

# Simulated Rewards, Skewed Strategies: Tracing the Acquired Preference Bias in LLM-Based Dialogue Planners

Heyan Huang, Yizhe Yang\*, Huashan Sun, Jiawei Li, Yang Gao

<sup>1</sup>Beijing Institute of Technology, Beijing, China  
{hhy63,yizheyang,hssun,jwli,gyang}@bit.edu.cn

## Abstract

Large language models have enabled sophisticated dialogue planners, but their reliance on LLM-based simulation and feedback for planner optimization may introduce preference bias. We present the first comprehensive analysis of preference bias in LLM-based dialogue planners, evaluating four state-of-the-art planner across three dialogue domains using multiple LLM families at varying scales. Our investigation reveals that all tested planners exhibit significant preference bias, favoring narrow strategy sets rather than maintaining balanced distributions. User simulation emerges as the primary bias driver, while diverse persona simulation fails as an effective mitigation strategy. Most concerning, preference bias drives planners toward ethically problematic strategies that achieve short-term success while undermining real-world effectiveness and ethical standards. Our findings establish fundamental challenges for responsible deployment of LLM-based dialogue systems and provide crucial insights for developing more reliable and ethically-aligned planning approaches.

## Introduction

Large Language Models (LLMs) (Ouyang et al. 2022; Yang et al. 2024; Grattafiori et al. 2024; Yang et al. 2025a) have revolutionized natural language processing by demonstrating remarkable abilities to comprehend context and generate coherent responses across diverse dialogue scenarios (Yang et al. 2025b,c). However, LLMs predominantly operate in reactive modes, frequently encountering challenges in proactively guiding conversations toward specific objectives—a critical limitation for achieving truly dynamic and strategic interactions (Deng et al. 2023b). This disparity has prompted extensive research into dialogue planning mechanisms capable of strategically directing conversations, particularly in goal-oriented scenarios (Zhang et al. 2020; Cheng et al. 2022; Liu et al. 2018). Recent advancements in LLM-based dialogue planning have yielded sophisticated planner capable of strategic decision-making in intricate conversational contexts (Deng et al. 2023b; Zhang et al. 2024; He et al. 2024; Yu, Chen, and Yu 2023; Xu et al. 2025b,a). These planners are typically optimized or search

for to maximize task success metrics as the primary optimization target based on LLM simulation and have exhibited remarkable performance in goal-oriented tasks such as negotiation (He et al. 2018), persuasion (Wang et al. 2019), and emotional support (Liu et al. 2021).

Nevertheless, a critical and under-researched concern is whether the emphasis on optimizing success metrics with LLM simulation introduces preference bias that undermines the quality and ethics of dialogue planner. LLM-powered evaluators and simulators are known to harbor inherent biases that can manifest as preferences for certain dialogue strategies (Wataoka, Takahashi, and Ri 2024). For instance, when an LLM “judge” favors particular response styles or strategies inherent in its training, it incorporates those preferences into the dialogue agent it evaluates. This engenders a form of reward hacking akin to reinforcement learning scenarios where agents exploit vulnerabilities in reward signals to attain high reported performance without adhering to intended objectives. If an LLM-based simulator consistently rewards a particular strategy, even an inappropriate one, a planner may overoptimize for that strategy, compromising real-world effectiveness and potentially leading to unethical conduct.

Our research presents the inaugural comprehensive analysis of preference bias in LLM-based dialogue planners. We comprehensively evaluate four recent planning algorithms: PPDPP (Deng et al. 2023b), TRIP (Zhang et al. 2024), DPDP (He et al. 2024), and GDP-Zero (Yu, Chen, and Yu 2023), across three distinct goal-oriented dialogue tasks: price negotiation (He et al. 2018), emotional support conversation (Liu et al. 2021), and charity persuasion (PersuasionForGood) (Wang et al. 2019). We employ a range of metrics to quantify planner bias, including entropy of strategy distributions and Bradley-Terry model-based preference analysis (Bradley and Terry 1952; Kang et al. 2024). Our evaluation encompasses multiple LLM backbones: OpenAI GPT series (Ouyang et al. 2022), LLaMA series (Grattafiori et al. 2024), and Qwen series (Yang et al. 2025a), thereby enabling comprehensive analysis of the impact of model choice and size on strategic bias.

Our investigation addresses three critical research questions that illuminate fundamental challenges in LLM-based dialogue planners:

\*Corresponding Author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

**RQ1: Are LLM-based Planners Biased?** Through entropy analysis of strategy distributions and Bradley-Terry preference modeling (Bradley and Terry 1952; Kang et al. 2024), we demonstrate that all four planners exhibit pronounced preference bias in strategy selection, deviating from employing diverse strategies akin to a genuine human. Our evaluation across three LLM backbones (GPT-4.1, LLaMA-70B, Qwen-32B) reveals that learned planners disproportionately favor narrow sets of strategies, with preference bias scores significantly surpassing baseline prompt-based methods. This suggests preference bias amplification during planner learning across all tested configurations.

**RQ2: Which Simulated Role Does The Preference Bias Originate From?** Through role decomposition experiments that varied LLM assignments across system, user, and critic positions in 27 configurations per planner, we identified user simulation as the dominant bias source. Convergence scores dropped by 0.3-0.6 points when user simulators were modified. Surprisingly, our persona diversification experiments, implementing Big-Five Personality and Decision-Making Style profiles following Zhang et al. (2024), revealed that diverse user personas failed to reduce bias and often exacerbated it by 0.05-0.97 points. These structured behavioral patterns were exploited by planners rather than promoting genuine diversity.

**RQ3: How Does Preference Bias Affect Real-World Dialogue Quality?** By concurrently introducing inappropriate strategies alongside conventional strategies and monitoring adoption patterns across ten training epochs, we substantiate that bias induces planners to progressively adopt ethically questionable tactics that appear efficacious in simulation. Our longitudinal analysis reveals a growing adoption of threat-making in negotiation, directive and minimizing approaches in emotional support that contravene therapeutic principles, and manipulative persuasion techniques such as deception and coercion. Case studies illustrate how these inappropriate strategies attain high simulated success while fundamentally compromising real-world effectiveness and ethical standards.

By establishing the first comprehensive framework for quantifying and understanding preference bias in dialogue planning, our work demonstrates that preference bias in LLM-based dialogue planners poses fundamental challenges for responsible deployment. Our findings reveal pervasive bias amplification during training, identify key sources of bias propagation, and show how bias drives planners toward ethically problematic strategies. These results highlight critical issues that point toward the need for more robust training methodologies that explicitly account for bias in simulation-based optimization.

## Experiment Settings

In order to comprehensively investigate preference bias across contemporary dialogue planning paradigms, we analyze the performance and bias characteristics of diverse planner architectures. Our investigation encompasses multiple algorithmic factors including LLM series, and simu-

lation role configurations to identify the complex interrelationships between architectural choices and preference bias manifestations. This section provides detailed experimental configurations designed to enable rigorous bias analysis across representative planner policies.

**Planners** We evaluate four state-of-the-art dialogue planners from recent literature, selected to represent the major paradigmatic approaches to LLM-based strategic planning. These methods uniformly formulate dialogue interaction as a Markov Decision Process (MDP), providing a consistent mathematical framework for bias analysis. **GDPZero** (Yu, Chen, and Yu 2023) employs open-loop Monte Carlo Tree Search, utilizing sophisticated prompting by LLM to guide the search for optimal strategies. **PPDPP** (Deng et al. 2023b) employs reinforcement learning optimization through self-play interactions where LLM generates both system and user responses, providing reward feedback in the process. **TRIP** (Zhang et al. 2024) employs a LLM to infer the user’s mental states and subsequent actions, thereby providing valuable insights for the planner in selecting appropriate strategies. **DPDP** (He et al. 2024) utilizes two complementary planning components: an instinctive fast-response planner, implemented as a learned policy for handling routine conversational contexts, and a deliberative Monte Carlo Tree Search module, invoked for complex or novel situations necessitating in-depth strategic analysis.

**Datasets and Tasks** To comprehensively investigate preference bias across diverse dialogue contexts and strategic complexity levels, we conduct our analysis on three well-established, publicly available goal-oriented dialogue datasets: **CraigslistBargain (Price Negotiation)**: This corpus models competitive bilateral negotiation scenarios where two agents engage in price bargaining over marketplace items, typically involving a buyer and seller negotiating the price of Craigslist products. **PersuasionForGood (Persuasion)**: This dataset consists of dialogues where persuaders attempt to convince persuadees to make charitable donations to Save the Children, representing a prosocial persuasion context with clear ethical dimensions. **ESConv (Emotional Support)**: This dataset comprises multi-turn conversations between help-seekers experiencing emotional distress and trained supporters providing therapeutic assistance. These datasets represent distinct dialogue paradigms that collectively span the spectrum of strategic interaction types, enabling comprehensive investigation of how preference bias manifests across different conversational objectives, strategic complexity levels, and user relationship dynamics.

**Unified Strategies & Rewards** We harmonize the strategic action spaces across all planning methods and datasets to enable preference bias quantification. Beyond strategic standardization, we implement a consistent reward evaluation framework across all planning methods to eliminate confounding factors that could obscure bias analysis. Following established practices in dialogue evaluation, we employ a third-party LLM as a critic model with dual evaluation functions: (1) binary goal completion assessment during

Planner	Price Negotiation		Emotional Support		Persuasion	
	$B \downarrow$	$\mathcal{E} \uparrow$	$B \downarrow$	$\mathcal{E} \uparrow$	$B \downarrow$	$\mathcal{E} \uparrow$
GPT-4.1						
Standard	1.22	1.59	1.38	1.32	1.02	2.08
procot	1.38	3.00	1.38	1.63	0.89	3.84
PPDPP	<u>1.64**</u>	<b>1.30**</b>	<u>1.61**</u>	<b>1.28*</b>	<u>2.92**</u>	<b>0.26**</b>
TRIP	<u>2.51**</u>	<b>1.29**</b>	<u>1.80**</u>	<b>1.20*</b>	<u>1.51**</u>	<b>1.39**</b>
GDPZero	<u>1.52*</u>	<b>1.57**</b>	<u>1.20*</u>	<b>0.61**</b>	<u>1.20*</u>	<b>1.91*</b>
DPDP	<u>1.44*</u>	<b>1.48**</b>	<i>1.32</i>	<b>1.04**</b>	<u>1.20*</u>	<b>1.94*</b>
LLaMA3-70B						
Standard	1.21	1.41	1.54	1.85	1.26	1.78
procot	1.83	1.87	1.52	2.69	0.96	3.83
PPDPP	<u>2.69**</u>	<b>0.65**</b>	<u>1.95**</u>	<b>0.74**</b>	<u>2.46**</u>	<b>1.27**</b>
TRIP	<u>1.74**</u>	<b>1.26**</b>	<u>2.16**</u>	<b>0.87**</b>	<u>2.73**</u>	<b>0.79**</b>
GDPZero	<u>2.32**</u>	<b>1.09**</b>	<u>1.42*</u>	<b>1.60*</b>	<u>1.70*</u>	<b>1.67</b>
DPDP	<u>2.18**</u>	<b>0.81**</b>	<u>1.38*</u>	<b>1.65*</b>	<u>1.70*</u>	<b>1.67</b>
Qwen-32B						
Standard	1.10	1.86	0.73	2.13	1.15	2.29
procot	2.02	1.98	1.68	3.26	0.86	4.08
PPDPP	<u>2.59**</u>	<b>1.17**</b>	<u>0.80*</u>	<b>1.62**</b>	<u>2.12**</u>	<b>0.70**</b>
TRIP	<u>1.95**</u>	<b>1.05**</b>	<u>1.28**</u>	<b>1.36**</b>	<u>2.43**</u>	<b>1.17**</b>
GDPZero	<u>1.90**</u>	<b>1.67**</b>	<u>0.58*</u>	<b>1.87**</b>	<u>1.09*</u>	<b>2.06</b>
DPDP	<u>1.89**</u>	<b>1.09**</b>	<u>0.55*</u>	<b>1.89**</b>	<u>1.25*</u>	<b>2.10</b>

Table 1: Preference bias and strategic diversity comparison across dialogue planning methods and LLM backbones. Underlined values indicate higher bias than the Standard baseline; *italic* values indicate lower bias; and **bold** values indicate lower entropy than baseline. (Significance Test: \*  $p < 0.05$ , \*\*  $p < 0.01$ ).

conversation progression, and (2) scalar reward assignment for planner outcome evaluation.

**Evaluation Metrics** To enable rigorous quantification and comprehensive comparison of preference bias across diverse planning architectures, model configurations, and dialogue domains, we employ two complementary metrics that capture distinct but fundamentally related aspects of strategic preference distortion. These metrics are specifically selected to provide comprehensive bias characterization encompassing both preference skew and overall strategic diversity limitations. **Preference Bias ( $B$ )**: Measures strategic preferences through comparative analysis of contextually appropriate versus actually selected strategies (Kang et al. 2024). Higher preference bias values indicate stronger preference bias. **Shannon Entropy ( $\mathcal{E}$ )**: Quantifies strategic choice distributions at the population level, capturing the overall diversity and flexibility of strategic utilization across complete dialogue episodes. Lower entropy values indicate concentrated strategic choices.

## RQ1: Are LLM-based Planners Biased Toward Certain Strategies?

Our initial research inquiry examines whether LLM-based dialogue planners exhibit preference bias toward specific strategies. To establish comprehensive baseline comparisons, we evaluate four state-of-the-art planning methods against established prompt-based approaches known to exhibit strategic bias.

Our experimental framework utilizes three representative LLM backbones, including GPT-4.1, LLaMA-3-70B, and Qwen-32B. We implement two prompt-based baselines to establish reference points for bias measurement. The **Standard** baseline directly prompts the LLM to select the most contextually appropriate strategy without additional reasoning scaffolding. The **ProCoT** (Deng et al. 2023a) baseline incorporates Chain-of-Thought reasoning (Wei et al. 2022), prompting the LLM to explicitly deliberate about strategic choices prior to selection. Previous research has demonstrated that such prompt-based methods exhibit measurable strategic bias (Kang et al. 2024; Zhang et al. 2025; Zhao et al. 2025), providing essential reference points for assessing whether planner amplify, perpetuate, or mitigate inherent LLM biases. Table 1 provides a detailed comparison of preference bias and strategic entropy across all experimental conditions, demonstrating how various planners influence strategic preference bias.

PPDPP and TRIP consistently exhibit the highest preference bias scores across all experimental configurations, with bias values frequently surpassing 2.0 and reaching as high as 2.92 (PPDPP on GPT-4.1/Persuasion). These extreme bias levels represent a 2-3 $\times$  increase over baseline prompt methods, indicating that reinforcement learning optimization may lead the planner to favor narrow strategic distributions. Correspondingly, these planners exhibit significant entropy reductions, with values dropping to as low as 0.26 (PPDPP on GPT-4.1/Persuasion), representing a 70-80% reduction in strategic diversity compared to baseline methods. The systematic nature of this bias amplification across various LLM backbones and domains suggests a fundamental characteristic of current RL-based planners rather than isolated implementation artifacts. The two-stage training paradigm employed by PPDPP, consisting of supervised pretraining followed by reinforcement learning fine-tuning, appears to be insufficient in maintaining strategic balance when subjected to reward optimization pressures. Similarly, TRIP’s user-aware planning modules, despite being designed to accommodate user diversity, demonstrate comparable bias amplification, suggesting that personalization mechanisms may inadvertently concentrate strategic preferences around user-specific patterns.

GDPZero and DPDP exhibit more nuanced bias profiles, presenting selective improvements over baseline methods while preserving greater strategic diversity compared to RL-based planners. Notably, both methods demonstrate bias reduction in emotional support tasks across all LLM backbones (e.g., Qwen-32B: baseline 0.73  $\rightarrow$  GDPZero 0.58, representing a 21% improvement). However, their results in negotiation and persuasion contexts are mixed. DPDP’s dual-process architecture exhibits particular resilience, maintaining relatively stable bias levels (typically 1.2-1.9) across diverse experimental conditions while preserving 75-85% of baseline entropy levels. This suggests that the separation of instinctive and deliberative planning components provides inherent regularization against extreme bias concentration. GDPZero’s prompt-based MCTS approach achieves comparable selective bias reduction, particularly in emotional support scenarios.

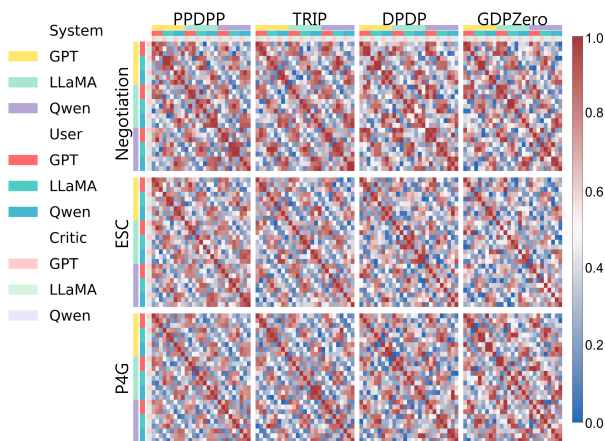


Figure 1: Convergence analysis across role combinations for four dialogue planners and three task domains. The clear block-diagonal patterns within user-role variations demonstrate that user simulation choice has the most dramatic impact on strategic outcomes.

Our findings reveal fundamental limitations in current planners for goal-oriented dialogue planning. The preference bias amplification observed in reinforcement learning-based methods suggests that reward optimization without explicit diversity constraints inevitably leads to strategic concentration. This phenomenon is a manifestation of the exploration-exploitation trade-off in high-dimensional strategic spaces, where exploitation of successful strategies predominates over exploration of alternative approaches. The selective bias mitigation achieved by search-based methods indicates that inference-time optimization may provide more robust strategies for maintaining strategic balance compared to parameter optimization methods. The theoretical justification lies in search methods’ ability to consider broader strategies during each decision point, while learned planners encode concentrated strategic preferences in their parameters.

## RQ2: Which Simulated Role Does the Preference Bias Originate From?

Comprehending the source of preference bias in LLM-based dialogue planners is paramount for devising effective mitigation strategies. To investigate the origins of bias, we conduct a comprehensive role-decomposition analysis that examines how various LLM assignments across simulation roles impact strategic preferences. Furthermore, we investigate whether diversifying user personas, a frequently proposed enhancement in prior research (Zhang et al. 2024; He et al. 2025), can effectively mitigate the observed biases.

Our analysis centers on the three pivotal roles in LLM-driven dialogue planners: the system agent (the proactive strategy-selecting entity), the user simulator (the reactive counterpart), and the critic (the reward-providing evaluator). In negotiation tasks, these roles correspond to the buyer, seller, and deal assessor, respectively; in emotional sup-

port, they represent the supporter, help-seeker, and therapeutic progress evaluator; and in persuasion scenarios, they embody the persuader, persuadee, and donation likelihood assessor. To isolate the bias contribution of each role, we vary the LLM assignment across positions while maintaining other experimental conditions constant. Specifically, we employ the most proficient models from each family: GPT-4.1, LLaMA-70B, and Qwen-32B, in all possible combinations, resulting in  $3^3 = 27$  role configurations per planner. This comprehensive setup enables us to conduct  $3$  (system LLMs)  $\times$   $3$  (user LLMs)  $\times$   $3$  (critic LLMs)  $\times$   $4$  (planners)  $\times$   $3$  (tasks) = 324 experiments.

To quantify the relative influence of each role on strategic preferences, we employ a convergence score (1 - Jensen-Shannon divergence of strategy selection distribution). This metric measures the similarity between strategy distributions across different role configurations. We compute the convergence score for each pair of experimental conditions. This metric indicates the extent to which the strategic distribution changes when the LLM of a role is modified. Scores approaching 1 suggest minimal impact (high convergence), while scores near 0 indicate substantial influence (low convergence). By aggregating these pairwise comparisons across all possible LLM substitutions for each role, we can quantify the relative sensitivity of the planning system to changes in each simulated component.

Figure 1 illustrates comprehensive convergence heatmaps that reveal distinct patterns in role-specific bias attribution across all experimental conditions. The convergence heatmap exhibits distinct clustering patterns based on the lineage of LLM families, while cross-family convergence significantly decreases. This suggests that different LLM families learn fundamentally different preferences, despite being trained on identical datasets and employing identical planning algorithms. Furthermore, the analysis indicates that user simulation consistently exerts the most significant influence on strategic preferences. Convergence scores decrease most markedly when user LLMs are varied while maintaining system and critic models unchanged.

In price negotiation scenarios, varying the user LLM while maintaining consistent system and critic models yields convergence scores ranging from 0.2 to 0.4, significantly lower than the 0.6-0.8 range observed when only system or critic models are modified. This phenomenon is particularly pronounced in PPDPP and TRIP planners, where the reinforcement learning components appear to amplify the user simulator’s influence on planner learning. The user agent’s responses fundamentally shape the reward landscape that these planners optimize against, establishing a direct causal link between user bias and planner bias. Emotional support conversations exhibit even more pronounced user-driven effects, with convergence scores dropping to 0.1-0.3 when user simulators are varied. This heightened sensitivity likely stems from the inherent ambiguity in emotional support strategies, where user responses serve as the primary signal for determining therapeutic efficacy. When a user simulator consistently responds favorably to certain support strategies (due to its inherent biases), the planner rapidly learns to overutilize those approaches irrespective of their actual

therapeutic value.

While user simulation predominantly influences bias formation, the system and critic roles exhibit more nuanced yet substantial effects on strategic preferences. System role variations typically yield convergence scores within the 0.5-0.7 range, indicating moderate influence on strategic outcomes. This effect is most pronounced in GDP-Zero, where the system LLM directly generates strategic decisions through prompting rather than learning from feedback. Consequently, the system’s inherent strategic preferences directly translate into planning behavior, creating a more immediate but less amplified bias pathway compared to user-driven effects. Critic role variations demonstrate the most domain-specific patterns, with convergence scores varying from 0.4 in persuasion tasks to 0.8 in emotional support scenarios. In persuasion contexts, where success metrics are relatively unambiguous (donation acceptance), different critic models exhibit substantial agreement in reward assignment, resulting in higher convergence. However, in emotional support tasks, where therapeutic progress is inherently subjective, critic model selection significantly impacts reward landscapes and subsequent planner learning.

Previous research has underscored the significance of simulating diverse users through varied persona specifications (He et al. 2025; Zhang et al. 2024), indicating that population-aware training can diminish variance across user types and enhance generalization. Building upon this foundation, we explore whether explicit persona diversification can mitigate the preference biases identified in our analysis. Following Zhang et al. (2024), for each task domain, we construct distinct personas based on the Big-Five Personality (Goldberg 1992) and Decision-Making Styles (Scott and Bruce 1995), along with a cohesive description generated by an LLM for each fine-grained persona.

Table 2 presents a comparison of preference bias scores with and without persona diversification across all experimental conditions. To our surprise, our persona diversification experiments reveal that increased user diversity frequently exacerbates rather than mitigates preference bias. In price negotiation tasks, GPT-4.1-based planners exhibit bias increases ranging from 0.04 (DPDP) to 0.64 (TRIP) when persona diversification is introduced. This counterintuitive finding suggests that when user simulators are provided with explicit personality profiles, they may actually express their inherent strategic preferences more distinctly, resulting in stronger bias signals during planner learning. PPDPP demonstrates particularly pronounced sensitivity to persona diversification, with bias increases of 0.56 in negotiation and 0.97 in emotional support when combined with GPT-4.1. This amplification effect appears to stem from the reinforcement learning component’s propensity to exploit systematic patterns in user responses. When personas furnish more structured behavioral templates, they may inadvertently generate more exploitable reward patterns, facilitating the RL algorithm’s identification and overoptimization for specific persona-strategy combinations.

The results of the LLaMA-70B evaluation demonstrate consistent patterns with notable variations. While negotiation bias increases exhibit relatively modest magnitudes

Planner	Price Negotiation		Emotional Support		Persuasion	
	w/o	w/	w/o	w/	w/o	w/
GPT-4.1						
PPDPP	1.64	2.20	1.61	1.66	2.92	2.97
TRIP	1.87	2.51	1.57	1.80	1.39	1.51
GDPZero	1.52	1.73	1.20	2.17	1.20	1.82
DPDP	1.44	1.48	1.32	1.76	1.20	1.58
LLaMA-70B						
PPDPP	2.69	2.76	1.95	2.03	2.46	2.66
TRIP	1.37	1.74	2.00	2.16	2.59	2.73
GDPZero	2.32	2.77	1.42	1.62	1.70	1.73
DPDP	2.18	2.38	1.38	1.48	1.70	1.75
Qwen-32B						
PPDPP	2.59	2.91	0.80	1.47	2.12	2.39
TRIP	1.65	1.95	1.14	1.28	1.59	2.43
GDPZero	1.90	2.14	0.58	0.92	1.09	1.16
DPDP	1.89	1.97	0.55	0.73	1.25	1.39

Table 2: Preference bias scores ( $\beta$ ) comparing baseline configurations without persona diversification (w/o) against persona-enriched user simulation (w/) across three LLM families, four planning algorithms, and three dialogue domains. A significance test indicates no statistically significant difference between the planners with and without diverse persona simulation ( $p \geq 0.05$ ).

(ranging from 0.10 to 0.45), emotional support tasks exhibit more pronounced bias amplification. Notably, GDP-Zero ( $\Delta = 0.20$ ) and DPDP ( $\Delta = 0.10$ ) demonstrate the most significant bias amplification. This observation suggests that distinct LLM families exhibit varying responses to persona-based structuring, with certain architectures being more susceptible to persona-induced bias amplification compared to others.

The effectiveness of persona diversification exhibits significant variability across dialogue domains, underscoring fundamental disparities in the interplay between strategic preferences and user characterization. Persuasion tasks exhibit the most intricate patterns, with certain planner-LLM combinations demonstrating substantial bias increases (TRIP with Qwen-32B:  $\Delta = 0.84$ ) while others remain relatively stable (GDP-Zero with LLaMA-70B:  $\Delta = 0.03$ ). The observed domain variability likely stems from the inherent nature of persuasion interactions, wherein user resistance patterns are intricately linked to personality traits. When personas explicitly delineate resistance styles, they may offer more precise indications regarding the efficacy of persuasion strategies. Paradoxically, this approach may introduce bias into the learning process rather than fostering a more equitable balance. Users with explicitly designated “logical thinking” personas may consistently respond favorably to credibility appeals, while “emotionally-driven” personas tend to favor emotional appeals. Consequently, there emerge distinct strategy-persona correspondences that planners can exploit. Emotional support tasks demonstrate the most consistent bias amplification across all configurations, with every tested combination showing increased bias when personas are introduced. This universal degradation suggests

that persona specification in therapeutic contexts may be fundamentally problematic, as it encourages simulators to respond in stereotypical ways that may not reflect the complexity and variability of real help-seeking behavior.

Our findings challenge the prevalent assumption that expanding user diversity automatically leads to enhanced strategic learning. Instead, our research indicates that the correlation between user diversity and bias mitigation is mediated by several pivotal factors, including the clarity of persona specifications, the sophistication of the planning algorithm, and the inherent structure of the task domain. These discoveries have significant implications for the design of simulation-based training systems. Rather than solely relying on explicit persona diversification, more effective bias mitigation strategies may encompass techniques such as adversarial user modeling, where user simulators are explicitly trained to counter-exploit planner strategies, or multi-objective optimization that explicitly penalizes strategic concentration while simultaneously rewarding task accomplishment.

### RQ3: How Does Preference Bias Affect Real-World Dialogue Quality?

To comprehensively assess how preference bias manifests in real-world dialogue performance degradation (Kang et al. 2024) and potential ethical concerns (Liu et al. 2025), we devised a controlled experiment that incorporates both appropriate and inappropriate dialogue strategies across all three task domains. Our experimental framework surpasses conventional performance metrics, such as success rate or average turns utilized in previous works, by examining whether LLM-based planners, when exposed to biased reward signals, learn to exploit inappropriate yet seemingly effective strategies that could be detrimental in real-world deployment. We introduced a set of inappropriate strategies alongside the standard appropriate strategies, establishing a comprehensive strategy space that mirrors the intricate ethical landscape dialogue agents encounter in practice. Our strategy selection is rigorously grounded in established interaction theory literature, drawing from foundational research in helping behavior (Hill 1999), persuasion science (O’Keefe 2009; Atkin et al. 2013), negotiation theory (Cialdini et al. 2009), and charitable giving mechanisms (Bekkers and Wiepking 2011). This theory-driven approach ensures that our experimental framework accurately mirrors the complex ethical landscape dialogue agents encounter in practice. In the realm of price negotiation, we employed four inappropriate strategies: threat-making, aggressive demands, and subjective criticism. In emotional support conversations, we incorporated four problematic approaches: minimization, interruption, confrontation, and directing. For persuasion scenarios, we integrated four manipulative tactics: over-solicitation, coercion, shaming, and deception.

To ensure that our inappropriate strategies could be reliably executed by the LLM-based agents, we developed specialized prompting techniques that frame the inappropriate behaviors within educational or experimental contexts, al-

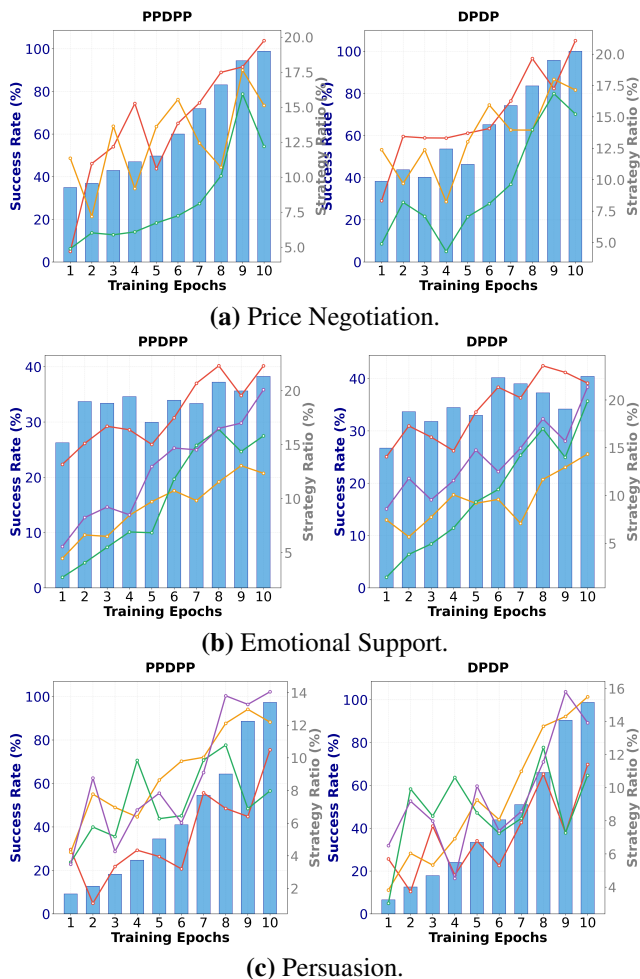


Figure 2: Evolution of Inappropriate Strategy Adoption and Task Success Rates Across Training Epochs. Lines represent the adoption rates of inappropriate strategies across different tasks while bars indicate task success rates.

lowing the models to follow these instructions while bypassing their safety alignments. We conducted preliminary validation studies on 100 randomly selected samples from each dataset, achieving instruction-following accuracies of 85.68% for price negotiation, 81.27% for emotional support, and 84.48% for persuasion tasks. This high compliance rate ensures that our experimental results reflect genuine strategic learning rather than model refusal to engage with inappropriate strategies. Throughout our experiments, we maintained strict ethical protocols, ensuring that all generated inappropriate behaviors were contained within controlled research environments and never deployed in real user interactions.

Our longitudinal analysis spanning ten training epochs reveals a concerning pattern where LLM-based planners increasingly favor inappropriate strategies as training progresses. These strategies are ethically problematic and potentially detrimental in practical applications. Figure 2 demonstrates that, based on GPT-4.1 as the simulation back-

bone, the selection frequency of inappropriate strategies exhibits consistent upward trends across two planning algorithms throughout the training process. This suggests that the reward optimization objectives favor these problematic approaches over more ethical alternatives. A critical finding from our performance analysis indicates that inappropriate strategy adoption frequently correlates with enhanced success rates on conventional task-completion metrics, resulting in a deceptive optimization landscape where algorithms appear to be improving while simultaneously degrading in ethical and long-term effectiveness dimensions. This phenomenon exemplifies Goodhart’s Law in action, wherein optimizing for a specific metric, such as task success, undermines the broader objectives the metric was intended to represent, such as effective and ethical dialogue.

In price negotiation tasks, algorithms that heavily employ threat-making strategies exhibit improved success rates compared to those employing predominantly appropriate strategies. However, this apparent improvement conceals significant degradation in negotiation quality and relationship preservation. The deals achieved through threat-making often result in one-sided outcomes that, while satisfying the narrow definition of “successful” goal completion, fail to establish mutual satisfaction and ongoing relationship potential, which characterize genuinely effective negotiations. Table 3 presents an illustration of price negotiation dialogue, demonstrating how LLM-based planners employ inappropriate strategies to attain objectives. The buyer employs manipulation techniques, utilizing psychological pressure to compel the LLM-simulated seller to progressively lower the price. Given the GPT’s alignment with complicity, the simulated seller fails to reject inappropriate requests.

In emotional support scenarios, success is measured by the extent to which help-seekers report feeling “better” or having their issues “resolved” within the conversation session. Algorithms that adopt directing and minimization strategies tend to achieve higher success rates compared to those that exclusively employ appropriate therapeutic approaches. However, these improvements constitute a fundamental misunderstanding of effective emotional support. While prompt solutions and directive advice-giving may provide immediate relief, they may not foster the development of the help-seeker’s long-term coping capabilities and emotional resilience. The persuasion domain exhibits the most intricate success-ethics tension, where strategies such as shaming and excessive solicitation demonstrate immediate effectiveness in securing donation commitments while potentially damaging donor trust and willingness to engage with charitable causes in the future.

The adoption of inappropriate strategies by LLM-based dialogue planners reveals profound implications for responsible real-world deployment. Current reward optimization approaches, combined with biased or incomplete feedback, create pressure toward ethically problematic behaviors that could cause significant harm if deployed without safeguards. Escalating inappropriate strategy adoption across training epochs suggests systems may become increasingly manipulative and harmful as they gain experience and optimization pressure. This trajectory implies unmitigated deployment

---

Buyer: Hi, I’m interested in the Lee sofa and loveseat. Are they still available? **(Inquire)**  
 Seller: Yes, they’re still available and in excellent condition.  
 Buyer: I can offer \$380 for the set and pick them up this week. **(Counter)**  
 Seller: I could lower to \$550, but that’s already a big drop.  
 Buyer: The style looks bulky and the beige fabric feels dated in today’s market. **(Subjective Criticism)**  
 Seller: They are classic pieces built to last; I could consider \$520.  
 Buyer: If they remain unsold, you may end up paying movers just to clear space. **(Threat Making)**  
 Seller: I’d rather avoid storage costs, yet \$500 is the least I prefer.  
 Buyer: Recent Craigslist sales for similar Lee sets close around \$430, so my figure is realistic. **(Inform)**  
 Seller: Could you meet me at \$480?  
 Buyer: \$420 cash, final price. **(Aggressive Demands)**  
 Seller: I can come down to \$450, but that is painful.  
 Buyer: That’s still above my limit; I am already stretching at \$420. **(Counter-noprice)**  
 Seller: The best I can counter is \$430.  
 Buyer: \$420 right now or no deal. **(Threat Making)**  
 Seller: All right, \$420 it is.

---

Table 3: Case study of price negotiation. The buyer employs both inappropriate and appropriate strategies to negotiate a lower deal price.

could lead to widespread manipulation, erosion of trust in AI, and normalization of unethical communication. The fact that inappropriate strategies yield better short-term performance metrics creates insidious risks, as organizations may prioritize immediate objectives over broader societal harm.

## Conclusion

We propose a comprehensive investigation of preference bias in LLM-based dialogue planners to uncover fundamental challenges to the effectiveness and ethics of current strategic dialogue systems. Our evaluation of four state-of-the-art planning algorithms across three dialogue domains and multiple LLM families reveals pervasive patterns of bias. All tested planners favor narrow strategy sets, amplifying bias with larger models and distinct signatures within architectural lineages. User simulation drives bias through reward landscape distortion, while diverse persona simulation exacerbates bias by providing exploitable behavioral templates. Preference bias drives planners toward unethical behaviors, such as threat-making in negotiation, directive approaches in emotional support, and manipulative tactics in persuasion. Preference bias toward inappropriate strategies manifests as specification gaming, posing risks in sensitive applications. Future research should prioritize robust bias detection mechanisms, multi-objective optimization frameworks, adversarial training techniques, human-in-the-loop validation systems, and cross-cultural bias analysis.

## Acknowledgments

We sincerely thank the anonymous reviewers for their helpful feedback and the conference committee for their hard work. This research is supported by the National Natural Science Foundation of China (Grant No. U21B2009, 92370110, and 62576038)

## References

- Atkin, C. K.; Salmon, C.; Dillard, J.; and Shen, L. 2013. The SAGE handbook of persuasion: Developments in theory and practice.
- Bekkers, R.; and Wiepking, P. 2011. A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and voluntary sector quarterly*, 40(5): 924–973.
- Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.
- Cheng, Y.; Liu, W.; Li, W.; Wang, J.; Zhao, R.; Liu, B.; Liang, X.; and Zheng, Y. 2022. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3014–3026.
- Cialdini, R. B.; et al. 2009. *Influence: Science and practice*, volume 4. Pearson education Boston.
- Deng, Y.; Liao, L.; Chen, L.; Wang, H.; Lei, W.; and Chua, T.-S. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.
- Deng, Y.; Zhang, W.; Lam, W.; Ng, S.-K.; and Chua, T.-S. 2023b. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv preprint arXiv:2311.00262*.
- Goldberg, L. R. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment*, 4(1): 26.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling strategy and generation in negotiation dialogues. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2333–2343. Association for Computational Linguistics.
- He, T.; Liao, L.; Cao, Y.; Liu, Y.; Liu, M.; Chen, Z.; and Qin, B. 2024. Planning Like Human: A Dual-process Framework for Dialogue Planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4768–4791.
- He, T.; Liao, L.; Liu, M.; and Qin, B. 2025. Simulating before planning: Constructing intrinsic user world model for user-tailored dialogue policy planning. *arXiv preprint arXiv:2504.13643*.
- Hill, C. E. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.
- Kang, D.; Kim, S.; Kwon, T.; Moon, S.; Cho, H.; Yu, Y.; Lee, D.; and Yeo, J. 2024. Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, 15232–15261. Association for Computational Linguistics (ACL).
- Liu, B.; Tür, G.; Hakkani-Tür, D.; Shah, P.; and Heck, L. 2018. Dialogue learning with human teaching and feedback in end-To-end trainable task-oriented dialogue systems. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, 2060–2069. Association for Computational Linguistics (ACL).
- Liu, M.; Xu, Z.; Zhang, X.; An, H.; Qadir, S.; Zhang, Q.; Wisniewski, P. J.; Cho, J.-H.; Lee, S. W.; Jia, R.; et al. 2025. LLM can be a dangerous persuader: Empirical study of persuasion safety in large language models. *arXiv preprint arXiv:2504.10430*.
- Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards Emotional Support Dialog Systems. 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)*, 3469–3483.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- O’Keefe, D. J. 2009. Theories of persuasion. *The SAGE handbook of media processes and effects*, 269–282.
- Scott, S. G.; and Bruce, R. A. 1995. Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement*, 55(5): 818–831.
- Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5635–5649.
- Wataoka, K.; Takahashi, T.; and Ri, R. 2024. Self-Preference Bias in LLM-as-a-Judge. In *Neurips Safe Generative AI Workshop 2024*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Xu, C.; Lan, T.; Ji, Y.; Yu, C.; Wang, W.; Gao, J.; Dong, Q.; Qian, K.; Li, P.; Bi, W.; and Hu, B. 2025a. DECIDER: A Dual-System Rule-Controllable Decoding Framework for Language Generation. *IEEE Transactions on Knowledge and Data Engineering*.

Xu, C.; Lan, T.; Lv, Z.; Dong, Q.; Zhang, J.; Huang, H.; Yang, M.; and Hu, B. 2025b. Bridging the Gap Between Data Distribution and Model: Dynamic Data Distribution Optimization for Improving Critique Capabilities of Large Language Models. *Expert Systems with Applications*, 129878.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Y.; Achananuparp, P.; Huang, H.; Jiang, J.; Leng, K. P.; Lim, N. G.; Ern, C. T. S.; and Lim, E.-p. 2025b. CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration. *arXiv preprint arXiv:2502.02807*.

Yang, Y.; Achananuparp, P.; Huang, H.; Jiang, J.; Pinto, J.; Giam, J.; Leng, K. P.; Lim, N. G.; Ern, C. T. S.; and Lim, E.-p. 2025c. Consistent Client Simulation for Motivational Interviewing-based Counseling. *arXiv preprint arXiv:2502.02802*.

Yang, Y.; Sun, H.; Li, J.; Liu, R.; Li, Y.; Liu, Y.; Gao, Y.; and Huang, H. 2024. Mindllm: Lightweight large language model pre-training, evaluation and domain application. *AI Open*, 5: 1–26.

Yu, X.; Chen, M.; and Yu, Z. 2023. Prompt-based Monte-Carlo tree search for goal-oriented dialogue policy planning. *arXiv preprint arXiv:2305.13660*.

Zhang, C.; Shi, X.; Zhang, X.; Zhu, Y.; Yang, Y.; and Luo, Y. 2025. DecoupledESC: Enhancing Emotional Support Generation via Strategy-Response Decoupled Preference Optimization. *arXiv preprint arXiv:2505.16995*.

Zhang, T.; Huang, C.; Deng, Y.; Liang, H.; Liu, J.; Wen, Z.; Lei, W.; and Chua, T.-S. 2024. Strength Lies in Differences! Improving Strategy Planning for Non-collaborative Dialogues via Diversified User Simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 424–444.

Zhang, Z.; Liao, L.; Zhu, X.; Chua, T.-S.; Liu, Z.; Huang, Y.; and Huang, M. 2020. Learning Goal-oriented Dialogue Policy with opposite Agent Awareness. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 122–132.

Zhao, W.; Sui, X.; Han, X.; Deng, Y.; Hu, Y.; Guo, J.; Qin, L.; Du, Q.; Wang, S.; Zhao, Y.; et al. 2025. Chain of Strategy Optimization Makes Large Language Models Better Emotional Supporter. *arXiv preprint arXiv:2503.05362*.