Enhancing Goal-oriented Proactive Dialogue Systems via Dynamic Multi-dimensional Consistency Optimization

Didi Zhang, Yaxin Fan, Peifeng Li*, and Qiaoming Zhu
School of Computer Science and Technology, Soochow University, Suzhou, China
{ddzhang2023, yxfansuda}@stu.suda.edu.cn, {pfli, qmzhu}@suda.edu.cn

Abstract

Previous work on goal-oriented proactive dialogue systems frequently failed to address the multi-dimensional consistency issue between generated responses and key contextual elements (e.g., user profile, dialogue history, domain knowledge, and subgoal). To address this issue, we propose a novel Dynamic Multi-dimensional Consistency Reinforcement Learning (DMCRL) framework, which adaptively measures the impact of each consistency dimension on overall dialogue quality and provides targeted feedback to improve response quality. Experimental results on two datasets demonstrate that our DMCRL significantly improves the consistency of generated responses.

1 Introduction

Goal-oriented proactive dialogue systems (GPDS) are designed to guide user interactions toward specific objectives by dynamically generating contextually relevant responses (Wang et al., 2024a, 2023a). Unlike traditional dialogue systems that passively respond to user queries (Touvron et al., 2023; Achiam et al., 2024), GPDS actively plan and steer conversations along predefined paths, ensuring smooth and coherent progression. They are widely used in applications such as personalized recommendations (Fu et al., 2020; Liu et al., 2020, 2021), customer support (Katragadda, 2024), and medical consultations (Xu et al., 2024b), where it is crucial to achieve specific results interactively.

Existing work on GPDS has primarily focused on optimizing goal-oriented dialogue planning (Wang et al., 2023b, 2024a,b; Zhang et al., 2024), such as generating optimal paths to achieve specific goals. These paths are typically composed of <Action, Topic> pairs. As illustrated in Figure 1, a dialogue goal pathway might start from <Greetings | NULL>, progress through a series of subgoals, and ultimately reach <Music Recommendations | *Housten Love*>. However, the above methods often overlook the critical issue of multi-dimensional consistency between generated responses and key contextual elements (e.g., user profile, dialogue history, domain knowledge, and subgoal).

Figure 1 illustrates an example of a GPDS interaction, where the system needs to generate responses that are consistent with various predefined contextual elements; otherwise, the dialogue may become disjointed and ineffective. For example, in response S_1 ($SystemResponse$), the system fails to incorporate the user profile, resulting in a lack of personalization. In S_3, the system deviates from the target movie *Crossing Hennessy* in its recommendation. In S_5, the system ignores the user's previously expressed intent in the dialogue history of not wanting to watch a movie and continues to recommend one. Moreover, in S_7, the system incorrectly applies a review of the song *A Chinese Ghost Story* to *Housten Love*, resulting in factual inaccuracy. These inconsistencies disrupt the coherence of the dialogue, hinder the system's ability to guide the user toward the intended goal, and ultimately degrade the user experience.

Although some studies on another task, persona-based dialogue generation (Song et al., 2021; Zhou et al., 2023), have attempted to understand and improve consistency, they typically focus on single-dimensional consistency or assume that all consistency dimensions hold equal importance throughout the dialogue. They overlook the fact that the significance of different consistency dimensions varies dynamically as the conversation progresses. As a result, they fail to effectively address the challenge of balancing multi-dimensional consistency. Figure 1 illustrates an interaction case where all consistency dimensions are treated as equally important ($w/Static$). In this case, U_1 and U_2 excessively apply the user profile, leading to an abrupt mention of the user's location and repetitive ref-

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25283–25296
November 4-9, 2025 ©2025 Association for Computational Linguistics

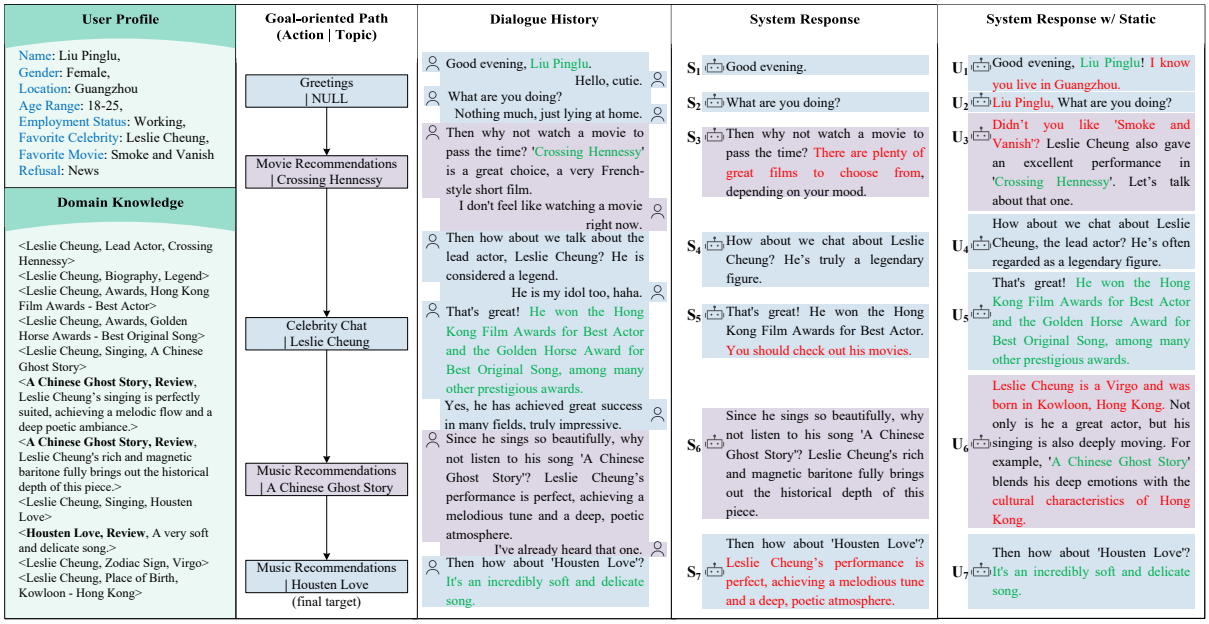


Figure 1: An example of GPDS from DuRecDial (Liu et al., 2020), where the system responses are generated by LLaMA3, with "w/ Static" indicating responses where all consistency dimensions are treated as equally important.

erences to the user's name. U_3 forcibly links the user's favorite movie to the recommended movie, even though they are unrelated. U_6 fails to prioritize the subgoal of recommending the current song and, instead, overly focuses on maintaining consistency with domain knowledge, providing excessive and unnecessary personal information about the singer, which leads to factual inaccuracies in the song description. While these dialogues perform well in terms of consistency across various dimensions, the overemphasis on treating all dimensions as equally important often disrupts the flow and coherence of the conversation.

To address these issues, we propose a Dynamic Multi-dimensional Consistency Reinforcement Learning (DMCRL) framework. DMCRL dynamically adjusts the attention weights assigned to different consistency dimensions as the dialogue progresses, enabling the system to maintain overall consistency while focusing on the most critical dimensions at each stage. Additionally, DMCRL introduces a counterfactual consistency reward mechanism, which simulates counterfactual scenarios by modifying predefined elements in the actual context, and updates the model on its ability to generate consistent and reasonable responses. Through this scenario contrast learning approach, DMCRL is able to more accurately identify multi-dimensional consistency discrepancies. The experimental results on two datasets show that our DMCRL significantly improves the consistency of responses across multiple dimensions.

2 Related Work

2.1 Goal-oriented Proactive Dialogue Systems

Previous GPDS guided dialogue generation by planning subgoal for each conversational turn, ensuring the conversation progresses towards a specific target. Most work focused on path planning and employed various techniques, such as CNN and biGRU-based methods (Liu et al., 2020), transformer-based goal-driven methods (Wang et al., 2022, 2024a), and Brownian bridge stochastic processes (Wang et al., 2023b). Additionally, they have explored unified Seq2Seq generation paradigms and prompt-based learning strategies (Deng et al., 2023). Graph-based models, including graph convolutional networks (Liu et al., 2023) and graph interaction methods (Zhang et al., 2024), as well as goal-constrained bidirectional planning (Wang et al., 2024b), have also been investigated. However, these methods often overlook the potential inconsistencies between the generated responses and key contextual elements, such as user profile, dialogue history, domain knowledge, and subgoal. Such inconsistencies can disrupt the flow of conversation or lead to task failure.

2.2 Consistency Generation

Previous studies in other tasks have attempted to address consistency issues in dialogue systems,

mainly focusing on single-dimensional consistency or treating all consistency dimensions equally. Zhou et al. (2023) improved coherence and consistency through over-sampling and post-evaluation, while Song et al. (2021) employed two BERT decoders to separately model response generation and consistency understanding. However, these methods are primarily aimed at persona-based dialogue generation and struggle to generalize to broader goal-oriented dialogue scenarios. Additionally, they overlook the fact that the importance of different consistency dimensions changes dynamically throughout the dialogue. Our work emphasizes the dynamic nature of multi-dimensional consistency, aiming to make the consistency optimization process more controllable.

3 Task Definition

Given a dialogue $D = \{U, H, K, G, R\}$, where U, H, K, G, and R represent the user profile, dialogue history, domain knowledge, goal-oriented path, and set of responses at each turn, respectively, GPDS consists of two primary subtasks: goal-oriented path planning and response generation.

Goal-oriented Path Planning This task aims to predict a sequence of subgoals $G = \{g_1, g_2, \ldots, g_T\}$ that guide the conversation toward achieving the final objective. At each turn t ($1 \leq t \leq T$), the model determines the next subgoal g_t based on the current dialogue state:

$$g_t = \arg\max_{g} P(g \mid U, K, H_{\leq t}, G_{<t}; \theta_{\text{plan}}) \quad (1)$$

where θ_{plan} represents the parameters of the planning model. In this stage, we adopt the same prediction model as TPNet (Wang et al., 2024a), ensuring effective goal progression. However, our main objective is to enhance the consistency of the generated responses.

Response Generation Given the user profile U, domain knowledge K, dialogue history $H_{\leq t}$, and current subgoal g_t, the response generation model at the t-th turn aims to produce a coherent and goal-aligned response r_t, forming part of the overall response set $R = \{r_1, r_2, \ldots, r_T\}$, as follows:

$$r_t = \arg\max_{r} P(r \mid U, K, H_{\leq t}, g_t; \theta_{\text{gen}}) \quad (2)$$

where θ_{gen} denotes model parameters. Our work primarily focuses on enhancing the alignment between the generated response and predefined elements (i.e., user profile, dialogue history, domain knowledge, and subgoal), thereby improving the overall performance of GPDS.

4 DMCRL Framework

To address the challenge of multi-dimensional consistency in GPDS, we propose the novel Dynamic Multi-dimensional Consistency Reinforcement Learning (DMCRL) framework. DMCRL dynamically balances the consistency requirements across four key dimensions: user profile, dialogue history, domain knowledge, and subgoal. As illustrated in Figure 2, the framework consists of two main stages: Supervised Fine-Tuning (SFT), and Dynamic Consistency Reinforcement Learning (DCRL) with an innovative reward function that incorporates counterfactual consistency rewards.

4.1 Supervised Fine-Tuning

In the SFT stage (Figure 2 Left), we fine-tune an open-source LLM on the dialogue dataset $S = \{D^i\}_{i=1}^{N}$, where each dialogue consists of multiple turns. The objective is to establish a strong baseline for response generation by minimizing the negative log-likelihood of the ground-truth responses at each turn as follows:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(U,K,H_{\leq t},g_t,r_t)\sim S} \sum_{t=1}^{T_i} \log P(r_t \mid U, K, H_{\leq t}, g_t; \theta_{\text{SFT}}) \quad (3)$$

where T_i denotes the number of turns in the i-th dialogue, r_t is the ground-truth response at the turn t, g_t is the subgoal at the turn t, and θ_{SFT} represents the parameters of the fine-tuned model.

4.2 Dynamic Consistency Reinforcement Learning

In the DCRL stage (Figure 2 Middle), we optimize the model to generate responses that are both consistent and well-balanced across multiple dimensions. The key innovations lie in the dynamic weight allocation and the counterfactual consistency rewards.

Dynamic Weight Allocation Given that the importance of consistency across dimensions (user profile, dialogue history, domain knowledge, and subgoal) can vary as the conversation progresses, the system must dynamically adjust the weight of each consistency dimension. This ensures that the generated response maintains overall consistency while also giving greater attention to the most relevant dimensions as the conversation progresses.

Leveraging ChatGPT's powerful annotation capability (Wang et al., 2023c; Xu et al., 2024a), we

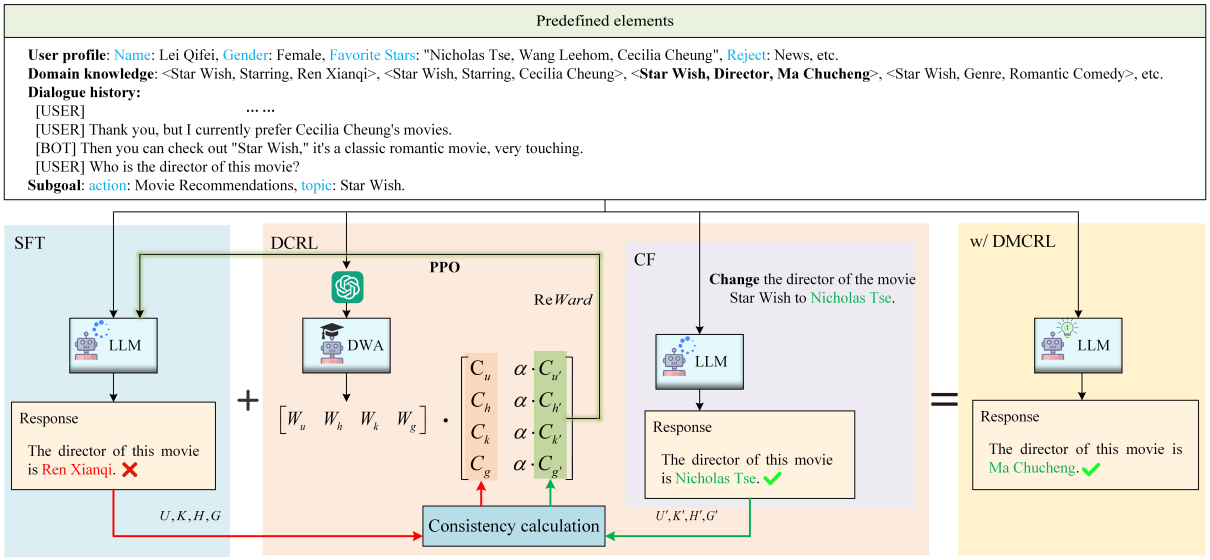


Figure 2: Overview of the DMCRL framework, where DWA, CF, and PPO represent Dynamic Weight Allocation, Counterfactual Scenario, and Proximal Policy Optimization, respectively.

automatically annotate turn-specific consistency weights using ChatGPT. The prompt used for annotation is in Appendix A.1, and the reliability of ChatGPT's annotations, based on human evaluations, is discussed in Appendix A.2. Then we train a Dynamic Weight Allocation (DWA) model that maps the predefined elements e, including user profile (u), dialogue history (h), domain knowledge (k), and subgoal (g), to a set of consistency weights $\mathbf{w} = \{w_u, w_h, w_k, w_g\}$ as follows:

$$\mathbf{w} = \text{DWA}(e; \theta_w) \tag{4}$$

where θ_w represents the parameters. DWA is trained by minimizing the mean squared error (MSE) loss between the predicted weights and the ChatGPT-annotated target weights $\hat{\mathbf{w}} = \{\hat{w}_u, \hat{w}_h, \hat{w}_k, \hat{w}_g\}$ on the annotated dialogue dataset $\tilde{S} = \{\tilde{D}^i\}_{i=1}^{N}$, where $\tilde{D}^i$ denotes a dialogue sample with annotated consistency weights.

$$\mathcal{L}_{\text{weight}} = \mathbb{E}_{(e,\hat{\mathbf{w}})\sim\tilde{S}} \sum_{d\in\mathcal{D}} \|w_d - \hat{w}_d\|^2 \tag{5}$$

where $\mathcal{D} = \{u, h, k, g\}$ represents the set of consistency dimensions.

Counterfactual Consistency Reward The counterfactual consistency reward C_{cf} improves the model's ability and stability in identifying and optimizing consistency biases by evaluating its ability to generate consistent responses under hypothetical scenarios and providing targeted feedback. The core idea is to simulate potential changes in the dialogue context and assess whether the model's responses remain consistent and reasonable under these perturbations. Specifically, we modify the predefined contextual elements e to construct a new counterfactual context e', such as adjusting the user's preferences, altering parts of the domain knowledge, or changing the current subgoal. Then, the model generates a new response r_t' based on e', and we evaluate whether it still maintains logical consistency, aligns with the overall context, and avoids introducing errors that violate global consistency due to local information changes.

This approach is inspired by causal inference (Pearl, 2010), where counterfactual analysis is used to understand how changes in input conditions affect outcomes. The counterfactual consistency reward is computed as follows:

$$C_{\text{cf}}(e', r_t') = \sum_{d\in\mathcal{D}} w_d \cdot C_d(e', r_t') \tag{6}$$

where w_d is the dynamically assigned weight for each consistency dimension d, and $C_d(e', r_t')$ denotes the consistency score between the generated response r_t' and the contextual element corresponding to dimension d, which is detailed in Section 4.3.

Adaptive Consistency Reward The reward function $C(e, r_t)$, the overall consistency score which integrates the dynamic weight allocation and counterfactual consistency evaluation, is defined as:

$$C(e, r_t) = \sum_{d\in\mathcal{D}} w_d \cdot C_d(e, r_t) + \alpha \cdot C_{\text{cf}}(e', r_t') \tag{7}$$

where $\alpha = 0.3$ is a scaling coefficient tuned on the development set that controls the impact of the counterfactual consistency score.

4.3 Consistency Score Computation

We design the consistency scoring method for each dimension to evaluate the system's consistency performance at different levels according to their definition, as detailed below:

User Profile The user profile consistency score measures whether the system response correctly incorporates relevant user profile information while avoiding content that the user has explicitly rejected. It is calculated as follows:

$$C_{\text{u}}(e, r_t) = \frac{\sum_{p_i \in P} \delta(r_t, p_i)}{|P|} - \lambda \cdot \frac{\sum_{n_j \in N} \delta(r_t, n_j)}{|N|} \tag{8}$$

where P is the set of key user profile elements (e.g., name, interests), and N is the set of content explicitly rejected by the user (e.g., movies). $\delta(r_t, x) = 1$ if the response r_t contains x, otherwise 0. λ is a penalty factor (set to 1), allowing C_{u} to range within $[-1, 1]$ so as to impose a stronger penalty when the response violates user rejections, unlike other scores which are bounded in $[0, 1]$.

Dialogue History The dialogue history consistency score measures whether the current response maintains semantic coherence with the previous conversation, avoiding abrupt or inconsistent transitions. We quantify the semantic coherence score using the TF-IDF method (Sparck Jones, 1972) combined with cosine similarity as follows:

$$C_{\text{h}}(e, r_t) = \frac{\mathbf{r_t} \cdot \mathbf{h}}{\|\mathbf{r_t}\| \|\mathbf{h}\|} \tag{9}$$

where $\mathbf{r_t}$ is the current response text, and $\mathbf{h}$ is the dialogue history.

Domain Knowledge The knowledge consistency score measures whether the system response correctly utilizes the knowledge base. We evaluate performance using the Knowledge F_1 score, as defined in Algorithm 1 in Appendix B.

Subgoal The subgoal consistency score evaluates whether the system response aligns with the current dialogue subgoal as follows:

$$C_{\text{g}}(e, r_t) = \mathbb{I}(t \subseteq r_t) \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which outputs 1 if the response r_t contains the current subgoal topic t, and 0 otherwise.

4.4 Optimization

To effectively optimize our consistency-driven reward under dynamic and potentially high-variance conditions, we adopt Proximal Policy Optimization (PPO) (Schulman et al., 2017) in combination with Generalized Advantage Estimation (GAE) (Schulman et al., 2018), balancing stability and sample efficiency. The objective is to maximize the expected cumulative reward as follows:

$$\mathcal{L}_{\text{RL}} = \mathbb{E}_{(e,r_t)\sim\pi}\Big[\hat{A}(e, r_t) - \beta \cdot \text{KL} \left(\pi(r_t \mid e) \| \pi_{\text{SFT}}(r_t \mid e)\right)\Big] \tag{11}$$

where $\hat{A}(e, r_t)$ is the advantage estimate obtained using GAE, π is the current policy, π_{SFT} is the SFT policy, and β is a hyperparameter controlling the KL divergence penalty.

5 Experimentation

5.1 Experimental Settings

Datasets We conduct experiments on two benchmark datasets: DuRecDial (Liu et al., 2020) and DuRecDial 2.0 (Liu et al., 2021). For data preprocessing and partitioning, we follow the same strategy as previous work (Wang et al., 2024a; Zhang et al., 2024) to ensure a fair comparison. The details of the datasets are provided in Appendix C.

Baselines We first compare our approach with two strong baselines: **MGNN** (Liu et al., 2023) effectively controls the dialogue flow by constructing a hierarchical goal graph and leveraging GCNs, while **GIGF** (Zhang et al., 2024) adopts a goal interaction graph framework based on heterogeneous graph attention networks to dynamically model multi-level goal relations. Additionally, to ensure a fair comparison, we evaluate Supervised Fine-Tuning baselines on five popular LLMs: **LLaMA3**, **Mistral**, **Phi-3.5**, **Qwen2.5** and **DeepSeek**. The specific model configurations can be found in Appendix D. This allows us to validate the effectiveness of DMCRL in a more competitive setting.

Evaluation Metrics We adopt the evaluation metrics used in previous work (Wang et al., 2024a). Word-level F_1 (W F_1) measures the exact word overlap between the generated and reference responses, reflecting the accuracy of the response. BLEU-2 (Papineni et al., 2002) evaluates the 2-gram overlap between the generated and reference responses. Distinct (Dist-2) (Li et al., 2016) measures the diversity of the generated responses. Knowledge F_1 (K F_1) (Liu et al., 2020) evaluates the ability of the model to effectively utilize domain knowledge. Goal Success Rate (Succ) (Wang et al., 2024a) reflects the model's ability to achieve the final target action and topic within the dialogue.

Method	W F_1	BLEU-2	Dist-2	K F_1	Succ
MGNN	43.50	0.274	0.064	45.00	-
GIGF	47.52	0.348	0.078	56.02	-
$\text{LLaMA3}_{(8B)}$	48.92	0.339	0.078	57.87	75.62
LLaMA3 w/ Static	$51.32_{\uparrow 2.40}$	$0.341_{\uparrow 0.002}$	$0.081_{\uparrow 0.003}$	$59.21_{\uparrow 1.34}$	$77.15_{\uparrow 1.53}$
LLaMA3 w/ DMCRL	$\mathbf{52.81}_{\uparrow 3.89}$	$0.349_{\uparrow 0.010}$	$0.082_{\uparrow 0.004}$	$\mathbf{60.76}_{\uparrow 2.89}$	$\mathbf{78.68}_{\uparrow 3.06}$
$\text{Mistral}_{(7B)}$	47.21	0.335	0.080	56.32	71.88
Mistral w/ DMCRL	$50.05_{\uparrow 2.84}$	$0.344_{\uparrow 0.009}$	$0.084_{\uparrow 0.004}$	$59.12_{\uparrow 2.80}$	$75.85_{\uparrow 3.97}$
$\text{Phi3.5}_{(3.8B)}$	46.63	0.326	0.084	55.29	73.12
Phi3.5 w/ DMCRL	$50.45_{\uparrow 3.82}$	$0.343_{\uparrow 0.017}$	$\mathbf{0.087}_{\uparrow 0.003}$	$58.19_{\uparrow 2.90}$	$76.14_{\uparrow 3.02}$
$\text{Qwen2.5}_{(7B)}$	48.72	0.342	0.082	55.39	72.91
Qwen2.5 w/ DMCRL	$51.56_{\uparrow 2.84}$	$\mathbf{0.350}_{\uparrow 0.008}$	$0.085_{\uparrow 0.003}$	$59.23_{\uparrow 3.84}$	$75.84_{\uparrow 2.93}$
$\text{DeepSeek}_{(1.5B)}$	46.26	0.331	0.084	55.47	74.03
DeepSeek w/ DMCRL	$49.47_{\uparrow 3.21}$	$0.342_{\uparrow 0.011}$	$0.085_{\uparrow 0.001}$	$58.66_{\uparrow 3.19}$	$77.13_{\uparrow 3.10}$

Table 1: Experimental results on the Chinese DuRecDial. The parameter sizes of the models are annotated as subscripts adjacent to the model names.

Method	W F_1	BLEU-2	Dist-2	K F_1	Succ
MGNN	36.75	0.194	0.073	31.32	-
GIGF	38.35	0.241	0.089	66.91	-
$\text{LLaMA3}_{(8B)}$	38.42	0.233	0.092	55.21	52.14
LLaMA3 w/ Static	$40.81_{\uparrow 2.39}$	$0.245_{\uparrow 0.012}$	$\mathbf{0.095}_{\uparrow 0.003}$	$63.54_{\uparrow 8.33}$	$53.48_{\uparrow 1.34}$
LLaMA3 w/ DMCRL	$41.17_{\uparrow 2.75}$	$0.247_{\uparrow 0.014}$	$\mathbf{0.095}_{\uparrow 0.003}$	$65.87_{\uparrow 10.66}$	$54.83_{\uparrow 2.69}$
$\text{Mistral}_{(7B)}$	39.09	0.237	0.090	58.89	50.85
Mistral w/ DMCRL	$\mathbf{41.85}_{\uparrow 2.76}$	$0.245_{\uparrow 0.008}$	$0.092_{\uparrow 0.002}$	$66.57_{\uparrow 7.68}$	$53.43_{\uparrow 2.58}$
$\text{Phi3.5}_{(3.8B)}$	40.64	0.245	0.094	54.56	53.47
Phi3.5 w/ DMCRL	$41.36_{\uparrow 0.72}$	$\mathbf{0.253}_{\uparrow 0.008}$	$\mathbf{0.095}_{\uparrow 0.001}$	$\mathbf{67.14}_{\uparrow 12.58}$	$54.13_{\uparrow 0.66}$
$\text{Qwen2.5}_{(7B)}$	38.25	0.239	0.087	56.20	53.14
Qwen2.5 w/ DMCRL	$40.97_{\uparrow 2.72}$	$0.249_{\uparrow 0.010}$	$0.088_{\uparrow 0.001}$	$63.85_{\uparrow 7.65}$	$\mathbf{55.77}_{\uparrow 2.63}$
$\text{DeepSeek}_{(1.5B)}$	38.11	0.237	0.091	58.38	51.03
DeepSeek w/ DMCRL	$41.08_{\uparrow 2.97}$	$0.241_{\uparrow 0.004}$	$0.093_{\uparrow 0.002}$	$66.37_{\uparrow 7.99}$	$54.01_{\uparrow 2.98}$

Table 2: Experimental results on the English DuRecDial 2.0.

Implementation Details Appendix E for details.

5.2 Experimental Results

The experimental results on the Chinese DuRecDial and English DuRecDial 2.0 are summarized in Tables 1 and 2. The results show that DMCRL consistently improves performance across different model architectures, achieving significant gains in key metrics such as W F_1, BLEU-2, K F_1, and Goal Success Rate, while maintaining competitive performance on Dist-2. These findings highlight the fact that DMCRL can enhance multi-dimensional consistency and overall dialogue quality.

We observe that the five LLMs demonstrate competitive performance compared to the SOTA baselines MGNN and GIGF, benefiting from their superior language understanding and generation capabilities. Although they are already capable of capturing the complex patterns of dialogues, our targeted consistency-enhancing method, DMCRL, can still further improve their performance.

Specifically, the improvements in W F_1 and BLEU-2 indicate that DMCRL can generate responses that better match reference responses in terms of lexical overlap and semantic relevance. The substantial gains in K F_1 highlight DMCRL's ability to enhance factual accuracy by reinforcing domain knowledge consistency. Furthermore, the increase in Goal Success Rate shows DMCR can guide the model toward achieving dialogue goals by addressing inconsistencies between responses and subgoals. Although the improvements in Dist-2 are relatively modest, they indicate that DMCRL maintains response diversity while enhancing consistency, ensuring that generated responses remain engaging and contextually appropriate.

To validate the idea that the importance of consistency across dimensions needs to be dynamically adjusted as the conversation progresses, we introduced a version of DMCRL without DWA, where all consistency dimensions are treated with equal importance, as a static consistency correction

Method	W F_1	BLEU-2	Dist-2	K F_1	Succ
DMCRL	52.81	0.349	0.082	60.76	78.68
w/o UP	52.47	0.344	0.081	60.23	78.05
w/o DH	52.13	0.342	0.082	59.92	77.57
w/o DK	51.23	0.341	0.081	58.48	76.95
w/o SG	51.72	0.342	0.080	59.89	75.64
w/o CF	50.37	0.344	0.081	59.31	77.17
w/o DWA	51.32	0.341	0.081	59.21	77.15

Table 3: Ablation results using LLaMA3 on DuRecDial, where UP, DH, DK, SG, CF, and DWA refer to User Profile, Dialogue History, Domain Knowledge, Subgoal, Counterfactual Reward, and Dynamic Weight Allocation, respectively.

method ($w/Static$) for comparison. The results for LLaMA3 show that, compared to $w/Static$, the full DMCRL outperforms across all metrics. This indicates that dynamically adjusting the importance weights of consistency dimensions helps the model adapt more flexibly to different conversational contexts, leading to more effective response generation. Similar trends are also observed on other models in Appendix F.

5.3 Ablation Study

We conducted an ablation study on DuRecDial to assess the contribution of different components to model performance and the results using LLaMA3 as backbone are shown in Table 3. Overall, the removal of any components led to a significant decrease in performance, indicating the critical importance of each consistency module.

Specifically, removing User Profile (w/o UP) and Dialogue History (w/o DH) both led to a decline, suggesting that ensuring generated responses align with the user's historical dialogue and preferences helps the model better understand user needs and generate appropriate responses. Additionally, removing Domain Knowledge (w/o DK) caused a significant drop in Knowledge F_1, highlighting the crucial role of enhancing domain knowledge consistency for factual accuracy in generated responses. Removing Subgoal (w/o SG) also led to a noticeable reduction in Goal Success Rate, implying that improving subgoal consistency can help the model make step-by-step progress in complex tasks and better achieve dialogue objectives.

Moreover, removing the counterfactual reward (w/o CF) led to a decline in all metrics, indicating that counterfactual scenarios help the model better identify and optimize consistency biases through contrastive learning. Finally, removing Dynamic Weight Allocation (w/o DWA) also resulted in performance degradation, which supports the idea that the importance of consistency across dimensions dynamically changes as the conversation progresses. Additionally, it can be observed that the change in the Distinct metric is relatively minor, reflecting that DMCRL can enhance consistency without compromising the diversity of the generated content.

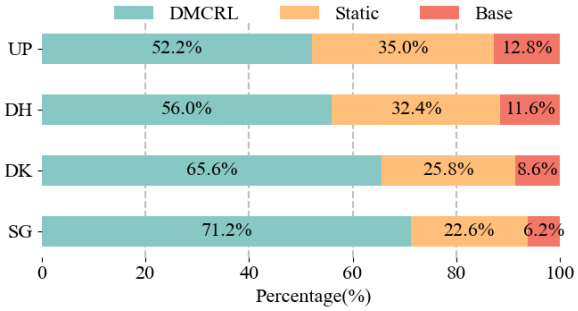


Figure 3: Three-way evaluation results for LLaMA3 w/ DMCRL vs. LLaMA3 w/ Static vs. LLaMA3

5.4 Human Evaluation on Consistency

To evaluate the effectiveness of DMCRL in enhancing multi-dimensional consistency, we conducted a human evaluation on DuRecDial using a three-way comparison. Specifically, we compared the consistency of responses generated by the baseline, a model with static consistency correction, and a model with DMCRL (dynamic consistency correction), focusing on their alignment with predefined elements such as user profile, dialogue history, domain knowledge, and subgoal. A total of 500 sets of responses were randomly selected for evaluation, with the constraint that the three responses in each set were mutually distinct. Annotators were instructed to choose the best response among the three anonymous options. The human evaluation criteria can be found in Appendix G. The evaluation results on LLaMA3 are shown in Figure 3.

The results show that the baseline still achieves a certain level of win rate, indicating it already shows a reasonable degree of consistency and can generate appropriate responses in some scenarios. However, the models with static consistency correction and DMCRL outperform the baseline across all dimensions, with DMCRL achieving the highest win rates. The evaluation results for other LLMs are shown in Appendix G, where similar trends are observed. This suggests that DMCRL, by imposing stronger dynamic consistency constraints, significantly improves response quality while maintaining dialogue coherence, enabling the model to

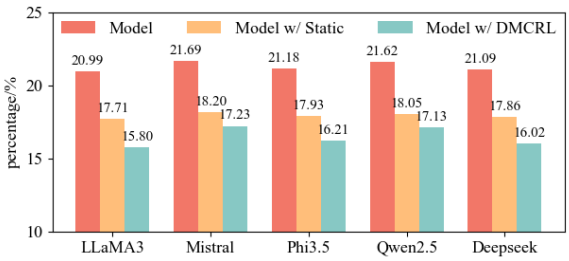


Figure 4: Subgoal failure rates on DuRecDial.

generate responses that better align with user needs and dialogue goals.

5.5 SubGoals Failure Analysis

A core requirement of GPDS is to progressively guide the conversation toward the final target by accomplishing intermediate subgoals. To evaluate whether the model exhibits strong subgoal consistency and guiding ability, we analyze its performance in subgoal completion. As shown in Figure 4, the results demonstrate that models without DMCRL often overlook key information related to the current subgoal. This lack of consistency may lead to fragmented dialogue flow, reducing task completion efficiency and degrading user experience. Although $w/Static$ improves consistency to some extent, its overall performance still lags behind the full DMCRL. In contrast, DMCRL dynamically adjusts consistency, allowing the model to more flexibly adapt to the goal requirements at different stages of the dialogue, significantly enhancing subgoal consistency and promoting the step-by-step achievement of the dialogue objective. Similar improvements are observed on DuRecDial 2.0, with details provided in Appendix H.

5.6 Case Study

To further analyze the impact of DMCRL on dialogue consistency, we conducted a case study comparing system responses before and after applying DMCRL. As shown in Table 4, DMCRL significantly enhances response consistency across multiple dimensions.

In terms of consistency with user profile ($Sample1$), the response without DMCRL lacks personalization, whereas DMCRL correctly incorporates the user's name, demonstrating a stronger awareness of user attributes. Regarding consistency with dialogue history ($Sample2$), the model without DMCRL overlooks the user's previously stated preference of not wanting to watch a movie. In contrast, DMCRL generates a response that aligns with the user's intent, ensuring contextual coherence and

Type	System Response
Sample 1: Consistency with the user profile	
w/o DMCRL	Good evening.
w/ DMCRL	Good evening, Liu Pinglu.
Sample 2: Consistency with the dialogue history	
w/o DMCRL	That's great! He won the Hong Kong Film Awards for Best Actor. You should check out his movies.
w/ DMCRL	That's amazing! He received the Hong Kong Film Awards for Best Actor and the Golden Horse Award for Best Original Song, along with numerous other prestigious honors.
Sample 3: Consistency with the domain knowledge	
w/o DMCRL	Then how about *Housen Love*? Leslie Cheung's performance is perfect, achieving a melodious tune and a deep, poetic atmosphere.
w/ DMCRL	Then how about *Housen Love*? It's a very soft and delicate song.
Sample 4: Consistency with the subgoal	
w/o DMCRL	Then why not watch a movie to pass the time? There are plenty of great films to choose from, depending on your mood.
w/ DMCRL	Then why not watch a movie to pass the time? You could check out *Crossing Hennessy*, a beautifully crafted short film with a touch of French style.

Table 4: A case generated by LLAMa3.

avoiding contradictions. For consistency with domain knowledge ($Sample3$), the response without DMCRL mistakenly applies a review of a different song to the current recommendation. After applying DMCRL, the model correctly retrieves and utilizes factual knowledge specific to the recommended song. In terms of consistency with subgoal ($Sample4$), the model without DMCRL provides a generic suggestion without explicitly recommending *Crossing Hennessy*, whereas DMCRL generates a more precise recommendation, effectively guiding the user toward the intended goal.

The appendix I provides a comparative case study with the static consistency correction method, further illustrating the advantages of DMCRL's dynamic consistency correction.

6 Conclusion

In this paper, we propose a Dynamic Multi-dimensional Consistency Reinforcement Learning framework DMCRL. By introducing dynamic weight allocation and counterfactual consistency verification mechanisms, DMCRL effectively improves and balances consistency across various dimensions. Extensive experiments on DuRecDial and DuRecDial 2.0 demonstrate that DMCRL significantly enhances response quality, factual accuracy, and goal success rate.

Limitations

Our proposed DMCRL framework has two primary limitations. First, relying on ChatGPT-based automatic annotation for consistency weight allocation may introduce potential biases inherent to large language models. Although we have demonstrated a high level of agreement between ChatGPT and human annotators, model-generated annotations may still overlook subtle aspects of importance that human evaluation could better capture. Second, while DMCRL dynamically balances consistency dimensions, it does not explicitly address potential conflicts between dimensions (e.g., when satisfying user profile information contradicts dialogue history). Future work could explore conflict resolution mechanisms to further enhance the robustness of the framework.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems. *ACM Trans. Inf. Syst.*, 41(3).

Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2020. Tutorial on Conversational Recommendation Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, page 751–753.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the Tenth International Conference on Learning Representations*, pages 1–13.

Vamsi Katragadda. 2024. Leveraging intent detection and generative ai for enhanced customer support. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5(1):109–114.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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, San Diego, California. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

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.

Zeming Liu, Ding Zhou, Hao Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, Ting Liu, and Hui Xiong. 2023. Graph-Grounded Goal Planning for Conversational Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4923–4939.

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 of the Association for Computational Linguistics*, pages 311–318.

Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment*, pages 39–58.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. High-dimensional continuous control using generalized advantage estimation. *Preprint*, arXiv:1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data. 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 167–177, Online. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023a. Target-oriented Proactive Dialogue Systems with Personalization: Problem Formulation and Dataset Curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143.

Jian Wang, Dongding Lin, and Wenjie Li. 2022. Follow Me: Conversation Planning for Target-driven Recommendation Dialogue Systems. *Preprint*, arXiv:2208.03516.

Jian Wang, Dongding Lin, and Wenjie Li. 2023b. 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. 2024a. A Target-Driven Planning Approach for Goal-Directed Dialog Systems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10475–10487.

Jian Wang, Dongding Lin, and Wenjie Li. 2024b. Target-constrained Bidirectional Planning for Generation of Target-oriented Proactive Dialogue. *ACM Trans. Inf. Syst.*, 42(5).

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *Proceedings of the Twelfth International Conference on Learning Representations*, pages 1–22.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024b. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6796–6814.

Xiaotong Zhang, Xuefang Jia, Han Liu, Xinyue Liu, and Xianchao Zhang. 2024. A Goal Interaction Graph Planning Framework for Conversational Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19578–19587.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. SimOAP: Improve Coherence and Consistency in Persona-based Dialogue Generation via Over-sampling and Post-evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9945–9959.

A ChatGPT-based Weight Annotation

A.1 Prompt for Annotation

We use ChatGPT [1] to annotate consistency weights at the turn level in dialogues. The goal is to evaluate the importance of consistency across the predefined elements such as user profile, dialogue history, domain knowledge, and subgoal for each assistant response. The prompt is as follows:

> You are given four contextual elements: a user profile, dialogue history, relevant domain knowledge, and a current subgoal, along with an assistant's response. Your task is to evaluate how important it is for the assistant's response to remain consistent with each of the four elements.
> Please assign an importance score from 1 (least important) to 5 (most important) to each dimension: User Profile, Dialogue History, Domain Knowledge, and Subgoal.
> Think step-by-step:
> 1.Identify what kind of information is necessary to generate a helpful and coherent response.
> 2.Determine which elements are most crucial for maintaining consistency at this specific turn.
> Focus only on the relative importance of each dimension's consistency. Do not assess the quality or correctness of the response itself.
> Return your output as a JSON object with four keys: "User Profile", "Dialogue History", "Domain Knowledge", and "Subgoal".
>
> [Start of Predefined Elements]
> $ {User Profile}
> $ {Dialogue History}
> $ {Domain Knowledge}
> $ {Subgoal}
> [End of Predefined Elements]
>
> [Start of the Assistant's Response]
> $ {Response}
> [End of the Assistant's Response]

[1] The version used is gpt-4o-2024-11-20.

A.2 Reliability Evaluation of ChatGPT Annotation

To assess the reliability of the consistency weights annotated by ChatGPT, we conducted a human evaluation on the turn-level annotation results. Specifically, we randomly sampled 500 annotated instances from the dataset and invited three graduate students specializing in natural language processing to independently assign importance scores across four predefined consistency dimensions: User Profile, Dialogue History, Domain Knowledge, and Subgoal. During the annotation process, the human annotators strictly followed the same guidelines used by ChatGPT (refer to Appendix A.1). We employed two complementary strategies to quantitatively measure the agreement between ChatGPT and human annotations:

First, we computed the **average absolute difference** (AvgDiff) between ChatGPT's scores and human scores. For each instance i, we calculated the absolute difference for each dimension d, averaged across the three annotators, and then averaged across all four dimensions:

$$\text{AvgDiff}_i = \frac{1}{4}\sum_{d=1}^{4}\left(\frac{1}{3}\sum_{h=1}^{3}|\text{GPT}_{i,d} - \text{Human}_{i,d,h}|\right) \quad (12)$$

where $\text{GPT}_{i,d}$ denotes ChatGPT's score for the instance i and dimension d, and $\text{Human}_{i,d,h}$ denotes the score given by the h-th human annotator. Based on this calculation, 84.2% (421/500) of the instances have an AvgDiff_i no greater than 1, indicating a high level of agreement in absolute scoring.

Second, considering that our focus is on the **relative importance across dimensions** rather than the absolute scores, we further computed the Spearman rank correlation coefficient between ChatGPT's and human annotators' importance rankings for each instance. The Spearman correlation measures the consistency of the ranking order among the four dimensions. The average Spearman correlation across the 500 samples is $\rho = \mathbf{0.87}$, demonstrating strong alignment in ranking preferences between ChatGPT and human annotators.

These results confirm that ChatGPT can reliably assess the turn-level relative importance of consistency dimensions, providing a sound basis for automated consistency weight annotation in reinforcement learning frameworks.

Dataset	Annotators	#Dialogue	#Utterance
DuRecDial	Crowd workers	8,004	126,186
DuRecDial 2.0	Human experts	6,080	98,719

Table 5: Statistics of DuRecDial (Chinese) and DuRecDial 2.0 (English).

B Knowledge Consistency Computation

Algorithm 1 outlines the computation of the Knowledge F_1 score used to assess knowledge consistency. This metric is based on the overlap between knowledge triples mentioned in the generated response and those in the reference response. It computes precision, recall, and their harmonic mean to quantify how well the system utilizes domain knowledge.

Algorithm 1 Knowledge F_1 Computation

SET:
Domain knowledge triples: $T\_triples$
Triples in generated responses: $G\_triples$
Triples in reference responses: $R\_triples$

1: Compute hit count:

$$\text{hit} = |G\_triples \cap R\_triples| \quad (13)$$

2: Compute precision:

$$\text{P} = \frac{\text{hit}}{|G\_triples|} \quad (14)$$

3: Compute recall:

$$\text{R} = \frac{\text{hit}}{|R\_triples|} \quad (15)$$

4: Compute Knowledge F_1 score:

$$Knowledge\ F_1 = 2 \times \frac{(\text{P} \times \text{R})}{\text{P} + \text{R}} \quad (16)$$

5: **return** Knowledge F_1;

C Datasets

We conduct experiments on two benchmark datasets: DuRecDial (Liu et al., 2020) and DuRecDial 2.0 (Liu et al., 2021). Both datasets are designed for goal-oriented proactive dialogue systems and provide structured annotations. Dataset statistics are shown in Table 5.

D Model Configurations

To ensure a fair comparison and fully validate the effectiveness of our method, we selected five

Method	W F_1	BLEU-2	Dist-2	K F_1	Succ
Mistral w/ Static	49.63	0.339	0.082	58.69	74.33
Mistral w/ DMCRL	$50.05_{\uparrow 0.42}$	$0.344_{\uparrow 0.005}$	0.084	$59.12_{\uparrow 0.43}$	$75.85_{\uparrow 1.52}$
Phi3.5 w/ Static	49.02	0.339	0.087	57.74	75.67
Phi3.5 w/ DMCRL	$50.45_{\uparrow 1.43}$	$0.343_{\uparrow 0.004}$	0.087	$58.19_{\uparrow 0.45}$	$76.14_{\uparrow 0.47}$
Qwen2.5 w/ Static	50.14	0.346	0.085	58.81	75.38
Qwen2.5 w/ DMCRL	$51.56_{\uparrow 1.42}$	$0.350_{\uparrow 0.004}$	0.085	$59.23_{\uparrow 0.42}$	$75.84_{\uparrow 0.46}$
DeepSeek w/ Static	48.71	0.335	0.085	57.20	76.88
DeepSeek w/ DMCRL	$49.47_{\uparrow 0.76}$	$0.342_{\uparrow 0.007}$	0.085	$58.66_{\uparrow 1.46}$	$77.13_{\uparrow 0.25}$

Table 6: Experimental results on the Chinese DuRecDial.

Method	W F_1	BLEU-2	Dist-2	K F_1	Succ
Mistral w/ Static	40.47	0.241	0.093	64.23	52.14
Mistral w/ DMCRL	$41.85_{\uparrow 1.38}$	$0.245_{\uparrow 0.004}$	0.092	$66.57_{\uparrow 2.34}$	$53.43_{\uparrow 1.29}$
Phi3.5 w/ Static	41.02	0.249	0.097	63.85	53.81
Phi3.5 w/ DMCRL	$41.36_{\uparrow 0.34}$	$0.253_{\uparrow 0.004}$	0.095	$67.14_{\uparrow 3.29}$	$54.13_{\uparrow 0.32}$
Qwen2.5 w/ Static	39.63	0.244	0.087	60.52	54.46
Qwen2.5 w/ DMCRL	$40.97_{\uparrow 1.34}$	$0.249_{\uparrow 0.005}$	0.088	$63.85_{\uparrow 3.33}$	$55.77_{\uparrow 1.31}$
DeepSeek w/ Static	40.27	0.240	0.093	63.14	53.88
DeepSeek w/ DMCRL	$41.08_{\uparrow 0.81}$	$0.241_{\uparrow 0.001}$	0.093	$66.37_{\uparrow 3.23}$	$54.01_{\uparrow 0.13}$

Table 7: Experimental results on the English DuRecDial 2.0.

widely used open-source large language models (LLMs) for Supervised Fine-Tuning, including: LLaMA3[2], Mistral[3], Phi-3.5[4], Qwen2.5[5], and DeepSeek[6]. These models have demonstrated strong performance across various tasks and datasets, and they represent the state-of-the-art in the field of large language models.

E Implementation Details

All experiments were conducted using the LLaMA-Factory framework[7] to fine-tune the LoRA (Low-Rank Adaptation) modules (Hu et al., 2022) of the models. The rank r and scaling parameter α for LoRA were set to 8 and 16, respectively. We used a learning rate of 1.0×10^{-4}, trained for 6 epochs, and employed a batch size of 4 with gradient accumulation set to 32. The maximum sequence length was set to 2048 tokens. All experiments were performed on two NVIDIA A100 GPUs.

[2] https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat

[3] https://huggingface.co/shenzhi-wang/Mistral-7B-v0.3-Chinese-Chat

[4] https://huggingface.co/microsoft/Phi-3.5-mini-instruct

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[6] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

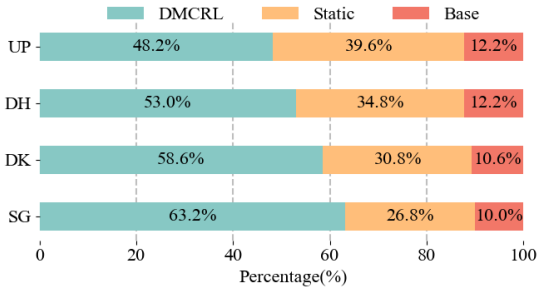
[7] https://github.com/hiyouga/LLaMA-Factory

F Comparison of Dynamic and Static Consistency Correction

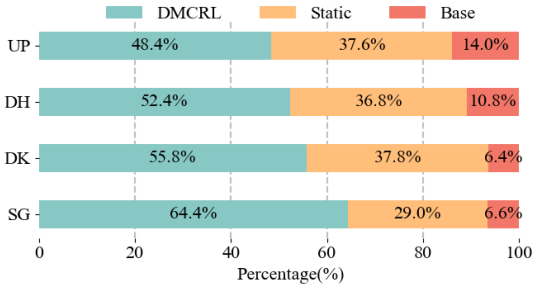
To further support the idea that the importance of consistency across dimensions needs to be dynamically adjusted as the dialogue progresses, we extend the comparison between the full DMCRL and its static variant (w/o DWA) to additional LLMs beyond LLaMA3. The results in Tables 6 and 7 consistently demonstrate that, compared to the static variant, the full DMCRL achieves superior performance across key metrics such as W F_1, BLEU-2, K F_1, and Goal Success Rate. This consistent advantage highlights the effectiveness of dynamically adjusting consistency weights, enabling the model to better adapt to varying subgoal demands, contextual shifts, and user preferences throughout the dialogue.

G Human Consistency Evaluation

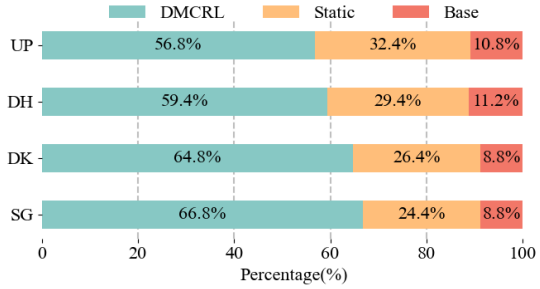
We randomly selected 500 sets of system responses, each comprising outputs from (1) the baseline model, (2) the model with static consistency correction (i.e., DMCRL w/o DWA, where all consistency dimensions are treated with equal importance), and (3) the model with DMCRL, with the constraint that the three responses in each set were mutually distinct. Three graduate students specializing in natural language processing were invited



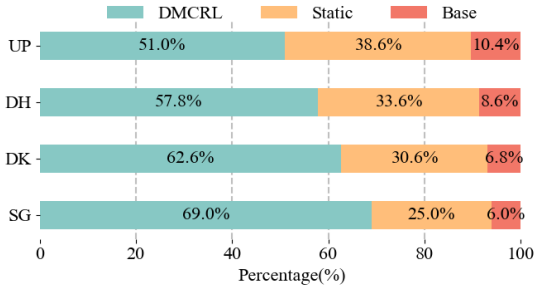
(a) Mistral w/ DMCRL vs. Mistral w/ Static vs. Mistral



(b) Phi3.5 w/ DMCRL vs. Phi3.5 w/ Static vs. Phi3.5



(c) Qwen2.5 w/ DMCRL vs. Qwen2.5 w/ Static vs. Qwen2.5



(d) DeepSeek w/ DMCRL vs. DeepSeek w/ Static vs. DeepSeek

Figure 5: Three-way evaluation results for Mistral, Phi3.5, Qwen2.5, and DeepSeek. Abbreviations: UP (user profile), DH (dialogue history), DK (domain knowledge), SG (subgoal).

to independently annotate these samples. During the evaluation, the annotators were asked to select the best response based on its multi-dimensional consistency from three anonymous options. The final results were obtained through majority voting.

The evaluation focused on the following dimensions:

Consistency with User Profile: Evaluators checked whether each response aligned with the user's profile (e.g., gender, interests, age), and selected the one that best matched.

Consistency with Dialogue History: Evaluators assessed whether each response was logically coherent with previous dialogue turns and properly understood the user's intent, avoiding contradictions or irrelevant content.

Consistency with Domain Knowledge: Evaluators judged whether the response accurately applied factual or domain-specific knowledge relevant to the context.

Consistency with Subgoal: Evaluators examined whether the response effectively advanced the current subgoal in the dialogue.

The three-way evaluation results for different models are shown in Figure 5. As can be seen, models with static consistency correction and DMCRL outperform the baseline across all dimensions, with DMCRL achieving the highest win rates. This

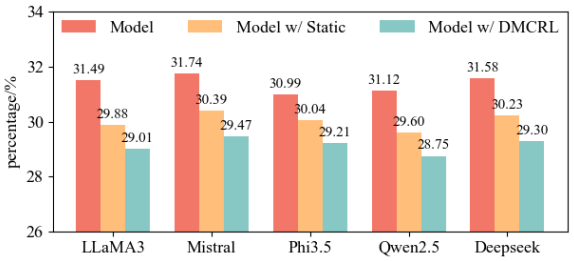


Figure 6: Subgoal failure rates on DuRecDial2.0.

suggests that DMCRL can dynamically enhance multi-dimensional consistency while maintaining coherence and improving overall response quality. These findings confirm the generalizability and effectiveness of DMCRL across different model architectures.

H Results of Subgoals Failure

Figure 6 presents the subgoal achievement comparison on the DuRecDial 2.0 dataset. Consistent with the observations on DuRecDial, the DMCRL-enhanced model significantly reduces the subgoal failure rate throughout the dialogue. This demonstrates the generalizability of DMCRL in enhancing subgoal consistency across different datasets and languages, thereby supporting more coherent and effective goal-oriented dialogue generation.

Type	System Response
Sample 1: Overemphasizing consistency with the UP	
w/ Static	U_1. Good evening, Liu Pinglu! I know you live in Guangzhou. U_2. Liu Pinglu, What are you doing? U_3. Didn't you like *Smoke and Vanish*? Leslie Cheung also gave an excellent performance in *Crossing Hennessy*. Let's talk about that one.
w/ DMCRL	U_1. Good evening, Liu Pinglu. U_2. What are you doing? U_3. Then why not watch a movie to pass the time? You could check out *Crossing Hennessy*, a beautifully crafted short film with a touch of French style.
Sample 2: Overemphasizing consistency with the DK	
w/ Static	Leslie Cheung is a Virgo and was born in Kowloon, Hong Kong. Not only is he a great actor, but his singing is also deeply moving. For example, *A Chinese Ghost Story* blends his deep emotions with the cultural characteristics of Hong Kong.
w/ DMCRL	Since his singing is so enchanting, why not enjoy his song *A Chinese Ghost Story*? Leslie Cheung's performance is perfect, achieving a melodious tune and a deep, poetic atmosphere.

Table 8: A case generated by LLAMa3.

I Case Study

To further evaluate the impact of DMCRL's dynamic characteristics on dialogue consistency and response quality, we conducted a case study comparing system responses generated with DMCRL and static consistency correction. As shown in Table 8, DMCRL significantly enhances dialogue fluency and coherence by dynamically balancing consistency across different dimensions, without compromising the consistency correction.

In the static consistency correction ($w/Static$) condition, in Example 1, U_1 and U_2 excessively apply the user profile, resulting in abrupt mentions of the user's location and repeated references to the user's name. U_3 forcibly links the user's preferred movie to the recommended one, even though the two are unrelated. In Example 2, the model fails to prioritize the subgoal of recommending the current song, instead overemphasizing consistency with domain knowledge, providing excessive and unnecessary personal information about the singer, which leads to factual errors in the song description. In contrast, with DMCRL applied (w/ DMCRL), the response in Example 1 avoids overusing the user profile, offering a more natural conversation flow. In Example 2, the model better prioritizes the subgoal of recommending the song, and when applying domain knowledge, it avoids irrelevant information, thereby maintaining both accuracy and conciseness in the response.