HALLUCINATION-CORRECTED ACTION COMPRESSION VIA LLM-RL COORDINATION FOR TASK-ORIENTED DIALOGUE POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Task-oriented dialogue policy optimization remains challenging due to the large and highly structured action space. Large language models (LLMs) can provide state-conditioned action priors to compress this space, but their hallucinations may introduce task-irrelevant, constraint-violating, or infeasible actions. Such unreliable actions not only undermine action-space compression but also bias downstream action-value estimation and policy learning. To address this issue, we propose **H**allucination-corrected **A**ction **C**ompression (**HAC**), a two-stage framework that corrects the LLM prior before using it to guide dialogue policy optimization. Stage I performs hallucination-induced correction on the original LLM prior. Specifically, an offset LLM is induced to amplify hallucination tendencies, and its predictions are contrasted with those of the original LLM to suppress unreliable preferences. This yields a factually corrected action prior, from which a compressed action set is constructed—producing a more trustworthy action space than directly using the original LLM. Stage II then performs stable RL-based policy optimization solely over this compressed set. A distributional Q-network estimates return distributions for each action, and a categorical cross-entropy objective aligns predicted and target distributions, enabling stable and discriminative learning within a reduced and reliable action space. Experiments on multiple task-oriented dialogue and factuality benchmarks show that HAC achieves superior task performance, improved factual reliability, and strong generalization.

1 INTRODUCTION

Task-oriented dialogue (TOD) systems rely on an effective dialogue policy (DP) to select actions that guide multi-turn interactions toward user goals (Algherairy & Ahmed, 2025). Optimizing DPs is challenging because TOD action spaces are large and highly structured. Semantically similar actions may differ substantially in task requirements, constraint satisfaction, or execution feasibility (Zhao et al., 2024a). Consequently, reinforcement learning (RL) agents must explore among many unreliable actions. This not only increases the exploration burden but also complicates action-value estimation by introducing low-quality transitions into policy learning (Arumugam & Griffiths, 2025; Ma et al., 2025; Zhao et al., 2019). Improving exploration efficiency while maintaining the reliability of the explored actions thus remains a central challenge in TOD policy optimization.

Existing methods for improving exploration efficiency mainly fall into two lines: RL-centric and LLM-empowered exploration paradigms. RL-centric exploration paradigms improve exploration from the RL perspective (Lin et al., 2025; Han et al., 2022; Mnih et al., 2015), but operate over the full action space and thus cannot fundamentally reduce the search burden in TOD (Zhao et al., 2024a). By contrast, LLM-empowered exploration paradigms compress the action space using original LLM priors (Yi et al., 2026; Zhao et al., 2025; Yan et al., 2025), which is preferred over fine-tuning as it preserves the model's language understanding and semantic reasoning. However, this compression process is constructed from original LLM priors, unreliable actions may be introduced into policy optimization and distort action-value estimation (Pan et al., 2024). Overall, existing approaches either improve exploration in the original action space while leaving its unreliability untouched, or compress the search space with hallucination-prone LLM priors. This unresolved tension between compression efficiency and action reliability motivates a closer examination of how

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

unreliable LLM preferences arise in dialogue-action generation and how they can be corrected before policy optimization.

To explore why LLMs generate unreliable actions in TOD, we conduct token-level attention analysis. As shown in Fig. 1, high attention is allocated to repeated action templates and slot names (e.g., `depart time`, `origin city`, `destination city`), whereas constraint tokens and slot values (e.g., `19:30`, `Beijing`, `Shanghai`) receive less attention. This attention drift causes the model to prioritize frequent tokens over task-critical constraints, producing actions that are syntactically valid but constraint-inconsistent. When such actions are used to construct a compressed action space, their unreliability is further propagated into downstream exploration and value learning.

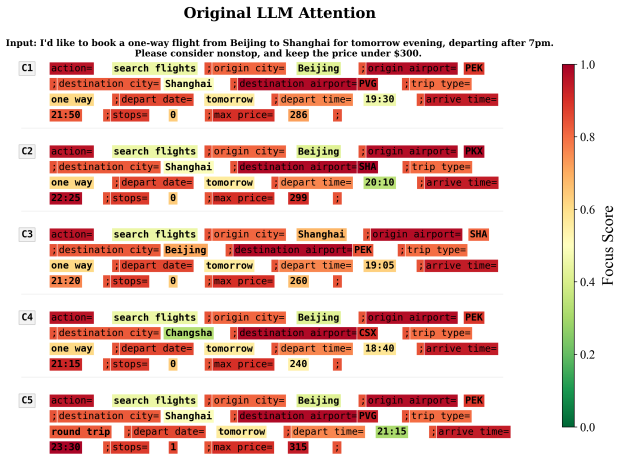

Figure 1: Token-level attention visualization of the original LLM during dialogue-action generation. Warmer colors indicate higher focus scores. Attention over five flight-booking actions (C1–C5) reveals a bias toward repetitive templates and slot names over task-critical constraints.

To this end, we propose **H**allucination-corrected **A**ction **C**ompression (**HAC**), a two-stage framework that coordinates reliable action-space construction with RL-based DP optimization. In stage I, HAC constructs a hallucination-corrected action prior by an induced-offset contrastive mechanism. An offset LLM is induced to amplify hallucination-prone preferences, and its predictions are contrasted with those of the original LLM to suppress unreliable action preferences. State-conditioned proposals generated from the corrected prior are then projected onto the executable dialogue action space, yielding a compact and more reliable action subset for each dialogue state. In stage II, this subset serves as the restricted decision space for policy optimization. Because the corrected prior indicates which actions are plausible but cannot determine their long-term task utility, a distributional Q-network is employed to model the return distribution of each action and thereby provide a more discriminative evaluation under stochastic dialogue outcomes. A categorical cross-entropy objective is then used to minimize the discrepancy between the predicted and target return distributions, which supports stable value learning and effective action selection. In this way, the stage I improves action reliability and reduces the exploration burden, while the stage II identifies the action that best supports long-term task completion.

In summary, our main contributions are:

- The cause of hallucinations in LLM dialogue actions is analyzed and identified for the first time as token-level attention drift, where critical constraint tokens receive less focus than frequently repeated templates.
- A two-stage framework is proposed in which hallucination-induced prior correction is performed before action-space compression, thereby constructing a more reliable action space for DP optimization.
- Extensive experiments on multiple TOD benchmarks are conducted, and the effectiveness of HAC in improving DP performance and exploration efficiency is demonstrated.

2 Related Work

Our work targets exploration-efficient TOD policy optimization. Existing studies follow two main lines: RL-centric and LLM-empowered exploration paradigms (Zhang et al., 2024).

RL-centric exploration paradigms: Efforts to improve RL exploration efficiency follow three main lines. **1) Randomized Exploration:** EPSILON_DQN (Mnih et al., 2015) adopts the ϵ-greedy strategy, which improves exploration by injecting random actions during DP optimization. Although simple and widely used, this strategy still explores over the full action space and cannot distinguish informative actions from invalid ones (Ecoffet et al., 2021; Osband et al., 2016). **2) Intrinsic-Motivation Exploration:** ICM_DQN (Lin et al., 2025) augments DP optimization with an intrinsic curiosity module, encouraging the agent to visit novel or uncertain states via additional reward signals. However, it improves coverage rather than reducing the search space, remaining inefficient in large TOD spaces (Pathak et al., 2017). **3) Parameterized Noise Exploration:** NOISY_DQN (Han et al., 2022) replaces heuristic exploration with learnable noisy networks, enabling state-dependent exploration via parameter perturbation. This strategy improves exploration adaptivity over hand-crafted randomization, but still operates on the full action space and thus cannot fundamentally address the large-space search problem (Fortunato et al., 2018). Overall, RL-based methods improve exploration behavior, but not the search space itself, so the replay buffer is still populated with many low-quality transitions.

LLM-empowered exploration paradigms: A second line of work leverages LLMs to improve exploration efficiency. **1) Direct Generative Decision Making:** representative methods include LLM_DA (Yi et al., 2026), which uses an LLM to generate dialogue actions, and LLM_Word (Yi et al., 2026), which extends to end-to-end response generation. These methods benefit from the strong semantic capability of LLMs and outperform RL-centric methods in the early stage. However, because they rely on direct generation rather than policy refinement, their performance typically saturates early (Liu et al., 2026b). **2) Prior-Guided Action Filtering:** LLMSAP (Zhao et al., 2025) uses an LLM prior to compress the action space into a smaller action subset, making exploration more targeted. However, its effectiveness depends on the correctness of the LLM-generated actions (Liu et al., 2026a; Wen et al., 2025). **3) Prior-Regularized Policy Optimization:** LLM_PRIOR_CQL (Yan et al., 2025) combines the LLM-induced prior with Q-learning, enabling actions in the compressed space to be scored rather than only filtered. However, because optimization is still built on an original LLM prior, unreliable actions continue to bias action-value estimation and limit policy quality. Although LLM-based prior methods reduce action-space size, they do not resolve the mismatch between the LLM objective and DP optimization, since LLMs are trained for next-token prediction rather than long-horizon reward maximization (Elizabeth et al., 2024; Steindl et al., 2025). As a result, direct reuse of original LLM priors improves search efficiency, but does not necessarily yield a reliable action space or an effective optimization objective for DP optimization.

Our method improves exploration efficiency by correcting the LLM prior before RL. It offers three key advantages: **1) Action Reliability:** it constructs a relatively reliable compressed action space via induced-offset contrastive correction, alleviating the hallucination in original LLM-generated actions; **2) Search Efficiency:** it reduces the exploration burden by restricting DP optimization to a compact action subset rather than the original full action space; and **3) Optimization Stability:** it performs policy optimization over the corrected action space with distributional Q-learning and a categorical cross-entropy objective, providing more stable value estimation under noisy and non-stationary dialogue returns.

3 METHODOLOGY

As shown in Fig. 2, **HAC** consists of two stages: 1) *LLM-based action space compression*, which performs hallucination-induced correction to obtain a corrected LLM and to compress the original action space into a small subset; 2) *RL-based policy optimization*, which conducts distributional Q-based action decision over the compressed action subset and optimizes the DP via categorical cross-entropy.

3.1 LLM-BASED ACTION SPACE COMPRESSION

Stage I improves action quality before RL by constructing corrected priors and compressing the original action space. It consists of two parts: *Hallucination-Induced Correction* and *Action Space Compression*, as detailed in Alg. 1.

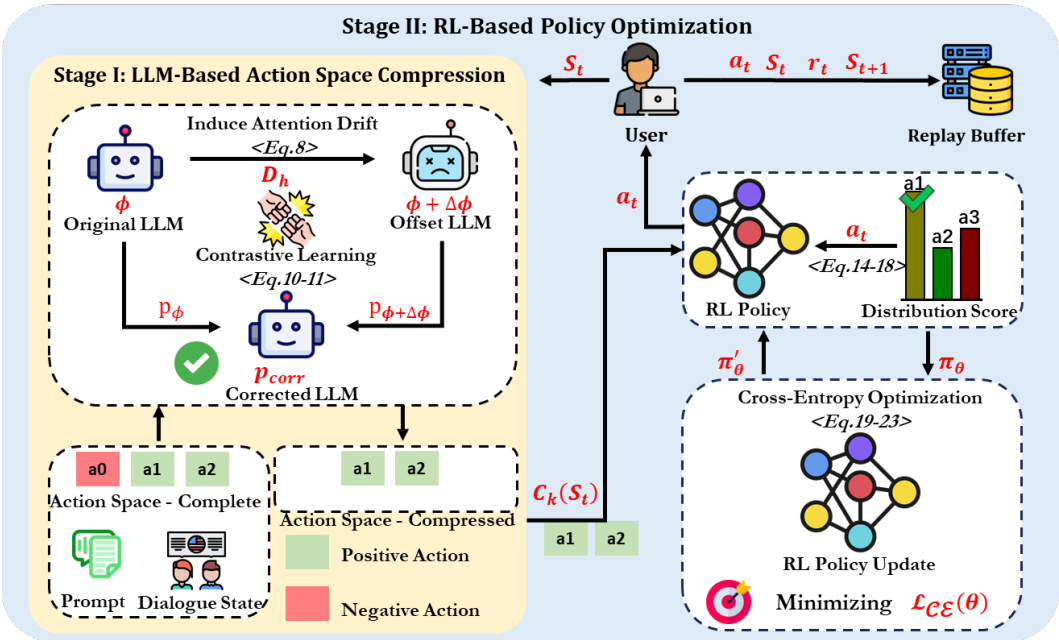

Figure 2: Overall workflow of HAC. Given the dialogue state s_t, HAC first contrasts the original LLM p_ϕ with the offset LLM $p_{\phi+\Delta\phi}$ to obtain a corrected action prior p_{corr}. The p_{corr} projects actions into a compact executable subset $C_k(s_t)$. RL optimization is performed over $C_k(s_t)$: the distributional Q-network evaluates the compressed actions and selects a_t, while the transition (s_t, a_t, r_t, s_{t+1}) is stored in the replay buffer, and the categorical cross-entropy loss $\mathcal{L}_{CE}$ is used to optimize the DP.

Algorithm 1 LLM-Based Action Space Compression

Input: Hallucination-induction dataset $\mathcal{D}_h$; original LLM parameters ϕ; contrast coefficient β; action number k
Output: Corrected action prior p_{corr}; action subset $C_k(\cdot)$
1: Obtain the offset LLM by Eq. 1
2: Freeze the original and offset LLMs
3: Construct the corrected token distribution by Eqs. 2–4
4: Project free-form outputs into executable actions by Eq. 5
5: Build $C_k(s_t)$ by Eq. 6

3.1.1 Hallucination-Induced Correction

Following an induce-then-contrast paradigm, we construct an offset LLM by fine-tuning the original model on a non-factual dataset $\mathcal{D}_h = \{(s_i, u_i, o_i)\}_{i=1}^{M}$:

$$\Delta\phi^* = \arg\min_{\Delta\phi} \sum_{i=1}^{M} -\log p(o_i \mid s_i, u_i; \phi + \Delta\phi), \tag{1}$$

where ϕ and $\phi + \Delta\phi$ denote the original and offset parameters, respectively.

For auto-regressive generation, the original LLM predicts the next token according to

$$p_\phi(x_t \mid x_{<t}) = \text{softmax}\left(\text{logit}_\phi(x_t \mid x_{<t})\right). \tag{2}$$

We suppress hallucination-prone preferences by contrasting the predictions of the two models:

$$F_t = \beta \log p_\phi(x_t \mid x_{<t}) - \log p_{\phi+\Delta\phi}(x_t \mid x_{<t}), \tag{3}$$

$$p_{\text{corr}}(x_t \mid x_{<t}) = \text{softmax}(F_t), \tag{4}$$

where $\beta > 0$ controls the contrast strength. The corrected LLM is used as an action prior rather than as the final DP.

3.1.2 ACTION SPACE COMPRESSION

Given state s_t and the prompt, the corrected LLM generates a free-form proposal o. A projection function $g(o) \to a$ maps the proposal into the admissible action space:

$$p_{\text{corr}}(a \mid s_t) = \sum_o \mathbb{I}[g(o) = a]\, p_{\text{corr}}(o \mid s_t), \tag{5}$$

where $\mathbb{I}[\cdot]$ is the indicator function. We then construct a compact executable subset:

$$\begin{aligned} C_k(s_t) &= \{a_t^1, \ldots, a_t^k\}, \\ a_t^i &\sim p_{\text{corr}}(a|s_t), \\ C_k(s_t) &\subseteq \mathcal{A}_t. \end{aligned} \tag{6}$$

3.2 RL-BASED POLICY OPTIMIZATION

Stage II selects action and estimates value from the compressed action subset. It consists of two parts: *Distributional Q-Based Action Decision* and *Cross-Entropy Policy Optimization*, as detailed in Alg. 2.

Algorithm 2 RL-Based Policy Optimization

Input: Main Q-network parameters θ; target Q-network parameters $\bar{\theta}$; candidate subset $C_k(\cdot)$; discount factor γ; temperature α; Gaussian width σ; atom width ς; soft update coefficient τ
Output: Optimized policy π_θ

```
for each training iteration do
    Observe s_t and construct C_k(s_t) by Alg. 1
    Compute Q(s_t, a; θ) by Eqs. 7–9
    Form π_θ(· | s_t) and sample a_t by Eqs. 10–11
    Execute a_t, obtain (r_t, s_{t+1}), and store the transition in R
    Sample a minibatch from R
    for each sampled transition (s_i, a_i, r_i, s_{i+1}) do
        Construct C_k(s_{i+1}) and compute y_i by Eq. 12
        Construct target distribution p^tar(·) by Eqs. 13–14
    end for
    Update θ by minimizing Eq. 15
    Update θ̄ by Eq. 16
end for
```

3.2.1 DISTRIBUTIONAL Q-BASED ACTION DECISION

We model the return by a categorical distribution over m atoms:

$$Z(s_t, a; \theta) = \sum_{i=1}^{m} \hat{p}_i(s_t, a; \theta)\delta_{z_i}, \tag{7}$$

where δ_{z_i} is a point mass at z_i. The atom probabilities are

$$\hat{p}_i(s_t, a; \theta) = \frac{\exp(l_i(s_t, a; \theta))}{\sum_{j=1}^{m} \exp(l_j(s_t, a; \theta))}, \tag{8}$$

and the scalar action value is given by their expectation:

$$Q(s_t, a; \theta) = \sum_{i=1}^{m} z_i \hat{p}_i(s_t, a; \theta). \tag{9}$$

The policy over the compressed action set is

$$\pi_\theta(a|s_t) = \frac{\exp(Q(s_t, a; \theta)/\alpha)}{\sum_{a' \in C_k(s_t)} \exp(Q(s_t, a'; \theta)/\alpha)}, \tag{10}$$

where $\alpha > 0$ controls the smoothness of the policy distribution. The executed action is sampled as

$$a_t \sim \pi_\theta(\cdot|s_t). \tag{11}$$

3.2.2 CROSS-ENTROPY POLICY OPTIMIZATION

After interaction, each transition is stored in $\mathcal{R}$. For a sampled transition, the target value is computed over the next-state action set:

$$y_t = r_t + \gamma \max_{a' \in C_k(s_{t+1})} Q(s_{t+1}, a'; \bar{\theta}). \tag{12}$$

We project y_t onto the categorical support by HL-Gauss:

$$Y_t \sim \mathcal{N}(y_t, \sigma^2), \tag{13}$$

whose probability mass on atom z_i is

$$p_i^{\text{tar}} = F_{Y_t}(z_i + \varsigma/2) - F_{Y_t}(z_i - \varsigma/2), \tag{14}$$

where σ is the Gaussian width, ς is the atom width, and F_{Y_t} is the cumulative distribution function of Y_t.

The main Q-network minimizes the categorical cross-entropy between the predicted and target distributions:

$$\mathcal{L}_{\text{CE}}(\theta) = -\mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{R}} \left[\sum_{i=1}^{m} p_i^{\text{tar}} \log \hat{p}_i(s_t, a_t; \theta) \right]. \tag{15}$$

Finally, the target network is softly updated as

$$\bar{\theta} \leftarrow \tau\theta + (1 - \tau)\bar{\theta}, \tag{16}$$

where τ is the update coefficient.

4 EXPERIMENTS

The experiments aim to: 1) show that hallucinations in LLM dialogue actions are caused by attention drift and are mitigated by HAC (Sec. 4.4); 2) assess the overall effectiveness of HAC in simulated (Sec. 4.5) and realistic human interactions (Sec. 4.6); 3) verify whether HAC improves factuality on both discrimination-based and generation-based benchmarks (Sec. 4.7); Further analysis on model scaling, generalization capability and ablation study (Sec. 4.8).

4.1 DATASETS

We evaluate the HAC on two public, widely used groups of benchmarks. The first includes three TOD benchmarks from the Microsoft Dialogue Challenge (Li et al., 2018): movie-ticket, restaurant, and taxi booking. These benchmarks assess DP performance. The second includes two factuality benchmarks, *TruthfulQA* (Lin et al., 2022) and *FACTSCORE* (Min et al., 2023). These benchmarks evaluate the reliability of corrected action priors. Detailed benchmark statistics and evaluation metrics are provided in App. A.

4.2 BASELINES

We compare HAC with seven representative baselines. First, we consider three RL-centric baselines: EPSILON_DQN (Mnih et al., 2015), which adopts the standard ϵ-greedy exploration strategy; ICM_DQN (Lin et al., 2025), which introduces an intrinsic curiosity module to encourage exploration; and NOISY_DQN (Han et al., 2022), which replaces heuristic exploration with parameterized noisy networks. Second, we include two direct generative LLM baselines: LLM_DA (Yi et al., 2026), which replaces the DP module with a LLM to generate dialogue actions, and LLM_Word (Yi et al., 2026), which directly generate end-to-end responses. Third, we consider two LLM-guided RL baselines: LLM_PRIOR_CQL (Yan et al., 2025), which combines an LLM-induced action prior with conservative Q-learning, and LLMSAP (Zhao et al., 2025), which employs an LLM-based action-prior mechanism to guide action selection in the compressed action space.

4.3 Experimental Settings

We select Qwen-7B as the backbone for HAC, as it offers a trade-off between action-prior quality and computational cost. Detailed model-scale results are provided in App. B. All baselines are trained for 500 epochs. The maximum dialogue length is $L = 40$. The replay buffer size is 10000, and the batch size is 16. For RL-based policy optimization, the discount factor is $\gamma = 0.99$. Both the main and target Q-networks are two-layer MLPs with 80 hidden units per layer. The Q-network is trained with a learning rate of 10^{-3}, and the target network is updated with coefficient $\tau = 10^{-2}$. For action space compression, the action number is $k = 5$, and the contrast coefficient is $\beta = 1.0$. For action selection over the compressed action set, the temperature parameter is $\alpha = 0.1$. For distributional value learning, we use m=51 atoms with width $\varsigma = (v_{\max} - v_{\min})/m$, and Gaussian width is $\sigma = 0.75\varsigma$. All results are averaged over five runs with different random seeds.

4.4 Qualitative Analysis

To show how HAC improves the LLM-induced action prior, we visualize token-level attention before and after correction. Specifically, 5 actions are generated from the user utterance: "I'd like to book a one-way flight from Beijing to Shanghai for tomorrow evening, departing after 7pm. Please consider nonstop, and keep the price under $300." Fig. 1 and Fig. 3 compare the original and corrected LLMs, where warmer colors indicate higher focus scores.

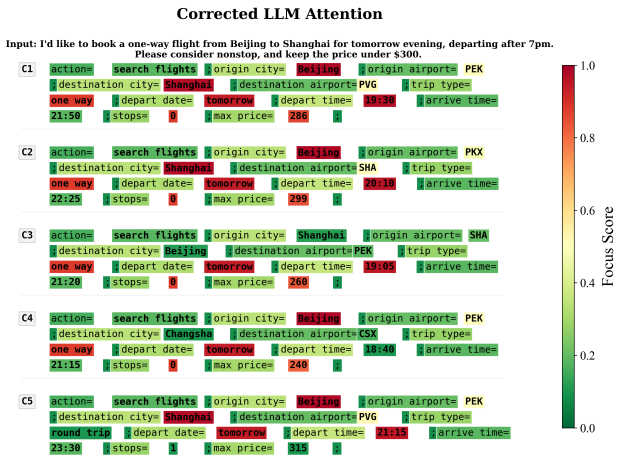

Figure 3: Token-level attention visualization of the LLM with our HAC-driven correction during action generation.

The original LLM overly attends to template tokens or slot names, such as `destination city=`, which preserve action structure but do not determine action validity. As a result, the original prior may generate syntactically plausible yet execution-invalid actions. In contrast, the corrected LLM focuses on constraint-bearing tokens or slot values such as `Beijing`, `Shanghai` and `tomorrow`, mitigating token-level attention drift by reducing template bias and emphasizing task-critical constraints. Based on attention correction, HAC achieves better DP performance, as shown in Sec. 4.5.

4.5 Main Results

Tab. 1 reports results on the three TOD domains. HAC achieves the strongest overall performance, with higher Success and Reward and fewer Turns. The advantage is clear on Restaurant and Taxi, where accurate action selection depends more heavily on slot consistency and execution validity.

The RL-centric baselines, EPSILON_DQN, ICM_DQN, and NOISY_DQN, remain limited across all domains. Their weakness suggests that improving exploration is insufficient for DP optimization under large and structured action spaces. Although these methods enlarge action coverage, they explore over the full action space and thus cannot suppress invalid actions during training. As a result, the policy is updated with many low-quality transitions, which slows convergence and limits final performance. The direct generative LLM baselines [1], LLM_DA and LLM_Word, show stronger early performance, especially LLM_DA, indicating that LLMs provide a useful semantic prior for dialogue decision making. However, their performance remains nearly unchanged across epochs, which shows that direct LLM generation can improve initialization but cannot support continual

[1] LLM_DA and LLM_Word are non-RL baselines that directly generate actions from prompted LLM without iterative policy optimization. Therefore, only post-convergence results (Epoch=500) are reported for fair comparison with RL-based methods.

Table 1: Results of different baselines on three datasets. Top performers in each column are bolded. All results are statistically significant (t-test, $p < 0.05$). Epochs denote early (50), mid (250), and post-convergence (500) stages. The results with variance are detailed in E.

Domain	Agent	Epoch=50			Epoch=250			Epoch=500		
		Success↑	**Reward↑**	**Turn↓**	**Success↑**	**Reward↑**	**Turn↓**	**Success↑**	**Reward↑**	**Turn↓**
Movie	EPSILON_DQN	0.019	-52.49	31.62	0.406	-4.29	27.98	0.600	21.73	22.56
	ICM_DQN	0.002	-47.39	31.12	0.203	-22.68	26.18	0.417	0.55	23.22
	NOISY_DQN	0.004	-44.87	26.68	0.255	-18.38	26.70	0.539	11.74	21.29
	LLM_DA	–	–	–	–	–	–	0.385	-6.56	27.06
	LLM_Word	–	–	–	–	–	–	0.234	-30.69	33.34
	LLM_PRIOR_CQL	**0.055**	-48.44	30.67	**0.586**	**10.62**	23.95	0.779	**46.67**	18.21
	LLMSAP	0.037	**-40.32**	**25.29**	0.490	6.02	22.06	0.785	37.57	16.43
	HAC	0.020	-42.77	26.18	0.505	8.19	**21.96**	**0.870**	46.23	**14.84**
Rest.	EPSILON_DQN	0.000	-41.73	25.48	0.143	-29.32	26.46	0.386	-5.76	22.97
	ICM_DQN	0.005	-43.56	30.10	0.027	-41.61	30.12	0.044	-40.41	30.73
	NOISY_DQN	**0.036**	**-39.63**	27.81	0.285	-16.81	26.83	0.407	-5.52	26.28
	LLM_DA	–	–	–	–	–	–	0.321	-7.93	20.75
	LLM_Word	–	–	–	–	–	–	0.180	-23.60	34.28
	LLM_PRIOR_CQL	0.034	-39.88	27.87	0.319	-13.26	25.96	0.519	5.90	23.65
	LLMSAP	0.019	-40.61	**26.58**	0.330	**-11.53**	**24.47**	0.584	12.73	21.72
	HAC	0.025	-40.30	27.10	**0.336**	-11.63	25.74	**0.654**	**19.26**	**21.19**
Taxi	EPSILON_DQN	0.047	-42.19	29.08	0.298	-12.31	26.04	0.513	11.27	22.34
	ICM_DQN	0.018	-43.74	30.69	0.087	-37.54	30.80	0.147	-30.74	29.57
	NOISY_DQN	0.012	-45.60	31.37	0.115	-35.83	30.50	0.251	-22.06	28.97
	LLM_DA	–	–	–	–	–	–	0.339	-10.01	25.31
	LLM_Word	–	–	–	–	–	–	0.243	-18.12	28.66
	LLM_PRIOR_CQL	0.019	-51.38	32.61	0.365	-10.83	29.49	0.608	20.37	23.34
	LLMSAP	**0.067**	**-41.99**	**28.73**	0.520	9.42	25.10	0.710	31.94	21.18
	HAC	0.058	-48.69	33.30	**0.625**	**24.74**	**22.42**	**0.802**	**48.41**	**17.75**

policy refinement. The LLM-guided RL baselines, LLM_PRIOR_CQL and LLMSAP, substantially outperform both RL-centric and direct generative LLM baselines, confirming the effectiveness of incorporating LLM priors into RL. Nevertheless, both methods are inferior to HAC. This gap suggests that directly reusing the original LLM prior is insufficient for robust DP optimization, since the resulting actions may still contain task-irrelevant proposals. In contrast, HAC first corrects hallucination-induced priors and then restricts policy optimization to a compressed action space. As a result, the distributional Q-network provides more accurate action-value estimation over the compressed action set. Combined with the categorical cross-entropy objective, this improves DP optimization under non-stationary targets.

4.6 Human Evaluation

To evaluate robustness in realistic interactions, we conduct human evaluations following the protocols of Zhao et al. (2024b) and Liu et al. (2021). 56 evaluators (28 domain experts and 28 general users) interact with trained agents. After each dialogue, evaluators rate Success Rate (SR, task completion) and Human Score (HS, 1-5 for naturalness and coherence). As shown in Tab. 2, HAC outperforms the baselines, aligning with the simulation results. The gains in SR and HS show that HAC improves task completion and dialogue quality in multi-turn interactions. Compared with RL-centric baselines, HAC reduces ineffective exploration; Compared with direct generative LLM baselines, it maintains task progress rather than relying on locally plausible generation; Compared with the LLM-guided RL baselines, it avoids directly relying on uncorrected LLM priors.

Table 2: Human evaluation results of different baselines.

Model	**Movie**		**Restaurant**		**Taxi**	
	SR↑	**HS↑**	**SR↑**	**HS↑**	**SR↑**	**HS↑**
EPSILON_DQN	0.442	2.62	0.295	2.24	0.379	2.59
ICM_DQN	0.129	1.79	0.052	1.49	0.081	1.35
NOISY_DQN	0.174	2.47	0.303	2.42	0.174	2.07
LLM_DA	0.267	2.24	0.136	1.83	0.242	2.31
LLM_Word	0.152	1.55	0.113	1.76	0.138	1.63
LLM_PRIOR_CQL	0.536	2.78	0.359	2.47	0.433	2.64
LLMSAP	0.585	3.19	0.463	2.81	0.520	2.93
HAC	**0.617**	**3.32**	**0.533**	**3.18**	**0.588**	**3.26**

Table 3: Results on TruthfulQA (MC1/2/3) and FactScore.

Model	TruthfulQA			FactScore		
	MC1 ↑	MC2 ↑	MC3 ↑	% Respond	# Facts	Score ↑
DeepSeek-Chat	52.19	69.32	40.53	82.6	54.3	67.1
GPT-4o	54.82	70.48	43.96	85.2	55.9	70.8
GPT-5.2	**63.76**	**80.17**	**51.02**	**96.6**	**63.4**	**78.9**
HAC+						
Qwen2.5-7B-Instruct	39.79	55.83	32.16	30.8	40.1	57.7
	49.23	**67.27**	**39.63**	33.7	44.3	**64.2**
Llama2-7B-Chat	36.55	48.91	30.79	36.7	43.1	62.9
	42.73	57.16	33.28	35.5	44.6	**65.7**
Mistral-7B-Instruct	38.63	51.37	30.65	37.4	35.4	53.3
	48.24	64.89	35.17	37.9	38.7	**60.8**
Gemma2-2B-it	30.26	43.81	21.73	44.3	26.1	43.7
	36.04	52.67	27.52	45.6	29.3	**46.2**

4.7 Factuality Evaluation

To assess whether HAC improves factuality across different evaluation protocols, we conduct experiments on discrimination and generation benchmarks. We use four open-source LLM backbones: Qwen2.5-7B-Instruct (Yang et al., 2024), Llama2-7B-Chat (Touvron et al., 2023), Mistral-7B-Instruct (Jiang et al., 2023), and Gemma2-2B-it (Team, 2024), and compare each original model with its HAC-enhanced counterpart (HAC+). In addition, DeepSeek-Chat (DeepSeek-AI, 2024), GPT-4o (OpenAI, 2024), and GPT-5.2 (OpenAI, 2025) are included as strong reference models.

HAC improves truthfulness on TruthfulQA. Tab. 3 shows that HAC improves MC1, MC2, and MC3 across all backbones. It indicates that the induce-offset contrastive learning effectively directs models toward truthful answers. The gains are observed across different model families, suggesting that the mechanism is not backbone-specific. It confirms that contrasting the original model with a hallucination-amplified offset model effectively suppresses untruthful decoding predictions.

HAC reduces hallucinations in open-ended generation on FACTSCORE. Tab. 3 verifies the effectiveness of HAC on generation-based factuality evaluation. Across all backbones, HAC improves the FACTSCORE, showing that the offset contrastive learning remains effective beyond multiple-choice settings. Importantly, the response ratio and the number of extracted facts are largely preserved, while factual precision increases. This suggests that the method does not improve factuality by becoming conservative, but by suppressing hallucinated content in the generated output.

4.8 Supplementary experiments

Model Scale Analysis (App. B): We analyze model scale and find that scaling from 3B to 7B yields clear gains, whereas larger models provide only marginal improvements, supporting 7B as a favorable balance between action-prior quality and computational cost. **Generalization Analysis (App. C)**: We examine HAC across different LLM backbones and show that its effectiveness is not backbone-specific. **Ablation Study (App. D)**: We remove key components individually and show that each removal degrades performance, confirming their complementary roles in HAC.

5 Conclusion

We present HAC, a two-stage framework that improves DP optimization by building a corrected action prior for reliable action-space compression. By correcting the original LLM prior via hallucination-induced correction, HAC constructs a compact action set that reduces unreliable actions. Based on the action space, HAC combines distributional Q-based action evaluation with a categorical cross-entropy objective to enable discriminative action assessment and stable policy optimization. Experiments on multiple benchmarks validate the effectiveness of HAC. The factuality evaluation, human evaluation, model-scale and generalization analysis further confirm its value.

ETHICS STATEMENT

We affirm that this work fully adheres to the ICLR Code of Ethics. We do not collect, use, or release any new human subject data. All experiments are conducted on established and publicly available datasets, which have been widely used in prior academic research.

REFERENCES

Atheer Algherairy and Moataz Ahmed. Prompting large language models for user simulation in task-oriented dialogue systems. *Comput. Speech Lang.*, 89:101697, 2025. doi: 10.1016/J.CSL.2024.101697. URL https://doi.org/10.1016/j.csl.2024.101697.

Dilip Arumugam and Thomas L. Griffiths. Toward efficient exploration by large language model agents. *CoRR*, abs/2504.20997, 2025. doi: 10.48550/ARXIV.2504.20997. URL https://doi.org/10.48550/arXiv.2504.20997.

DeepSeek-AI. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. *Nat.*, 590(7847):580–586, 2021. doi: 10.1038/S41586-020-03157-9. URL https://doi.org/10.1038/s41586-020-03157-9.

Michelle Elizabeth, Morgan Veyret, Miguel Couceiro, Ondrej Dusek, and Lina Maria Rojas-Barahona. Do large language models with reasoning and acting meet the needs of task-oriented dialogue? *CoRR*, abs/2412.01262, 2024. doi: 10.48550/ARXIV.2412.01262. URL https://doi.org/10.48550/arXiv.2412.01262.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Rémi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rywHCPkAW.

Shuai Han, Wenbo Zhou, Jiayi Lu, Jing Liu, and Shuai Lü. NROWAN-DQN: A stable noisy network with noise reduction and online weight adjustment for exploration. *Expert Syst. Appl.*, 203:117343, 2022. doi: 10.1016/J.ESWA.2022.117343. URL https://doi.org/10.1016/j.eswa.2022.117343.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *CoRR*, abs/1807.11125, 2018. URL http://arxiv.org/abs/1807.11125.

Jie Lin, Yuhao Ye, Shaobo Li, Hanlin Zhang, and Peng Zhao. Improving exploration in deep reinforcement learning for incomplete information competition environments. *IEEE Trans. Emerg. Top. Comput. Intell.*, 9(5):3665–3676, 2025. doi: 10.1109/TETCI.2025.3555250. URL https://doi.org/10.1109/TETCI.2025.3555250.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Jiacheng Liu, Mayi Xu, Qiankun Pi, Wenli Li, Ming Zhong, Yuanyuan Zhu, Mengchi Liu, and Tieyun Qian. Format as a prior: Quantifying and analyzing bias in llms for heterogeneous data. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 32060–32068. AAAI Press, 2026a. doi: 10.1609/AAAI.V40I38.40477. URL https://doi.org/10.1609/aaai.v40i38.40477.

Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. LLM collaboration with multi-agent reinforcement learning. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 32150–32158. AAAI Press, 2026b. doi: 10.1609/AAAI.V40I38.40487. URL https://doi.org/10.1609/aaai.v40i38.40487.

Sihong Liu, Jinchao Zhang, Keqing He, Weiran Xu, and Jie Zhou. Scheduled dialog policy learning: An automatic curriculum learning framework for task-oriented dialog system. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Findings of ACL, pp. 1091–1102. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.FINDINGS-ACL.94. URL https://doi.org/10.18653/v1/2021.findings-acl.94.

Runyu Ma, Jelle Luijkx, Zlatan Ajanovic, and Jens Kober. Explorllm: Guiding exploration in reinforcement learning with large language models. In *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, pp. 9011–9017. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127622. URL https://doi.org/10.1109/ICRA55743.2025.11127622.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015. doi: 10.1038/NATURE14236. URL https://doi.org/10.1038/nature14236.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2026-04-22.

OpenAI. Update to gpt-5 system card: Gpt-5.2. https://openai.com/index/gpt-5-system-card-update-gpt-5-2/, 2025. Accessed: 2026-04-22.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4026–4034, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8d8818c8e140c64c743113f563cf750f-Abstract.html.

Yan Pan, Davide Cadamuro, and Georg Groh. Exploring hallucinations in task-oriented dialogue systems with narrow domains. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim (eds.), *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, PACLIC 2024, Tokyo, Japan, December 7-9, 2024*, pp. 609–618.

Tokyo University of Foreign Studies, 2024. URL https://aclanthology.org/2024.paclic-1.58/.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pp. 2778–2787. PMLR, 2017. URL http://proceedings.mlr.press/v70/pathak17a.html.

Sebastian Steindl, André Kestler, Ulrich Schäfer, and Bernd Ludwig. An improved, strong baseline for pre-trained large language models as task-oriented dialogue systems. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 11388–11398. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.610/.

Gemma Team. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024. doi: 10.48550/ARXIV.2408.00118. URL https://doi.org/10.48550/arXiv.2408.00118.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Xiangyu Wen, Jianyuan Zhong, Zhijian Xu, and Qiang Xu. Guideline compliance in task-oriented dialogue: The chained prior approach. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Findings of ACL, pp. 6750–6776. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.377. URL https://doi.org/10.18653/v1/2025.findings-naacl.377.

Xue Yan, Yan Song, Xidong Feng, Mengyue Yang, Haifeng Zhang, Haitham Bou-Ammar, and Jun Wang. Efficient reinforcement learning with large language model priors. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=e2NRNQ0sZe.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Comput. Surv.*, 58(6):148:1–148:38, 2026. doi: 10.1145/3771090. URL https://doi.org/10.1145/3771090.

Ye Zhang, Wang Zhao, Jingyu Wang, and Yuan Yuan. Recent progress, challenges and future prospects of applied deep reinforcement learning : A practical perspective in path planning. *Neurocomputing*, 608:128423, 2024. doi: 10.1016/J.NEUCOM.2024.128423. URL https://doi.org/10.1016/j.neucom.2024.128423.

Tiancheng Zhao, Kaige Xie, and Maxine Eskénazi. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 1208–1218. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1123. URL https://doi.org/10.18653/v1/n19-1123.

Yangyang Zhao, Mehdi Dastani, Jinchuan Long, Zhenyu Wang, and Shihan Wang. Rescue conversations from dead-ends: Efficient exploration for task-oriented dialogue policy optimization. *Trans. Assoc. Comput. Linguistics*, 12:1578–1596, 2024a. doi: 10.1162/TACL_A_00717. URL https://doi.org/10.1162/tacl_a_00717.

Yangyang Zhao, Ben Niu, Mehdi Dastani, and Shihan Wang. Bootstrapped policy learning for task-oriented dialogue through goal shaping. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 4566–4580. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.EMNLP-MAIN.263. URL https://doi.org/10.18653/v1/2024.emnlp-main.263.

Yangyang Zhao, Ben Niu, Yuxuan Tan, Shihan Wang, and Libo Qin. Semantic-aware action space compression via LLM-DRL synergy for efficient task-oriented dialogue policy exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 17808–17820. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.968/.

Table 4: Statistics of the three single-domain TOD datasets.

Dataset	Dialogues	Intentions	Slots
Movie	2,890	11	29
Restaurant	4,103	11	30
Taxi	3,094	11	29

A Datasets

We evaluate the proposed method on two groups of benchmarks. The first group consists of three public and widely used single-domain TOD datasets derived from the Microsoft Dialogue Challenge (Li et al., 2018): movie-ticket, restaurant, and taxi booking. Detailed statistics of the three domains are summarized in Tab. 4. For dialogue policy (DP) evaluation, we report three standard metrics: success rate (Success), average cumulative reward (Rewards), and average dialogue turns (Turns). A better dialogue policy achieves a higher success rate and reward while requiring fewer dialogue turns.

The second group consists of two factuality benchmarks for LLMs: *TruthfulQA* (Lin et al., 2022) and *FACTSCORE* (Min et al., 2023). For discrimination evaluation, we adopt the widely used *TruthfulQA* benchmark and report its multiple-choice metrics, including MC1, MC2, and MC3. MC1 measures whether the model assigns the highest score to the correct answer, MC2 evaluates the normalized probability mass assigned to all correct answers, and MC3 examines whether each correct answer receives a higher score than all incorrect answers. For generation evaluation, we use *FACTSCORE*, which assesses the factual precision of biography generation by decomposing model outputs into atomic facts and verifying them against reference knowledge sources. Following the standard evaluation protocol, we report the response ratio (% response), the average number of atomic facts per response (# facts), and the factual precision score (score). A better LLM obtains higher MC1, MC2, and MC3 on *TruthfulQA*, and achieves higher factual precision on *FACTSCORE* while maintaining a reasonable response ratio and generating informative atomic facts.

B Model Scale Analysis

To evaluate LLM scale impact, we conduct action-prior generation with backbones of different sizes. As shown in Fig. 4, performance improves markedly when scaling from 3B to 7B, indicating that a stronger backbone provides a more reliable action prior for DP optimization. However, the gains become marginal beyond 7B. The 14B and 32B models only yield slight further improvements over the 7B backbone. This suggests that, once the backbone reaches a sufficient semantic and factual reliability, the benefit of additional scale for action-prior quality quickly saturates.

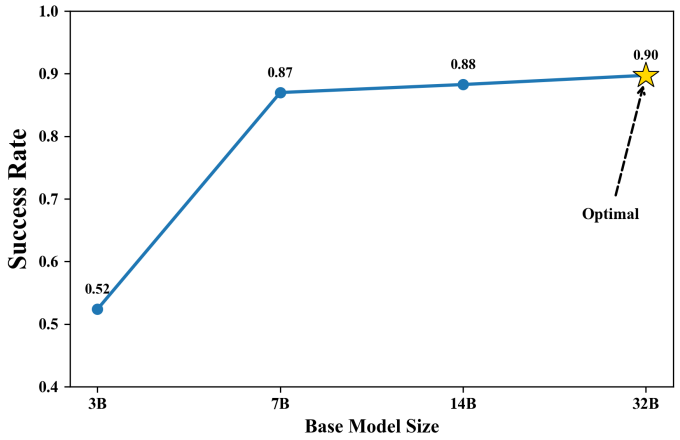

Figure 4: Performance on Qwen with Different Scales.

These results indicate that the proposed framework does not require large LLMs to achieve strong performance. Instead, a medium-scale backbone delivers the optimal trade-off between effectiveness and efficiency. We thus adopt the 7B-scale model as the default backbone in experiments.

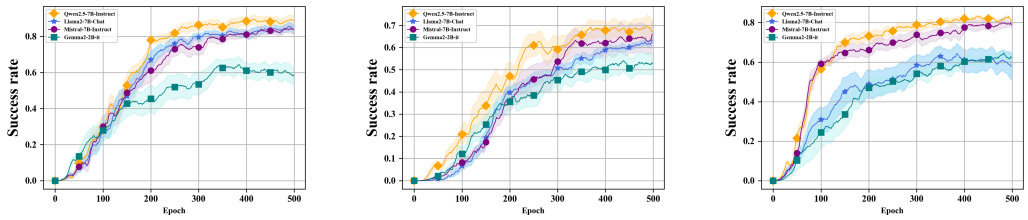

Figure 5: Performance of HAC across different LLM backbones.

Table 5: Ablation results on three dialogue domains. The best result in each column is shown in bold. Epochs 50, 250, and 500 denote the early, intermediate, and post-convergence stages, respectively.

Domain	Method	Epoch = 50			Epoch = 250			Epoch = 500		
		Success↑	**Reward↑**	**Turn↓**	**Success↑**	**Reward↑**	**Turn↓**	**Success↑**	**Reward↑**	**Turn↓**
Movie	RL	0.0082	-53.10	33.42	0.3325	-8.74	28.16	0.6027	22.16	22.43
	RL + H	0.0168	-46.35	28.74	0.4562	3.85	23.47	0.8125	41.36	16.72
	RL + Q	0.0127	-49.28	30.15	0.4018	-0.96	25.62	0.7418	34.57	18.65
	RL + H + Q	**0.0204**	**-42.77**	**26.18**	**0.5046**	**8.19**	**21.96**	**0.8699**	**46.23**	**14.84**
Rest.	RL	0.0041	-46.82	34.16	0.1487	-30.64	31.27	0.3875	-6.84	27.62
	RL + H	0.0193	-42.17	29.46	0.2948	-16.52	27.11	0.5986	13.72	22.05
	RL + Q	0.0136	-44.05	31.08	0.2475	-21.38	28.64	0.5481	8.96	22.41
	RL + H + Q	**0.0250**	**-40.30**	**27.10**	**0.3360**	**-11.63**	**25.74**	**0.6540**	**19.26**	**21.19**
Taxi	RL	0.0095	-57.26	38.12	0.2516	-18.43	31.85	0.5142	11.68	22.73
	RL + H	0.0476	-51.32	35.41	0.5738	17.63	24.76	0.7489	39.82	19.24
	RL + Q	0.0319	-54.07	36.88	0.4682	8.54	27.33	0.6827	29.63	20.46
	RL + H + Q	**0.0580**	**-48.69**	**33.30**	**0.6246**	**24.74**	**22.42**	**0.8024**	**48.41**	**17.75**

C Generalization Analysis

Fig.5 reports the generalization results across four LLM backbones on the three TOD domains. All backbones improve steadily during training, showing that the proposed framework is not tied to a specific model family and can be combined with different open-source LLMs. Qwen2.5-7B-Instruct achieves the highest success rates, Llama2-7B-Chat and Mistral-7B-Instruct remain competitive, whereas Gemma2-2B-it saturates earlier. This suggests that stronger backbones provide higher-quality action priors, leading to more effective action compression and policy optimization.

D Ablation Study

We conduct ablation experiments to evaluate the effectiveness of the two main components in HAC. Here, RL denotes the basic reinforcement learning policy, H denotes hallucination-corrected action compression, and Q denotes distributional Q-based policy optimization. The complete method is therefore denoted as RL + H + Q.

As shown in Tab. 5, both RL + H and RL + Q consistently outperform the basic RL model across the Movie, Restaurant, and Taxi domains. The improvements obtained by RL + H demonstrate that correcting the LLM prior and compressing the action space can reduce ineffective exploration and accelerate policy learning. Meanwhile, the gains achieved by RL + Q verify that distributional action-value estimation and categorical cross-entropy optimization improve the stability and effectiveness of dialogue policy learning.

The complete RL + H + Q model achieves the highest Success and Reward and the fewest dialogue Turns at the early, intermediate, and post-convergence stages. These results demonstrate that the two components play roles: H improves the reliability of the candidate action space, whereas Q provides more discriminative action evaluation and more stable policy optimization within the compressed space. You may include other additional sections here.

E MAIN RESULTS

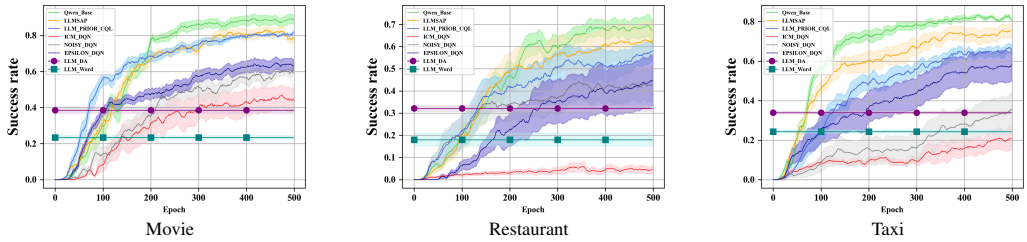

Figure 6: The learning curves of different baselines on three datasets with variance.