On the Limit of Language Models as Planning Formalizers

- Cassie Huang & Li Zhang, Drexel University (2024)
- Bhavya Sri Sangireddy

Motivation

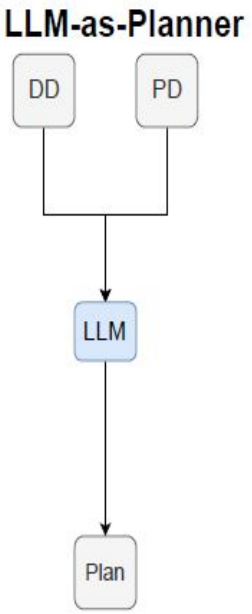
- LLMs often generate plans that are not executable or verifiable in real environments.
- LLM-as-Formalizer: convert natural language to PDDL for deterministic planning.
- LLM-as-Planner: directly generates plan steps and they are less reliable.

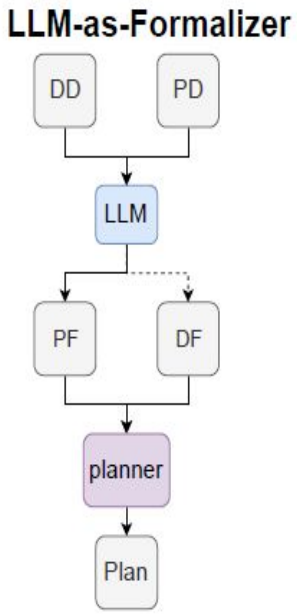
Two Approaches

LLM-as-Planner: LLM outputs sequence of steps and they are often incorrect.

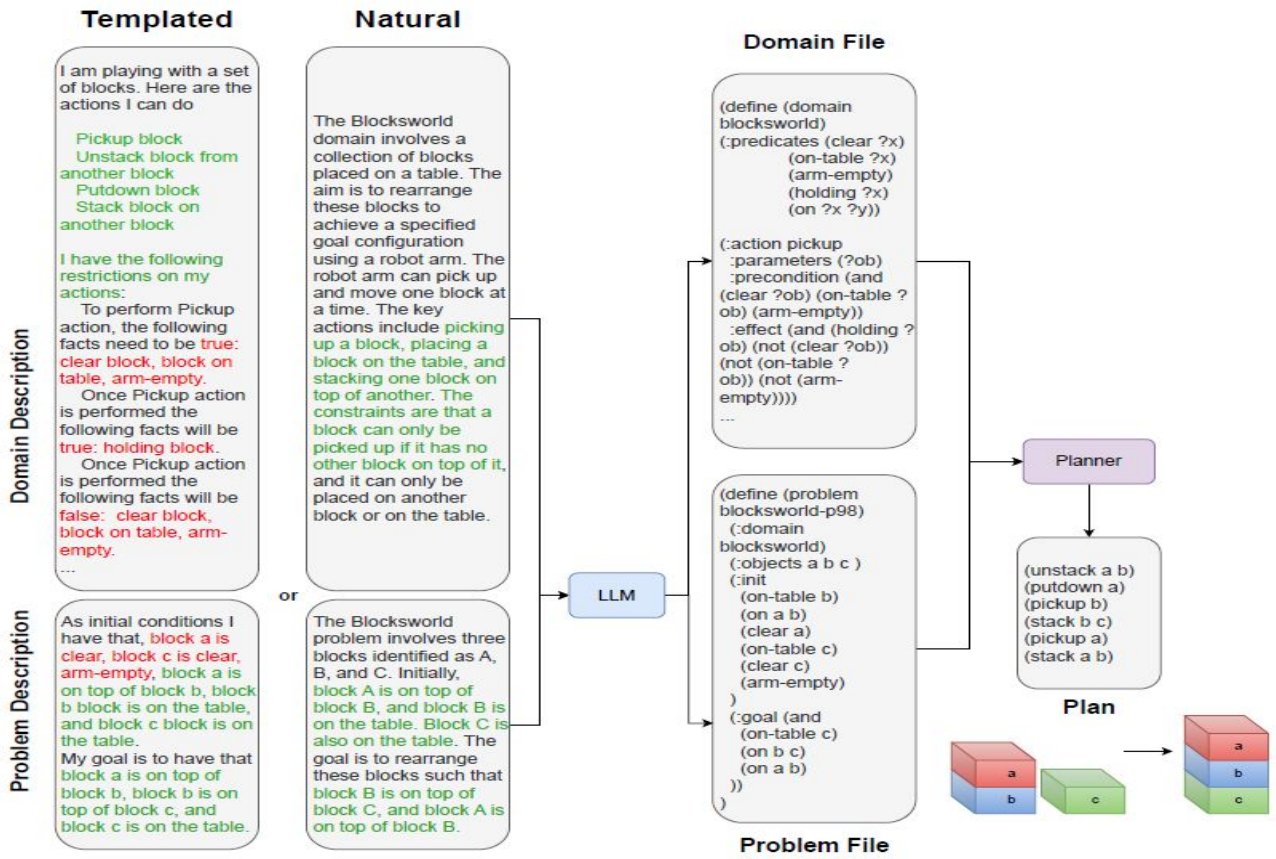
LLM-as-Formalizer: LLM translates text to PDDL and then solver computes correct plan.

Goal: Evaluate if LLMs can generate complete PDDL from natural descriptions.





BLOCKS WORLD EXAMPLE



Methodology

- Datasets: BlocksWorld, Logistics, Barman, Mystery BlocksWorld.
- Two Metrics:

1. Solvability: Applies only to the LLM-as-formalizer approach. Measures the percentage of predicted PDDL domain + problem files that can be parsed and solved by the planner). Even if the plan isn't correct, if the solver can find any plan, it counts as solvable.
2. Correctness: Applies to both LLM-as-formalizer and LLM-as-planner. Measures the percentage of valid plans that successfully achieve the goal state according to the ground-truth PDDL

Example of Formalization

- Natural description: “The robot arm can pick up and move one block at a time.”
- Formalized PDDL:
- (:action pickup : parameters (?b - block) : precondition (and (clear ?b)(on-table ?b)(arm-empty)) : effect (and (holding ?b)(not (on-table ?b))(not (arm-empty))))
- Solver/Planner generates valid executable plan.

Three levels of naturalness for Descriptions (PD and DD)

Heavily Templated

To perform Pickup action, the following facts need to be true: clear block, block on table, arm-empty.

Once Pickup action is performed the following facts will be true: holding block.

Once Pickup action is performed the following facts will be false: clear block, block on table, arm-empty.

Moderately Templated

I can only pick up or unstack one block at a time.

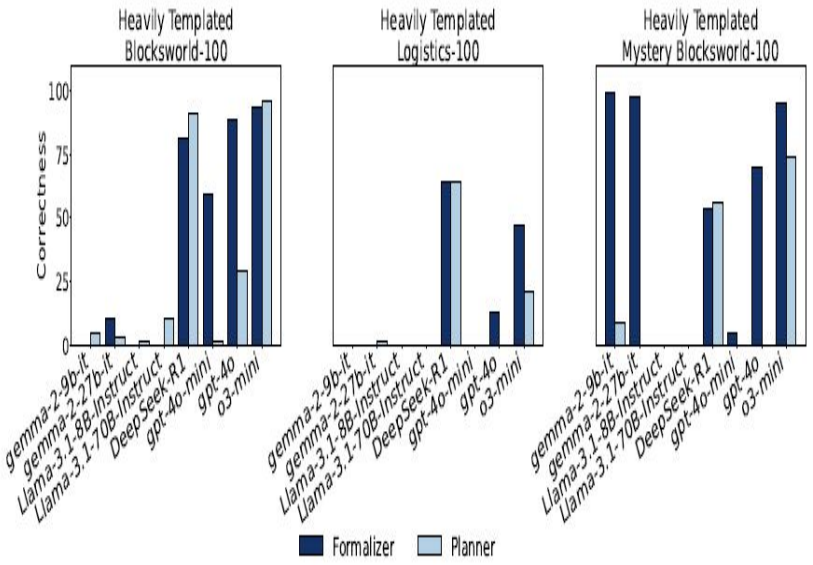
I can only pick up or unstack a block if my hand is empty.

I can only pick up a block if the block is clear. A block is clear if the block has no other blocks on top of it and if the block is not picked up.

Natural

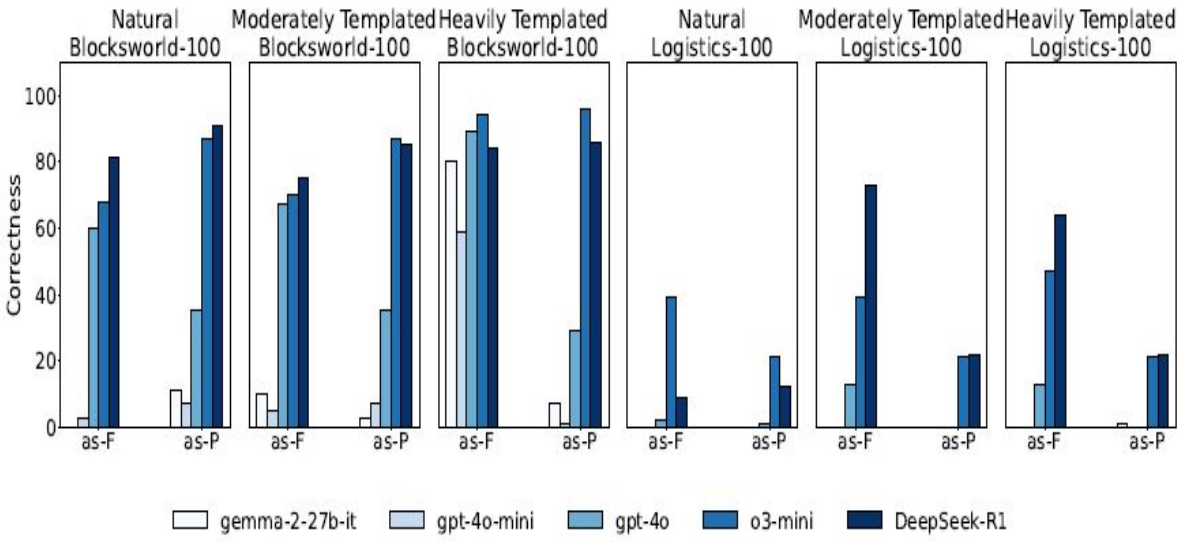
The robot arm can pick up and move one block at a time from one position to another. It is only able to move the top block from any stack or table, and have only one block held by the robot arm at a time. The main actions available are 'pick up', ...

Results with respect to LLM's



- DeepSeek, GPT-4o-mini, GPT-4o, o3-mini models performed better than Llama and Gemma

Performance of LLM as Planner and Formalizer across different naturalness levels.



Key Findings

- LLM-as-formalizer > LLM-as-planner for most models and datasets.
- GPT-4o, O3-mini, DeepSeek-R1 produce the most accurate PDDL.
- Open-source models (Llama, Gemma) struggle with syntax .
- Performance drops as descriptions become more natural. Heavily templated are easy to parse and natural ones may leave out common sense.

Strengths of the paper

- Performs better than the LLM-as-Planner. LLM as a formalizer deterministically finds a plan that satisfies all the given preconditions, goals and effects. And the plan is then validated by VAL.
- As the LLM as formalizer doesn't skip some steps and never make impossible actions it gives the executable plan and the LLM as the formalizer has high correctness across the datasets.

Strengths

- LLM as formalizer is robust to long-tail lexical distribution meaning model still performs well even when the words are rephrased or uncommon. Model focuses on underlying meaning and structure rather than memorizing exact words.
- Most errors are syntax and semantic errors in generated PDDL (either in PF or DF). It makes debugging easier. If PDDL is missing precondition, we can debug the Domain file.

Weaknesses

- Uses small toy domains and can't represent real world scenarios. These are simplified ones and they are fully observable and nothing is uncertain. Real world scenarios are complex and there also exists uncertainty. PDDL may not fully model the uncertain conditions.
- Performance drops sharply with natural inputs.
- Slower due to solver dependency.

Takeaways

- Formalization improves reliability and executability compared to direct planning.
- LLMs can translate natural text into structured logic but struggle with ambiguity.

Q&A