Constrained Human Preference Alignment for Natural Language Planning with LLMs

Yu Zhou
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR, China
zy-yu.zhou@connect.polyu.hk

Haokai Hong*
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR, China
haokai.hong@connect.polyu.hk

Ran Cheng
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR, China
ranchengcn@gmail.com

Kay Chen Tan
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hong Kong SAR, China
kctan@polyu.edu.hk

***Abstract*—Recent advances in large language models (LLMs) have established them as promising candidates for natural language planning tasks. However, existing approaches often fail to address two critical challenges: 1) the effective alignment of LLM-generated plans with human preferences, and 2) the dynamic enforcement of diverse constraints inherent in planning scenarios. To bridge these gaps, we propose a constraint-aware human-preference alignment framework for natural language planning. Our contributions are threefold. First, we design a process reward model that aligns LLM outputs with human preferences through step-by-step feedback, facilitating efficient and interpretable preference learning. Second, we develop a constraint-aware mechanism integrated into the rewriting strategy, which dynamically penalizes violations of task-specific constraints at each reasoning step. Third, we introduce a unified adaptive metric enabling a multifaceted assessment of planning quality. We validate our framework through experiments on planning benchmarks, demonstrating improvements in success rate with constraints and human preference alignment over baselines.**

***Index Terms*—LLM, planning, preference alignment, constraint**

I. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks [1]. Recent advances have positioned LLMs as promising solutions for natural language planning, where the goal is to generate a sequence of actions to achieve a desired objective [2]. However, current LLM-based planning methods often fall short in two critical aspects: efficiently aligning LLM outputs with nuanced human preferences and systematically handling the diverse constraints inherent in real-world planning scenarios. Specifically, aligning the behavior of LLMs with human preferences remains a core challenge in artificial intelligence. While reinforcement learning from human feedback [3] has proven effective in preference alignment, its application to complex planning tasks is often limited by sparse rewards and the difficulty of providing holistic feedback on long-term outputs. Additionally, real-world planning further requires adherence to diverse constraints, LLM-generated plans risk infeasibility or logical inconsistency [4].

To address these gaps, we propose a novel constraints-aware human-preference alignment framework for natural language planning tasks. Our approach consists of three key components: First, we introduce a process reward model designed to align plans with human preferences through stepwise feedback. This model enables more efficient and interpretable preference learning by providing intermediate rewards based on the progress toward the goal. Second, we integrate a constraint-aware mechanism into the rewriting strategy, which dynamically enforces adherence to task-specific constraints. Third, recognizing the limitations of existing evaluation metrics, we propose a unified indicator that quantifies performance across environment, commonsense, and hard constraints. This metric enables a holistic assessment of planning capabilities under multifactorial requirements. Our work advances the integration of LLMs into real-world planning systems by bridging critical gaps in preference alignment and constraint-aware reasoning.

II. Method

In this paper, we propose a novel framework based on evolutionary computation [5] for efficiently searching solutions in natural language planning tasks. Each candidate solution is represented as a natural-language sequence. As illustrated in Fig. 1, this framework follows a generate–evaluate–refine structure where an LLM first produces an initial population of N candidate solutions, then preference-guided evaluation and constraint-aware rewriting strategies are executed repeatedly until a maximal fitness evaluation is reached. This iterative

This work was supported in part by National Natural Science Foundation of China (Grant No. U21A20512), Research Grants Council of the Hong Kong SAR (Grant No. C5052-23G, PolyU15229824, PolyU15218622, PolyU15215623), and The Hong Kong Polytechnic University (Project IDs: P0053758, P0051130, P0052694).

* Corresponding author

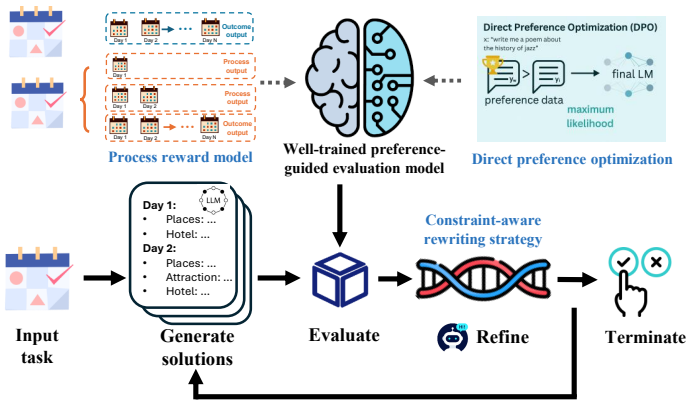

Fig. 1: The proposed method follows a generate–evaluate-refine evolutionary process. In this process, candidate solutions are generated by an LLM, and then evaluated and refined via the proposed evaluation model and constraint-aware strategy.

procedure improves population quality while enforcing task-specific constraints and preference alignment.

1) Preference-Guided Evaluation: To align candidate plans with human preferences and accelerate the search process, we adopt a process reward model based on a fine-tuned LLM about planning tasks as an evaluation model. Considering a plan x as a multi-step sequence, this PRM assigns a step-wise score to each step and aggregates them into a global successful score/rate with preference denoted as $R(x) = \sum_{t=1}^{D} r_t(x)$. This score is the fitness value of a candidate solution. This process feedback mitigates sparse-reward issues and accelerates the convergence of the evolutionary process. The PRM is trained offline via direct preference optimization. After being well-trained, the evaluated model with PRM supplies step-dense rewards with preference.

2) Constraint-Aware Structure-Preserving Rewriting Strategy: In each generation, constraint-aware structure-preserving rewriting (CSPR) improves candidate solutions through four stages: identify, feedback, rewrite, and generate. The process begins by segmenting each candidate solution into fragments, where any fragment that satisfies may have a suboptimal preference score or violate constraints is flagged as an error. For each error segment, an LLM then generates structured rewriting suggestions by integrating feedback from reward and penalty values. Based on these suggestions, an LLM is used to rewrite the selected fragments to produce improved segments and form a revised solution. All revised candidate solutions are added to the population in the next iteration. The rewriting modes of CSPR consist of self-refinement, where each solution modifies its segments, and pairwise crossover, which merges high-quality fragments from two solutions into a hybrid one.

3) Novel Evaluation Metric Design: After that, we propose a novel evaluation metric based on the successful rate with preference and constraints (i.e., environmental feasibility, commonsense consistency, and hard constraint), which can denoted as: $f_{\text{metric}}(x) = R(x) - \gamma_1 \cdot g_{\text{env}}(x) - \gamma_2 \cdot g_{\text{com}}(x) - \gamma_3 \cdot g_{\text{hard}}(x)$.

TABLE I: Performance comparison across different methods

Method	Successful rate with constraints (% ↑)			Human preference alignment (↑)		
	Travel Planner	Neural Plan Trip Task	Neural Plan Meeting Task	LLaMA 8B-RM	Gemma2 2B-RM	Qwen2.5 7B-RM
1-Pass	3.56	18.19	18.82	86.17	82.36	82.58
GPT-0 1-Pass	9.44	35.81	41.80	85.88	81.49	83.92
Best-of-N	55.28	74.77	67.76	86.87	81.57	84.22
Sequential-Revision+	80.52	72.01	61.91	87.30	82.93	80.85
CoT	45.72	40.49	51.78	86.19	82.38	83.00
ToT	49.76	41.80	52.67	85.43	83.93	84.59
EvoAgent	75.70	52.55	64.30	86.62	82.61	83.28
Our Method	**84.31**	**87.11**	**80.74**	**89.03**	**86.44**	**86.15**

III. Experiment

We select two representative natural language planning tasks, namely TravelPlanner [2] and Natural Plan to evaluate the effectiveness of our approach. We compare our method with the following baselines: 1-Pass, GPT-0 1-Pass, Best-of-N, Sequential-Revision+, Chain-of-Thought (CoT) [6], Tree-of-Thought (ToT) [7], and EvoAgent. The experimental results are provided in Table I. In detail, the left part of Table I presents the success score of different methods across three planning tasks, and our method demonstrates superior adaptability and optimization capability across different tasks. The right part of Table I reports alignment results under three reward models (i.e., Llama-8B-RM, Gemma2-2B-RM, and Qwen2.5-7B-RM). Our method outperforms all baselines across these models.

IV. Conclusion

This work proposed a constraint-aware human-preference alignment framework for LLM-based natural language planning, integrating a PRM with stepwise feedback for preference learning, a constraint mechanism enforcing compliance, and a unified adaptive metric for multifaceted evaluation. Experiments on planning benchmarks demonstrate improvements over baselines in success rate and human preference alignment.

References

[1] X. Wu, S.-H. Wu, J. Wu, L. Feng, and K. C. Tan, "Evolutionary computation in the era of large language model: Survey and roadmap," *IEEE Transactions on Evolutionary Computation*, vol. 29, no. 2, pp. 534–554, 2025.

[2] J. Xie, K. Zhang, J. Chen, T. Zhu, R. Lou, Y. Tian, Y. Xiao, and Y. Su, "Travelplanner: A benchmark for real-world planning with language agents," in *Proceedings of the 41st International Conference on Machine Learning*, pp. 54590–54613, 2024.

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 27730–27744, 2022.

[4] K.-H. Lee, I. Fischer, Y.-H. Wu, D. Marwood, S. Baluja, D. Schuurmans, and X. Chen, "Evolving deeper LLM thinking," *arXiv preprint arXiv:2501.09891*, 2025.

[5] Y. Huang, S. Wu, W. Zhang, J. Wu, L. Feng, and K. C. Tan, "Autonomous multi-objective optimization using large language model," *IEEE Transactions on Evolutionary Computation*, 2025. early access, doi: 10.1109/TEVC.2025.3561001.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 24824–24837, 2022.

[7] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: deliberate problem solving with large language models," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 11809–11822, 2023.