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 recommendation system (CRS) tasks. Advanced LLMs (*e.g.*, ChatGPT) are limited in CRS tasks for 1) generating grounded responses with recommendation-oriented knowledge, or 2) proactively guiding users through different dialogue goals. In this work, we first analyze those limitations through a comprehensive evaluation to assess LLMs' intrinsic capabilities, showing the necessity of incorporating 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 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. Experimental results on two 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 ten-fold enhancement in recommendation accuracy over LLM-based CRS[1].

1 Introduction

Conversational Recommender Systems (CRS) integrate conversational and recommendation system (RS) technologies, facilitating users in achieving recommendation-related goals through conversations (Jannach et al., 2021). In contrast to a traditional RS which is evaluated on single recommendations, CRS focuses on multi-round interaction tasks such as 1) control of the dialogue flow and goals (*goal planning*), 2) retrieving knowledge from knowledge resources (*knowledge retrieval*), 3) response generation (*response generation*) and 4) item recommendation (*recommendation*), as parts of a holistic system (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)

In existing CRS research, the prevalent methodology employs general language models (LMs; e.g., DialoGPT) as the foundational architecture for conversational tasks (Zhou et al., 2020a; Deng et al., 2023b; Wang et al., 2022). This approach, however, overlooks the domain-specific nature of CRS tasks (e.g., "movie recommendation" or "chatting about movie stars"), resulting in a notable mismatch. The incorporation of domain-specific knowledge or goal-oriented guidance is essential to bridge this gap (Wang et al., 2021; Zhang et al., 2021). For instance in Figure 1, lacking specific knowledge like *"Jimmy's Award"* limits the CRS's ability to provide pertinent recommendations and lacking dialogue goals like *"Recommendation"* perpetuating discussions irrelevant to the dialogue's objective, which adversely affects the language's informativeness and proactivity (Deng et al., 2023b).

The emergence of large language models that are significantly more proficient in response generation has reduced the reliance on supplementary knowledge or manual intervention (for ease of reference,

[1]Our code is publicly available at Anonymous4ChatCRS

we term such models as Large LMs (LLMs); e.g., ChatGPT). This development leads to a natural inquiry: *(RQ1) Can LLMs independently function as effective conversational recommenders?* Our analysis (§3) conclusively answers ***"No"***. Despite being trained on extensive datasets, LLMs are primarily tailored for broad applications, facing challenges in proactively guiding users towards recommendation-specific objectives or in retaining detailed external knowledge, such as *the awards for certain movie stars* (He et al., 2023; Deng et al., 2023a).

Recognizing the inherent constraints of LLMs before integrating external knowledge or goal guidance is essential for crafting effective CRS systems (Li et al., 2023). This leads to two additional RQs: *(RQ2) To what extent is external knowledge and goal guidance necessary for successful LLM-based CRS?* and *(RQ3) What are efficient methods to integrate external knowledge and goal guidance into LLM-based CRS?* To address these inquiries, our study first evaluates the baseline capabilities of LLMs in key CRS tasks: response generation and recommendation, for both open- and closed-source LLMs (*cf. RQ1*). Subsequently, we continue to examine the performance of LLMs in CRS tasks with added external knowledge or goal guidance, aiming for an **empirical analysis** *(RQ2)*, as depicted in Figure 2a. Our results highlight LLMs' limitations on these tasks due to the absence of goal guidance and external knowledge, and its enhancement potential through such integration. Leveraging these insights, we introduce a novel **ChatCRS modelling framework** that decomposes CRS tasks into manageable sub-tasks. These sub-tasks are delegated to specialized agents for goal planning or knowledge retrieval, all managed by an LLM-based conversational agent. This arrangement ensures the framework's adaptability across different LLMs without needing model fine-tuning (Figure 2b, *cf. RQ3*). Our contributions can be summarised as:

- We present the first comprehensive evaluation of LLMs on CRS tasks as a holistic system, including response generation, recommendation, goal-planning, and knowledge retrieval. Our empirical analysis underscores the challenges that LLMs face in executing CRS tasks.
- Leveraging these insights, we develop the ChatCRS framework that decomposes the CRS task into three distinct sub-tasks, unifying goal planning and knowledge retrieval into LLM-based CRS with two specialized small agents.
- Experimental findings validate the efficacy and efficiency of ChatCRS in all CRS tasks, establishing a new benchmark for state-of-the-art performance. Furthermore, our analysis elucidates how external inputs contribute to LLM-based CRS.

2 Related Work

Conversational approaches in CRS. Existing research in CRS has been categorised into two different approaches (Gao et al., 2021; Li et al., 2023), named *attribute-based* and *conversational* approaches. In attribute-based approaches, the system aims to improve the recommendation accuracy by exchanging item attribute information with users and there is no conversation involved during the interaction (Lei et al., 2020; Zhang et al., 2018). In conversational approaches, the system interacts with users through real conversations and guides users through the recommendation-related goals (Li et al., 2018; Chen et al., 2019). CRS in conversational approaches mostly adopts language models (LMs) for fundamental dialogue operations and subsequent studies incorporate external knowledge or goal guidance to enhance their performance but they fail to analyse the inherent capability of LMs with or without external knowledge or goal guidance (Deng et al., 2023b; Wang et al., 2022). In addition, those methods also require full fine-tuning to integrate the external knowledge or goal guidance for the final generation.

Multi-agent and tool-augmented LLM. The advent of LLMs has transformed traditional LMs into conversational agents capable of actively pursuing specific conversational goals rather than just generating replies (Wang et al., 2023). This is achieved by decomposing complex tasks into manageable subtasks handled by specialized agents and invoking additional tools (i.e., tool-augmented generation), such as KB retrieval that accesses external knowledge bases (KBs) (Yao et al., 2023; Wei et al., 2023; Yang et al., 2023; Jiang et al., 2023). This approach enhances LLMs' reasoning capabilities and their ability to engage with a broader KB.

In contrast to existing methodologies, ChatCRS distinguishes itself by integrating goal-planning and tool-augmented knowledge retrieval agents in a unified approach. This framework leverages the inherent abilities of LLMs in language modelling and reasoning, without the necessity for comprehensive fine-tuning.

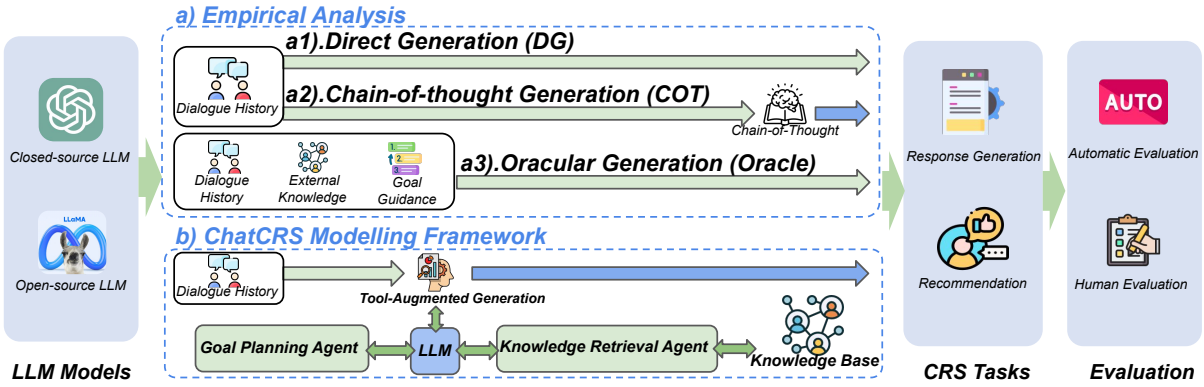


Figure 2: a) Empirical analysis of LLMs in CRS tasks with DG, COT& Oracle; b) System design of ChatCRS framework using LLMs as a conversational agent to control the goal planning and knowledge retrieval agents.

General Instructions: *You are an excellent conversational recommender that helps user..., please generate your response in the format of [...].*

Ins: Given the ***dialogue history***, your task is to generate the ***next system response*** and ***recommendation items.***

Input:
✔Dialogue History: ***

Output:
✔System Response: ***
✔Recommendation Items: ***

a) DG Prompt

Ins: Given the ***dialogue history***, your task is to first predict the *<next dialogue goal> or <knowledge triple>*, and then generate the ***next system response*** and ***recommendation items.***

Input:
✔Dialogue History: ***

Output:
✔*Predicted <Dialogue Goal> or <knowledge Triple>: ****
✔System Response: ***
✔Recommendation Items: ***

b) COT Prompt

Ins: Given the ***dialogue history*** and the ***<next dialogue goal> or <knowledge> or <both>***, your task is to generate the ***next system response*** and ***recommendation items.***

Input:
✔Dialogue History: ***
✔***<Dialogue Goal> or <Knowledge Triple> or <Both>: ****

Output:
✔System Response: ***
✔Recommendation Items: ***

c) Oracle Prompt

Figure 3: ICL prompt design for empirical analysis, detailed examples are shown in Appendix A.1.

3 Preliminary: Empirical Analysis

Without loss of generality, we restrict our consideration to the scenario where a system (denoted by $system$) interacts with a user u. Each dialogue contains T turns of conversations, denoted as $C = \{s_j^{system}, s_j^u\}_{j=1}^T$, where each turn contains a single utterance from the system and its associated response from the user. 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 some methods, knowledge K or dialogue goals G are given as input to facilitate the recommendation and response generation. So, at the j^{th} turn, given the user's contextual history, $system$ generates recommendation results i and system response s_{j+1}^{system} (Eq. 1).

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

3.1 Empirical Analysis Approaches

Building on the advancements of LLMs over general LMs in language generation and reasoning, we explore LLM-based CRS task performance to assess their inherent response generation and recommendation capabilities, with and without integrating external knowledge or goal guidance. We design tasks where LLMs 1) directly generate system responses and recommendations *(Direct Generation; DG)*. Also, to evaluate the necessity of external inputs, we configure the LLM to 2) **internally** reason with its built-in knowledge and goals for response and recommendation *(Chain-of-Thought; COT)* or 3) leverage gold-standard labelled **external** knowledge and goals to enhance oracular performance *(Oracular Generation; Oracle)*, as illustrated in Figure 2a.

The primary experimental approach involves in-context learning (ICL) on the DuRecDial dataset (Liu et al., 2021), with an overview of ICL prompts and examples provided in Figure 3 and Appendix A.1, respectively. Experiment details and metrics are detailed in § 5. Evaluations focus on content preservation ($bleu\text{-}n$, $F1$) and diversity ($dist\text{-}n$) for response generation, and top-K ranking accuracy ($NDCG@k, MRR@k$) for recommendation tasks. We outline each experiment and its testing objective 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 3a).

LLM	*Approach*	*K/G*	*bleu*1	*bleu*2	*bleu*	*dist*1	*dist*2	*F*1	$Acc_{G/K}$
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 1: Empirical analysis for response generation task (K/G: Knowledge or goal guidance; $Acc_{G/K}$: Accuracy of knowledge or goal predictions; **Red**: Best result for each model; Underline: Best results for all).

- ***Chain-of-thought Generation (COT).*** With dialogue history as input, COT generates knowledge 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 3b).
- ***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 3c).

3.2 Empirical Analysis Findings

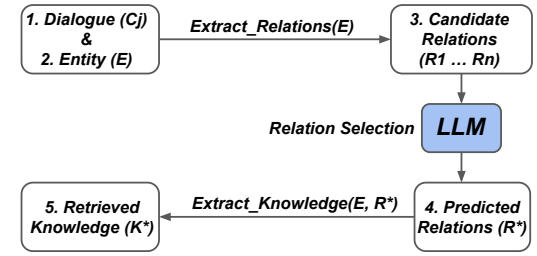
We summarize our three main findings given the results of the response generation and recommendation tasks shown in Tables 1 and 2.

Finding 1*: The Necessity of External Knowledge and Goal Guidance in LLM-based CRS.* Inclusion of external knowledge and goal guidance 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 input *(cf. RQ 1&2)*. Remarkably, the Oracle approach yields over a tenfold improvement in recommendation tasks, compared to DG and COT methods, as shown in Table 2. Although utilizing internal knowledge and goal guidance (COT) marginally benefits both tasks across various LLMs, we see that the low accuracy of internal predictions adversely affects performance, particularly in response generation.

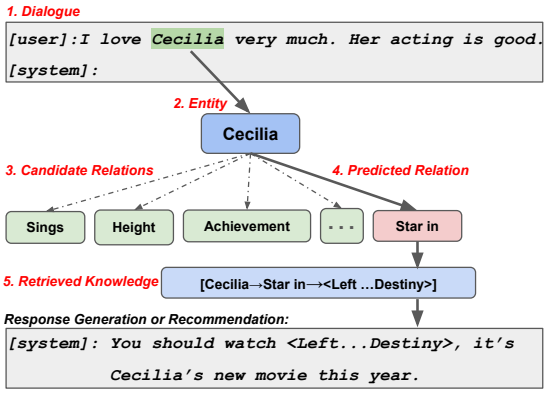
LLM	*Task*	*NDCG*@10/50	*MRR*@10/50
ChatGPT	DG	0.024/0.035	0.018/0.020
	COT-K	0.046/0.063	0.040/0.043
	Oracle-K	**0.617/0.624**	**0.613/0.614**
LLaMA-7b	DG	0.013/0.020	0.010/0.010
	COT-K	0.021/0.029	0.018/0.020
	Oracle-K	**0.386/0.422**	**0.366/0.370**
LLaMA-13b	DG	0.027/0.031	0.024/0.024
	COT-K	0.037/0.040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

Table 2: Empirical analysis for recommendation task (K: Knowledge; **Red**: Best result for each model).

Finding 2*: Improved Internal Knowledge or Goal Planning Capability in Advanced LLMs.* Table 1 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. Furthermore, the enhanced goal prediction accuracy further corroborates this finding. Nonetheless, the application of external knowledge and goal guidance continues to enhance performance across all LLM variants, contributing to state-of-the-art (SOTA) outcomes.



(a) Knowledge retrieval from external KB.



(b) An example of knowledge retrieval in ChatCRS

Figure 4: ChatCRS knowledge retrieval agent.

Finding 3*: Differential Impact of External Inputs on LLM Performance in CRS Tasks.* Table 1 indicates that open-source LLMs gain more from external goal guidance in generation tasks, whereas closed-source LLMs, like ChatGPT, improve significantly with external knowledge. For recommendation tasks (Table 2), external knowledge benefits all LLM types, underlining its critical role in supplementing LLMs' inherent lack of domain-specific information, required for accurate recommendations.

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

We showed that a CRS benefits from engaging with external KBs to supplement domain-specific and recommendation-oriented knowledge. However, training LLMs to memorize an entire KB is impractical due to computational demands and input token length constraints (Wei et al., 2021). Therefore, we employ a method that starts from entity E within the dialogue C_j and retrieves a knowledge triple "entity–relation–entity" by traversing along E's relations R (Moon et al., 2019).

In line with Jiang et al., our knowledge retrieval agent interfaces the LLM with the external KB to select appropriate relations. This agent first gathers all relations adjacent to entity E from the KB (denoted as $F1$), upon which the LLM is instructed to predict the most pertinent relation R^* given the dialogue history C_j. This agent then acquires the corresponding knowledge tuple K^* using entity E and predicted relation R^* (denoted as $F2$), formulated in Eq 2 and shown in Figure 4 (a). An example depicted in Figure 4 (b) demonstrates the process using the dialogue *"I love Cecilia..."* and the entity *[Cecilia]*. The system first extracts all potential relations for *[Cecilia]*, from which the LLM selects the most relevant relation, *[Star in]*. Knowledge retrieval then fetches the complete knowledge triple *[Cecilia–Star in–<Left...Destiny>]*.

We implement N-shot ICL to guide LLMs in choosing knowledge relations via a knowledge retrieval agent. This approach feeds entities from the dialogue history into the LLM, deliberately omitting the target recommendation entity to ensure the relevance of the retrieved knowledge (Moon et al., 2019; Jiang et al., 2023).

$$\begin{aligned} R_1, .., R_n &= F1\ (\ C_j, E, KB) \\ R^* &= LLM\ (\ C_j, E, R_1, .., R_n) \\ K^* &= F2\ (\ E, R^*, KB) \end{aligned} \quad (2)$$

4.2 Goal Planning agent

CRS datasets feature diverse dialogue goals, including "Greeting", "Movie Recommendation", and "Asking Questions", each necessitating specific system responses. Accurately classifying these goals is crucial for effective dialogue planning and proactive response generation in CRS. Utilizing goal annotations 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 input C_j, the model is trained to predict the next dialogue goal G^*, opti-

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
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 3: Results of response generation task on DuRecDial and TG-Redial datasets.

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

mizing the loss function outlined in Eq 3, with θ representing the trainable parameters of LoRA.

$$L_g = -\sum_i^n \log P_\theta \left(G^* | \, C_j\right) \tag{3}$$

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 4. Given that both goal planning and knowledge retrieval are engineered to produce text outputs, any LLM can serve as the final generation mechanism. Furthermore, the conversational agents are guided by N-shot ICL prompts, enabling LLMs to effectively execute CRS-related tasks.

$$i, \, s_{j+1}^{system} = LLM(\, C_j, K^*, G^*) \tag{4}$$

5 Experiments

5.1 Experimental Setups

Datasets We conduct the experiments on two multi-goal CRS benchmark datasets, namely DuRecDial2 (Liu et al., 2021) in English and TG-ReDial (Zhou et al., 2020b) in Chinese with statistics presented in Table 11. 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.

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 5: 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).

Baselines We compare our model with ChatGPT and LLaMA-7b/13b (Touvron et al., 2023) in few-shot settings. We also compare fully-trained UniMIND (Deng et al., 2023b), MGCG (Liu et al., 2020), SASRec (Kang and McAuley, 2018) baselines (all previous CRS and RS SOTA models).

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

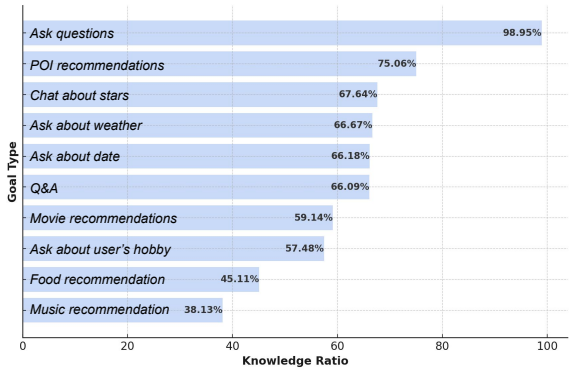


Figure 5: Knowledge ratio for each goal type on DuRecDial. (X-axis: Knowledge Ratio ; Y-axis: Goal type)

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 human evaluation process and each criterion are discussed in § A.2.

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 QLora as a parameter-efficient way to fine-tune LLaMA-7b (Dettmers et al., 2023; Deng et al., 2023b). 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 checkpoints and experimental setup are discussed in § A.3.

5.2 Experimental Results

ChatCRS significantly improves LLM-based conversational systems for CRS tasks, outperforming state-of-the-art baselines in response generation in both datasets, enhancing content preservation and language diversity (Table 3). ChatCRS sets new SOTA benchmarks on both datasets using 3-shot ICL prompting by incorporating external knowledge and goal direction. In recommendation tasks (Table 4), LLM-based approaches under few-shot ICL 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 3-shot ICL, without full-data fine-tuning.

Model	Knowledge				
	N-shot	Acc	P	R	F1
ChatGPT	3	0.095	0.031	0.139	0.015
LLaMA-13b	3	0.023	0.001	0.001	0.001
ChatCRS-L	3	0.503	0.307	0.341	0.302
ChatCRS-C	3	**0.560**	**0.583**	**0.594**	**0.553**

Table 6: Results for knowledge retrieval on DuRecDial. (L/C stands for baseline LLM of LLaMA/ChatGPT)

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
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 7: Results of goal planning task.

Human evaluation highlights ChatCRS's impact on enhancing CRS-specific language quality. Table 5 shows the results of human evaluation on both ChatCRS and component ablations, showing that 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 terms of 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

External knowledge is pivotal for LLM-based CRS, addressing the gap between universal pre-training and domain-specific CRS tasks. Despite the effectiveness of integrating knowledge into ChatCRS, we aim to further explore its impact on enhancing CRS performance across tasks. By an-

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', 'Star 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.

alyzing dialogues from the DuRecDial datasets, categorized by goal types, we calculated a "Knowledge Ratio" to measure the necessity of relevant knowledge in CRS task completion.

Our analysis, depicted in Figure 5, shows that recommendation tasks rank highly in terms of knowledge necessity, with "POI recommendation" dialogues requiring pertinent knowledge in 75% of cases. Contrasting with traditional RS which relies on user data for collaborative recommendation, CRS only depends on dialogue history for content-based recommendation. This shift underscores the limitations of LLMs in harnessing internal knowledge, a challenge highlighted by our analysis of knowledge retrieval accuracy (Table 6). ChatCRS overcomes these limitations by interfacing LLMs' to reason over external KBs.

Furthermore, a case study on the "Asking questions" goal type with the highest knowledge ratio, demonstrates the advantage of 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. This case study highlights ChatCRS's ability to leverage external knowledge, significantly improving CRS accuracy and informativeness.

Goal guidance enhances the task-specific language quality of LLMs in CRS applications. In contrast to the role of knowledge, goal guidance contributes more to the linguistic quality of CRS by managing the dialogue flow. To examine the goal planning proficiency of ChatCRS versus other LLM-based methods, we showcase goal planning outcomes in Table 7. LLM-based solutions often struggle in scenarios involving multiple CRS goals due to a deficiency in task-specific capabilities. 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. This underscores goal guidance's critical role in fostering proactive and effective engagement in CRS tasks.

Therefore, we address RQ3 by concluding that ***ChatCRS's efficiency originates from utilizing LLMs' inherent strengths in generating responses and reasoning, coupled with the strategic deployment of smaller agents on knowledge retrieval and goal-planning to enhance CRS implementations.***

6 Conclusion

This paper conducts an empirical investigation into the LLM-based CRS, emphasizing 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.

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., 2020b), 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

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023a. 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. 2023b. 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.

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*.

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.

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.

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. 2018. 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.

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.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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, 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. 2023. 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.

Jun Zhang, Yan Yang, Chencai Chen, Liang He, and Zhou Yu. 2021. KERS: A knowledge-enhanced framework for recommendation dialog systems with multiple subgoals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1092–1101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020a. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1006–1014.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. 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.

A Appendix

A.1 ICL Prompt Examples for CRS Tasks

In Section § 3.1, 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. We provide sample instructions within the prompts in Table 9. Furthermore, we detail the input and output formats below, with actual input-output examples presented in Table 10.

♠ 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 9: Example of instruction in prompt design

A.2 Human Evaluation

We selected 100 dialogues from the DuRecDial dataset to evaluate the performance of four methodologies: ChatGPT[2], LLaMA-13b[3], 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

[2]OpenAI API: gpt-3.5-turbo

[3]Hugging Face: LLaMA2-13b-hf

♠ 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 10: Example of input and output format in prompt design

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 5, 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.3 Experiment Settings

In our Empirical Analysis and Modelling Framework, we implement few-shot learning across vari-

ous Large Language Models (LLMs) such as ChatGPT[4], LLaMA-7b[5], and LLaMA-13b[6] 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., 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 knowledge-guided generation due to the fixed dialogue goal of "Recommendation".

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., 2023b). 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., 2020b) comprises only Chinese conversations, a pre-trained Chinese LLaMA model is used for inference[7]. 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. Statistic of two experimented datasets are shown in Table 11.

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

Table 11: Statistics of datasets

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