM baselines, underscoring the insufficiency of LLMs alone in CRS tasks, highlighting the indispensable role of external inputs. Remarkably for the *recommendation* task, the Oracle approach with external knowledge yields over a tenfold improvement compared to DG and COT (Table 1). Although utilizing internal knowledge and goal guidance (COT) marginally benefits LLMs, they are still insufficient for both CRS tasks.

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

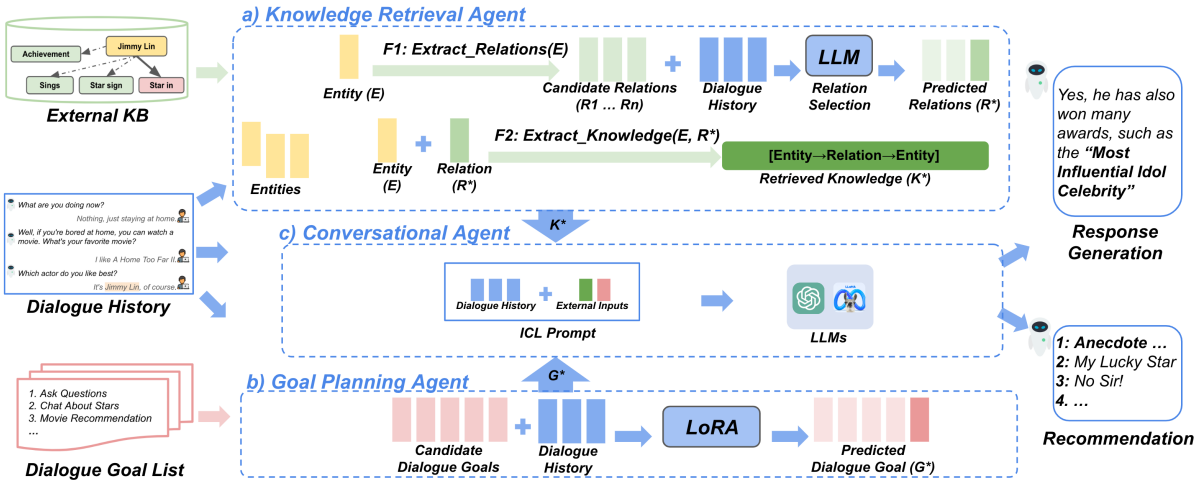

Figure 3: Overall ChatCRS system design including a) Knowledge retrieval agent that interfaces and reasons over external KB; b) Goal planning agent and c) Conversational agent generate final results for both CRS tasks.

advanced counterparts. Nonetheless, such internal knowledge or goal planning schemes are still insufficient for CRS in domain-specific tasks because the integration of more accurate external 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 CRS 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 LLMs to target domains more effectively. Consequently, we jointly leverage both types of knowledge in designing our ChatCRS framework.

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 3). 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

We employ a path-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). Our methods enable the interface between LLMs and external KB through two functions named $F1$ and $F2$, where $F1$ extracts all the neighbouring relations of a given entity and $F2$ extracts the knowledge triples in the KB given the entity and selected relations (Jiang et al., 2023). Firstly, **entities** for each utterance is directly provided by extracting entities in the KB 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 and potential 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 3a. Given the extracted entity *[Jimmy]*, the system first employs $F1$ to extract all candidate relations for *[Jimmy]*, from which the LLM selects the most relevant relation $R*$ based on dialogue history. Subsequently, the knowledge retrieval agents will apply $F2$ to fetch the complete item-based knowledge triples (e.g., *[Jimmy–Stars in–<movie 1, movie 2, ..., movie n>]*) as the final result. We perform the knowledge retrieval individually when there are multiple entities in one utterance. In the scenario

where there are multiple item-based knowledge triples, we randomly selected K item-based knowledge due to input token length limitations. We implement N-shot ICL to guide LLMs in predicting knowledge relations relevant to the context (We show the detailed ICL prompts in Table 13).

4.2 Goal Planning Agent

Accurately predicting the dialogue goals is crucial for 1) proactive response generation (Deng et al., 2023b,a) and 2) balancing recommendations versus conversations in CRS. For example, the conversation in Figure 1 follows the dialogue goals of *"greeting"→"ask question"→"chat about stars"→"recommendation"*, which proactively lead the conversation and gather the information for the final recommendation outputs. Utilizing goal annotations for each dialogue utterance from CRS datasets, we leverage an LLM, adjusting it for goal generation by incorporating a Low-Rank Adapter (Hu et al., 2021; Dettmers et al., 2023). LoRA 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 instruction fine-tuned to generate the dialogue goal G^* for the next utterance using the prompt of dialogue history, optimizing the loss function in Eq. 2 where θ represents LoRA's trainable parameters (Table 14 shows the actual prompt).

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

4.3 LLM-based Conversational Agent

In ChatCRS, the knowledge retrieval and goal-planning agents serve as two intermediate tools for CRS tasks, while LLMs function as controlling agents that utilize these tools independently or jointly to accomplish primary CRS objectives. Upon receiving a new dialogue history C_j, the LLM-based conversational agent employs these agents to determine the dialogue goal G^* and relevant knowledge K^*, which then instruct the generation of either system responses s_{j+1}^{system} or item recommendation i through ICL prompting scheme, formulated in Eq 3 and shown in Figure 2c.

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

Notably, we have designed the system in a modular way, allowing each agent to function independently which enables the integration of new LLMs, even with personalized agents to complete the task.

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

Table 4: Dataset statistics.

5 Experiments

5.1 Setup

Datasets. We conduct the experiments on two multi-goal and human-annotated CRS benchmark datasets: a) DuRecDial (Liu et al., 2021) which collects knowledge and goal-guided CRS dialogues in English and Chinese, and b) TG-ReDial (Zhou et al., 2020) which collects topic-guided dialogue in Chinese (statistics in Table 4). Both datasets are fully annotated for goal guidance for each dialogue turn, while only DuRecDial contains knowledge annotation; thus, an external KB–CN_DBpedia (Zhou et al., 2022) is used for TG-Redial.

Baselines. To validate the efficacy of ChatCRS, we select both LLM-based and training-based baselines. **ChatGPT**[2] and **LLaMA 2-13b** (Touvron et al., 2023) in few-shot settings are adopted as the LLM baselines. For the training-based baselines, except for **UniMIND** (Deng et al., 2023c), all the other baselines only focus on single CRS tasks. For the *response generation* task, **MGCG** (Liu et al., 2020), **MGCG-G** (Liu et al., 2023b), **TPNet** (Wang et al., 2023a) are adopted. For the *recommendation* task, **GRU4Rec** (Liu et al., 2016) and **SASRec** (Kang and McAuley, 2018) are adopted. Specifically, **UniMIND** (Deng et al., 2023c) are adopted for both CRS tasks. A detailed description of the baselines is given in § A.1.

Automatic & Human Evaluation. For response generation evaluation, we adopt $BLEU$, $F1$ and $Dist$. For recommendation evaluation, we use $NDCG@k$ and $MRR@K$. For the knowledge and goal agent, we adopt Accuracy (Acc), Precision (P), Recall (R) and $F1$ to evaluate the accuracy. For human evaluation, we randomly sampled 100 dialogues from DuRecDial and evaluate 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 (Deng et al., 2023c); details in § A.2.

Implementation Details. For a fair comparison

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

Model	N-shot	DuRecDial		TG-Redial	
		$NDCG@10/50$	$MRR@10/50$	$NDCG@10/50$	$MRR@10/50$
GRU4Rec	$Full$	0.219 / 0.273	0.171 / 0.183	0.003 / 0.006	0.001 / 0.002
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-13b	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.

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-13b	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 6: Results of response generation task on DuRecDial and TG-Redial datasets.

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 7: 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).

with the LLM baselines, we adopt the same **ChatGPT**[3] and **LLaMA 2-13b**[4] (Touvron et al., 2023) models in few-shot settings using an N-shot ICL prompt (Dong et al., 2022; Sanner et al., 2023). We set ChatGPT's temperature to 0 to yield replicable output given the same input. For the goal planning agent, we adopt QLora as a parameter-efficient way to structure tune a smaller version of LLaMA 2-7b with a batch size of 8 and a learning rate of $1x10^{-4}$ using Adam. Details are given in § A.2.

5.2 Experimental Results (RQ1)

ChatCRS significantly improves LLM-based conversational systems for both CRS tasks as a holistic system. For the recommendation task (Table 5), despite the limitation of the training process and domain-specific information related to the item space, ChatCRS using few-shot methods shows comparable results with fully-finetuned baselines like UniMIND by leveraging external knowledge. Compared to LLM baselines in few-shot settings, ChatCRS significantly increased recommendation accuracy, specifically a tenfold improvement on both the DuRecDial and TG-Redial datasets, which demonstrates the huge potential of applying knowledge in enhancing the LLM-based CRS.

For the response generation task (Table 6), ChatCRS outperformed existing fully-finetuned baselines in fluency and language diversity. In terms of content preservation, as demonstrated by the $F1$ score, baseline models fine-tuned on the training data performed better due to their familiarity with the language style in the original dataset. However, automatic evaluation methods, which compare generated responses with ground-truth responses, do not fully capture language quality. Therefore, we present human evaluation results in Table 7. The human evaluation confirmed that LLMs generally excel in language proficiency and coherence compared to previous smaller LMs, aligning with the results in Table 6. For CRS-specific language quality, such as informativeness and proactivity, ChatCRS showed significant improvement over all baselines, highlighting the importance of incorporating external inputs.

Furthermore, a holistic CRS requires both response generation and recommendation capabilities. However, among previous fine-tuned baselines, only UniMIND can perform both tasks, while others are limited to a single task (e.g., MGCG & TPNet for response generation, and GRU4Rec & SASRec for item recommendation), limiting their application in real-world scenarios. Building on

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

[4]https://huggingface.co/models?search=llama2

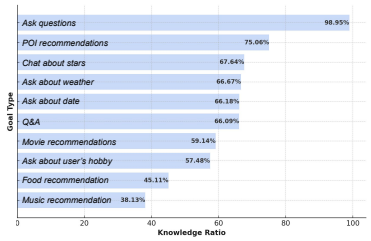

Figure 4: Knowledge ratio for each goal type on DuRecDial dataset.

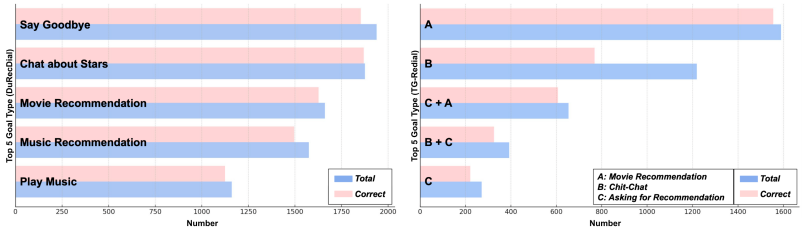

Figure 5: Results of ChatCRS goal predictions with different goal types on DuRecDial (left) and TG-Redial (right) datasets.

LLMs, our methods leverage the language proficiency and reasoning capability of LLMs, integrating external knowledge bases and goal-guiding mechanisms for enhanced performance. Additionally, our methods are backbone-agnostic and can be applied to any LLM.

5.3 Ablation Study (RQ2)

Human evaluation highlights ChatCRS's enhancement in CRS-specific language quality. Table 7 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 responses more aligned with the dialogue goal. Significant enhancements in CRS-specific language quality, particularly in informativeness (quantifies depth of knowledge) and proactivity (assesses how effective the responses anticipate the underlying goals), underscore the value of integrating external knowledge and goals. Ablation studies, removing either agent, demonstrate a decline in scores for informativeness and proactivity respectively, confirming the efficacy of both external inputs for CRS-specific language quality.

5.4 Detailed Analysis (RQ3)

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. For example, for the dialogue goal of "Asking questions", 98% of the utterances require a necessary knowledge to facilitate the response generation. Our analysis, depicted in Figure 4, also shows that all recommendation-related goals fall in the top 10 knowledge ratios, with "POI recommendation" ranking the highest in terms of knowledge necessity, requiring pertinent knowledge in 75% of cases to make a proper recommendation to the users.

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

However, the internal knowledge of LLMs is insufficient for handling domain-specific CRS tasks, which is reaffirmed by the analysis of knowledge retrieval accuracy in Table 8. Contrasting with traditional RS which relies on user data for collaborative recommendations, CRS mainly depends on context/content-based recommendation, which recommends based on the description of users' needs or target items (He et al., 2023). This shift highlights the limitations of 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 of existing knowledge-enhanced methods in LLM-based CRS tasks, (Zhang, 2023; Di Palma, 2023),

Model	Knowledge Retrieval (DuRecDial)				
	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 8: Results for knowledge retrieval agent.

Model	Goal Planning					
	DuRecDial			TG-RecDial		
	P	R	F1	P	R	F1
MGCG	0.76	0.81	0.78	0.75	0.81	0.78
UniMIND	0.89	0.94	0.91	**0.89**	**0.94**	**0.91**
ChatGPT	0.05	0.04	0.04	0.14	0.10	0.10
LLaMA-13b	0.03	0.02	0.02	0.06	0.06	0.05
ChatCRS	**0.97**	**0.97**	**0.97**	0.82	0.84	0.81

Table 9: Results for goal planning agent.

Case Study 1: Factual Knowledge Answer
Dialogue history (Dial_ID 39-1): ***[Target Factual Knowledge:*** ***('Jiong He'–'Zodiac sign'–'Taurus')]***
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.
Case Study 2: Proactive Response Generation
Dialogue history (Dial_ID 89-14): ***[Target Dialogue Goal: Movie Recommendation]***
System Responses: **[✓] UniMIND**: How about <Left Right...Destiny>, 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 for 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?

Table 10: Case study for ChatCRS with baselines.

we anticipate future studies to further explore such approaches in CRS.

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 overall goal prediction accuracy in Table 9 and the detailed breakdown of the top 5 goal types in each dataset in Figure 5. ChatCRS show SOTA performance in the goal prediction in the DuRecDial dataset because it contains a better-balanced distribution 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 compared with the "Chit-Chat" goals. Additionally, it contains more multi-goal utterances (e.g., "Chit-Chat and Asking for Recommendation"), making the goal predictions more challenging.

5.5 Case Study (RQ3)

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 10). Standard LLMs produce responses with fabricated content, but ChatCRS accurately retrieves and integrates external knowledge, ensuring factual and informative responses.

Dialogue goals guide LLMs towards a proactive conversational recommender. For a clearer understanding, we present a scenario in Table 10 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 work presents ChatCRS, a novel framework that enhances LLM-based conversational recommender systems (CRS) by integrating external knowledge retrieval and goal-guided planning. Our empirical analysis highlights the limitations of LLMs in domain-specific CRS tasks and demonstrates the necessity of knowledge-grounded and goal-directed guidance. By leveraging a multi-agent architecture, ChatCRS efficiently decomposes CRS tasks, significantly improving both recommendation accuracy and conversational fluency. The framework provides a scalable, model-agnostic solution, reducing reliance on expensive fine-tuning while maintaining adaptability across domains. Future work can explore more advanced planning mechanisms and self-improving retrieval strategies to further enhance CRS performance.

Limitations

Our 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. We anticipate that future datasets will address this issue by providing detailed annotations.

Ethical Considerations

Our study involved a human evaluation (§ 5.1) protocol, which have been vetted by our Institutional Review Board through its IRB Exemption process. 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 a fair wage for such part-time work under our local jurisdiction.

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.

Huy Dao, Yang Deng, Dung D. Le, and Lizi Liao. 2024. Broadening the view: Demonstration-augmented prompt learning for conversational recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 785–795. ACM.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pages 6583–6591. ijcai.org.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1431–1441. ACM.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023b. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023c. 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 Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018a. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9748–9758, Red Hook, NY, USA. Curran Associates Inc.

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

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Qiang Liu, Shu Wu, Diyi Wang, Zhaokang Li, and Liang Wang. 2016. Context-aware sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1053–1058. IEEE.

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. 2022. Graph-grounded goal planning for conversational recommendation. pages 1–1. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

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.

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.

Peixin Qin, Chen Huang, Yang Deng, Wenqiang Lei, and Tat-Seng Chua. 2024. Beyond persuasion: Towards conversational recommender system with credible explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4264–4282. Association for Computational Linguistics.

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.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1929–1937, New York, NY, USA. Association for Computing Machinery.

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.

An Zhang, Yang Deng, Yankai Lin, Xu Chen, Ji-Rong Wen, and Tat-Seng Chua. 2024. Large language model powered agents for information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 2989–2992. ACM.

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 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., 2023c).
- ***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., 2023c).

In our Empirical Analysis and Modelling Framework, we implement few-shot learning across various Large Language Models (LLMs) such as ChatGPT[5], LLaMA-7b[6], and LLaMA-13b[7] 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, 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., 2023c). 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[8]. 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

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

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 4.

A.2 Human Evaluation

We selected 100 dialogues from the DuRecDial dataset to evaluate the performance of four methodologies: ChatGPT[9], LLaMA-13b[10], 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 7, 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.

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

♠ 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 11: 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 12: Example of input and output format in prompt design for Empirical Analysis.

♠ 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 13: 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 14: Example of prompt design in Goal Planning Agent.