Thoughts to Target: Enhance Planning for Target-driven Conversation

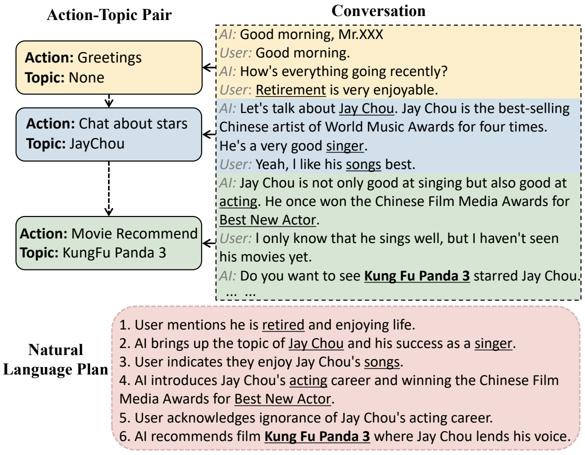
Zhonghua Zheng[1*] **Lizi Liao**[2] **Yang Deng**[2] **Ee-Peng Lim**[2] **Minlie Huang**[3] **Liqiang Nie**[1]

[1] Harbin Institute of Technology, Shenzhen
[2] Singapore Management University
[3] Tsinghua University

polang1999@gmail.com, {lzliao,ydeng,eplim}@smu.edu.sg, aihuang@tsinghua.edu.cn, nieliqiang@gmail.com

Abstract

In conversational AI, large-scale models excel in various tasks but struggle with target-driven conversation planning. Current methods, such as chain-of-thought reasoning and tree-search policy learning techniques, either neglect plan rationality or require extensive human simulation procedures. Addressing this, we propose a novel two-stage framework, named **EnPL**, to improve the LLMs' capability in planning conversations towards designated targets, including (1) *distilling natural language plans* from target-driven conversation corpus and (2) *generating new plans* with demonstration-guided in-context learning. Specifically, we first propose a filter approach to distill a high-quality plan dataset, ConvPlan[1]. With the aid of corresponding conversational data and support from relevant knowledge bases, we validate the quality and rationality of these plans. Then, these plans are leveraged to help guide LLMs to further plan for new targets. Empirical results demonstrate that our method significantly improves the planning ability of LLMs, especially in target-driven conversations. Furthermore, EnPL is demonstrated to be quite effective in collecting target-driven conversation datasets and enhancing response generation, paving the way for constructing extensive target-driven conversational models.



AI: Good morning, Mr.XXX
User: Good morning.
AI: How's everything going recently?
User: Retirement is very enjoyable.
AI: Let's talk about Jay Chou. Jay Chou is the best-selling Chinese artist of World Music Awards for four times. He's a very good singer.
User: Yeah, I like his songs best.
AI: Jay Chou is not only good at singing but also good at acting. He once won the Chinese Film Media Awards for Best New Actor.
User: I only know that he sings well, but I haven't seen his movies yet.
AI: Do you want to see **Kung Fu Panda 3** starred Jay Chou.
... ...

1. User mentions he is retired and enjoying life.
2. AI brings up the topic of Jay Chou and his success as a singer.
3. User indicates they enjoy Jay Chou's songs.
4. AI introduces Jay Chou's acting career and winning the Chinese Film Media Awards for Best New Actor.
5. User acknowledges ignorance of Jay Chou's acting career.
6. AI recommends film **Kung Fu Panda 3** where Jay Chou lends his voice.

Figure 1: The structured plan (e.g., Action-Topic Pairs) generated by traditional dialogue planning methods hinders both human and LLMs understanding.

1 Introduction

Unlike task-oriented conversations that encompass a broader range of tasks, goal-driven conversations focus on reaching a specific goal or objective, such as recommending a target movie (Dao et al., 2023; Liao et al., 2023). The dialogue systems are required to lead the conversation to the target flexibly and coherently. Due to its purpose and flexibility, target-driven dialogue agents have a broad-based demand, e.g., conversational recommendation (Kang et al., 2019; Dao et al., 2024), psychotherapy (Sharma et al., 2020; Fei et al., 2024), and education (Clarizia et al., 2018; Guo et al., 2024). These conversations, usually characterized by defined user requirements, rely on precise planning capabilities, making it crucial to build autonomous conversational AI.

In traditional target-driven conversation methods, many studies control dialogue generation through next-turn transition prediction (Tang et al., 2019), subgoal generation (Kishinami et al., 2022; Deng et al., 2023b), and knowledge path reasoning (Wu et al., 2022). To accomplish this task, effective conversation planning is crucial (He et al., 2024), which requires reasonable actions to guide the dialogue topics to targets smoothly. Different from summarizing a conversation, the process of planning requires not only capturing the key content but also ensuring logical coherence and naturalness. However, previous studies have employed greedy strategies with single-round topic prediction mechanisms that lack global planning of the conversation process (Yang et al., 2022). These approaches

---

[*] Work was done during an internship at SMU.

[1] Resources of this paper can be found at https://github.com/pandazzh2020/ConvPlan

tend to be short-sighted and lead to incoherent topic cues. The generated plan is also too structured (e.g., a sequence of entities or action topic pairs) and not conducive to human understanding. This inherent rigidity prompts a shift in focus toward emergent conversational frameworks, a realm dominated by Large Language Models (LLMs).

Recent advancements have propelled LLMs to the forefront of conversational AI due to their exceptional generation capabilities (Aher et al., 2023). However, LLMs fall short of proactively planning the conversation process (Zheng et al., 2024; Deng et al., 2023a), making it insufficient in handling target-driven conversation. This is because target-driven conversations aim to achieve a global target that often cannot be explicitly defined as a subtask. Conversation agents are required to be able to direct the conversation to the target flexibly and the process must be coherent.

Nevertheless, to enhance the planning and reasoning ability of LLMs, many researchers have investigated Chain-of-thought (CoT) (Wei et al., 2023) and Tree of Thoughts approach (ToT) Yao et al. (2023), known as reasoning chains or rationales, to eventually lead to the final answer. However these works usually only apply to some well-defined tasks (such as Game of 24), focusing on the evaluation of the final task and neglecting the measurement of the rationality of the plan. In addition, many works use the tree-search approach to improve planning capabilities of LMs (Zhang et al., 2023; Yu et al., 2023). For example, Yu et al. (2023) treat policy planning as a stochastic game and use prompting for every stage of an open-loop tree search. However, when these methods are faced with the complexity of real-world applications, they require a lot of user simulation.

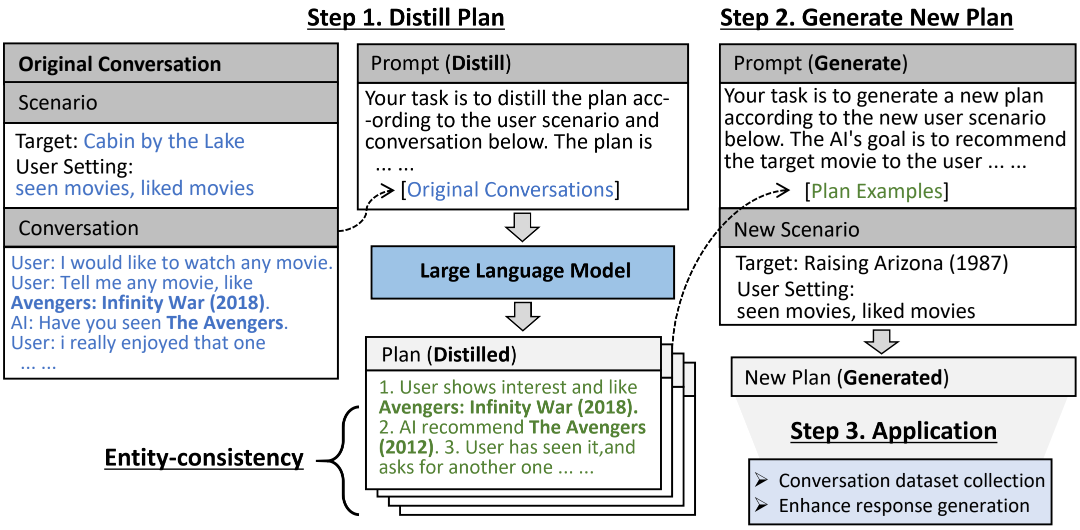
In this paper, we aim to improve the constrained planning ability of LLMs in the task of target-driven conversation. LLMs have strong comprehension and generation capability but weak planning capability (Yuan et al., 2023; Xie et al., 2023). As illustrated in Figure 1, the structured plan could be difficult to understand by both human and LLMs. To mitigate this issue, we propose a novel two-stage planning construction framework, named Enhance Planning framework (EnPL). EnPL first leverages the existing manually collected conversation dataset to distill natural language plans describing the conversation process through LLMs. We propose a filtering approach, which calculates the entity consistency score between the distilled plans and the conversations, to select high-quality plans for constructing a target-driven conversation plan dataset, named ConvPlan. It consists of 12K high-quality plans with targets, user settings, and plans. Given a new user setting and target, the distilled plans can then serve as demonstrations for generating a new plan as thought to the target with the exceptional in-context learning capability of LLMs. We fully verify the rationality and intelligence of the newly generated plan and reveal that these plans can further guide conversation collection and enhance response generation, pointing out feasible directions for constructing large-scale target-driven conversation datasets and model training.

Our contributions are summarized as follows: (1) We propose a novel two-stage framework, named EnPL, to improve the LLMs' capability in planning conversations towards designated targets, including distilling natural language plans from target-guided dialogue corpus and generating new plans with demonstration-guided in-context learning. (2) We propose a filter approach to select high-quality plans distilled by LLMs and introduce a novel evaluation metric, named EntityCov, based on entity-coverage for plan validation. (3) Based on EnPL, we first create a high-quality plan dataset (ConvPlan) for constrained language planning. By leveraging the ConvPlan, we validate that the generated plans play a guiding role in collecting large-scale datasets and enhancing response generation.

2 Related Work

2.1 Target-driven Conversation

Target-driven conversation systems focus on how to naturally lead users to accept the designated targets gradually through conversations. Previous research has explored various approaches for using keywords and topics as guided targets (Tang et al., 2019). The advancement of research in this field was catalyzed by the emergence of several datasets such as DuRecDial (Liu et al., 2021), GoRecDial (Kang et al., 2019), TG-ReDial (Zhou et al., 2020), and INSPIRED (Hayati et al., 2020). Additionally, external commonsense knowledge graphs were used to facilitate keyword transition (Wu et al., 2019; Ma et al., 2021) and response retrieval using GNNs (Zhong et al., 2020; Liang et al., 2021). These datasets typically feature structured plans comprising sequences of keywords or action-topic pairs. While methodical, these structures lack interpretability and miss crucial conversational details,



User: I would like to watch any movie.
User: Tell me any movie, like **Avengers: Infinity War (2018)**.
AI: Have you seen **The Avengers**.
User: i really enjoyed that one
… …

Your task is to distill the plan acc-ording to the user scenario and conversation below. The plan is … …

1. User shows interest and like **Avengers: Infinity War (2018).** 2. AI recommend **The Avengers (2012)**. 3. User has seen it,and asks for another one … …

Your task is to generate a new plan according to the new user scenario below. The AI's goal is to recommend the target movie to the user … …

Figure 2: Detailed overview of our proposed two-stage framework (EnPL). Step 1: a large language model is prompted to distill plans (blue) from the existing dataset. Step 2: (green) The distilled plans are used to compose a prompt comprised of other descriptions. The prompt and a new scenario will guide LLM to generate new plans. Step 3: The generated plans can be used for applications such as data collection and enhance response generation.

posing challenges for both human users and LLMs. To address this, there is an increasing emphasis on generating plans in natural language, offering greater clarity and ease of understanding.

2.2 Goal-oriented Planning Script Generation

Prompting in the field of LLM research has seen significant developments towards generating more flexible and efficient outputs. Many researchers have investigated Chain-of-thought (CoT) prompting (Wei et al., 2023; Wang et al., 2023c) and Tree of Thoughts approach (ToT) (Yao et al., 2023). However, these efforts focus on improving the reasoning power of LLMs, while neglecting to measure the rationality of the plan, and are not suitable for planning dialogue process. In order to improve the planning capabilities of LLMs, many previous works have investigated how to perform content planning (such as selecting key entities and arranging their sequence) for text generation (Puduppully et al., 2019; Su et al., 2021). Currently, multiple planning frameworks have been proposed for complex generation tasks (Hua et al., 2021; Li et al., 2022). Our work is more relevant to dialogue generation planning (Kishinami et al., 2022; Yang et al., 2022; Cohen et al., 2022). Wang et al. (2023a) introduced the COLOR model to guide goal-oriented dialogue generation using Brownian bridge processes to generate dialogue-level planning. However, this approach is susceptible to error propagation, and when the model fails to plan an appropriate dialogue path, the performance of dialogue generation significantly deteriorates. Our proposed EnPL framework is a novel method to enhance the planning capabilities of large models and can be used to guide target-driven conversation generation.

2.3 LLM for Dialogue Generation

The field of LLMs for dialogue generation has seen remarkable progress. Recent studies highlight its potential across various dialogue applications, such as conversational question-answering (Ye et al., 2022; Xu et al., 2023), emotional support dialogues (Zheng et al., 2024; Chu et al., 2024), open-domain social dialogues (Chen et al., 2023; Kim et al., 2022), tutoring dialogues (Macina et al., 2023), and more. Despite the remarkable quality of LLM-synthetic dialogue data, this type of data inevitably inherits the limitation of LLMs in handling proactive dialogues, such as inappropriate content, limited understanding of user intent, inability to clarify uncertainty, limited ability to make strategic decisions and plans, etc. In target-driven dialogues, there is a need for the system to proactively plan the conversation process, set targets, and take actions (Wang et al., 2023b), that goes beyond the current capabilities of LLMs. So our approach aims to enhance the planning ability of LLMs.

3 EnPLAN Framework

As illustrated in Figure 2, the proposed framework can be decomposed into two stages: (1) plan distillation and (2) plan generation. In stage 1, aiming at the existing LLMs with weak planning capability but strong comprehension and generation capability, we use the existing manually collected conversation dataset DuRecDial (Liu et al., 2021)[2] to distill plans describing the conversation process through LLMs. In stage 2, we employ the distilled plans as examples. Then, given a new user setting and target, we can select the plan examples in different ways and generate a new plan as thoughts to target by combining the powerful in-context learning capability of LLMs.

3.1 Distill Plan from Existing Conversation

3.1.1 Problem Formulation

Denote $D = (s_i, c_i)^N$ to be a dataset with N training instances, where s_i is a scenario which is a tuple of user setting and target item (u_i, t_i) and c_i is the corresponding target-driven conversation. Also, we have a handful of human-written instances $E = (s'_i, c'_i, p'_i)^M$, where p'_i is a free-text plan to describe the conversation plan sketch to the target item and $(s'_i, c'_i)^M \in D$ with $M << N$ (we set $M = 30$ in our experiments). Our goal is to fully leverage LLM with E as examples to distill reasonable plans p_i for all (s_i, c_i), where $1 \leq i \leq N$, so that we can utilize these distilled plans from LLM to enhance planning for new scenarios.

3.1.2 Filter Plan with Entity-consistency

We further utilize entity-consistency to improve the quality of the distilled plans. The main idea is to filter high-quality ones from multiple distilled plans. Based on the examples E given, we explain to ChatGPT what a plan is and specify the criteria for distilling the plan by referring to the Chain of Thought (CoT) approach (Wang et al., 2023c). We then guided ChatGPT to distill plans (prompts are shown in Appendix A).

We first extract the set of key entities from the distilled plan K_{plan} and the original conversation K_{conv} using TextRank algorithm (Mihalcea and Tarau, 2004). Then, we calculate the consistency score between the plan and the original conversation using the Levenshtein distance algorithm[3]. Unlike the original Levenshtein distance algorithm, we treat key entities as the smallest units instead of individual characters. The Levenshtein distance between K_{plan} and K_{conv} (of length i and j respectively) is given by $Leven_{p,c} = L(i, j)$:

$$L(i,j) = \begin{cases} \max(i,j), & \text{if} \min(i,j) = 0 \\ s, & \text{otherwise} \end{cases} \tag{1}$$

where K_{plan} and K_{conv} are noted as p and c, respectively, for simplicity. Then s is computed by

$$\begin{aligned} s = \min\{\mathrm{L}(i-1, j) + 1, \mathrm{L}(i, j-1) + 1, \\ \mathrm{L}(i-1, j-1) + 1_{(p_i \neq c_j)}\} \end{aligned} \tag{2}$$

We calculate the consistency score via:

$$consistency = 1 - \frac{\mathrm{L}(i,j)}{\max\,(i,j)} \tag{3}$$

An example is shown in Appendix B. The Levenshtein distance directly reflects the degree of difference between the distilled plan and the original conversation, considering the order of entity occurrences. We filter out the top 2 plans with the highest consistency scores from the 10 distilled plans in each round to form the plan repository (ConvPlan).

3.2 Demonstrated Planning for New Scenario

We construct new scenarios each including a user setting and a target item $s_j = (u_j, t_j)$, and then select (s_i, p_i) as an example from the distilled plans. Our goal is to give new s_j under the guidance of example (s_i, p_i) to generate new plan p_j.

For better guiding LLM to generate new plans, it is important to select examples for new user scenarios. We explore three different strategies for selecting examples.

Random-based. Randomly select scenarios and plans as example (s_i, p_i) in ConvPlan. This setup does not consider the similarity and diversity between the new user scenario s_j and the user scenarios s_i in existing plans.

Similarity-based. Based on the similarity, we select the similar user scenarios and plans as example s_i, p_i. Specifically, we select the plan with the largest overlap (max $(|s_j \bigcap s_i|)$) between the movie in the current user scenarios s_j and the movie contained in s_i.

Diversity-based. We use K-means++ clustering (Chang et al., 2021) to select the most representative and diverse plan samples, which will maximize the possibility of maximizing the large models to generate diverse plans. We first map each data point into a vector, then cluster the vectors with the

---

[2] Note that our framework also can be applied to other target-driven conversation datasets.

[3] `https://en.wikipedia.org/wiki/Levenshtein_distance`

K-means algorithm. The objective is the sum of the squared errors (SSE), called cluster inertia:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{i,j} ||x^i - \mu^j||_2^2, \quad (4)$$

where μ^j is the centroid of the j-th cluster, x^i is the embedding vector of U_i, and $w_{i,j} = 1$ if x^i belongs to the cluster j and 0 otherwise. We optimize the objective function with the EM algorithm (Dempster et al., 1977) which iteratively assigns each data point to its closest cluster centroid. The initial centroid points are chosen based on the K-means++. The first cluster center is chosen uniformly at random from the data points, after which each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster center. This means we maximize the chance of spreading out the K initial cluster centers. We use 50 random seeds for selecting initial centers and the clustering with the minimum SSE is chosen.

3.3 Applications: Usage of Generated Plan

Guide Conversation Dataset Collection. Utilizing the EnPL-generated plans as demonstrated in Figure 2 allows for the delineation of a coherent and logical dialogue pathway, facilitating the step-by-step achievement of targeted conversational objectives. We regard each plan as a natural language prompt to guide LLM (like ChatGPT) to generate complete conversations and compare with human-annotated methods to verify the ability of our plans to augment conversation data (Wang et al., 2022, 2023a) (Table 5).

Enhance Response Generation. Our planning can also be used for response generation enhancement. Following previous studies (Wang et al., 2023a), we perform self-play simulations, to simulate multi-turn conversations and compute the success rate of generating the target keyword within 8 turns on TGConv (Yang et al., 2022) dataset (Table 6). We also use the plan as a natural language prompt for response generation and compare with keyword-based prompt methods (Yang et al., 2022; Wang et al., 2023a) (Appendix C.3).

4 Evaluating Step 1: Distill Plan

4.1 Baselines

We explore prompting for three different ways of distilling plans (Appendix A).

GPT4-abs. GPT4-abs (Liu et al., 2024) is a method that utilizes GPT4 for text summarization and quality assessment.

Direct Prompt. Directly gives the LLM instructions to generate a plan describing the conversation process, including zero-shot and one-shot settings. The one-shot demonstration is randomly selected from 30 manually constructed plan examples.

CoT+Prompt. Based on the manual examples given, explain to LLM what a plan is and specify the criteria for generating the plan by referring to the Chain of Thought (CoT) method (Wang et al., 2023c), also including zero-shot and one-shot settings.

4.2 Proposed Evaluation Metrics

Entity-centered Protocol The quality and rationality of the plan can be measured and verified through the correspondence of the conversation data and the support of the related knowledge base. Referring to (Mihalcea and Tarau, 2004), we designed the entity-coverage evaluation metric **EntityCov**. First, the text is divided into nodes $V_1, V_2, \ldots, V_n$, and the edges $E(i,j)$ between nodes are constructed to represent the association strength between nodes. Initially, the weight of each node is $W(i) = 1$. Then, TextRank uses an iterative method to calculate the node's weight. Taking into account the correlation between nodes, the formula is as follows:

$$W(i) = (1-d)+d \cdot \sum_{j} \left(\frac{W(j) \cdot W(i,j)}{\sum_k W(k)} \right), \quad (5)$$

where j is the neighbor node of node i, and d is the damping coefficient (usually 0.85). Iteratively calculating weight values until convergence, this process enables the identification of the most important words or phrases in the conversation as keywords. Then extract the first 20 keywords K_{conv} based on the final weight value of the node. On this basis, we take the union of the keywords K_{user} and K_{conv} in user information and get $K_{conv+user} = K_{user} \bigcup K_{conv}$. We then use the above principle to get the keyword list K_{plan} in the plan, and calculate the entity-coverage score:

$$EntityCov = \frac{|K_{plan} \bigcap K_{conv+user}|}{|K_{conv+user}|}. \quad (6)$$

Human-centered Protocol The best method for evaluating such texts is human evaluation, where human annotators assess the generated plans' quality. This evaluation can be done from different

Methods	EntityCov	BERTScore	BARTScore	Coherence
GPT4-abs	0.4385	0.5676	-3.610	0.3485
Direct Prompt	0.3961	0.6143	-3.586	0.3986
w/ *example*	0.4657	0.5874	-3.395	0.4252
CoT+Prompt	0.4551	0.6197	-3.384	0.4167
w/ *example*	0.5142	0.6251	**-3.282**	0.4348
EnPL	**0.5509**	**0.6630**	-3.3559	**0.4597**

Table 1: Automatic evaluation of plan distillation. Results in bold indicate significant superiority over others.

Methods	Coh.	Rel.	Int.	Con.	Ove.
GPT4-abs	2.02	2.45	2.31	1.97	2.07
Direct Prompt	1.95	2.46	2.23	2.39	2.22
w/ *example*	2.24	2.40	2.42	2.51	2.41
CoT+Prompt	2.13	2.47	2.35	2.40	2.38
w/ *example*	2.15	2.54	2.51	2.42	2.43
EnPL	**2.30**	**2.63**	**2.74**	**2.55**	**2.58**
κ	0.45	0.35	0.33	0.47	0.42

Table 2: Human evaluation results in plan distillation. The scores (from 0 to 3) are averaged over all the samples rated by five annotators. κ denotes Fleiss' Kappa (Fleiss, 1971), indicating fair or moderate inter-annotator agreement ($0.2 < \kappa < 0.6$).

perspectives, and we propose a few common varieties: (1) **Coherence (Coh.)**: Is the overall logic of the plan coherent and clear? (2) **Relevance (Rel.)**: Can the plan capture the key information and discussion process of the original conversation? (3) **Intelligence (Int.)**: whether the plan to guide the conversation process to target is smart. (4) **Concise (Con.)** Is the language of the plan concise? (5) **Overall (Ove.)**: Which version do you prefer overall?

Other Metrics To evaluate the performance of plans distilled, we adopt **BERTScore** (Zhang et al., 2019) and **BARTScore** (Yuan et al., 2021) to measure the semantic similarity between the plan and the original conversation. Following (Yang et al., 2022), we also use **Coherence** as another global evaluation metric. BERTScore calculates the cosine similarity between two sentences based on BERT model. BARTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence. Coherence is a global evaluation metric, that measures the average contextual semantic similarity between the last utterance in the context and generated utterance.

4.3 Quality Analysis for Distilled Plans

To demonstrate the effectiveness of distilled plans within our EnPL framework, we carried out both automatic evaluation compared to other methods and human evaluation involving five master's students. We randomly selected 50 distilled plans from ConvPlan for comparative analysis. For human evaluation, participants were prompted with the questions in Section 4.2. The comparison outcomes presented in Table 1 and Table 2 reveal the following findings: (a) Our method demonstrates a capacity to include more key entities and clearer logical structures compared to directly summarizing dialogues. (b) We find that the Direct Prompt lacks comprehensive examples and guidance, leading LLM to struggle in understanding the task of plan distillation, resulting in unsatisfactory responses and formatting inconsistencies. (c) Compared to CoT+Prompt, under similar examples and guidance, the plans we distilled closely resemble the original conversations due to our utilization of entity-consistency, filtering the distilled plans to ensure their quality. Overall, our approach effectively guides LLMs in distilling dialogue plans and efficiently filters them, affirming the high quality and practicality of our ConvPlan (distilled plans).

5 Evaluating Step 2: Generate New Plan

In this section, we fully verify the rationality and intelligence of the newly generated plan and reveal that generated plans can further guide the generation of target-driven conversations.

5.1 New Scenarios Setting

To create a scenario similar to the real case, we use the 2k scenarios in the DuRecDial testset (Liu et al., 2021) as new scenarios to guide LLM to generate new plans. These scenarios include target movie, user profile, and knowledge graph. The user profile contains personal information (e.g. name, gender, age, residence city, occupation, etc.) and his/her preference And the knowledge graphs include star, movie, music, news, food, and so on. LLMs could generate more realistic and content-rich plans with the assistance of this information.

5.2 Baselines

For plan generation, our baselines include:
Direct prompting (Brown et al., 2020) is a standard method of prompting that makes a request directly to the LLM, including ChatGPT (175B) and LLaMA2 (70B).
CoT prompting (Liu et al., 2024) use a new CoT prompting paradigm of text summarization

Baselines	EntityCov	BERTScore	BARTScore	Coherence
LLaMA2	0.2556	0.3743	-3.675	0.3137
Direct prompting	0.2125	0.4823	-3.652	0.3169
CoT prompting	0.3273	0.5017	-3.506	0.3809
TopKG-Plan	0.2753	0.4362	-3.771	0.2802
COLOR	0.2976	0.5145	-3.545	0.2731
EnPL w/o filtering	0.3304	0.5198	-3.453	0.4465
EnPL	**0.3882**	**0.5535**	**-3.215**	**0.4584**

Table 3: Automatic evaluation results in plan generation.

that considers LLMs as the reference on commonly used summarization datasets such as the CNN/DailyMail dataset (Liu et al., 2024).

COLOR (Wang et al., 2023a) uses the Brownian bridge stochastic process to plan dialogue process, which models global coherence and incorporates user feedback in goal-directed dialogue planning.

Our variations. We analyze the following variants of our method: (1) w/ Random, which randomly selects context examples in ConvPlan; (2) w/ Similarity, which selects plans with similar scenarios; (3) w/ Diversity, which uses K-means++ clustering to select diverse and representative examples.

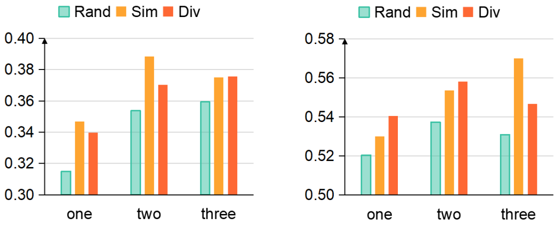
5.3 Evaluation Results for Plan Generation

Automatic Evaluation. Our EnPL demonstrates superior performance over other models in generating new plans, as shown in Table 3. EnPL outshines baselines across most metrics, notably showing that Direct prompting with ChatGPT (175B) slightly exceeds the performance of LLaMA2 (70B), likely due to ChatGPT's larger generative capacity and comprehension. EnPL excels in similarity-based metrics like BERTScore and BARTScore, producing longer, more detailed content with a wider inclusion of key entities. This suggests that precise scenario prompts enable the LLM to utilize its extensive knowledge to generate diverse content. Traditional plan generation methods used by COLOR and TopKG-Plan yield less coherent plans compared to EnPL, which significantly enhances plan coherence. EnPL's two-stage process not only refines a quality plan dataset, ConvPlan, but also effectively uses selected examples to guide LLMs in crafting comprehensive and coherent new plans.

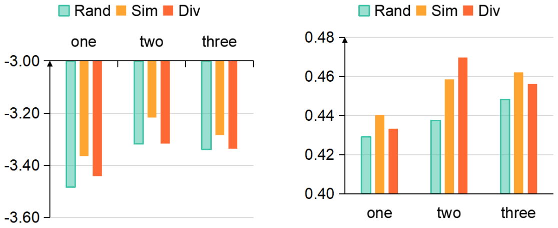
Human Evaluation. We further conduct a human evaluation on the generated plans with five annotators. The outcomes (shown in Table 4) reveal several findings: (1) LLaMA2 slightly underperforms compared to our EnPL, which is understandable considering our method builds upon ChatGPT, offering a larger generation space and better comprehension. (2) The COLOR's performance in plan generation is unsatisfactory. We observed that COLOR, relying on an external knowledge graph, lacks the capability for comprehensive planning, resulting in lower scores. (3) Detailing to explain the plan proves crucial; otherwise, the LLM lacks an understanding of the task's goal. Direct prompting may provide ambiguous guidance, leading to struggles in generating plans, thereby affecting scores in Clarity and Intelligent metrics. Overall, the results align with those of the automatic evaluation, which reveals that our method adeptly guides LLMs in generating reasonable new plans.

Baselines	Coh.	Rel.	Int.	Con.	Ove.
LLaMA2	2.03	2.21	2.03	2.32	2.11
Direct prompting	2.18	2.59	2.51	2.74	2.46
CoT prompting	2.37	2.76	2.56	2.67	2.64
TopKG-Plan	1.66	2.27	1.63	2.29	2.03
COLOR	1.72	2.07	1.72	2.35	2.13
EnPL w/o filtering	2.45	2.79	2.51	2.58	2.67
EnPL	**2.46**	**2.81**	**2.56**	**2.78**	**2.71**
κ	0.42	0.37	0.35	0.40	0.41

Table 4: Human evaluation results in plan generation.



(a) EntityCov (b) BERTScore



(c) BARTScore (d) Coherence

Figure 3: The impact of the number of examples (one, three, and five) and selection strategy on our framework. We select the best version EnPL w/ similarity giving 3 examples for subsequent experiments.

5.4 Effect of Demonstration Selection

We analyze the impact of selection strategies and example quantity on LLMs' plan generation capabilities, shown in Figure 3. The scenario similarity strategy, which selects plans from ConvPlan based on scenario closeness, outperforms the diversity-based strategy and random selection, evidenced by higher BERTScore and BARTScore metrics. This strategy's effectiveness highlights the value of tailored examples in enhancing plan generation. Our findings also reveal that using three examples strikes the optimal balance between learning comprehensiveness and plan refinement, with diminishing returns observed when increasing to five examples due to input length constraints and cost considerations. Consequently, we adopt the similarity-based strategy with three examples for further experiments, confirming its efficiency in guiding LLMs to generate more accurate and contextually relevant plans.

6 Evaluating Step 3: Applications

We further validate the effectiveness of applying the plans generated by EnPL on two applications: 1) Guide Conversation Dataset Collection, and 2) Enhance Response Generation.

	Appr.	Info.	Proact.	Coh.	Succ.
DuRecDial 2.0	2.54	**2.64**	**2.61**	2.77	2.83
Our EnPL	**2.65**	2.62	2.58	**2.85**	**2.95**
κ	0.48	0.43	0.39	0.52	0.37

Table 5: Human evaluation of conversation quality. The scores (from 0 to 3) are averaged over all the samples rated by five annotators.

Guide Conversation Dataset Collection. As shown in Table 5, we conduct human evaluation on the collected conversations that are generated by using EnPL-generated plans. We find that our EnPL exhibits advantages over the manually constructed DuRecDial 2.0. Although DuRecDial 2.0 slightly outperforms us in informativeness, the difference is negligible. Our approach enables the generation of more contextually appropriate dialogues. Additionally, our EnPL attains higher scores in coherence and target success rate, possibly because manually crafted conversations often involve abbreviated or omitted discourse, leading to reduced coherence. Our plans effectively steer conversations toward their goals while maintaining coherence, offering a feasible approach for large-scale data collection, considering the high cost and limited scale of manually constructed datasets.

Model	Easy Target		Hard Target	
	Succ.	Coherence	Succ.	Coherence
GPT-2†	22.3	0.23	17.3	0.21
DialoGPT†	32.3	0.30	23.8	0.25
TopKG†	48.9	0.31	27.3	0.33
COLOR† **w/ D**	66.3	0.36	30.1	0.35
EnPL w/ D	69.5	0.37	52.8	0.33
EnPL w/ C	**96.3**	**0.44**	**87.1**	**0.41**

Table 6: Automatic evaluation results of dialog-level response generation on TGConv dataset. C and D are short for ChatGPT and DialoGPT, respectively. Models marked with † are reported from Wang et al. (2023a).

Enhance Response Generation. We conduct both dialogue-level (Table 6) and turn-level (Table 7) automatic evaluations on the improvement of response generation. Results are presented in Appendix C. We observe that our planning can flexibly and coherently lead conversations to the target. By incorporating rich entities, our planning guides the LLM to generate diverse results, showcasing the advantage of planning in natural language forms. Our plan can chart the course of the next dialogue steps based on the context, guiding the LLM to generate responses at each step. Under the guidance of planning, the model gains a better understanding of when and what to discuss, facilitating proactive conversation advancement and successful target achievement. Our guided planning lays the foundation for constructing more robust and intelligent conversational agents.

7 Conclusion

This paper introduces a novel two-stage enhanced planning framework to overcome challenges in target-driven conversation planning via LLMs. Our method involved harnessing the generative capabilities of LLM in distilling plans from existing human-curated datasets. We filter the over-generate plans and introduce comprehensive methods for plan validation. We further guide LLM to generate plans according to new user scenarios and targets via in-context learning. Our approach not only advances the capabilities of LLMs in planning target-driven conversations but also provides a scalable strategy for generating large-scale datasets. Consequently, this is a significant step towards building sophisticated target-driven conversational models. Future research will focus on refining the plan generation and validation process for even greater processing efficiency and accuracy.

Limitations

Our framework significantly advances LLM-based conversation planning but faces limitations inherent to LLMs, such as biases in training data and tendencies to produce incorrect information (Deng et al., 2024). While we enhance LLMs' planning capabilities, our focus isn't on modifying the model architecture itself, and our reliance on automatic evaluation metrics might lead to overestimation or underestimation, despite attempts to balance these with human evaluations. Currently, our ConvPlan dataset is limited to English and pure text, restricting multilingual and multimodal (Liao et al., 2018) applicability. A notable area we will explore shortly is the dynamic generation of conversation plans mid-dialogue, which would address our framework's current limitation of only generating plans at the conversation's outset and significantly enhance adaptability in real-time interactions.

Ethical Considerations

We protect the privacy rights of crowd-sourced workers and pay them above the local minimum wage (pay at a rate of $7 per hour). We acknowledge that constructing datasets from large language models may suffer from toxic language and cause severe risks for social society (Weidinger et al., 2021; Baldini et al., 2022). Factuality, Toxicity and Biases We recognize that the factuality of generated content is crucial, especially in high-stakes scenarios. Therefore, we ask the annotators to discard the offensive and harmful data when reviewing the ConvPlan. They also assess and revise the content to minimize hallucinations, factual errors, and any inappropriate or misleading information. However, there may still be prejudicial data in our final dataset that goes unnoticed. We highlight that our ConvPlan dataset is not intended for safety-critical applications or as a substitute for expert advice in such domains. Significant further progress needs to be made in areas like debiasing, grounding in actuality, and efficient serving before we can safely deploy this type of system in a production setting.

Acknowledgment

This research was supported by the Google South Asia & Southeast Asia research award and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 23-SISSMU-010).

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 337–371. PMLR.

I Elaine Allen and Christopher A Seaman. 2007. Likert scales and data analyses. *Quality progress*, 40(7):64–65.

Ioana Baldini, Dennis Wei, and Karthikeyan Natesan Ramamurthy. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners.

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural text generation.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, and Seokhwan Kim. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2024. Towards multimodal emotional support conversation systems. *arXiv preprint arXiv:2408.03650*.

Fabio Clarizia, Francesco Colace, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2018. Chatbot: An education support system for student. In *International Conference on Cryptography and Security Systems*.

Deborah Cohen, Moonkyung Ryu, Yinlam Chow, Orgad Keller, Ido Greenberg, Avinatan Hassidim, Michael Fink, Yossi Matias, Idan Szpektor, Craig Boutilier, et al. 2022. Dynamic planning in open-ended dialogue using reinforcement learning. *arXiv preprint arXiv:2208.02294*.

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*, pages 785–795.

Huy Dao, Lizi Liao, Dung Le, and Yuxiang Nie. 2023. Reinforced target-driven conversational promotion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12583–12596.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

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, Singapore, December 6-10, 2023*, pages 10602–10621.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.

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):77:1–77:25.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. Empathyear: An open-source avatar multimodal empathetic chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Shasha Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: Methods, applications, and prospects. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8038–8047.

Shirley Anugrah Hayati, Dongyeop Kang, and Qingxiaoyang Zhu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4768–4791.

Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. 2021. DYPLOC: Dynamic planning of content using mixed language models for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6408–6423.

Dongyeop Kang, Anusha Balakrishnan, and Shah. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. 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 1951–1961.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Choi Yejin. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization.

Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. 2022. Target-guided open-domain conversation planning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 660–668.

Mike Lewis, Yinhan Liu, Naman Goyal, and Marjan Ghazvininejad. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiwei Li, Michel Galley, Chris Brockett, and Jianfeng Gao. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Qintong Li, Piji Li, and Wei Bi. 2022. Event transition planning for open-ended text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3412–3426.

Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2019. Towards deep conversational recommendations.

Zujie Liang, Huang Hu, Can Xu, and Jian Miao. 2021. Learning neural templates for recommender dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7821–7833.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Richard Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics, pages 8639–8656.

Zeming Liu, Haifeng Wang, and Zheng-Yu Niu. 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.

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

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *ArXiv*, abs/2305.14536.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6908–6915.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909.

Jianheng Tang, Tiancheng Zhao, and Chenyan Xiong. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.

Jian Wang, Dongding Lin, and Wenjie Li. 2022. Follow me: Conversation planning for target-driven recommendation dialogue systems. *arXiv preprint arXiv:2208.03516*.

Jian Wang, Dongding Lin, and Wenjie Li. 2023a. Dialogue planning via brownian bridge stochastic process for goal-directed proactive dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 370–387.

Jian Wang, Dongding Lin, and Wenjie Li. 2023b. A target-driven planning approach for goal-directed dialog systems. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023c. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634.

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.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, and Jonathan Uesato. 2021. Ethical and social risks of harm from language models.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, and Hua Wu. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*, 25:3113–3124.

Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. 2023. Translating natural language to planning goals with large-language models.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data.

Zhitong Yang, Bo Wang, Jinfeng Zhou, and Yue Tan. 2022. TopKG: Target-oriented dialog via global planning on knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 745–755.

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.

Chenchen Ye, Lizi Liao, Suyu Liu, and Tat-Seng Chua. 2022. Reflecting on experiences for response generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5265–5273.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning.

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Robert Jankowski, Yanghua Xiao, and Deqing Yang. 2023. Distilling script knowledge from large language models for constrained language planning.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *ArXiv*, abs/2106.11520.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. 2023. Planning with large language models for code generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yizhe Zhang, Siqi Sun, Michel Galley, and Yen-Chun Chen. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345.

Peixiang Zhong, Yong Liu, Hongya Wang, and Chunyan Miao. 2020. Keyword-guided neural conversational model. In *AAAI Conference on Artificial Intelligence*.

Kun Zhou, Yuanhang Zhou, and Wayne Xin Zhao. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139. International Committee on Computational Linguistics.

A Prompt Details

A.1 Distill Plan (Stage 1)

A.1.1 GPT4-abstract

Conversation: ${**Conversation**}
Please summarize the conversation.
Summary:

A.1.2 Direct Prompt

Here is an example:
Conversation: ${**Conversation**}
Distilled Plan:${**Plan**}

Please distill the plan according to the target-driven conversation below. The plan shows the process of the conversation AI recommending the target movie to the user.
Conversation: ${**Conversation**}
Plan:

A.1.3 EnPL Prompt (CoT Prompt)

Here is an example:
Conversation: ${**Conversation**}
Distilled Plan: ${**Plan**}

Your task is to distill the plan according to the target-driven conversation below. The AI's goal is to recommend the target movie to the user. The plan shows the process of the conversation AI recommending the target movie to the user. The conversation between recommendation AI and the user is target-driven, gradually shifting the topic to the target movie. And the plan should be as short as possible to reflect the focus of the conversation. Attention to entities mentioned in the reservations dialogue. Only return the plan.
The following is the conversation you need to use in distilling plan:
Conversation: ${**Conversation**}
Plan:

A.2 Generate New Plan (Stage 2)

A.2.1 Direct Prompting

Please generate a conversation plan according to the "Target" and "User Setting" below. The AI's goal is to recommend the target movie to the user. The plan shows the process of the conversation AI recommending the target movie to the user.
Target: ${**Target**}
User Setting: ${**User Setting**}
Plan:

A.2.2 EnPL Prompt (CoT Prompting)

Examples:
Target: ${**Target**}
User Setting: ${**User Setting**}
Plan: ${**Plan**}

Your task is to generate a conversation plan according to the "Target" and "User Setting" below. The AI's goal is to recommend the target movie to the user. The plan shows the process of the conversation AI recommending the target movie to the user. The conversation process between conversation AI and the user is target-driven, gradually shifting the topic to the target movie. You can expand on the information you know to make the conversation process richer. You can refer to the Example above. Only return the plan. The following are the "Target" and "User Setting" you need to use in generating a new plan:
Target: ${**New Target**}
User Setting: ${**New User Setting**}
Plan:

A.3 Usage of Generated Plan

A.3.1 Prompt of Conversation Generation

The following is the prompt template we use the generated plan to guide ChatGPT to generate target-driven conversations. Table 10 shows an example of this process.

Here is an example:
Target: ${**Target**}
Plan: ${**Plan**}
Generated conversation: ${**Conversation**}

Your task is to create a movie recommendation conversation between a user and an AI recommender according to the Plan below. The AI's goal is to recommend the target movie to the user. Generate a conversation with as many topic changes as possible to generate more rounds of dialogue. Switch the topic to the target during the chat with the user. Make the conversation more like a real-life chat and be specific. In the example above, where User/AI represents whether the speaker is a User or an AI. Below is the Target and Plan you need to refer to generate conversation.
Target: ${**Target**}
Plan: ${**Plan**}
Generated conversation:

A.3.2 Prompt of Response Generation

The following is the prompt template we use the generated plan to guide ChatGPT to generate next-turn response. During self-chat simulation, we use our EnPL framework to generate plan turn by turn.

Your task is to generate the next-turn response according to the Plan and Context above. The Context is a part of movie recommendation conversation between a user and an AI recommender. The AI's goal is to recommend the target movie to the user. Generate a conversation with as many topic changes as possible to generate more rounds of dialogue. Switch the topic to the target during the chat with the user. Make the conversation more like a real-life chat and be specific. In the example above, where User/AI represents whether the speaker is a User or an AI.
Target: ${**Target**}
Context: ${**Context**}
Plan: ${**Plan**}
Next-turn response:

B An Example of Entity-consistency

Figure 4 shows the workflow of entity-consistency to filter distilled plans. The K_{plan} and K_{conv} are the lists of key entities extracted from the distilled plan and the original conversation using TextRank (Mihalcea and Tarau, 2004). Then, we calculate the consistency score between the plan and conversation using the Levenshtein distance algorithm.

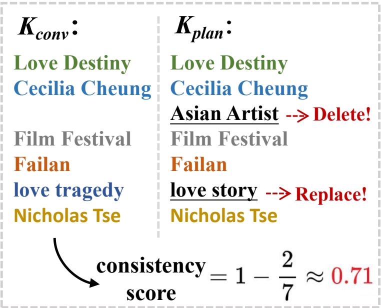


Figure 4: The workflow of entity-consistency to filter distilled plans.

C Details of Response Generation

C.1 Experimental Setup

Dataset We choose the DuRecDial 2.0 (Liu et al., 2021) dataset as appropriate for our experiments, which is a crowdsourced dataset of human-to-human dialogues in recommendation-oriented scenarios. The significant reason for using DuRecDial is that this dataset contains rich auxiliary information, such as movies or celebrities that users like, and even food preferences. This information can assist LLM in generating high-quality plans. In fact, we conducted comprehensive experiments on the ReDial dataset (Li et al., 2019), but due to the lack of auxiliary information, the results were not as expected. Another reason is the scarcity of manually constructed high-quality datasets in the target-driven dialogue domain, but our method provides a solution to address this issue. Additionally, we conducted evaluation experiments using the TGConv (Yang et al., 2022) dataset for multi-turn self-play simulations. The TGConv dataset contains high-quality open-domain dialogues on a variety of commonsense topics. Each dialogue is designed to direct the conversation towards a specific keyword or topic through coherent keyword transitions, which are categorized as either easy-to-reach or hard-to-reach based on their difficulty level.

Baselines For conversation generation, our baselines include: GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020), BART (Lewis et al., 2020), TCP-Dial (Wang et al., 2022), COLOR (Wang et al., 2023a), and TopKG (Yang et al., 2022). We choose these methods because they are highly relevant to our problem setting, and COLOR is currently the state-of-the-art model in our knowledge. In addition to guiding ChatGPT to generate conversations, we also conduct experiments on DialogGPT to make a fair comparison.

Evaluation Metrics Inspired by (Wang et al., 2023a), we adopt the same evaluation metrics, including perplexity (PPL), distinct (D-1/2) (Li et al., 2016), BLEU-n (B-1/2) (Papineni et al., 2002), wordlevel F1 and knowledge F1 (Know. F1) (Liu et al., 2020). To evaluate models' goal-directed performance, we use the goal success rate (Succ.) as the global evaluation metric. In DuRecDial 2.0 dataset, Succ. measures the proportion of correct target topic generation within the target turn and the two adjacent turns in the test set, as per Wang et al. (2023a). Additionally, we also use Coherence (Section 4.2) as another global evaluation metric, which measures the average contextual semantic similarity between the last utterance in the context and generated utterances.

C.2 Dialog-level Response Generation on TGConv

For the TGConv dataset, we perform self-play simulations, following Wang et al. (2023a); Yang et al. (2022), to simulate multi-turn conversations and compute the success rate of generating the target keyword within 8 turns.

As shown in Table 6, we find that guiding conversations to reach the target seemed challenging in all baseline open-domain chat environments. However, our EnPL w/ G achieved substantial improvements, generating more coherent discourse and

Model	**PPL (↓)**	**F1**	**B-1/2**	**D-1/2**	**Know. F1**	**Succ.**
GPT-2†	5.33	36.86	0.314 / 0.222	0.024 / 0.081	43.62	41.80
DialoGPT†	5.26	38.12	0.324 / 0.252	0.023 / 0.076	44.71	46.46
BART†	6.46	36.11	0.279 / 0.181	0.030 / 0.096	43.33	58.40
TCP-Dial†	5.88	34.46	0.293 / 0.201	0.027 / 0.091	45.75	60.49
COLOR† **w/ D**	5.22	43.14	0.371 / 0.277	0.024 / 0.073	57.89	73.20
EnPL w/ D	6.28	42.45	0.364 / 0.251	0.026/ 0.089	62.72	77.81
EnPL w/ C	**8.97**	**47.26**	**0.407 / 0.318**	**0.033/ 0.098**	**66.41**	**96.25**

Table 7: Automatic evaluation results of turn-level response generation on DuRecDial 2.0 dataset. Models marked with † are reported from Wang et al. (2023a). C and D are short for ChatGPT and DialoGPT, respectively.

shifting the topic to the target with a higher success rate. Under the guidance of our natural language planning, we can utilize LLM's rich domain knowledge and understanding ability to perform complex reasoning on the dialogue process to achieve targets. Other baselines, besides being limited by the generation space, make it difficult for keyword-based planning to describe a clear dialogue path, further reducing Coherence metrics.

C.3 Turn-level Response Generation on DuRecDial

Table 7 shows the results in DuRecDial 2.0. We can observe that plans in natural language form (our EnPL) have significant advantages over keyword-based plans in terms of the number of relevant entities and clarity. Firstly, our EnPL w/ ChatGPT exhibits a significant improvement in global success rate because our plan describes a complete path to achieve the target, rather than a few separate keywords. And, except for EnPL w/ ChatGPT, both BART and TCP-Dial outperform other models in D-1/2, as they generate fewer repeated words, resulting in more diversified utterances. Additionally, EnPL and COLOR achieve higher knowledge F1 scores because they are more likely to generate utterances with correct knowledge. In contrast, our approach outlines a clear and logically strong path, describing how to achieve the target step by step, making it easier for the model to generate high-quality conversations. Overall, our method shows significant improvement across all metrics. It indicates that, under the guidance of planning, LLM can better connect domain knowledge, dialogue scenarios, and targets, knowing when to discuss what content, thus guiding to achieving the target.

Category	**Number**
Plans	12,147
Utterances	8.7
Avg. length of Plans	78.5
Avg. length of Utterances	9
Entities	98,390
Unique entities	10,253
Avg. entities of Plans	8.1
Avg. entities of Utterances	0.93

Table 8: The statistics of our ConvPlan dataset.

D Details of Human Evaluation

We recruited 5 master students to serve as annotators for this project. We randomly selected 50 dialogue examples conversations guided by EnPL w/ ChatGPT on DuRecDial 2.0 and TGConv datasets, respectively. And we select 50 more examples from the DuRecDial 2.0 dataset. At least two different annotators rated each dialogue example. For a fair comparison, the examples were randomly renamed as "example-1", "example-2", and so forth. Referring to (Liu et al., 2020), we adopted the following metrics to evaluate the quality of dialogues: (1) appropriateness (Appr.), which measures whether the utterance responds to the dialogue context appropriately, and (2) informativeness (Info.), which measures whether the utterance is informative by making full use of the grounded knowledge. (3) proactivity (Proact.), which measures whether a model proactively leads the dialogue; (4) coherence (Coh.), which manually examines whether the whole dialogue is fluent, coherent, and smoothly transited; (5) target success (Succ.), which estimates whether a model effectively reaches the target. All the metrics are rated with the four-level Likert scale (Allen and Seaman, 2007) ranging from 0 to 3, with higher scores indicating better quality. Appendix A.3 shows the prompts used to guide LLM to generate dialogue, and Table 10 shows an example of this process.

E Details of ConvPlan dataset

Table 8 shows statistics on the average length of plans, number of utterances, etc.

F Case Study

Table 9 shows some cases generated from LLaMA2, COLOR, and our EnPL.

User Scenario	
Target Movie: Who Am I **User Setting:** liked movies: The Foreigner, Jackie Chan: My Story, Miracle seen movies: Chop Socky: Cinema Hong Kong liked music: Piano in the Attic rejected movies: Bruce Lee	
Methods	**Generated Plan**
LLaMA2	1. Ask the user about their favorite movie. 2. User discusses different genres of movies with AI. 3. AI mentions Jackie Chan as a popular action star. 4. Users express interest in Jackie Chan but lack understanding. 5. AI mentions that Jackie Chan received an accolade in a film festival. 6. User agrees and praises the actor. 7. AI recommends "Who Am I" as a movie that showcases Jackie Chan's martial arts skills. 8. User shows interest in the recommended movie. 9. Provide more details about "Who Am I" to persuade the user to give it a try. 10. User expresses gratitude.
COLOR	1. `[Action:]` Respond Q&A `[Topic:]` Jackie Chan: My Story `[Action:]` Chat about stars `[Topic:]` Jackie Chan `[Action:]` Movie recommendation `[Topic:]` Who Am I 2. `[Action:]` Chat about stars `[Topic:]` Jackie Chan `[Action:]` Movie recommendation `[Topic:]` Who Am I 3. (same as 2.) 4. `[Action:]` Movie recommendation `[Topic:]` Who Am I 5. (same as 4.) (too structured, hard to understand)
EnPL	1. User inquires about the star of the movie "Jackie Chan: My Story". 2. AI confirms that Jackie Chan is the star. 3. User expresses admiration for Jackie Chan. 4. AI mentions that Jackie Chan had won the Golden Horse Award for Best Actor. 5. User shows interest in Jackie Chan's achievements. 6. AI recommends another Jackie Chan movie, "Who Am I". 7. User shows interest in the recommended movie. 8. AI describes the recommended kung fu movie. 9. User finds the description interesting and decides to watch the movie. (more related entities and coherence)

Table 9: Generated plan examples in a new scenario. The plans are generated by different methods, including (a) LLaMA2, (b) COLOR, and (c) Ours (EnPL). The topics and topic-related entities that also appear in the domain knowledge are marked with underlines.

Plan
1. User asks about movie "Fly Me to Polaris".
2. AI replies that Cecilia Cheung stars the movie.
3. User asks for more information about Cecilia Cheung.
4. AI praises Cecilia Cheung's talent and mentions her awards in the Hong Kong Film Awards and Golden Bauhinia Awards.
5. AI recommends a movie called "Failan" starred by Cecilia Cheung, which is a touching story of a woman who marries a man she never met before.
6. User expresses interest but mentions that he has already seen it.
7. AI recommends another movie "The Bullet Vanishes" starring Nicholas Tse, who is also one of the user's favorite celebrities, describing it as a thrilling detective story set in 1930s Shanghai.
8. User agrees to watch.
9. AI encourages the user to watch it and expresses confidence that they will enjoy it
Generated Conversation
User: Who was the lead in the movie "Fly Me to Polaris"?
AI: That would be Cecilia Cheung.
User: Oh, Cecilia Cheung! Tell me more about her.
AI: She's an incredibly talented and versatile actress, known not just for her skills but also for her beauty. She's been recognized with awards at the Hong Kong Film Awards and the Golden Bauhinia Awards.
User: Impressive! Any particular movie you'd recommend?
AI: Definitely! There's a touching one called "Failan" where she plays a woman who marries a stranger. It's a really heartfelt story.
User: That sounds interesting, but I have watched it.
AI: Ah, got it. How about "The Bullet Vanishes"? It stars Nicholas Tse, someone you're a fan of, right? It's a thrilling detective story set in 1930s Shanghai.
User: Nicholas Tse! I'm definitely intrigued. I'll consider watching that.
AI: Great choice! I'm confident you'll enjoy it.
User: I'm going to grab some food now.
AI: Sure thing. Enjoy your meal! Feel free to come back anytime for more recommendations.

Table 10: Example of using the generated plan to guide LLM to generate a target-driven conversation. The target movie is marked with underlines.