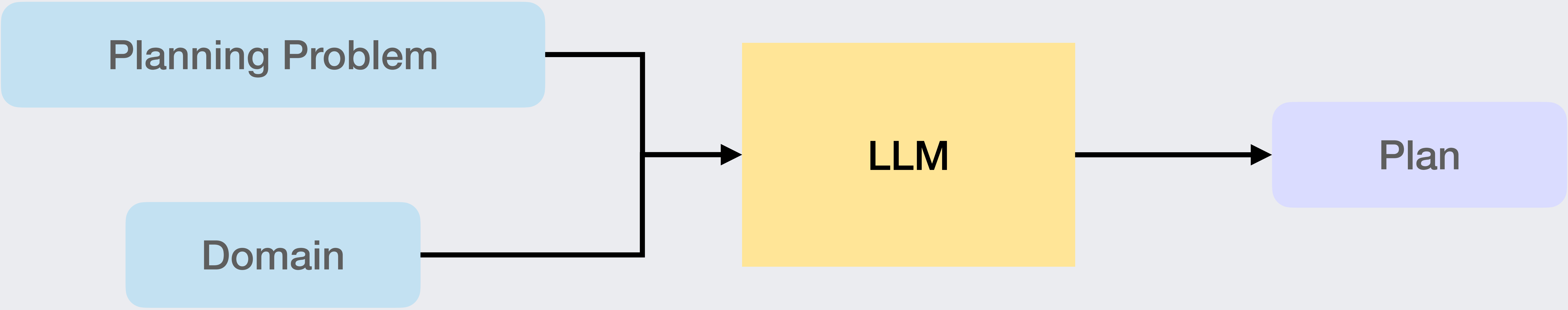
Towards a More Rigorous PDDL Generation Benchmark,
or,
`(:goal (benchmark pddl_generators))`

Max Zuo*, Francisco Piedrahita-Velez*, Xiaochen Li, Michael Littman, Stephen Bach

BROWN

LLMs for Planning

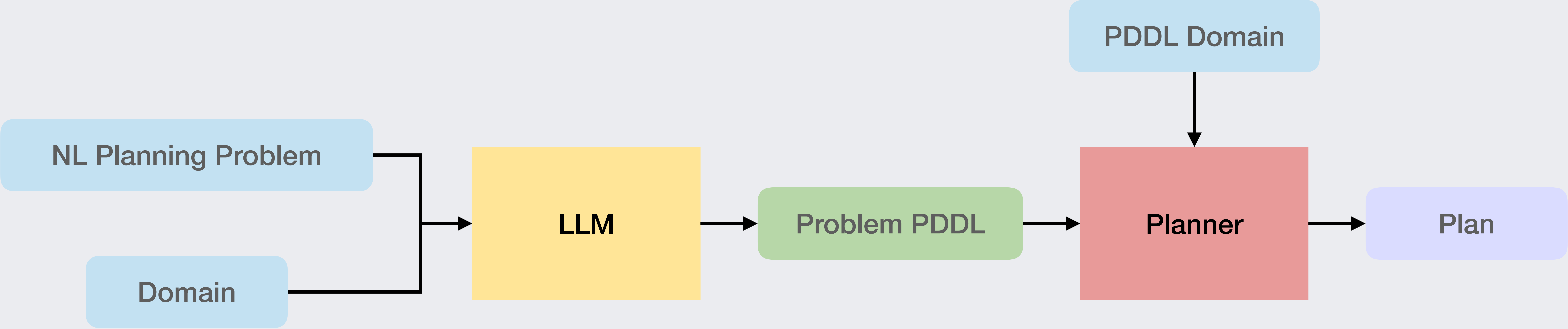
LLMs as Planners



Valmeekam, Karthik, et al. "On the planning abilities of large language models-a critical investigation." Advances in Neural Information Processing Systems 36 (2023): 75993-76005.

Bohnet, Bernd, et al. "Exploring and Benchmarking the Planning Capabilities of Large Language Models." arXiv preprint arXiv:2406.13094 (2024).

LLMs with Planners

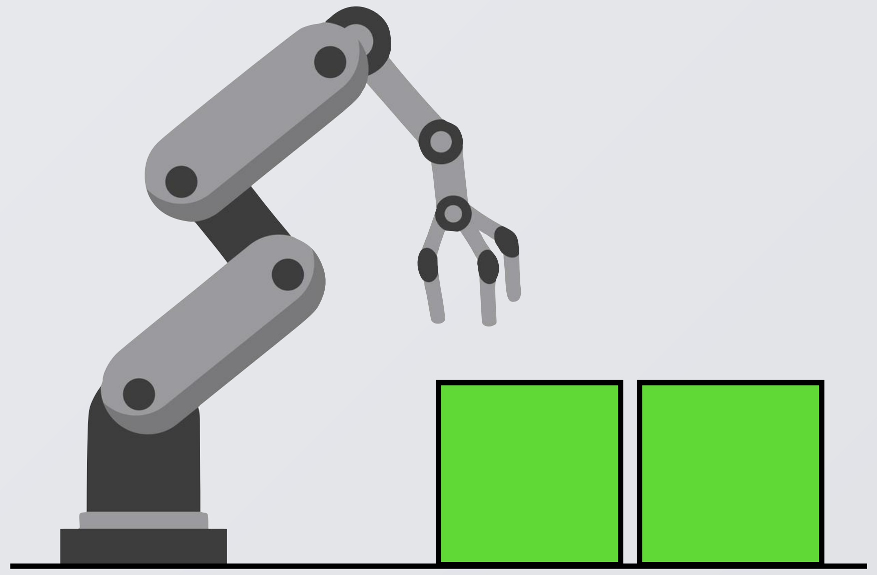


Liu, Bo, et al. "Llm+ p: Empowering large language models with optimal planning proficiency." *arXiv preprint arXiv:2304.11477* (2023).

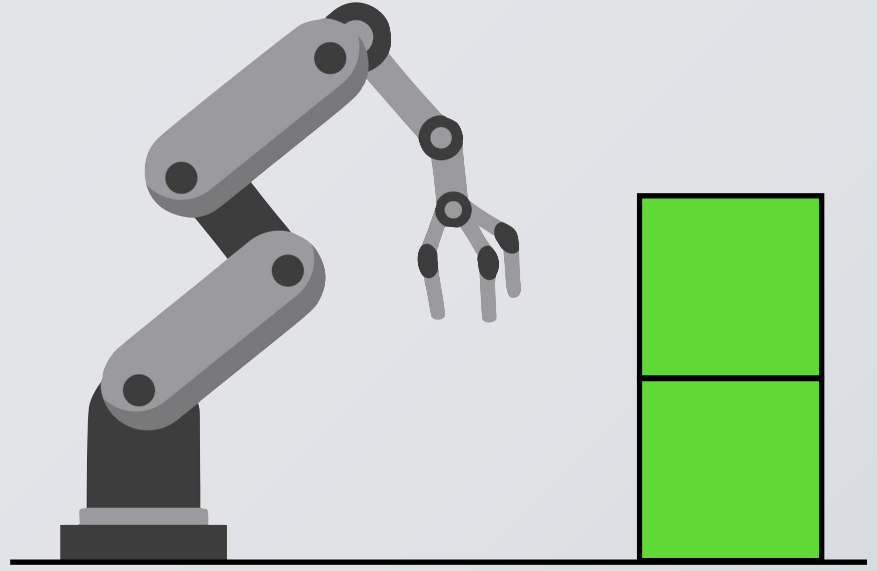
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Prompt

We have two blocks.

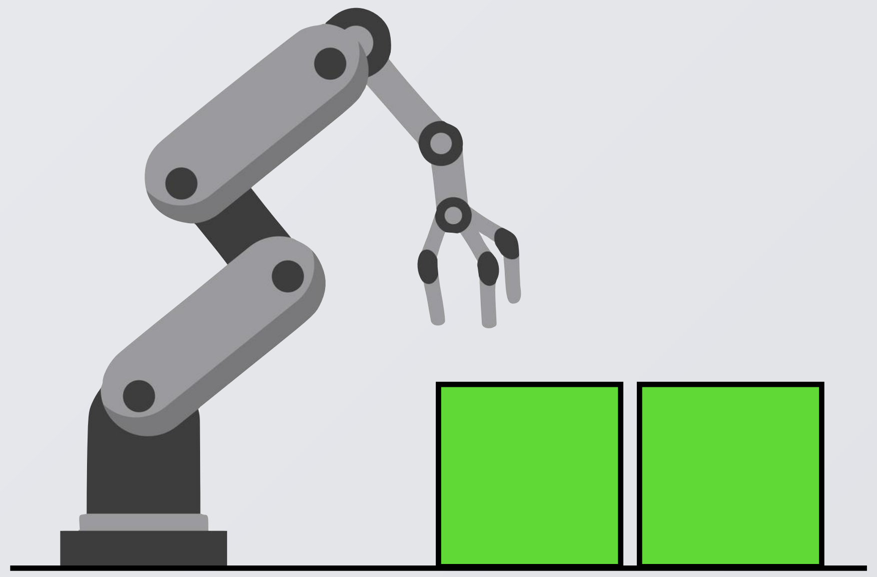
They both start on the table.

The goal is to have them be stacked.

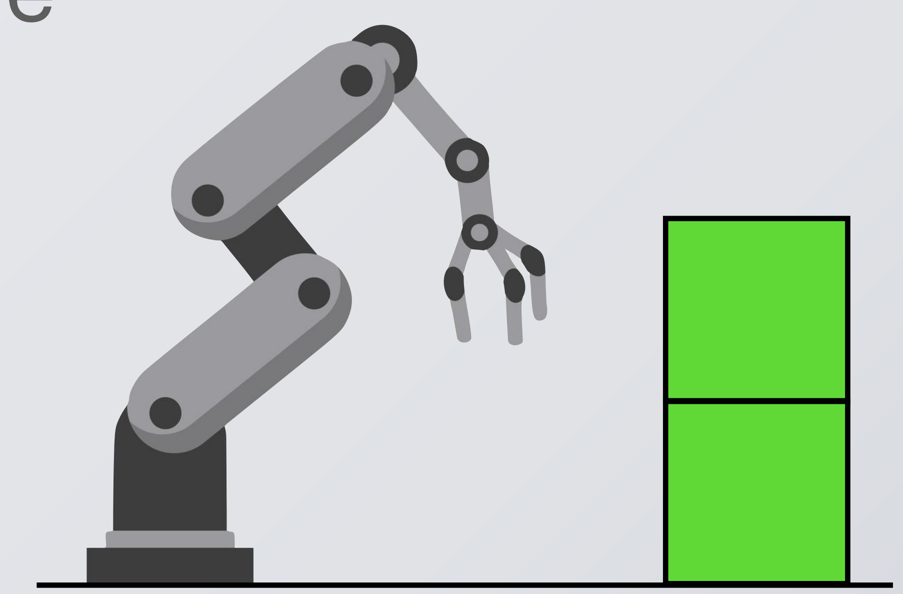
Evaluating PDDL

Problem

Initial Scene



Goal Scene



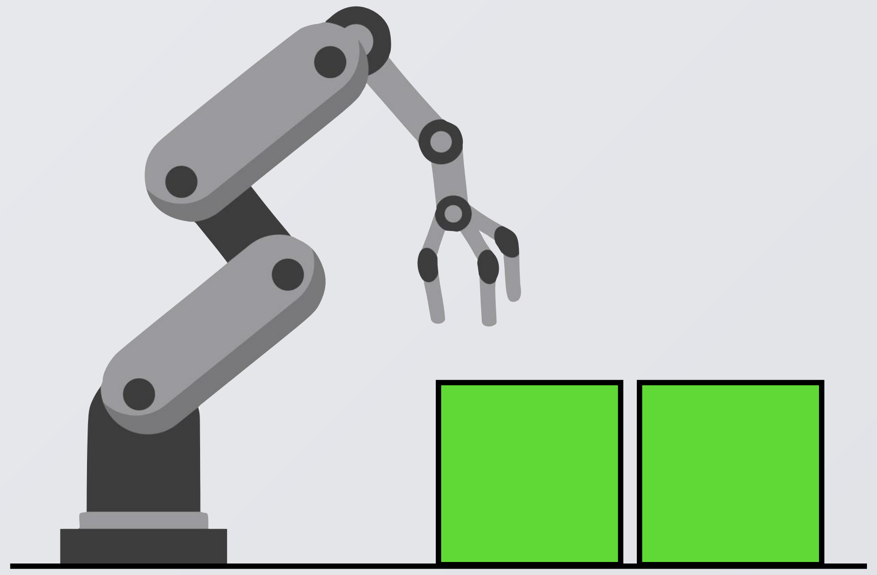
Goal PDDL

```
(:goal
  (and
    (arm-empty)
    (clear A)
    (on A B)
    (on-table B)
  )
)
```

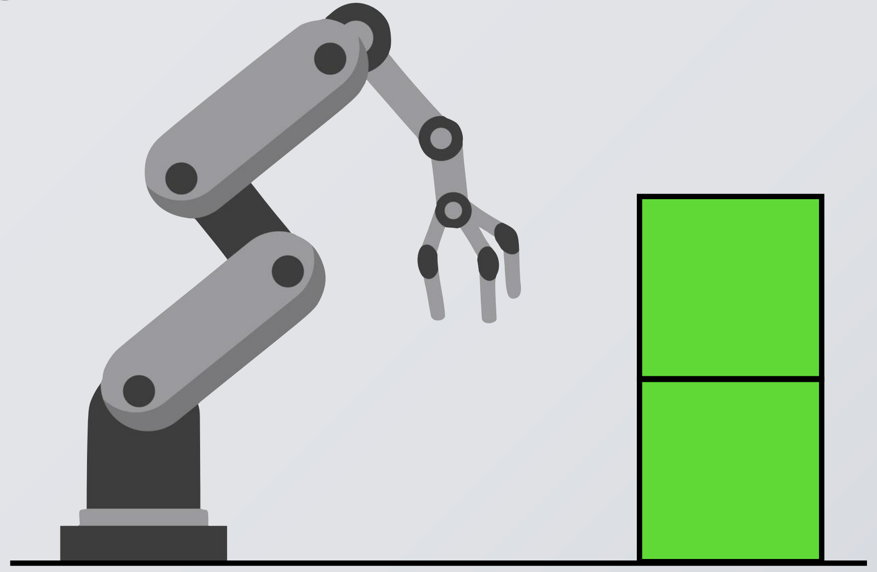
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

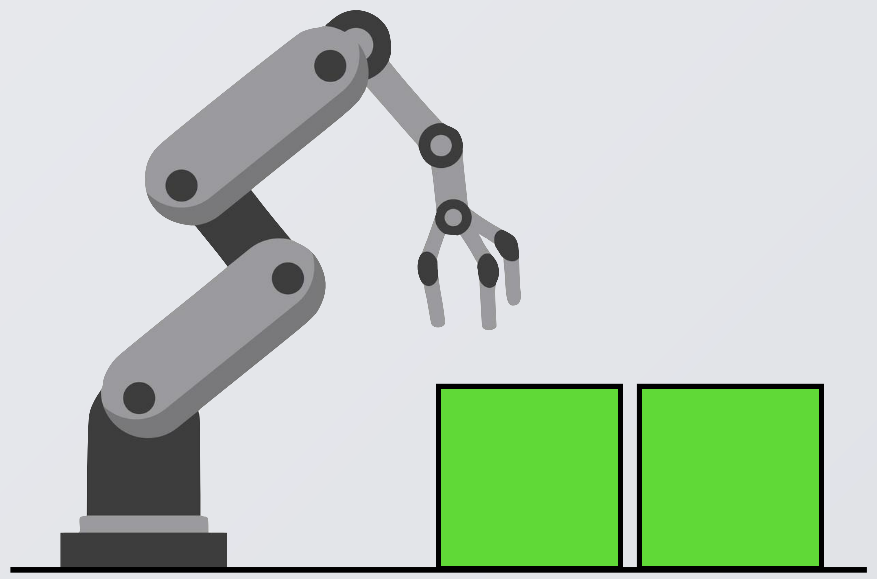
```
(:goal
  (and
    (arm-empty)
    (clear A)
    (on A B)
    (on-table B)
  )
)))
```

Not parsable – Incorrect Syntax

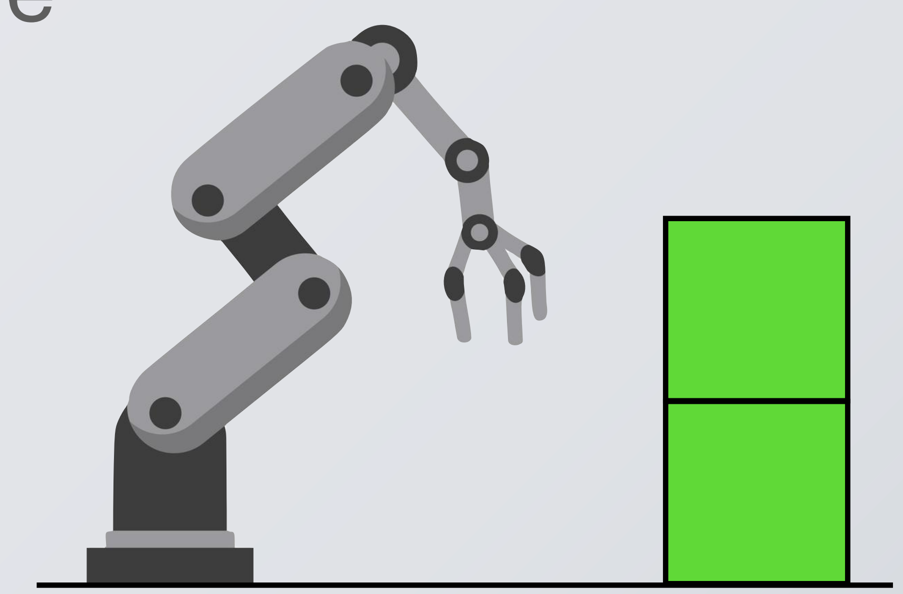
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

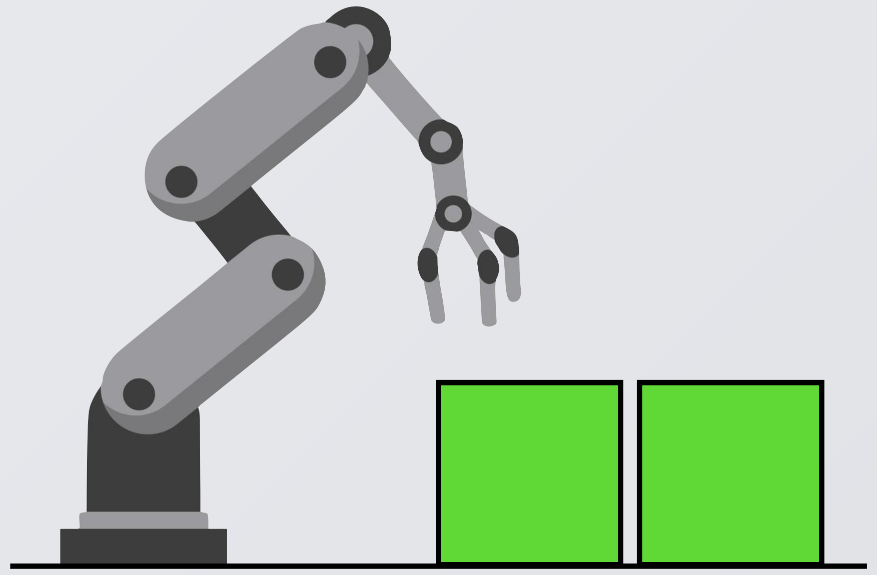
```
(:goal
  (and
    (arm-empty)
    (on B A)
    (on A B)
    (on-table B)
  )
)
```

Parsable – Not Solvable

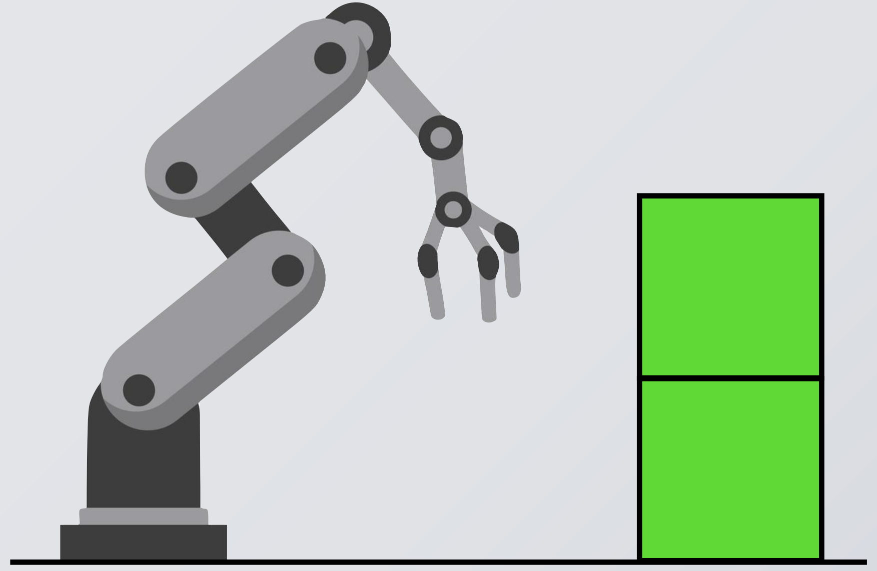
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

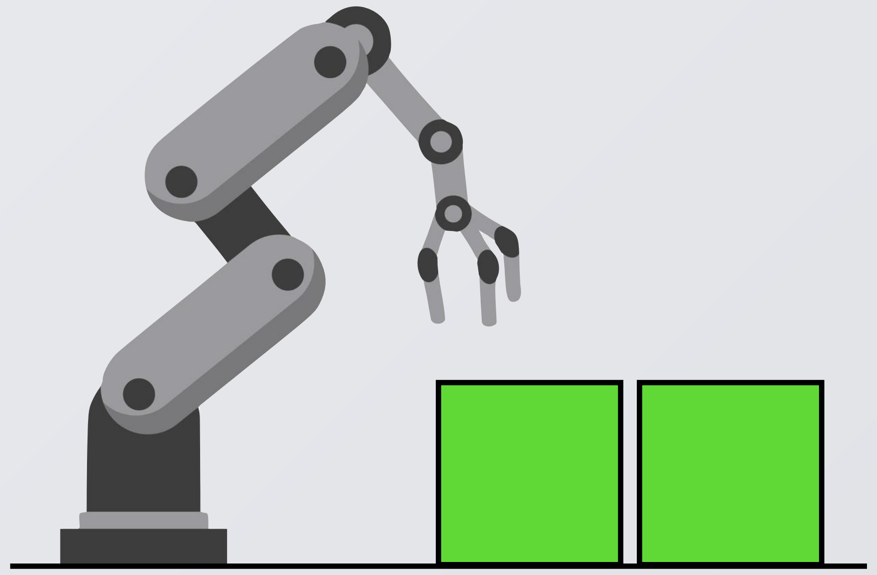
```
(:goal
  (and
    (arm-empty)
    (clear A)
    (on-table A)
    (on-table B)
  )
)
```

Solvable – Incorrect

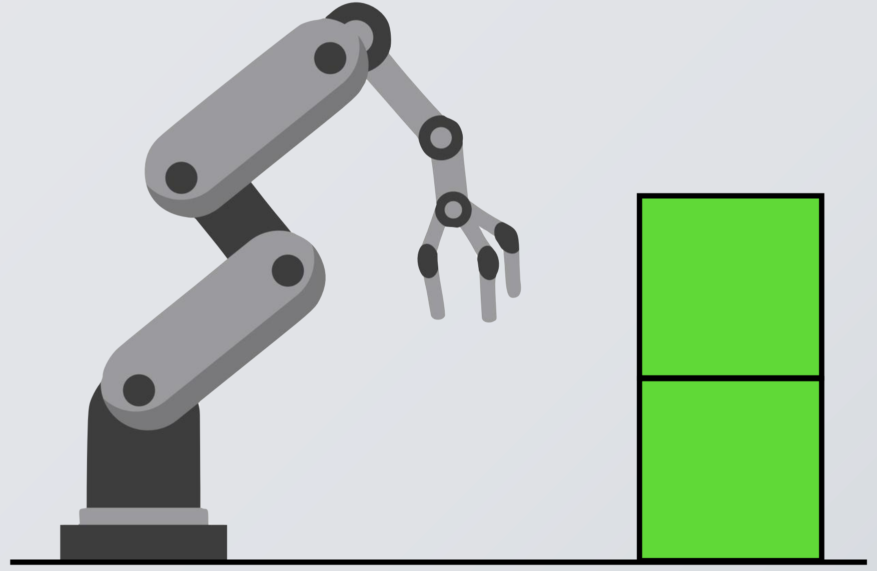
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

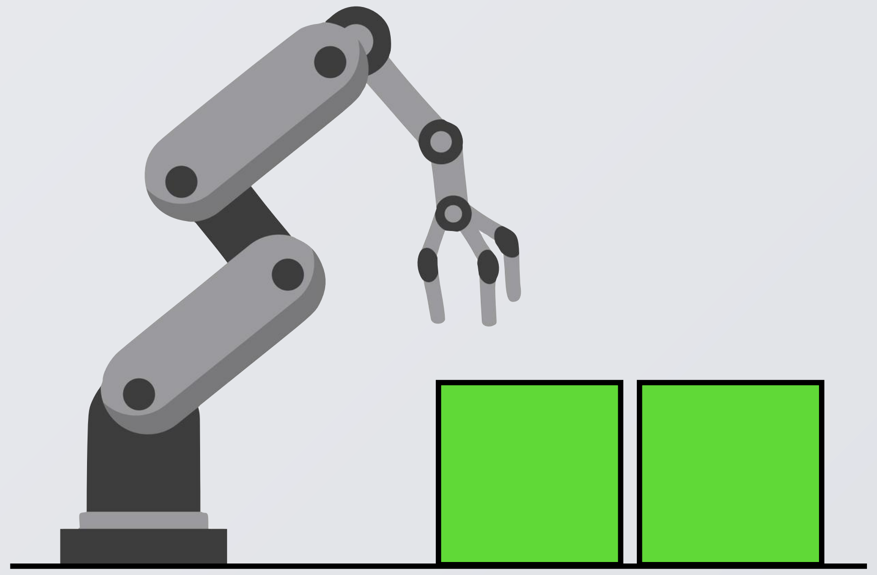
```
(:goal
  (and
    (arm-empty)
    (clear A)
    (on A B)
    (on-table B)
  )
)
```

Correct

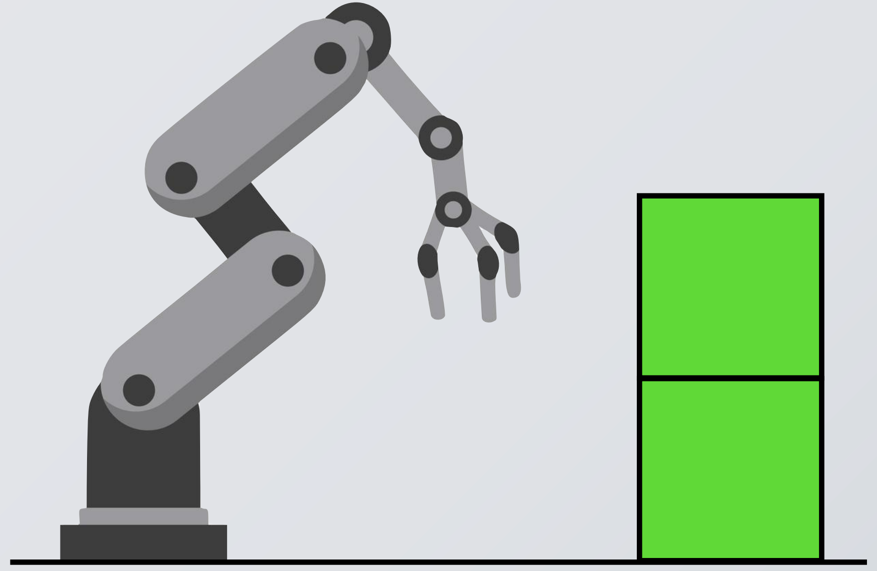
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

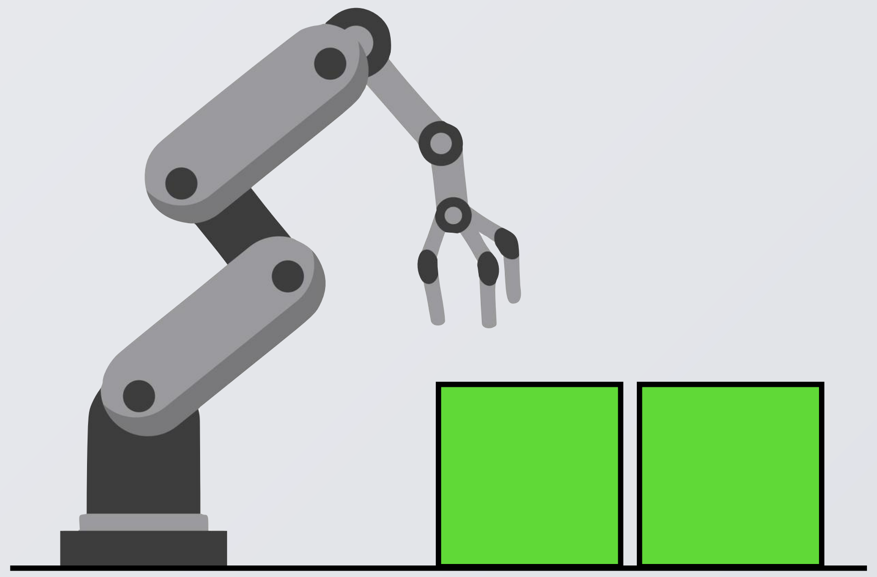
```
(:goal
  (and
    (on-table B)
    (on A B)
    (arm-empty)
    (clear A)
  )
)
```

Correct

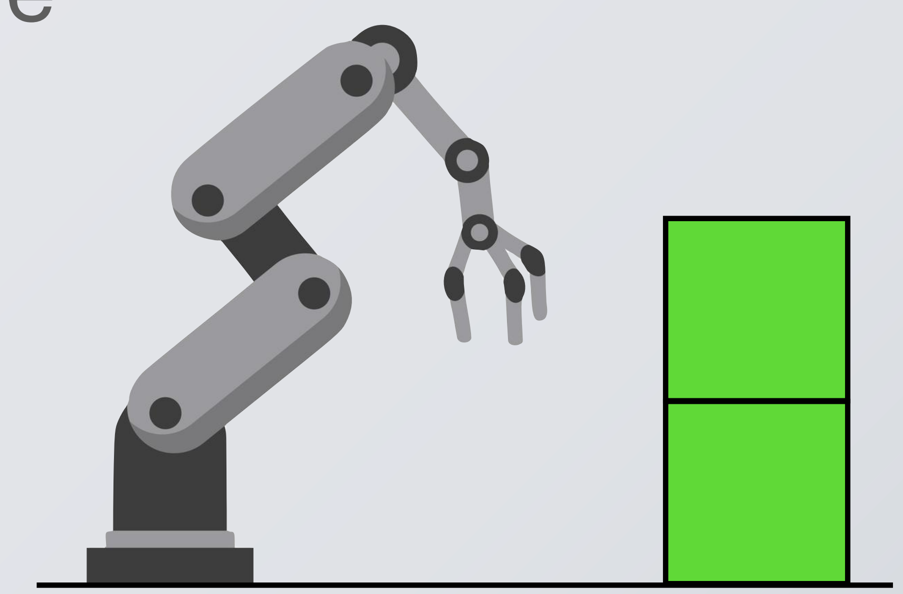
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

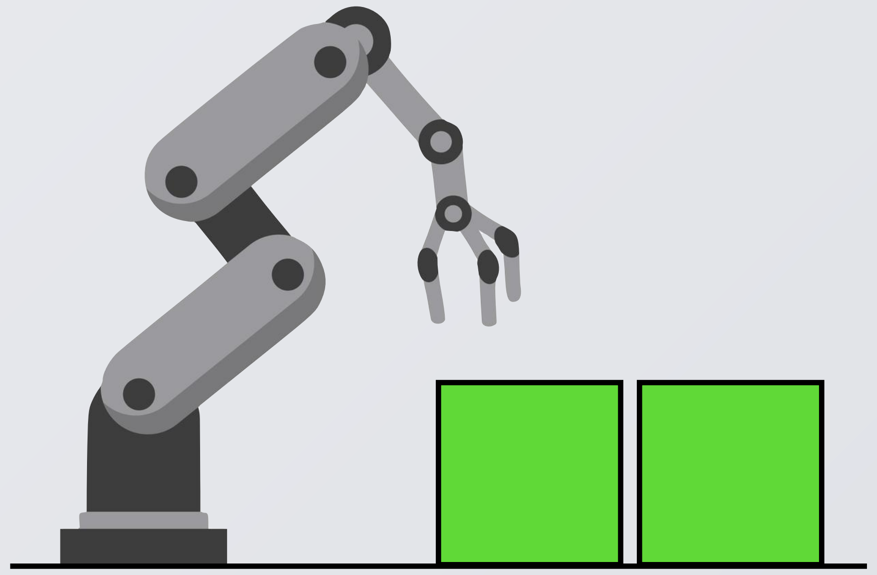
```
(:goal
  (and
    (arm-empty)
    (clear o1)
    (on o1 o2)
    (on-table o2)
  )
)
```

Correct

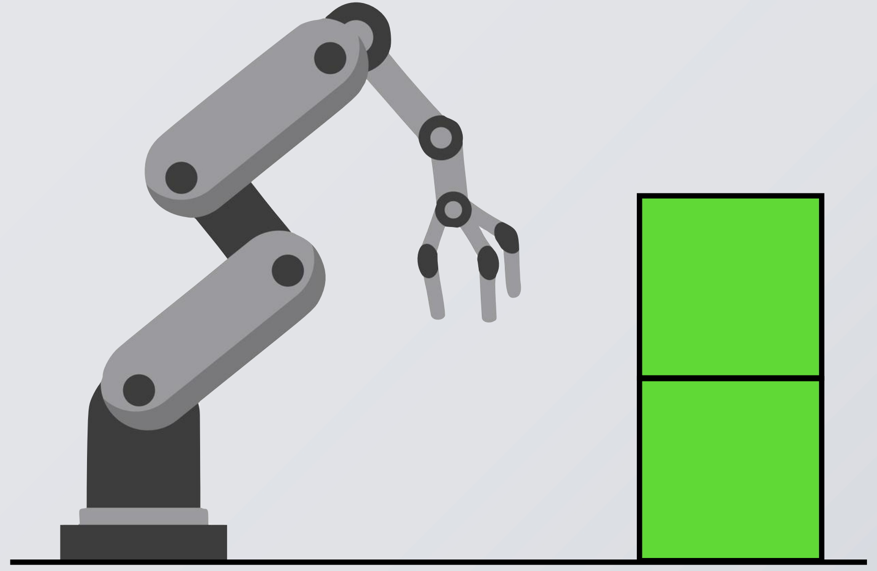
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

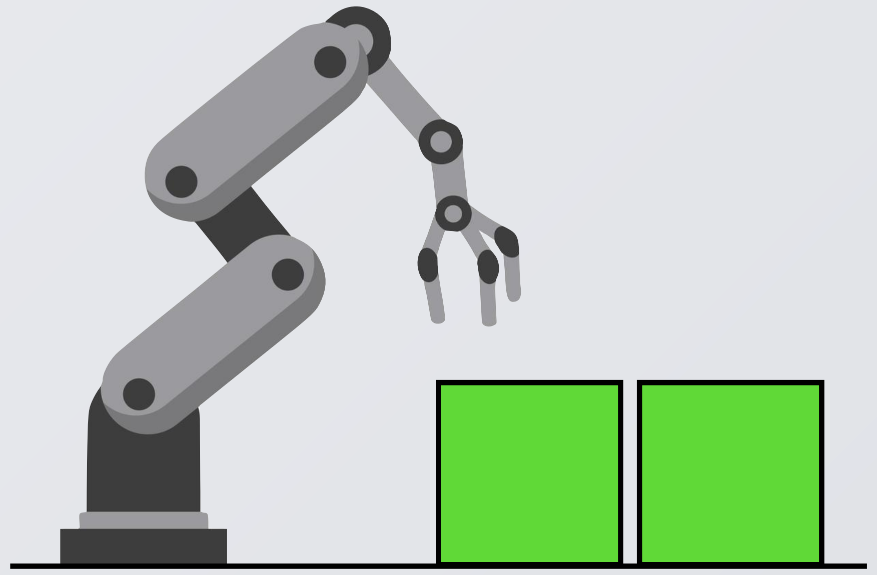
```
(:goal
  (and
    (arm-empty)

    (on A B)
    (on-table B)
  )
)
```

Correct

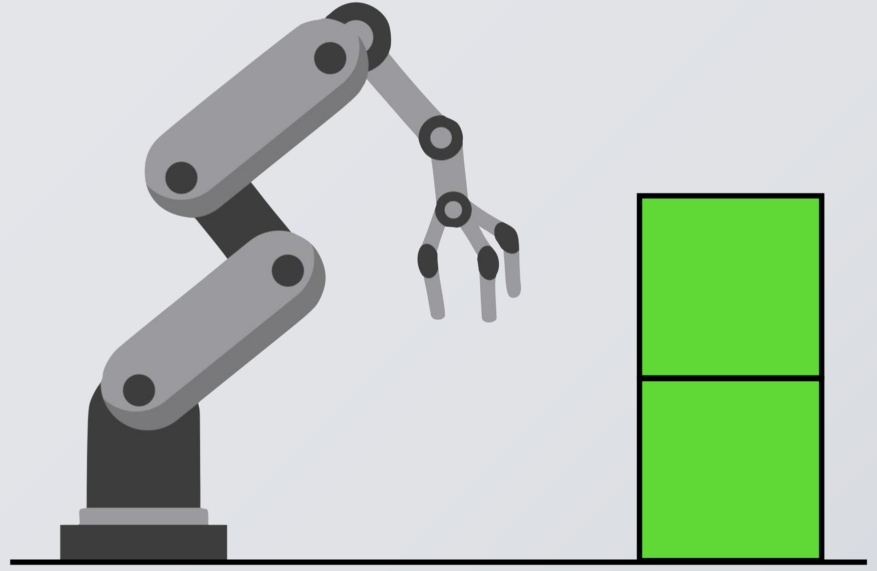
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

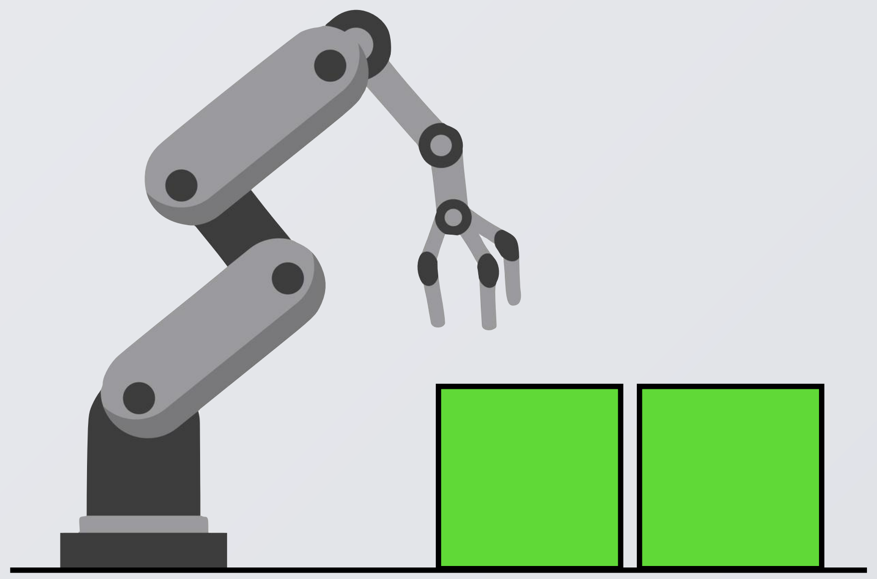
```
(:goal
  (and
    (arm-empty)

    (on A B)

  )
)
```

Correct

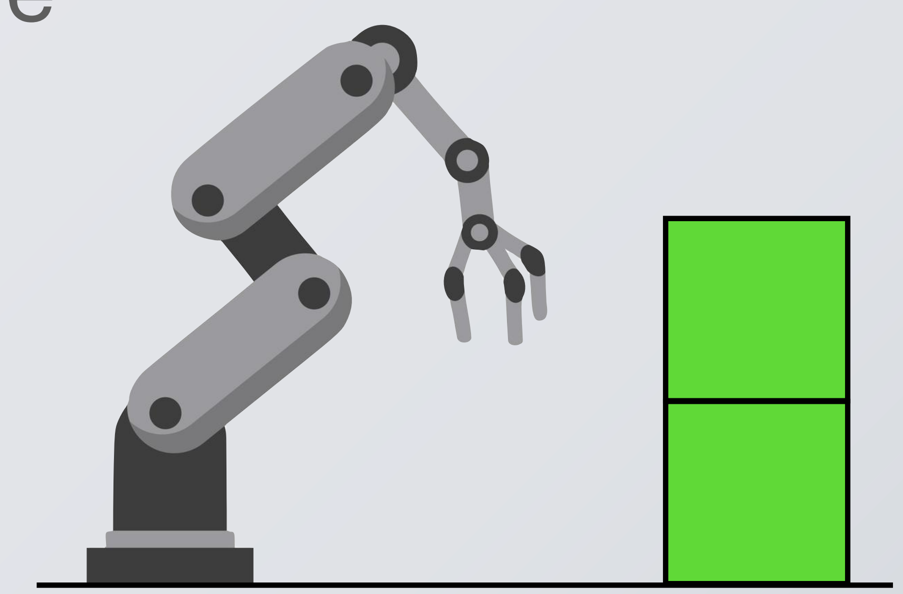
Evaluating PDDL

Problem

Initial Scene



Goal Scene



Goal PDDL

```
(:goal
  (and


    (on A B)


  )
)
```

Correct

Evaluating PDDL

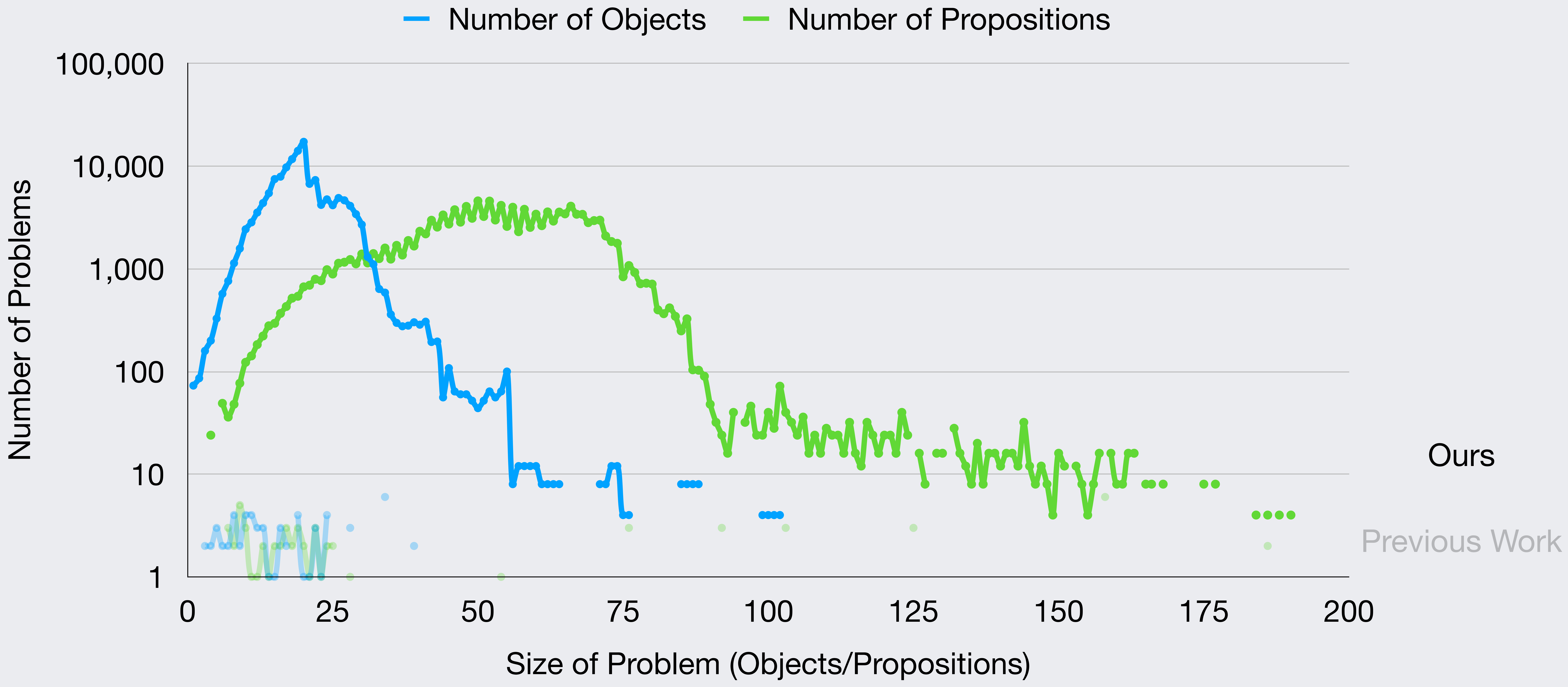
- PDDL generation is important
- PDDL correctness:
  - > Parsability
  - > Solvability
  - No one ground truth “text”

Planetarium 🪐

Planetarium 🪐

- Dataset
  - 145k natural language/PDDL problem pairs
  - Blocks World, Gripper, & Floor Tile domains
- Evaluation Algorithm
  - Convert PDDL into problem graphs
- Benchmark

Dataset

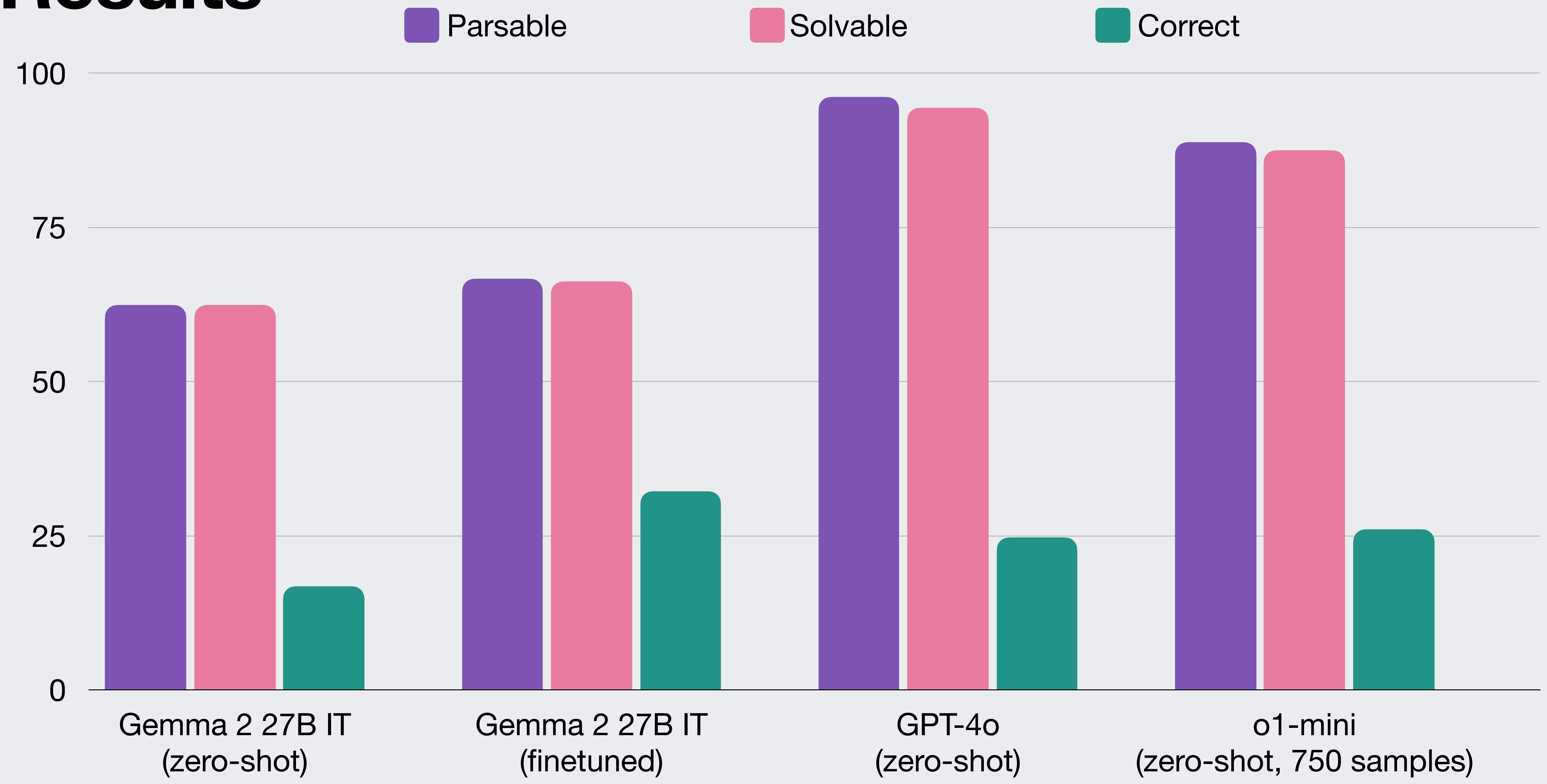


(For Blocks World, Gripper, and Floor Tile)

Experimental Setup

- Measure Parsability, Solvability, and Correctness
- Models:
    - Gemma 2 IT (zero-shot/fine-tuned)
    - GPT-4o
    - o1-mini
- Test set of ~16k problems
    - Blocksworld/Gripper: Held out tasks
    - Floor Tile: Entire domain held out

Results


Parsable
Solvable
Correct
100
75
50
25
0
Gemma 2 27B IT
(zero-shot)
Gemma 2 27B IT
(finetuned)
GPT-4o
(zero-shot)
o1-mini
(zero-shot, 750 samples)

Summary

- PDDL generation and evaluation is key

- As important as it is, it hasn’t been done properly in the past

- Empirically, LLMs don’t perform well on this

Thank You!

HF Dataset

GitHub

See you at NAACL!