

PROC2PDDL: Open-Domain Planning Representations from Texts

Tianyi Zhang^{1*} Li Zhang^{1*} Zhaoyi Hou³
 Ziyu Wang¹ Yuling Gu² Peter Clark²
 Chris Callison-Burch¹ Niket Tandon²

¹University of Pennsylvania ²Allen Institute for Artificial Intelligence
³University of Pittsburg
 {zty|zharry|ccb}@upenn.edu

Abstract

Planning in a text-based environment continues to be a major challenge for AI systems. Recent approaches have used language models to predict a planning domain definition (e.g., PDDL) but have only been evaluated in closed-domain simulated environments. To address this, we present PROC2PDDL, the first dataset containing open-domain procedural texts paired with expert-annotated PDDL representations. Using this dataset, we evaluate state-of-the-art models on defining the preconditions and effects of actions. We show that PROC2PDDL is highly challenging, with GPT-3.5’s success rate close to 0% and GPT-4’s around 35%. Our analysis shows both syntactic and semantic errors, indicating LMs’ deficiency in both generating domain-specific programs and reasoning about events. We hope this analysis and dataset helps future progress towards integrating the best of LMs and formal planning.¹

1 Introduction

Planning is the task of finding a sequence of actions to achieve a goal in a given environment (Fikes and Nilsson, 1971; LaValle, 2006). In real life, the environment is often described with natural language texts. To enable text-based, automated planning, recent work has used language models (LMs) to *generate plans* (Valmeekam et al., 2023a; Stein and Koller, 2023). However, this approach is found to fall short with regard to both performance and interpretability (Valmeekam et al., 2023c,b). Alternatively, another recent line of work has instead used LMs to *translate* the natural language description of environments to planning domain definition language (PDDL) (Ghallab et al., 1998). This symbolic representation can then be solved by a planner in a plan (Collins et al., 2022; Lyu et al.,

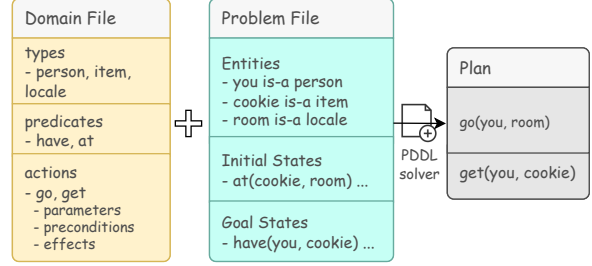


Figure 1: A PDDL solver produces a plan based on a minimal domain file and problem file. Previous work assumes the domain file as given, while we predict the action definitions in the domain file.

2023; Liu et al., 2023; Xie et al., 2023; Wong et al., 2023). Despite of success of such a neurosymbolic method, all the above work has only been evaluated in **closed-domains** simulated environments such as a household (e.g., ALFRED (Shridhar et al., 2020)) or discrete object placement (e.g., BlocksWorld (Valmeekam et al., 2024)) (see Table 1).

To enable **open-domain**, text-based planning, we propose PROC2PDDL, a dataset to evaluate models’ ability to generate PDDL given procedural texts. PROC2PDDL consists of 27 pairs of open-domain procedures and PDDL representations. Each PDDL representation include a domain file $\mathbb{D}\mathbb{F}$ that models the types, predicates, and actions, and a problem file $\mathbb{P}\mathbb{F}$ that models the entities, initial states, and goal states (see Figure 1). Because PROC2PDDL is not bound to any simulation, the PDDL representations are manually annotated by experts trained on this task to ensure validity, resulting in 27 domain files and 81 problem files.

Using this dataset, we study the task of action modeling (Lindsay et al., 2017) formulated as follows. The input is some natural language texts and the *header* of a $\mathbb{D}\mathbb{F}$ (i.e., types, predicates, and names of actions). The output is the *definitions* of actions in the $\mathbb{D}\mathbb{F}$ (i.e., parameters, preconditions, and effects) (see Figure 2). During evaluation, the

*Equal contribution.

¹Our resources can be found at <https://github.com/zharry29/nl-to-pddl-action>.

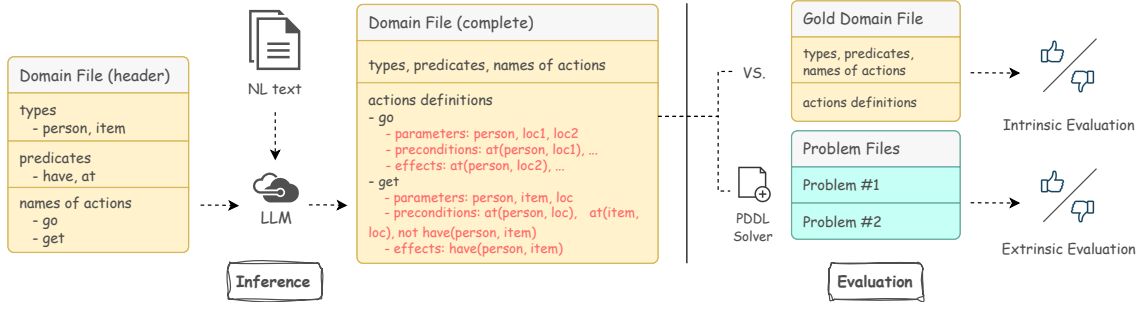


Figure 2: Our formulation of the $\mathcal{DF}$ -action prediction task. Given the domain file header that specifies the ontology, a model predicts action definitions including parameters, preconditions, and effects based on textual descriptions. During evaluation, the predicted $\mathcal{DF}$ is both compared to a reference and used to solve corresponding $\mathcal{PF}$ s.

predicted $\mathcal{DF}$ and a ground-truth $\mathcal{PF}$ are provided to a PDDL solver that may or may not successfully produce a plan. We then evaluate the existence and correctness of such a plan. In this formulation, our assumption of the $\mathcal{DF}$ header is necessary to ensure the consistency of semantics between the $\mathcal{DF}$ and the $\mathcal{PF}$ for evaluation. It is also empirically motivated; for example, a kitchen robot may have access to the types like ‘ingredients’ and predicates like ‘diced’ via some information extraction system given descriptive texts, but it may still need to predict, for “swinging a knife”, the precondition that it is only safe to do so to the ‘ingredients’ and the effect that they will become ‘diced’.

Through our experiment, we show that the task of action modeling in PROC2PDDL is highly challenging to state-of-the-art LMs, where GPT-3.5 almost fails completely and GPT-4 can only generate exactly matching $\mathcal{DF}$ s 16% of the time and solvable $\mathcal{DF}$ s 33% of the time. By devising a summarize-extract-translate pipeline that prompt LMs to modularly generate PDDL, we improve both metrics by 2-3%. In our analysis, the syntactic errors indicate LMs’ weakness in generating low-resource and domain-specific programming languages (Cassano et al., 2023) like PDDL, while the semantic errors suggest LMs’ inaccuracies to reason about actions and environments.

2 Problem Formulation

The problem of predicting a planning domain definition in a text-based environment can be seen as translating natural language texts to PDDL, which consists of a domain file $\mathcal{DF}$ and one or more problem files $\mathcal{PF}$ s.

A $\mathcal{DF}$ defines the following elements:

- a header H , which consists of
 - types of entities (e.g., *object*, *location*, *player*)

	# $\mathcal{DF}$	Datasets
Ours	27	PROC2PDDL
(Wong et al., 2023)	2	MineCraft, ALFRED
(Lyu et al., 2023)	1	SayCan
(Xie et al., 2023)	2	Blocksworld, ALFRED
(Liu et al., 2023)	7	Blocksworld, etc.
(Huang et al., 2023)	1	Tabletop
(Huang et al., 2022)	1	VirtualHome
(Silver et al., 2022)	18	Blocksworld, etc.
(Valmeekam et al., 2022)	2	Blocksworld, Logistics

Table 1: Our work proposes and evaluates models using PROC2PDDL which is open-domain and based on procedural texts, while past work has relied on closed-domain benchmarks which can be expressed with a singular $\mathcal{DF}$ with a fixed set of actions, based on some simulation.

- predicates (e.g., if object is *at* a location)
- names of possible actions (e.g., *boil water*)
- definitions of actions A , which consist of
 - parameters (e.g., water, pot) as a list of typed variables
 - preconditions (e.g., water and pot belongs to player; water is not treated) as a conjunctive normal form of predicates
 - effect (e.g., water is treated) as a conjunctive normal form of predicates

A $\mathcal{PF}$ defines the following elements:

- entities and their type (e.g., rainwater is water)
- initial states (e.g., bucket is empty)
- goal states (e.g., bucket is filled with rainwater; rainwater is treated)

We say that a $\mathcal{DF}$ and a $\mathcal{PF}$ can be solved if there exists a sequence of actions $A_1, \dots, A_n$ that results in a transition from the initial state to the goal state.

Traditionally, the task of text-based action modeling involves predicting $\mathcal{DF}$ based on text T , where a successfully generated $\mathcal{DF}$ can thus solve provided $\mathcal{PF}$ s defined accordingly. In this paper we tackle a simplified formulation where the types, predicates, and names of actions (H , domain

header) are provided; a model needs to predict the parameters, preconditions, and effects of each action (A , action definitions). As the information specified by H is guaranteed to be consistent with that of the PFs, the evaluation is well-defined.

3 Experiments

We provide some first attempts towards our formulation of action modeling in open-domain texts.

Dataset We introduce the PROC2PDDL dataset of 27 different $\mathbb{T}$ - $\mathbb{DF}$ -PFs tuples, drawing procedural texts from wikiHow articles of various topics (see Appendix A). A class of graduate students in a U.S. university with prior knowledge of PDDL are each given a wikiHow article and annotate a $\mathbb{DF}$ and multiple corresponding PFs from the article, each with a gold plan to solve it. On average, there are 13.33 defined actions in a $\mathbb{DF}$ and 8.07 instantiated actions in a gold plan. During prediction, we treat $\mathbb{T}$ as an annotated one-line summary of all annotated actions, though results using more variations of $\mathbb{T}$ can be found in Appendix C. On average, it takes several hours to train each human annotator, and another several hours to produce a $\mathbb{T}$ - $\mathbb{DF}$ -PFs tuple. We partition the 27 examples into a 5:6:16 train-development-test splits. In this work, the train split is unused as all our methods are zero-shot; only the development set is used for error analysis; the test set is strictly held out for evaluation. Some sample data of PROC2PDDL can be found in Appendix B.

Models We focus on zero-shot prompting LMs as PDDL for any domain is extremely costly to annotate, leading to hardly any training data. To predict action definitions A in $\mathbb{DF}$ based on the header H and some $\mathbb{T}$, we prompt an LM in a zero-shot manner (Brown et al., 2020) by describing the task and providing the input. Note that few-shot prompting is prohibitively costly due to the excessive length of a resulting $\mathbb{DF}$. We also incorporate the chain-of-thought (CoT) technique (Wei et al., 2023) that explicitly prompts an LM to perform three essential sub-tasks:

- Summarization: describe each action, including the expected preconditions and effects;
- Extraction: list the involved entities and their states before and after;
- Translation: based on the information above, convert $\mathbb{T}$ to PDDL.

The verbatim prompt wordings can be found in Appendix D. We experiment with two large LMs,

Model %	Intrinsic action acc.	Extrinsic	
		PF solve	exact plan
gpt-3.5	0.2	1.0	1.0
gpt-4	15.9	33.7	4.2
gpt-4 + CoT	18.1	35.8	6.3
gold	100.0	100.0	100.0

Table 2: LMs are greatly challenged by PROC2PDDL, where gpt-3.5 has close to zero performance, gpt-4 has non-trivial performance, while gpt-4 with chain-of-thought is slightly better.

Model %	Parameter	Precondition	Effect
gpt-4	36.7	31.1	53.0
gpt-4 + CoT	42.2	29.7	48.1

Table 3: LMs (both GPT-4 models) are worse at predicting preconditions than effects.

gpt-3.5-turbo-16k and gpt-4-32k dated June 2023 with max tokens of 8192², temperature of 0.5, and default hyperparameters otherwise.

4 Evaluation and Analysis

Now that a model generates the parameters, preconditions, and effects for each action, we have a complete $\mathbb{DF}$. We evaluate it in two ways (Figure 2). *Intrinsically*, we semantically compare the predicted A with the ground-truth provided by our PROC2PDDL and report an action-wide accuracy. Equivalence of two action definitions does not depend on the naming of variables nor on the order within conjunctions. *Extrinsically*, to measure actions’ coherence, we use a BFS-based PDDL solver³ to attempt to solve ground-truth PFs with the predicted $\mathbb{DF}$ and report a success rate. An unsolved PF is caused by (1.) no plan can be found, or (2.) the solver runs for more than 30 seconds, or (3.) the solver returns an error (usually a syntax error in generated PDDL).

The intrinsic and extrinsic results are shown in Table 2. gpt-3.5-turbo which achieves impressive performance on many tasks has a close-to-zero performance. In contrast, its predecessor gpt-4 performs significantly better with 18% action prediction accuracy and 36% solve rate of PFs based on the predicted $\mathbb{DF}$. Still, the performance far worse than ideal, showing that even a simplified

²Due to the need for very long input and output, the choice of open-source models is limited. We are in progress of implementing Mixtral-8x7B.

³<https://github.com/pucrs-automated-planning/pddl-parser>

		Unsolved		Solved	
	Syntactic Error	Bad Action	Good Action	Bad Plan	Good Plan
gpt-4	3	7	2	0	3

Table 4: A small-sample inspection shows that models make both syntactic and semantic errors.

open-domain planning formulation is challenging to state-of-the-art LMs.

CoT is helpful overall since it explicitly spells out many implicit entities and state changes in the extraction stage which are critical to predicting preconditions and effects. In most situations, the model summarizes the action and extracts the entity states correctly, though sometimes missing a few implicit entities. However, CoT’s bottleneck lies in the translation stage, during which there are mainly three types of errors.

1. mismatched predicates: the model uses (at ?loc ?item) instead of (inventory ?item);
2. hallucinated predicates: the model creates a new predicate (soaked ?item) while neglecting the existing (submerged ?item);
3. complicated predicates: the model adds unnecessary predicates (inventory ?submerged_item - item) when already has (inventory ?item).

To address these, we leave to future work to demonstrate and standardize the translation process by clearly describing when an entity change or stage change is and is not needed, while also encouraging the model to strictly match the given predicates. Finer-grained evaluation results are shown in Table 3 to tease out the performance regarding such component within an action. It is clear that the LM is worse at predicting preconditions than at predicting effects. This is understandable as procedural texts like wikiHow tend to be less explicit about predictions than about effects (e.g., from *bake for 10 minutes* it is obvious that the food will be baked, but it is unclear what state it had been in).

To provide deeper insights into model performance, we manually inspect the model output of gpt-4 on all 6 examples (15 PFs) in the development set. We consider the following scenarios.

Unsolved Whenever the predicted $\mathbb{DF}$ cannot solve a $\mathbb{PF}$, either a syntactic or a semantic error has occurred. For a **syntactic error**, the output may contain illegal expressions that cannot be parsed. For example, (inventory ?player (clean ?strips)) is unacceptable because the

arguments to a predicate must be atomic types, not another predicate. For a **semantic error** (namely, a ‘bad action’), we identify the first problematic action that differs with the ground-truth. For example, if the action `cut_plant` misses a critical effect of (inventory ?player ?stalk), then other actions such as `graft_stalk` requiring it cannot be executed. At times, there could be false negatives where the predicted action definitions are in fact reasonable but still cannot lead to a solution (namely, a ‘good action’).

Solved Even when the predicted $\mathbb{DF}$ solves a $\mathbb{PF}$, the plan may be different from the gold plan. It is naturally possible that the predicted plan is a fluke made possible by under-specified preconditions or over-exaggerated effects, as well as loopholes in the $\mathbb{PF}$ leading to unreasonable shortcuts. For the example in Figure 1, a model could *cheat* by defining the action `get` by not requiring the person and object to be in the same location; thus, the predicted plan would unreasonably omit the action `go`. However, at times, the predicted plan could also be a reasonable alternative.

The statistics of these errors are shown in Table 4. When no solution can be found, true negative is highly likely as the model indeed makes aforementioned mistakes during action prediction. When some solution is found, false positive is still possible as the predicted plan may be unreasonable. See attached materials for a complete error analysis of these examples. Our aforementioned future pipeline that separates summarization and translation would likely mitigate these errors.

5 Conclusion

We present PROC2PDDL, the first open-domain dataset that juxtaposes natural language and planning domain definition language. Our experiments show that LMs find it challenging to model the precondition and effects of actions, leading to both syntactic and semantic errors. We hope our analysis and dataset helps future progress towards integrating the best of LM and formal planning.

6 Limitations

Any planning language, including PDDL which we consider in this work, is an approximation of planning in the real world and cannot accurately reflect its complexity. Due to the consideration for simplicity in the annotation process, we use the primitive version of PDDLs, with restricted

expressions and syntax, instead of newer versions of the planning language which extend its syntax in a variety of way.

Annotating PROC2PDDL is extremely costly as it requires knowledge of PDDL and much effort to translate procedural texts to PDDL. Thus, our dataset is relatively small with a limited range of topics. Due to the highly complex and subjective nature of the annotation process, each annotated example may reflect idiosyncratic thought processes and biases of the individual annotator.

As with many similar works, there is a known gap between high-level planning such as ours (with high-level actions like “boil”) and the actions used by present-day robots (with low-level motor functions like “move”). However, like similar works, we believe our efforts can see more practical application in the near future.

Our modeling efforts so far have mainly considered options of zero-shot prompting. There of course exists many other approaches including the few-shot setting, fine-tuning, and the model distillation paradigm, which we plan to experiment with in the future. Moreover, our evaluation is imperfect in that even a well-annotated DF-PF pair might have multiple successful plans. Manual inspection is still necessary to accurately gauge models.

Acknowledgements

This work is supported in part by the DARPA KAIROS Program (contract FA8750-19-2-1004), AFRL (contract FA8750-23-C-0507), the Office of the Director of National Intelligence (ODNI) via the IARPA HIATUS Program (contract 2022-22072200005), the NSF (Award 1928631), and gifts from Roblox and Salesforce. Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, NSF, AFRL, the U.S. Government, or of Roblox or Salesforce. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Carolyn Jane Anderson, Michael Greenberg, Abhinav Jangda, and Arjun Guha. 2023. Knowledge transfer from high-resource to low-resource programming languages for code llms. *arXiv preprint arXiv:2308.09895*.

Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. 2022. *Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks*.

Richard E Fikes and Nils J Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208.

Malik Ghallab, Adele Howe, Craig Knoblock, Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. 1998. PDDL - the planning domain definition language. Technical Report "CVC TR-98-003/DSC TR-1165", Yale Center for Computational Vision and Control.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. 2023. Grounded decoding: Guiding text generation with grounded models for robot control. *arXiv preprint arXiv:2303.00855*.

Steven M LaValle. 2006. *Planning algorithms*. Cambridge university press.

Alan Lindsay, Jonathon Read, Joao Ferreira, Thomas Hayton, Julie Porteous, and Peter Gregory. 2017. Framer: Planning models from natural language action descriptions. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 27, pages 434–442.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. *Llm+p: Empowering large language models with optimal planning proficiency*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. *Faithful chain-of-thought reasoning*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 10740–10749.

Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2022. Pddl planning with pretrained large language models. In NeurIPS 2022 Foundation Models for Decision Making Workshop.

Katharina Stein and Alexander Koller. 2023. Autoplanbench:: Automatically generating benchmarks for llm planners from pddl. arXiv preprint arXiv:2311.09830.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023a. 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.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. Advances in Neural Information Processing Systems, 36.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023b. On the planning abilities of large language models—a critical investigation. arXiv preprint arXiv:2305.15771.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can’t plan (a benchmark for llms on planning and reasoning about change). arXiv preprint arXiv:2206.10498.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023c. [Large language models still can’t plan \(a benchmark for llms on planning and reasoning about change\)](#).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).

Lionel Wong, Jiayuan Mao, Pratyusha Sharma, Zachary S Siegel, Jiahai Feng, Noa Korneev, Joshua B Tenenbaum, and Jacob Andreas. 2023. Learning adaptive planning representations with natural language guidance. arXiv preprint arXiv:2312.08566.

Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. 2023. [Translating natural language to planning goals with large-language models](#).

A Topics

Below are a list of the titles of wikiHow articles in PROC2PDDL, chosen per the requirement of a graduate-level university class.

- create secret society
- throw an anime party
- open a coconut
- calculate pi
- hack
- get out of quicksand
- make a detective kit
- lock picking
- make papyrus
- survive on a desert island
- survive in the jungle
- survive a war
- survive a comet hitting earth
- survive a nuclear attack
- survive in the woods
- survive deserted island
- survive shark attack
- survive emp attack

Each topic may have one or more annotated DFs representing different domains. The homogeneity of the last 7 topics is due to the class’ topic of interactive fictions.

B Sample Data: T, DF, and PF

To exemplify PROC2PDDL, below is an example procedural text $\mathbb{T}$ titled ‘survive in the jungle’, up to the third step, truncating the rest.

1. Collect rainfall from leaves and bamboo stalks. Look for large leaves that collect rainfall and bend them into a funnel to pour the water into a bottle or straight into your mouth. Bend bamboo stalks to let the water that collects in the compartments flow out into a container or break the bamboo compartment off at the line that goes across the stalk to use it as a water bottle. You could also look for rock formations that form natural pools and collect rainwater, but it is best to do this after a fresh rainfall to avoid pools that have been sitting for a long time and may be contaminated with bacteria. If you don't have a water bottle or other container to collect water, try to find other natural containers in the jungle such as a coconut shell or piece of wood shaped like a bowl. You can also leave these items out when it rains to collect the fresh water.
2. Boil water from streams to kill any bacteria. Look for running streams to find fresh water. Filter out any particles through a sock, shirt, or other fabric, then start a fire and boil the water to kill bacteria that can make you sick. If you don't have a pot to boil water in, then you can use a tin can, single-walled stainless steel water bottle, or any other metal container. If you have no way of making a fire or boiling the water, then you should avoid drinking water

from streams. It can be contaminated with many types of bacteria from animals that will make you very sick. Always avoid drinking water from stagnant pools as the water is likely contaminated.

3. Make a solar water still with a container and a plastic sheet. Dig a hole in an area that receives at least some direct sunlight and put a container, such as a water bottle or can, in the middle of the hole. Fill the space between the sides of the hole and the container with wet leaves. Place a plastic sheet over the top of the hole and put rocks or other heavy objects around the edges to hold it in place. Put a small stone in the middle of the sheet above the container. The plastic sheet will accumulate condensation that will drip down the underside of the sheet and into the container. This water is distilled and safe to drink. You can use natural containers such as bamboo or a coconut shell if you don't have a bottle or can. A solar still does not collect large amounts of water. It should be used as a supplemental source of water rather than a primary source.

.....

Below is a sample annotated **DF** of the above:

```
(define (domain survive_in_the_jungle)
  (:requirements :strips :typing)
  (:types
    stone wood bamboo_container water fire
    sos_sign fruit - item
    basecamp - location
    ill dehydrated hungry - condition
    player
    direction
  )
  (:predicates
    (has_bamboo ?loc - location) ; this
      location has bamboo to create a
      container
    (has_rainfall ?loc - location) ; this
      location has received rainfall to
      collect water
    (has_fruit ?loc - location) ; this location
      has fruits to pick
    (treated ?water - water) ; True if the
      water has been decontaminated by
      boiling it
    (is ?c - condition ?p - player) ; True if
      the player is under the specified
      condition
    (at ?obj - object ?loc - location) ; an
      object is at a location
    (inventory ?player ?item) ; an item is in
      the player's inventory
    (connected ?loc1 - location ?dir -
      direction ?loc2 - location) ; location
      1 is connected to location 2 in the
      direction
    (blocked ?loc1 - location ?dir - direction
      ?loc2 - location) ; the connection
      between location 1 and 2 is currently
      blocked
  )
)
```

```
(:action go ; navigate to an adjacent
  location
  :parameters (?dir - direction ?p - player ?
    l1 - location ?l2 - location)
  :precondition (and (at ?p ?l1) (connected ?
    l1 ?dir ?l2) (not (blocked ?l1 ?dir ?
    l2))))
  :effect (and (at ?p ?l2) (not (at ?p ?l1)))
)

(:action get ; pick up an item and put it in
  the inventory
  :parameters (?item - item ?p - player ?l1 -
    location)
  :precondition (and (at ?p ?l1) (at ?item ?
    l1))
  :effect (and (inventory ?p ?item) (not (at
    ?item ?l1)))
)

(:action get_bamboo_container; get a bamboo
  container using surrounding bamboo
  :parameters (?p - player ?loc - location)
  :precondition (and (at ?p ?loc) (has_bamboo
    ?loc))
  :effect (inventory ?p bamboo_container)
)

(:action collect_rain_water
  :parameters (?p - player ?loc - location)
  :precondition (and (at ?p ?loc) (inventory
    ?p bamboo_container) (has_rainfall ?
    loc))
  :effect (and (inventory ?p water) (not (
    treated water)))
)

(:action create_fire
  :parameters (?p - player ?loc - location)
  :precondition (and (at ?p ?loc) (inventory
    ?p stone) (inventory ?p wood))
  :effect (and (at fire ?loc) (not (inventory
    ?p stone)) (not (inventory ?p wood)))
)

(:action treat_water
  :parameters (?p - player ?loc - location)
  :precondition (and (inventory ?p water) (
    not (treated water)) (at fire ?loc))
  :effect (and (treated water))
)

(:action drink_water
  :parameters (?p - player)
  :precondition (and (inventory ?p water) (
    treated water))
  :effect (not (is dehydrated ?p))
)

(:action drink_untreated_water
  :parameters (?p - player)
  :precondition (and (inventory ?p water) (
    not (treated water)))
  :effect (is ill ?p)
)

(:action create_sos_sign
  :parameters (?p - player)
  :precondition (and (inventory ?p stone) (at
```

```

      ?p basecamp))
    :effect (and (not (inventory ?p stone)) (at
      sos_sign basecamp))
  )
  (:action pick_fruit
    :parameters (?p - player ?loc - location)
    :precondition (and (at ?p ?loc) (has_fruit
      ?loc))
    :effect (inventory ?p fruit)
  )

  (:action eat_fruit
    :parameters (?p - player)
    :precondition (and (is hungry ?p) (
      inventory ?p fruit))
    :effect (and (not (inventory ?p fruit)) (
      not (is hungry ?p)))
  )

  (:action escape
    :parameters (?p - player)
    :precondition (and (at ?p basecamp) (at
      sos_sign basecamp) (not (is hungry ?p))
      (not (is dehydrated ?p)) (not (is ill
        ?p)))
    :effect (not (at ?p basecamp))
  )
)

```

Below is an annotated $\mathbb{P}\mathbb{F}$ of the above:

```

(define (problem escape)
  (:domain survive_in_the_jungle)

  (:objects
    npc - player
    jungle bamboo_forrest basecamp - location
    in out north south east west up down -
      direction
    stone wood sos_sign - item
    ill dehydrated hungry - condition
  )

  (:init
    (at npc basecamp)
    (connected basecamp west bamboo_forrest)
    (connected bamboo_forrest east basecamp)
    (connected basecamp east jungle)
    (connected jungle west basecamp)

    (has_bamboo bamboo_forrest)
    (has_fruit jungle)
    (has_rainfall jungle)

    (at stone bamboo_forrest)
    (at stone basecamp)
    (at wood jungle)

    (is dehydrated npc)
    (is hungry npc)
  )

  (:goal (and (not (at npc basecamp)) (not (at
    npc bamboo_forrest)) (not (at npc jungle)
    ) ))
)

```

Model %	Intrinsic action acc.	Extrinsic	
		$\mathbb{P}\mathbb{F}$ solve	exact plan
w/o $\mathbb{T}$ (baseline)	13.7	26.3	3.2
$\mathbb{T}$ =sum	15.9	33.7	4.2
$\mathbb{T}$ =sum, CoT	18.1	35.8	6.3
$\mathbb{T}$ =map	11.8	13.7	2.1
$\mathbb{T}$ =map, CoT	8.9	26.3	1.1
$\mathbb{T}$ =rel	11.6	27.4	0.0
$\mathbb{T}$ =rel, CoT	12.2	21.1	4.2
$\mathbb{T}$ =all	12.1	28.4	0.0
$\mathbb{T}$ =all, CoT	12.1	31.6	0.0

Table 5: Performance of GPT-4 using different portions of text $\mathbb{T}$. Metrics include action-wide accuracy, average edit distance of action definitions, the proportion of $\mathbb{P}\mathbb{F}$ s that can be solved, and the proportion of generated plans that exactly match the gold plans.

C Choice of Input Text

We consider the following choice of wikiHow texts as $\mathbb{T}$.

Prompt without text (w/o $\mathbb{T}$) is an ablation baseline where the model predicts A solely based on H . Naturally, none of the three aforementioned stages are involved in this prompt condition.

Prompt with text (w/ $\mathbb{T}$) additionally provides the model with four different portions of $\mathbb{T}$, involving the three aforementioned stages, as follows:

($\mathbb{T}$ = **all**): All steps in a wikiHow article.

($\mathbb{T}$ = **rel**): In PROC2PDDL, each wikiHow article consists of step paragraphs that may or may not be used in defining the actions in the $\mathbb{D}\mathbb{F}$. Hence, a mapping between actions and steps is also annotated. This context includes relevant steps to all actions in a $\mathbb{D}\mathbb{F}$.

(e.g., Step 1. Find fresh water... Step 2. Collect food... Step 7. Set up camp...)

($\mathbb{T}$ = **map**): Each action mapped with steps based on the annotated mapping in PROC2PDDL.

(e.g., clean_water: Step 1. Find fresh water...)

($\mathbb{T}$ = **sum**): One-line summaries of actions annotated in PROC2PDDL.

(e.g., clean_water; boil water to clean it)

The four prompts are increasingly general. Distinguishing from the requirements, the first text conditions demand accurate information identification and extraction, while the last one clearly defines the action but still requires the model to infer implicit entity states. All prompts request an exact translation.

Result analysis As shown in Table 5, in w/o $\mathbb{T}$ setting, fully relying on its implicit knowledge, the model is already capable of inferring PDDL syn-

tactically and semantically. In w/ $\mathbb{T}$ settings, our model shows an ‘U’ performance in terms of the text length. Using a sentence-long description for each action provided by PROC2PDDL, the model achieves the best performance among all, showing a strong deduction ability with the limited but precise NL input. The **all** setting ensues, which requires the most extraction rather than inference. In contrast, the middle ones (**rel/map**) with decreasing signal-to-noise ratio lead to worse results, indicating its shortage of extraction-inference trade-off. The signals contain both the described entity states and step relations, explicitly and implicitly. This shortage may come less from the entity states (e.g., fish, spear in hunt_fish), but more from the relation between actions (e.g., make_spear to hunt_fish) which may be expressed in the **sum** and **all** settings.

D Prompts

For reproducibility, we provide the verbatim prompts that we used in the above experiments.

D.1 Prompt without CoT

Could you fill out the below PDDL actions with the predicates based on the text?

All fields: parameters, precondition and effect, should have predicates.

For each action, do NOT change the name and do NOT drop the action and do NOT add more actions.

The output should be in correct PDDL format.

<wikiHow text and domain header>

here are the actions I want:

<insert_action_names>

here are the types I have:

<insert_types>

here are the predicates I have:

<insert_predicates>

here are the texts containing steps to <insert_goal>:

<insert_text>

Example Completion:

(:action clean_water

 :parameters (?player - human ?water - water)

 :precondition (inventory ?player ?water)

 :effect (treated ?water)

)

D.2 Prompt with CoT

Could you fill out the below PDDL actions with the predicates based on the text? All fields: parameters, precondition and effect, should have predicates.

For each action, do NOT change the name and do NOT drop the action and do NOT add more actions and:

First, summarize the action in a few sentences based on the text and provide its requirements and its aims if it has.

Next, identify ALL the entities involved in the action and describe whether it changed, unchanged, added, removed in the action in natural language.

Last, translate it into PDDL format. Check all the related entities are in the ‘parameters’.

Please use this output format:

- action name: ...

- summarize action: ...

- what is needed to do the action: ...

- what is the result: ...

- entities:

- entity name: ...

- before: ...

- after: ...

- describe how to match it to relevant predicates step by step:

1. ...

2. ...

<wikiHow text and domain header>

here are the actions I want:

<insert_action_names>

here are the types I have:

<insert_types>

here are the predicates I have:

<insert_predicates>

here are the texts containing steps to <insert_goal>:

<insert_text>

Example Completion:

- action name: clean_water

- summarize action: The player cleans water in their inventory using heat from a fire.

- what is needed to do the action: The player must have untreated water in their inventory and be at a location with fire.
- what is the result: The player has treated water in their inventory.

- entities:
- entity name: player
- before: Having untreated water in inventory.
- after: Having treated water in inventory.
- entity name: water
- before: Untreated.
- after: Treated.

- describe how to match it to pddl relevant predicates step by step:

1. Check if the player has untreated water in their inventory.
2. Check if the player is at a location with a fire.
3. Replace untreated water with treated water in the player's inventory in the effect.

PDDL:

```
(:action clean_water
  :parameters (?player - human ?loc - location ?water - water)
  :precondition (and (at ?player ?loc) (inventory ?player ?water) (not (treated ?water)) (has_fire ?loc))
  :effect (treated ?water)
)
```

E Calculating Actions Equivalence

The distance between two actions can be divided to two parts:

1. The distance between parameters:

We do not need to consider the specific parameter names; we only need to consider the parameter types. For each parameter in Action1, we iterate over the parameter list of Action2 to find the first parameter in Action2 with the same type. We use two hash maps, p1 and p2, to record these two parameters and their corresponding types. We increment the counter by 1, remove that parameter from the parameter list of Action2, and break from the current loop. After the iteration, we obtain the number of matching parameters, n. The distance between parameters can be calculated as $|\text{number of parameters in Action1} - n| + |\text{number of parameters in Action2} - n|$.

2. The distance between preconditions/effects:

For each condition in Action1, we iterate over the condition list of Action2. The conditions can only match if they have the same predicate and the same number of parameters. We iterate over the parameters in these conditions and make the following judgments:

- If neither of the two current parameters has appeared before (in p1 and p2) and these parameters are not identical, they don't match.
- If the two parameters have different categories, they don't match.
- If the two parameters have the same categories and don't have an index, we consider them as the same parameter entity and give them the same index. We continue the iteration.
- If the two parameters already have indexes, we check if the indexes are equal. If they are not equal, they don't match. Otherwise, we continue the iteration.
- In any other case, they don't match.

If all parameters of the two conditions match, we increment n by 1. The distance between preconditions/effects can be calculated as $|\text{number of preconditions/effects in Action1} - n| + |\text{number of preconditions/effects in Action2} - n|$.