Towards a More Rigorous PDDL Generation Benchmark, *or*, `(:goal (benchmark pddl_generators))`

Max Zuo*, Francisco Piedrahita Velez*, Xiaochen Li, Michael L. Littman, Stephen H. Bach

[1] Department of Computer Science, Brown University
zuo@brown.edu, fpiedrah@brown.edu, xiaochen_li@brown.edu,
mlittman@cs.brown.edu, stephen_bach@brown.edu

Abstract

Planetarium is a benchmark that evaluates language models' ability to generate PDDL code from natural language descriptions of planning tasks. It introduces a novel PDDL equivalence algorithm that assesses the correctness of generated PDDL against ground truth. Additionally, *Planetarium* includes a dataset of 145,918 text-to-PDDL pairs spanning 73 unique types of initial state and goal condition combinations, offering varying levels of difficulty for evaluation. *Planetarium* aims to highlight that foundation models (FMs) don't need to perform every task. Hybrid approaches apply FMs alongside traditional AI methods and offer an alternative with compelling advantages.[1]

There has been growing interest in using large language models (LLMs) to solve planning problems (Valmeekam et al. 2023; Silver et al. 2022; Bohnet et al. 2024). However, while promising, this approach has demonstrated limited success; o1-Preview only achieves 23.65% accuracy on common planning problems (Gundawar et al. 2024). A key limitation of relying solely on LLMs for planning is their inability to guarantee correctness or optimality, rendering them unreliable for critical applications.

Despite these challenges, integrating LLMs into the planning pipeline remains an appealing prospect. Traditional planning systems often require extensive domain knowledge and expertise in modeling planning problems, which can act as significant barriers to adoption. By leveraging LLMs, it may be possible to alleviate these challenges, broadening access to planning technologies.

One promising research direction involves using LLMs to convert natural language prompts into structured planning languages, such as the Planning Domain Definition Language (PDDL) (Liu et al. 2023). Traditional symbolic planners can then process these structured representations (Fikes and Nilsson 1971; Helmert 2006), which are highly efficient and capable of producing correct solutions. Problem descriptions also allow users to leverage other planning tools like online replanners (Garrett et al. 2020). Early evidence suggests that LLMs are more effective in generating correct plan descriptions than in producing correct plans (Liu et al. 2023; Xie et al. 2023), highlighting their potential utility as tools for bridging the gap between human instructions and formal planning representations.

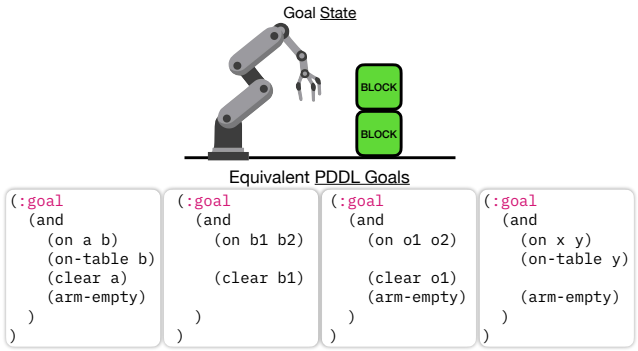

Figure 1: An example of how one planning goal can correspond to multiple PDDL goal definitions. All four PDDL definitions map to the same goal state.

Despite this approach's promise, the field lacks established techniques and benchmarks for evaluating the translation of natural language planning descriptions into PDDL. Existing work on generating PDDL with LLMs typically considers the generated PDDL correct if there exists a plan that can solve it (Liu et al. 2023). However, this criterion is insufficient. An LLM might produce a valid PDDL that is unrelated to the user's instructions yet still deemed correct under this definition, leading to false positives.

Rigorous evaluation of LLMs as PDDL generators requires a precise definition of correctness. We argue that PDDL generated from a textual description can only be considered correct if it faithfully represents the same underlying planning problem as the ground truth PDDL. This is a non-trivial task, as different PDDL instances can represent the same planning problem without being identical (Figure 1).

To address these shortcomings, we introduce *Planetarium*, a benchmark designed to assess the ability of LLMs to translate natural language descriptions of planning problems

*These authors contributed equally.

[1]Code to our benchmark and dataset is available at: https://github.com/BatsResearch/planetarium.

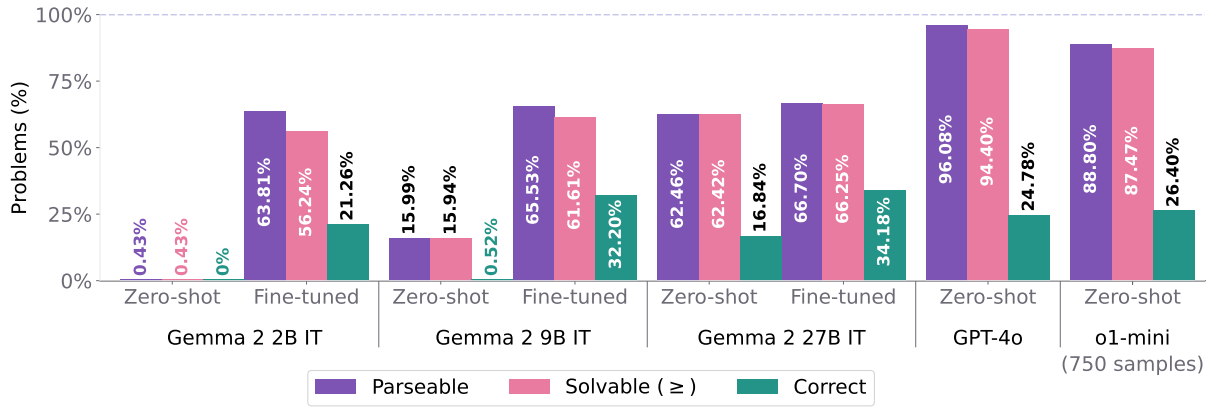

Figure 2: Performance of various models on the *Planetarium* test set. If a planner found a plan that solves a problem in the time allotted, it was marked solvable in the figure above. However, if a planner does not return a plan, a problem may still be "solvable."

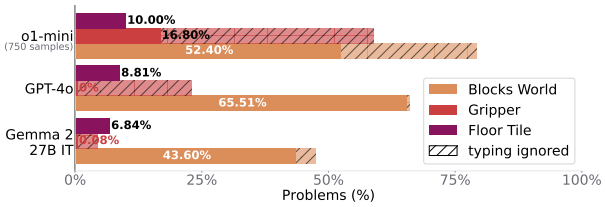

Figure 3: Breakdown of zero-shot performance by domain for Gemma 2 27B IT, GPT-4o, and o1-mini.

into PDDL.

A Rigorous Benchmark

Planetarium formally defines planning problem equivalence as two problem descriptions encoding isomorphic initial states and non-trivial goal conditions, and presents an algorithm to verify whether two STRIPS PDDL problems meet this definition (Zuo et al. 2024). This algorithm transforms PDDL code into scene graphs, computes an expansion of the goal states for both PDDL problems, and then checks if the two problems are graph isomorphic. Our method ensures two PDDL problems match if and only if they represent the same underlying planning task.

In this context, we define a generated PDDL as *parsable* if it adheres to correct PDDL syntax, *solvable* if there exists a plan that can solve it, and *correct* if it passes our equivalence algorithm. Of these three metrics, only our algorithm captures the semantics of the planning problem.

The Dataset

Planetarium includes a dataset of diverse planning tasks for evaluating LLMs on PDDL generation. Each instance consists of a textual prompt and a ground truth PDDL representation of the task. The tasks vary along two dimensions—abstractness and size—allowing for assessing the difficulty of PDDL generation. These tasks are built around three domains: Blocks World, Gripper, and Floor Tile. These domains are challenging for LLMs and represent the types of domains used in other studies evaluating LLMs on planning tasks (Valmeekam et al. 2023; Liu et al. 2023).

Experiments

We evaluate two API-access models, GPT-4o and o1-mini, and three open-weight models, Gemma 2 2B IT, Gemma 2 9B IT, and Gemma 2 27B IT (Gemma Team 2024), against the *Planetarium* test set. The percentage of parseable, solvable, and correct generated plans, averaged across all domains in both zero-shot and fine-tuned settings, are shown in Figure 2.

Our work reveals that language models struggle to generate semantically correct structured planning language descriptions. Models like GPT-4o can often produce valid, seemingly correct descriptions (94.4%) when, in reality, only a small fraction (24.8%) are semantically correct.

Interestingly, models consistently applied typing in their generated PDDL even though none of our domains required `:typing`, which we pass in context (Figure 3). We see this in the increase in the accuracy of several models when relaxing our algorithm to ignore object types during evaluation. This suggests that these models heavily rely on their semantic priors and pretraining to solve such problems and may often ignore pertinent information in their context.

Conclusion

We hope that *Planetarium* will drive progress on hybrid approaches combining LLMs and classic planners, setting a standard for evaluating such tasks.

Acknowledgements

This research is supported in part by the Office of Naval Research (ONR) award N00014-20-1-2115. We acknowledge support from Cisco, Cognex, and the Brown Computer Science Faculty Innovators Fund. Disclosure: Stephen Bach is an advisor to Snorkel AI, a company that provides software and services for data-centric artificial intelligence.

References

Bohnet, B.; Nova, A.; Parisi, A. T.; Swersky, K.; Goshvadi, K.; Dai, H.; Schuurmans, D.; Fiedel, N.; and Sedghi, H. 2024. Exploring and Benchmarking the Planning Capabilities of Large Language Models. *arXiv preprint arXiv:2406.13094*.

Fikes, R. E.; and Nilsson, N. J. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3): 189–208.

Garrett, C. R.; Paxton, C.; Lozano-Pérez, T.; Kaelbling, L. P.; and Fox, D. 2020. Online replanning in belief space for partially observable task and motion problems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 5678–5684. IEEE.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Gundawar, A.; Valmeekam, K.; Verma, M.; and Kambhampati, S. 2024. Robust Planning with Compound LLM Architectures: An LLM-Modulo Approach. arXiv:2411.14484.

Helmert, M. 2006. The fast downward planning system. *J. Artif. Int. Res.*, 26(1): 191–246.

Liu, B.; Jiang, Y.; Zhang, X.; Liu, Q.; Zhang, S.; Biswas, J.; and Stone, P. 2023. LLM + P: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Silver, T.; Hariprasad, V.; Shuttleworth, R. S.; Kumar, N.; Lozano-Pérez, T.; and Kaelbling, L. P. 2022. PDDL Planning with Pretrained Large Language Models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2023. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xie, Y.; Yu, C.; Zhu, T.; Bai, J.; Gong, Z.; and Soh, H. 2023. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*.

Zuo, M.; Velez, F. P.; Li, X.; Littman, M. L.; and Bach, S. H. 2024. Planetarium: A Rigorous Benchmark for Translating Text to Structured Planning Languages. arXiv:2407.03321.