Revealing the Barriers of Language Agents in Planning

Jian Xie[♠] Kexun Zhang[♡*] Jiangjie Chen[◇*†] Siyu Yuan[♠]
Kai Zhang[♣] Yikai Zhang[♠] Lei Li[♡] Yanghua Xiao[♠]
[♠]Fudan University [♡]Carnegie Mellon University
[◇]ByteDance Inc. [♣]The Ohio State University
{jianxie22, syyuan21, ykzhang22}@m.fudan.edu.cn, kexun@cmu.edu
jiangjiec@bytedance.com, zhang.13253@osu.edu, leili@cs.cmu.edu, shawyh@fudan.edu.cn

Abstract

Autonomous planning has been an ongoing pursuit since the inception of artificial intelligence. Based on curated problem solvers, early planning agents could deliver precise solutions for specific tasks but lacked generalization. The emergence of large language models (LLMs) and their powerful reasoning capabilities has reignited interest in autonomous planning by automatically generating reasonable solutions for given tasks. However, prior research and our experiments show that current language agents still lack human-level planning abilities. Even the state-of-the-art reasoning model, OpenAI o1, achieves only 15.6% on one of the complex real-world planning benchmarks. This highlights a critical question: ***What hinders language agents from achieving human-level planning?*** Although existing studies have highlighted weak performance in agent planning, the deeper underlying issues and the mechanisms and limitations of the strategies proposed to address them remain insufficiently understood. In this work, we apply the feature attribution study and identify two key factors that hinder agent planning: the **limited role of constraints** and the **diminishing influence of questions**. We also find that although current strategies help mitigate these challenges, they do not fully resolve them, indicating that agents still have a long way to go before reaching human-level intelligence. Resources are available on the GitHub.

1 Introduction

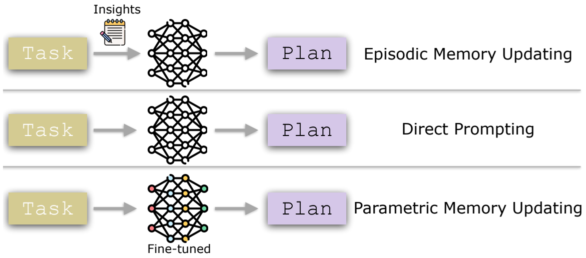

Figure 1: Memory updating strategies for language agents. Insights are learned from previous attempts.

Planning is the process of determining the sequence of actions needed to achieve a goal. It involves goal decomposition, constraint consideration, and foresight for simulating and predicting outcomes. In the development of artificial intelligence, this capability is considered the "Holy Grail" for achieving or even surpassing human intelligence (Kahneman, 2011; OpenAI, 2023b). However, the path to achieving autonomous planning is a long journey. Researchers have long focused on building custom systems tailored to specific tasks (Newell et al., 1959; McDermott, 1992; Silver et al., 2017). While these systems could deliver precise solutions through rigorous problem solvers, the extensive effort required for task-specific design prevents them from achieving universal problem-solving capabilities or general intelligence.

The advent of language agents (Weng, 2023; Su, 2023; Sumers et al., 2024), which are powered by large language models (LLMs; OpenAI (2022, 2023a); G Team et al. (2023); Dubey et al. (2024); Yang et al. (2024)), changes the landscape. Thanks to the flexibility of natural language, LLM-based language agents have shown strong potential to generalize to various planning tasks without relying on traditional curated, task-specific solvers written in domain-specific languages like Planning Domain Definition Language (PDDL). However, despite these language agents demonstrating impressive capabilities across various tasks (Yao et al., 2022, 2023; Zheng et al., 2024a; Gu et al., 2024), their performance in planning remains disappointing and is viewed as mere "*approximate retrieval*" (Kambhampati et al., 2024) rather than engaging in genuine reasoning. Specifically, even the most capable model, OpenAI o1 (OpenAI, 2024), which claims to surpass human PhD-level accuracy on several

[*] Equal Contribution

[†] Part of the work done while at Fudan University.

reasoning tasks, achieves only 15.6% in a real-world travel planning benchmark, TravelPlanner (see Figure 2), far below human-level planning abilities. To uncover the fundamental reasons behind the weak performance, we seek to answer the first research question in this paper: ***RQ1: Why do current language agents struggle with planning?***

In order to enhance language agents' performance in planning tasks, numerous strategies have been proposed recently, which can be categorized into three main branches, as shown in Figure 1: episodic memory updating through prompt optimization (Zhao et al., 2024; Shinn et al., 2024; Fu et al., 2024), parametric memory updating through model training (Zeng et al., 2023a; Song et al., 2024; Yin et al., 2024), and translating queries into formal planning languages, followed by resolution using external solvers (Liu et al., 2023; Dagan et al., 2023). Although these strategies have shown performance improvements across various tasks, their underlying mechanisms remain largely opaque. Moreover, these strategies still fall short of human-level intelligence (Valmeekam et al., 2024a,b; Stechly et al., 2024), particularly in complex real-world tasks (Xie et al., 2024b; Gundawar et al., 2024; Chen et al., 2024). Therefore, based on the findings from RQ1, this paper seeks to answer the research questions, ***RQ2: What happens during memory updating for language agents*** and ***RQ3: What hinders these strategies from achieving high-level planning abilities?*** Specifically, we focus on language agents' vanilla planning as well as planning following memory updating, which reflect the internal planning capabilities of language agents rather than the translation ability.

In this paper, we delve into the two main components of planning: constraints and questions, which serve as the foundational elements for planning tasks. Constraints refer to the rules that agents must adhere to when generating a plan, while questions represent the goals that drive the planning process. Understanding how agents handle these elements is crucial for improving their performance in complex planning tasks. Using Permutation Feature Importance (Breiman, 2001; Fisher et al., 2019) to analyze the feature attribution of constraints and questions, our investigation reveals several key findings: ***1)*** Language agents show a limited understanding of constraints, and the influence of the question weakens as the planning horizon increases. ***2)*** Episodic memory updating improves constraint understanding but relies on global understanding, and it's still difficult for agents to reference constraints in a fine-grained manner. ***3)*** Parametric memory updating enhances the question's impact on the final plan, but the diminishing influence of the question remains a challenge. ***4)*** Both strategies resemble "shortcut learning" and struggle with dynamic constraints in planning.

2 Related Work

2.1 Language Agent

The advent of large language models sparks widespread attention due to their remarkable abilities, such as mathematical reasoning, creative writing, and information retrieval (Gómez-Rodríguez and Williams, 2023; Zhang et al., 2023; Lou et al., 2024; Zhu et al., 2024). Building on these models, language agents expand their capabilities to engage with the real world, including utilizing tools (Gu et al., 2024), grounding environments (Zheng et al., 2024a), and even controlling real-world robotics (Zeng et al., 2023b), functioning as a "reasoning brain" beyond mere text generation. The conceptual framework of language agents includes: *1)* **Memory module** handles both long-term memory embedded in the model's parameters, such as commonsense (West et al., 2022), and short-term memory specific to tasks (Majumder et al., 2023). *2)* **Tool-use module** enables agents to utilize external tools to compensate for inherent limitations, such as calling a calculator for arithmetic tasks or retrieving up-to-date information from external databases (Lu et al., 2023; Xie et al., 2024a; Wu et al., 2024). *3)* **Planning module** controls the entire task process, including goal decomposition, action sequencing, and forward estimation, requiring comprehensive and advanced reasoning abilities (Weng, 2023; Sumers et al., 2024).

2.2 Planning in Language Agents

Planning, a hallmark of human intelligence, serves as a critical component in language agent systems, as it directly controls task execution and goal achievement. Improving an agent's planning abilities thus leads to overall improvements across various tasks. However, previous studies show that current agents still struggle with planning tasks, such as classical tasks like block manipulation (Valmeekam et al., 2024a) or real-world tasks like travel planning (Xie et al., 2024b; Zhang et al., 2024). While these studies highlight agents' weaker performance in planning, they mainly pro-