# PGPO: Enhancing Agent Reasoning via Pseudocode-style Planning Guided Preference Optimization

Zouying Cao[1,3,4], Runze Wang[2], Yifei Yang[1,3,4], Xinbei Ma[1,3,4], Xiaoyong Zhu[2], Bo Zheng[2,*], Hai Zhao[1,3,4,*]

[1]School of Computer Science, Shanghai Jiao Tong University, [2]Taobao & Tmail Group of Alibaba
[3]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University,
[4]Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3

## Introduction

- Due to the dominance of natural language (NL) in agent reasoning, existing researches mainly focus on generating NL plans. We investigate the effectiveness of **pseudocode-style plans (P-code Plan)** in agent reasoning, which are more concise and structured than NL plans. P-code Plan demonstrates its superiority in boosting the generalization ability of LLM agents.
- We further propose a pseudocode-style **P**lanning **G**uided **P**reference **O**ptimization method named **PGPO** which empowers LLM agents with enhanced reasoning capabilities under the guidance of pseudocode-style plans. Analyses reveal the advantage of PGPO in reducing action errors and omissions during reasoning.

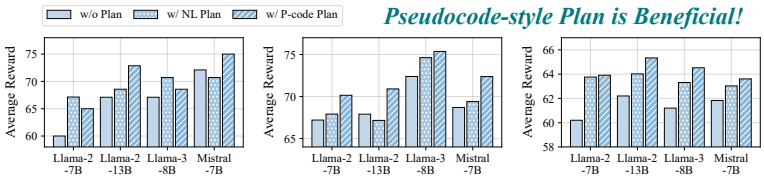



Figure 1: Comparison between w/ P-code Plan and w/o Plan, w/ NL Plan during the SFT process for LLM agents.

## P-code Plan Generation Pipeline

- **Thought Extraction**: extract agent thoughts from existing ReAct-style expert trajectories, which involves task-specific knowledge for planning.
- **Plan Distillation**: instruct powerful LLM to summarize the step-by-step plan following the predefined P-code Plan format via few-shot prompting strategy.
- **Plan Verification**: minor manual refinement is needed to guarantee the quality of the LLM-generated P-code Plans.

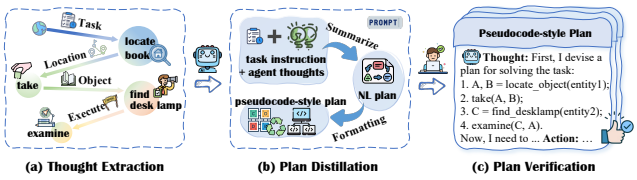
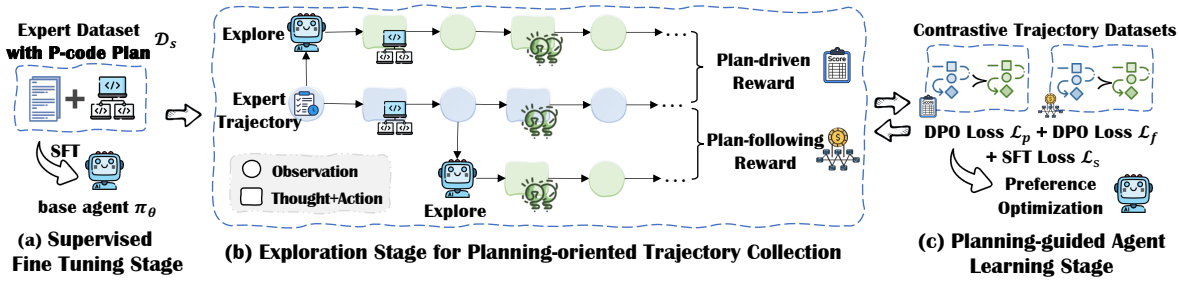



Figure 2: Overview of P-code Plan generation pipeline. We first extract the thought part from existing ReAct-style datasets. Then, we prompt GPT-4o to summarize the thought process into high-level plans. Pseudocode-style plans are finally structured with predefined formats, followed by manual verification to ensure accuracy.

## Methodology: P-code Plan-Guided Agent Learning




- **Contact us:**
zouyingcao@sjtu.edu.cn
yunze.wrz@alibaba-inc.com
zhaohai@cs.sjtu.edu.cn

- **Full paper**

## Experiments

### ① Main Results.

*new state-of-the-art performance*

| Method | Llama-2-7B ALFWorld Seen | Llama-2-7B ALFWorld UnSeen | Llama-2-7B WebShop | Llama-2-7B TextCraft | Llama-2-7B Avg. | Llama-2-13B ALFWorld Seen | Llama-2-13B ALFWorld UnSeen | Llama-2-13B WebShop | Llama-2-13B TextCraft | Llama-2-13B Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| SFT | 60.0 | 67.2 | 60.2 | 28.0 | 53.9 | 67.1 | 67.9 | 62.2 | 29.0 | 56.6 |
| ETO | 68.6 | 72.4 | 67.4 | 35.0 | 60.9 | 75.0 | 69.4 | 68.9 | 42.0 | 63.8 |
| IPR | 70.3 | 74.7 | 71.3 | 34.0 | 62.6 | 75.0 | 76.9 | 72.2 | 39.0 | 65.8 |
| PGPO | **76.4** | **76.9** | **72.2** | **43.0** | **67.1** | **77.1** | **77.6** | **73.7** | **48.0** | **69.1** |

| Method | Llama-3-8B ALFWorld Seen | Llama-3-8B ALFWorld UnSeen | Llama-3-8B WebShop | Llama-3-8B TextCraft | Llama-3-8B Avg. | Mistral-7B ALFWorld Seen | Mistral-7B ALFWorld UnSeen | Mistral-7B WebShop | Mistral-7B TextCraft | Mistral-7B Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| SFT | 67.1 | 72.4 | 61.2 | 20.0 | 55.2 | 72.1 | 68.7 | 61.8 | 31.0 | 58.4 |
| ETO | 72.1 | 73.1 | 66.2 | 36.0 | 61.9 | **75.0** | 72.4 | 66.2 | 38.0 | 62.9 |
| IPR | 72.9 | 73.9 | 72.0 | 38.0 | 64.2 | 73.6 | 73.1 | **69.6** | 36.0 | 63.1 |
| PGPO | **75.0** | **76.9** | **72.3** | **46.0** | **67.6** | **75.0** | **77.6** | 69.0 | **45.0** | **66.7** |

Table 1: Main results of PGPO compared to training-based baselines on ALFWorld, WebShop and TextCraft.

*alleviate the need for few-shot context achieve comparable or even better results*

| Method | ALFWorld Seen | ALFWorld UnSeen | WebShop | TextCraft |
|---|---|---|---|---|
| ReAct+GPT-4 | 42.9 | 38.1 | 63.2 | 28.0 |
| ReAct+GPT-3.5 | 7.9 | 10.5 | 62.4 | 20.0 |
| ADaPT+GPT-4 | 75.0 | 69.4 | 64.8 | **48.0** |
| ADaPT+GPT-3.5 | 70.3 | 71.6 | 62.7 | 26.0 |
| **PGPO**+Llama-2-7B | **76.4** | **76.9** | 72.2 | 43.0 |
| **PGPO**+Llama-3-8B | 75.0 | **76.9** | **72.3** | 46.0 |

Table 2: Comparative experiments vs. prompt-based baselines.

### ② Approach Ablations.

| | ALFWorld Seen | ALFWorld UnSeen | WebShop | TextCraft |
|---|---|---|---|---|
| **PGPO** | 76.4 | 76.9 | 72.2 | 43.0 |
| - *P-code* | 75.7↓ 0.7 | 71.6↓ 5.3 | 69.6↓ 2.6 | 40.0↓ 3.0 |
| - $L_f$ | 74.3↓ 2.1 | 75.4↓ 1.5 | 70.4↓ 1.8 | 41.0↓ 2.0 |
| - $L_s$ | 69.3↓ 7.1 | 68.7↓ 8.2 | 64.8↓ 7.4 | 35.0↓ 8.0 |

Table 3: Approach ablations of PGPO. - *P-code* represents using NL plans to replace P-code Plans. - $L_f$ denotes leaving out the estimation of plan-following reward, followed by the removal of $L_f$. - $L_s$ means the removal of SFT loss.

### ③ Ablation on Optimization Iterations.

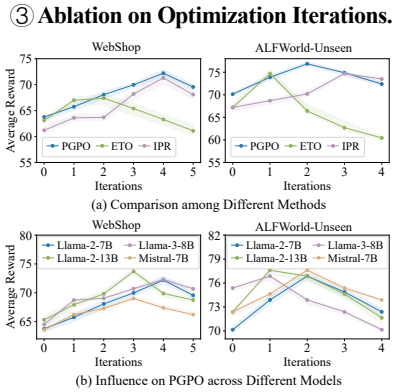



Figure 3: Ablation study on optimization iterations. (a) provides a comparison of the performance of PGPO against ETO and IPR across varying iterations. (b) shows the influence of increasing iterations on PGPO across different base models. iter=0 is the SFT stage.

## Analysis

### ④ Analysis on training time efficiency.

*PGPO delivers a 9% performance improvement while maintains reasonable training efficiency, requiring less than twice the time cost of ETO.*

### ⑤ P-code Plan guidance can reduce the incidence of action errors and omissions in reasoning.

*invalid action rate*

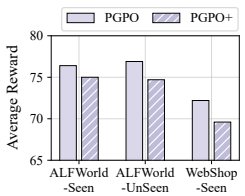
| Method | ALFWorld Seen | ALFWorld UnSeen |
|---|---|---|
| ETO | 34.28% | 32.83% |
| IPR | 30.71% | 28.35% |
| PGPO | **23.57%** | **26.86%** |

&

*success rate*

| Method | Webshop |
|---|---|
| ETO | 37.5 |
| IPR | 40.5 |
| PGPO | **41.0** |

### ⑥ Step-wise reward does not necessarily elicit better LLM agents.




*It is still challenging to accurately determine the contribution of the intermediate step, thus introducing step-wise reward instead plays a negative role in agent reasoning.*

## Conclusion

- P-code Plan can capture efficient structural logic of reasoning compared with NL plans, suitable for LLM agent's generalization to analogous agent tasks.
- With two planning-oriented rewards, PGPO further enhances LLM agents' ability to generate high-quality P-code Plans and subsequent reasoning.