

Planning with Language Models Through The Lens of Efficiency

Michael Katz, Harsha Kokel, Kavitha Srinivas, Shirin Sohrabi
IBM Research

Abstract

We analyse the cost of using LLMs for planning and highlight that recent trends are profoundly uneconomical. We propose a significantly more efficient approach and argue for a responsible use of compute resources; urging research community to investigate LLM-based approaches that upholds efficiency.

Introduction

Recent work has addressed the issue of planning in LLMs, spurred by their abilities in natural language tasks. The approaches vary widely from giving a planning problem to an LLM and asking it to output an entire sequence of actions, to asking an LLM to plan step by step, including backtracking. The purpose of our work is to examine the complexity of existing LLM based approaches applied to planning and search, because from a pragmatic perspective, the cost of LLM evaluations is significant, either in terms of GPU resources, or expensive API calls to hosted LLMs. In all explored cases, a proposed solution to the planning or search problem is neither sound nor complete when LLMs are involved, so the expense is an even bigger factor. Because LLMs are also non-deterministic, a key part of finding a solution involves different rollouts at each step, which further add to the inefficiency. We include this in our analysis of complexity, when it is built into the algorithm.

Having explored the existing approaches, we propose an alternative, an approach that exploits the strengths of LLM, while mitigates the weaknesses of the existing approaches, doing so in an efficient manner. We propose using the language models for deriving the symbolically represented search components that allow for performing the search itself without calling LLMs. That way, the search space correctness can be checked before the search is performed, allowing for soundness and completeness of the search algorithms imply these properties of the overall solution.

Complexity of Planning using LLM

We look at various approaches proposed for planning using LLMs and present their *LM Evaluation Complexity*; quantify the number of LLM requests required by that approach.

IO The Input-Output (IO) prompting is the straightforward use of LLMs to generate an output for a given input.

Here, the model is usually evaluated once, generating an output that may include a sequence of steps, making the complexity of this approach $O(1)$.

CoT The Chain-of-Thought approach (Wei et al. 2022) prompts the model with a predefined example of a chain of steps (thoughts) to resolve the question, in an attempt to make the model generate similar chains in its output. Here, still the model is evaluated once, so the complexity is $O(1)$.

ReAct The ReAct approach (Yao et al. 2023b) aimed at interleaving CoT with acting steps. Each acting step may result in a separate evaluation of the language model. Therefore, the number of evaluations is worst case linear in the number of steps n in the expected output, $O(n)$. This approach is akin to re-planning at each step; or treating LLM as a policy, referenced for each state in the trajectory.

ReWOO ReWOO (Xu et al. 2023) aims at tackling the inefficiency in the number of LLM evaluations, requesting the first model evaluation to plan all future steps, then each step is executed without model evaluation if possible (and with, if not), finalizing by a model evaluation with the concatenated output of previous states as the input. The best case complexity is therefore $O(1)$, with only two evaluations performed. The worst case complexity is however still $O(n)$, since each of the n external tool calls can be to an LLM.

RAP Reasoning via Planning (RAP) approach performs a search (e.g., MCTS) using the LLM for expansion (generate successors) and for heuristic or reward prediction of a state (Hao et al. 2023). Here, the complexity is the worst among the approaches explored so far, being linear in the size of the search space. MCTS is an incomplete algorithm, where the search tree size can be controlled by bounding the number of successors generated per state d , number of rollouts N , and their depth L . The overall complexity is $O(N \times d \times L)$.

ToT Tree of Thoughts (Yao et al. 2023a) approach is similar to that of RAP, where the “thought generator” is expanding a state by calling an LLM and the state evaluator is evaluating the k generated nodes. The evaluation can be done per state (one call per state), or across states (one call across all the current generated nodes). Then, a search is performed on the search tree. The worse case overall complexity is $O(N)$, where N is the number of nodes generated by the search algorithm. The authors use bounded incomplete versions of

the well-known BFS and DFS algorithms, with the depth bound T , branching bound k , and beam size b , restricting the complexity to $O(b \times k \times T)$.

GoT Graph of Thoughts (GoT) (Besta et al. 2024) is similar to tree of thought, except that it can connect the thought nodes in the tree above into a graph for more flexible representations through (a) aggregation of thoughts a , (b) improvement of a thought to refine it further i , (c) generation of a thought based on existing ones g . In GoT, in addition, repeat operation k defines how often a particular operation is repeated. The complexity is similar to that of the ToT, $O(N)$, where N is the number of nodes generated which are governed by k , a , i and g operations.

Policy generation with LLMs Another work takes a different approach. Focusing on problems where a general policy (or a generalized plan) that does not require any search exist, the authors propose querying LLMs to provide a policy implementation in python directly (Silver et al. 2024). The policy is then checked on a few held out problem instances from the domain, providing a feedback on problems that are not solved by the policy. The number of calls to the LLM is therefore $O(1)$ **per domain**, and the result can be amortised among all the problem instances of that domain.

Proposed Approach

We argue for an efficient and responsible use of compute resources, pleading for efficient use of LLMs. Instead of using an LLM at each expansion and evaluation, one could use it to derive a mechanism that allows performing search efficiently. This can be a symbolic model, like STRIPS or PDDL, from which all the search components can be computed, as is done by classical planners that perform heuristic search (Helmert 2006). This can also be directly producing the search components: a successor function, a goal test, and even a heuristic function (Hart et al. 1968).

While the former has been explored in the literature (Guan et al. 2023; Oswald et al. 2024), here we exemplify the latter approach with the 24 game (Yao et al. 2023a). The 24 game takes 4 integers as an input that can be manipulated through the four most common arithmetic operations: addition, subtraction, multiplication, and division. The goal of the game is to produce a formula (also representable by a sequence of operations) that evaluates to 24, if one exists. As mentioned before, Yao et al. (2023a) achieve that with an incomplete algorithm that expands 15 and generates 45 nodes in the worst case ($k = 3$, $b = 5$, and $T = 3$), and therefore calls the LLM 60 times per input. We propose instead to use the LLM to produce the code for two functions: *successor function* and *goal test*. We leave the heuristic function to future work, noting that even without it, the search space of the 24 game is rather small, up to $1 + 36 + 648 + 3888 = 4573$ states.

Listing 1 describes an example correspondence with the DeepSeek-Coder-33B model V20240126.0 (Guo et al. 2024), in a chat mode. We have conducted the experiment for the successor function 5 times, observing that the model required 2 interactions on average. We see that, with

Listing 1: 24 game example

User: The 24 Game is a mathematical card game in which the objective is to find a way to manipulate four integers so that the end result is 24. The game is played with a list of four numbers, and the player must use all four numbers exactly once, using any combination of addition, subtraction, multiplication, or division, to arrive at the number 24. If a state is defined by a list of numbers, how are successor states defined?

Response: In the context of the 24 Game, a successor state is a state that can be reached from the current state by applying one of the four basic arithmetic operations (addition, subtraction, multiplication, division) to any two numbers in the current state.

Here is a Python implementation of the successor states definition:

```
import itertools
def successor_states(state):
    # Generate all possible pairs of numbers in the state
    pairs = list(itertools.combinations(range(len(state)), 2))
    successors = []
    # For each pair, apply each operation to generate a successor state
    for pair in pairs:
        i, j = pair
        a, b = state[i], state[j]
        # Addition
        successor=state[:i] + state[i+1:j] + state[j+1:] + [a+b]
        successors.append(successor)
        # Subtraction (both orders)
        successor=state[:i] + state[i+1:j] + state[j+1:] + [a-b]
        successors.append(successor)
        successor=state[:i] + state[i+1:j] + state[j+1:] + [b-a]
        successors.append(successor)
        # Multiplication
        successor=state[:i] + state[i+1:j] + state[j+1:] + [a*b]
        successors.append(successor)
        # Division (both orders)
        if b != 0:
            successor=state[:i]+state[i+1:j] + state[j+1:] + [a/b]
            successors.append(successor)
        if a != 0:
            successor=state[:i]+state[i+1:j] + state[j+1:] + [b/a]
            successors.append(successor)
    return successors
```

User: How do you test if a state is a goal state? Please implement in python

Response: In the context of the 24 Game, a goal state is a state where all four numbers have been used exactly once and the result of the arithmetic operations is 24. To check if a state is a goal state, we can simply check if the result of the arithmetic operations is 24.

Here's a Python function that checks if a state is a goal state:

```
def is_goal_state(state):
    # Check if the result of the arithmetic operations is 24
    return abs(state[0] - 24) < 0.001
```

User: If the number of elements in a state decrease with each successor state, how many elements are in the goal state?

Response: If the number of elements in a state decreases with each successor state, the goal state will have a single element, which is the result of the arithmetic operations. This means that the goal state will have one number, and that number will be 24.

Here's a Python function that checks if a state is a goal state:

```
def is_goal_state(state):
    # Check if the state is a list with a single element that is 24
    return len(state) == 1 and abs(state[0] - 24) < 0.001
```

a minor help from a user, the model is able to produce correct successor and goal test functions. We further tested all 5 successor functions and goal test within a simple BFS implementation on a suite of 1362 instances from ToT. The total time to solve all 1362 problems varies over these 5 cases from 3.6s to 6.9s in our naive implementation, hinting that some successor functions can be more efficient than other. This is comparable to a **single** LLM evaluation time; which is $\approx 7s$ for GPT-4 Chat (OpenAI Dev. Forum 2024). For each of the 5 tested successor functions, BFS found a solution in the 1361 cases where a solution exists and report that no solution exists in the only one unsolvable case, while ToT reports success rate of 45% and 75% (breadth 1 and 5, respectively) on a subset of 100 medium difficulty problems. The significant differences in the use of computational resource and the performance measures clearly demonstrate that LLMs can be effectively employed for planning purposes without compromising on the efficiency or accuracy. We urge the research community to focus their efforts on such approaches.

References

- Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nycz, P.; and Hoefler, T. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *AAAI*, 17682–17690. AAAI Press.
- Guan, L.; Valmeekam, K.; Sreedharan, S.; and Kambhampati, S. 2023. Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS 2023)*.
- Guo, D.; Zhu, Q.; Yang, D.; Xie, Z.; Dong, K.; Zhang, W.; Chen, G.; Bi, X.; Wu, Y.; Li, Y. K.; Luo, F.; Xiong, Y.; and Liang, W. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence. arXiv:2401.14196 [cs.SE].
- Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; and Hu, Z. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*.
- Hart, P. E.; et al. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2): 100–107.
- Helmert, M. 2006. The Fast Downward Planning System. 26: 191–246.
- OpenAI Dev. Forum. 2024. Performance analysis of Assistants versus Chat completion. <https://community.openai.com/t/performance-analysis-of-assistants-versus-chat-completion-chat-completion-seems-somewhat-faster-for-complete-message-generation-streaming-taken-into-account/628368>.
- Oswald, J.; Srinivas, K.; Kokel, H.; Lee, J.; Katz, M.; and Sohrabi, S. 2024. Large Language Models as Planning Domain Generators. In Bernardini, S.; and Muise, C., eds., *Proceedings of the Thirty-Fourth International Conference on Automated Planning and Scheduling (ICAPS 2024)*. AAAI Press.
- Silver, T.; Dan, S.; Srinivas, K.; Tenenbaum, J.; Pack Kaelbling, L.; and Katz, M. 2024. Generalized Planning in PDDL Domains with Pretrained Large Language Models. In Dy, J.; and Natarajan, S., eds., *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2024)*. AAAI Press.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS 2022)*, 24824–24837.
- Xu, B.; Peng, Z.; Lei, B.; Mukherjee, S.; Liu, Y.; and Xu, D. 2023. Rewoo: Decoupling reasoning from observations for efficient augmented language models. arXiv:2305.18323 [cs.CL].
- Yao, S.; Yu, D.; Zhao, J.; Shafraan, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS 2023)*.
- Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafraan, I.; Narasimhan, K.; and Cao, Y. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net.