

# PLASMA: Making Small Language Models Better Procedural Knowledge Models for (Counterfactual) Planning

Faeze Brahman<sup>12</sup> Chandra Bhagavatula<sup>1</sup> Valentina Pyatkin<sup>1†</sup> Jena D. Hwang<sup>1†</sup>

Xiang Lorraine Li<sup>15</sup> Hirona J. Arai<sup>3</sup> Soumya Sanyal<sup>3</sup>

Keisuke Sakaguchi<sup>4</sup> Xiang Ren<sup>13</sup> Yejin Choi<sup>12</sup>

<sup>1</sup>Allen Institute for Artificial Intelligence <sup>2</sup>University of Washington

<sup>3</sup>University of Southern California <sup>4</sup>Tohoku University <sup>5</sup>University of Pittsburgh

[faezeb@allenai.org](mailto:faezeb@allenai.org)

## Abstract

Procedural planning, which entails decomposing a high-level goal into a sequence of temporally ordered steps, is an important yet intricate task for machines. It involves integrating common-sense knowledge to reason about complex contextualized situations that are often counterfactual, e.g. “scheduling a doctor’s appointment without a phone”. While current approaches show encouraging results using large language models (LLMs), they are hindered by drawbacks such as costly API calls and reproducibility issues. In this paper, we advocate planning using smaller language models. We present PLASMA, a novel two-pronged approach to endow small language models with procedural knowledge and (counterfactual) planning capabilities. More concretely, we develop *symbolic procedural knowledge distillation* to enhance the implicit knowledge in small language models and an *inference-time algorithm* to facilitate more structured and accurate reasoning. In addition, we introduce a novel task, *Counterfactual Planning*, that requires a revision of a plan to cope with a counterfactual situation. In both the original and counterfactual setting, we show that orders-of-magnitude smaller models (770M-11B parameters)<sup>1</sup> can compete and often surpass their larger teacher models’ capabilities.<sup>1</sup>

## 1 Introduction

Powered by massive scale, large language models (LLMs) excel on many downstream tasks that require commonsense. One such task is *procedural planning* [27], a task that involves decomposing a high-level **goal** into a sequence of coherent, logical, and goal-oriented steps (**plan**) (e.g. “see a movie” → “Look up movie showings”, “Choose a movie” . . .). Recent approaches model this task as a conditional text generation problem using LLMs [23, 11, 1]. Despite their reasonable performance on the task, their steep computational cost and inaccessibility hinder wider adoption of LLMs [24].

We present PLASMA (PLAN with SMALL models), a novel two-pronged framework to impart planning abilities in small LMs. We achieve this through *symbolic procedural knowledge distillation* to enhance the implicit knowledge in small LMs (Figure 1) and an *inference-time decoding algorithm* to enable structured reasoning (Figure 2). We formulate *symbolic procedural knowledge distillation* [41, 3] in two stages: (i) Knowledge verbalization to generate procedural knowledge from an LLM, and (ii) Knowledge distillation to transfer LLM-generated knowledge to a smaller LM.

<sup>†</sup>Authors contributed equally.

<sup>1</sup>We make our dataset and code publicly available at: <https://github.com/allenai/PlaSma>

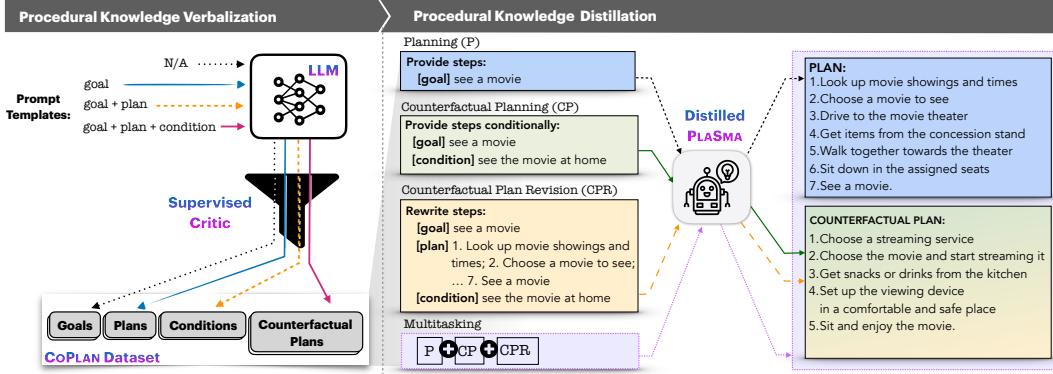


Figure 1: Symbolic Procedural Knowledge Distillation.

In addition to the standard planning task, we introduce and verbalize knowledge for novel task formulations under counterfactual settings: *Counterfactual planning* and *Revision*. These tasks enable a more realistic setting by requiring models to reason about contextually constrained situations in real-world applications; specifically, the model generates or revises a plan based on a given goal (e.g., "see a movie") while adhering to an additional **condition** (e.g., "at home"). Our knowledge verbalization process results in a large (counterfactual) procedural planning dataset, COPLAN, which is then used to train smaller models, PLASMA, using both task-specific and multi-task distillation.

We observe that the standard next-token prediction objective in auto-regressive LMs (applied during distillation) does not equip them with sufficient causal and temporal reasoning abilities to generate high-quality plans, or a mechanism to rectify their mistakes in earlier steps. To address this challenge, we develop a *verifier-guided step-wise beam search* to better leverage the multi-step structure of plans (resulting in PLASMA+). Concretely, we incorporate a step-wise verifier in our decoding process to guide PLASMA+ to generate more semantically coherent and temporally accurate plans.

Through experiments, we show that our approach is effective at endowing smaller LMs with planning abilities. For the standard planning task, smaller student models (of varying sizes) achieve 17.57% relative improvements, on average, over their teacher. The best student model is comparable even to GPT-3, a model 16 times the student's size. Furthermore, we, for the first time, distill counterfactual planning abilities in small-size models, achieving 93% validity rate according to human evaluation. In a simulated environment [29], our model significantly outperforms previous work based on GPT-3 [11] on executability (by 17%) and correctness (by 25%). Taken together, our framework including symbolic procedural distillation, decoding-time algorithm, and the proposed tasks and the accompanying COPLAN dataset provide valuable resource and direction for advancing research in the field of procedural planning.

## 2 Small Language Models as Procedural Knowledge Models

In this section, we discuss how to endow small students with procedural knowledge and (counterfactual) planning capabilities. We first describe our knowledge verbalization and distillation framework which we collectively refer to as Symbolic Procedural Knowledge Distillation (§2.1, §2.2). We then propose a strategy to enhance the reasoning capabilities of small students via a novel verifier-guided step-wise decoding algorithm (§2.3).

### 2.1 CoPLAN: Procedural Knowledge Verbalization from Large Teachers

Large language model can perform new tasks by adapting to a few in-context examples [4]. We thus leverage this emergent reasoning capabilities of LLM to circumvent the challenge of crowdsourcing supervised datasets at scale. We collect data targeting the following three tasks:

1. **Goal-based Planning (pl.)**, decomposing a high-level goal  $g$  into a sequence of temporally extended steps  $y = \{s_t\}_{t=1}^T$ .

2. **Counterfactual Planning (cp.)**, decomposing a high-level goal  $g$  into a sequence of temporally extended steps  $y = \{s_t\}_{t=1}^T$  while satisfying a given condition  $c$ .
3. **Counterfactual Plan Revision (cpr.)**, rewriting an initial plan  $y$  to a given goal  $g$  into a new plan  $y'$  in order to satisfy a given condition  $c$ .

Our knowledge verbalization pipeline shown in the left side of Figure 1 is a two-stage process: 1) instance generation through few-shot prompting, and 2) automatic data curation using a critic to filter out the low quality data. The process results in COPLAN, a quality dataset containing goals, plans, conditions, and counterfactual plans.

**Step 1. Data Generation** We start by generating a large pool of goals  $\mathcal{G}$  with a diverse range of topics in a bootstrapping fashion. We initiate the seed goal pool with 100 goals generated by GPT-3 (text-curie-001) along with 5 example goals provided by the authors. With the seed goal pool, we iteratively expand it by GPT-3 with randomly selecting example goals for prompting.

For each generated goal  $g \in \mathcal{G}$ , we few-shot prompt a teacher model  $\mathcal{M}$  to generate a set of ordered steps, as a plan  $y$  to achieve the goal. The input to  $\mathcal{M}$ , including instruction and few-shot examples, takes the format shown in Figure 7. Since LLMs can be sensitive to instruction, and/or few-shot examples [28, 21], we randomize the prompt by (i) manually creating a set of semantically similar instructions and each time randomly sample from the instruction set (ii) creating dynamic in-context examples for each input. We use a subset of the existing ProScript [34] and DeScript [39] datasets as our seed source to form in-context examples,  $\mathcal{P} = \{(g_j, y_j)\}_{j=1}^M$ :

$$y_i \sim \mathcal{M}(y_i|g_i, \mathcal{P})$$

The result is a pool of 140k pairs of goal and plans,  $(g, y)$ , generated from the teacher model.

For the counterfactual setting, we also obtain conditions  $c$ , and modified plans  $y'$  from a teacher model  $\mathcal{M}$  through few-shot prompting. We manually design our prompts  $\mathcal{P}$  to collect natural language conditions concerning the environment the task is performed in such as Location (“the store is closed”), Equipment (“you don’t have a sharp tool”), Safety (“the car breaks down”) or user’s specifications such as Physical Condition and Preference (“you have an injury”). For a given goal  $g_i$  and plan  $y_i$ , we sample conditions:

$$c_i \sim \mathcal{M}(c_i|g_i, y_i, \mathcal{P})$$

Next, we few-shot prompt  $\mathcal{M}$  to rewrite an initial plan  $y$  for a given goal  $g$  such that it satisfies the requirement of a condition  $c$ :

$$y'_i \sim \mathcal{M}(y'_i|g_i, y_i, c_i, \mathcal{P})$$

The prompting templates and examples of conditions are shown in Figure 8 and Table 6.

**Step 2. Automatic Data Curation** To retain high-quality data for planning under the original and counterfactual settings, we filter out generated samples from Step 1, i.e. generated plans, conditions and counterfactuals, that are invalid or of low quality. A plan  $y$  is considered invalid if it contains an *illogical order* of steps, is *off-topic* (w.r.t the goal) or *incomplete*. Whereas a counterfactual plan  $y'$  should not only satisfies these general criteria but should also adhere to the condition.

To this end, we train separate supervised critic models to judge the quality of generated samples of different types. We collect human annotations of *valid* vs. *invalid* samples on Amazon Mechanical Turk to train a RoBERTa-Large [17] as our critic models. All critics are binary classifiers which identify whether a tuple of either (goal, plan), (goal, plan, condition) or (goal, plan, condition, modified plan) is valid. We provide more details on annotation instructions, and hyper-parameter tuning in Appendix B.1 and B.2.

Naturally, there is a trade-off between dataset size and precision. Following West et al. [41], we test several confidence thresholds at which the critic rejects a pair and choose the best values (0.65, 0.76, 0.82)<sup>2</sup> according to precision-recall curves. After filtering out low quality data, our final COPLAN dataset consists of 2 main subsets including 57,794 (goal, plan) for the original **goal-based planning** task ( $\mathcal{D}^{pl.}$ ), and 43,690 (goal, plan, condition, modified plan) for the **counterfactual** settings, ( $\mathcal{D}^{cp.}$  and  $\mathcal{D}^{cpr.}$ ). On the original planning task, COPLAN is  $\times 11$  larger in scale than existing datasets [34, 39] while keeping the precision at 74%. On the proposed counterfactual settings, our dataset is to

<sup>2</sup>These values are for plan, condition and counterfactual plans, respectively.

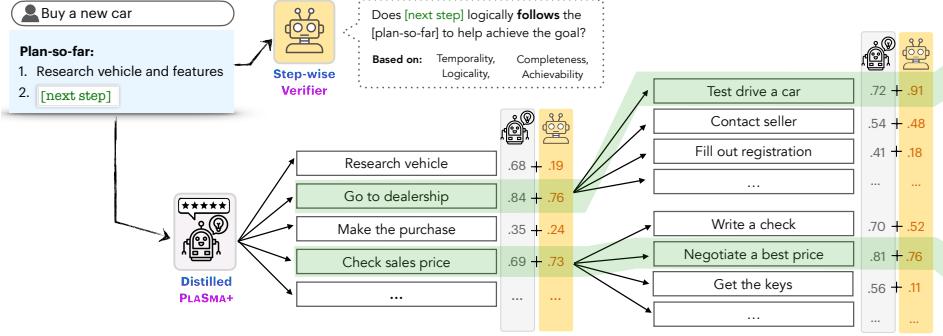


Figure 2: Verifier-guided Step-wise Beam Search. For brevity, we only showcase with  $N = 5$  and  $K = 2$  for the first step and  $N = 4$  and  $K = 2$  for the second step. The scores are for illustration purposes only.

the best of our knowledge the first large-scale counterfactual procedural planning dataset. Analyses show that the COPLAN includes a diverse array of topics covered by goals (§A.1) and conditions (§A.2).

## 2.2 PLASMA: Procedural Knowledge Distillation into Small Students

After obtaining our procedural planning data COPLAN, we use it to fine-tune student models on the three different tasks. We consider both task-specific and multi-task distillation objectives to transfer generated procedural knowledge into the student models:

**Task-specific Distillation.** Following the common practice, we use the standard autoregressive language modeling objective [32] to fine-tune separate student models for each task:

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y) \sim D^{task}} [-\log p_\theta(y|\mathcal{T}(x))], \quad \text{for } \text{task} \in \{pl., cp., cpr.\} \quad (1)$$

where  $\mathcal{T}(x)$  is a task-specific template for each task-specific input  $x$  (see right side of Figure 1).

**Multi-task Distillation.** We also aim to improve the generalization of the student model by exploiting the knowledge contained in the three related tasks as an inductive bias [33, 40]. We thus minimize the joint loss:

$$\begin{aligned} \mathcal{L}(\theta) = & \mathbb{E}_{(g,y) \sim D^{pl.}} [-\log p_\theta(y|\mathcal{T}(g))] \\ & + \mathbb{E}_{(g,c,y) \sim D^{cp.}} [-\log p_\theta(y|\mathcal{T}(g,c))] + \mathbb{E}_{(g,c,y,y') \sim D^{cpr.}} [-\log p_\theta(y'|\mathcal{T}(g,c,y))] \end{aligned} \quad (2)$$

We name this student PLASMA-Mul.

## 2.3 PLASMA+: Advancing Student with Verifier-guided Decoding

During inference, the student may generate logically and/or temporally ill-formed sequence of steps  $y = \{s_t\}_{t=1}^T$  as it is only trained to maximize the next-token probability. For example, in Figure 2, it may generate “write a check” at step 3 with relatively high confidence due to a spurious correlation between “sales price” and “check”. We mitigate this issue via step-wise guided decoding. Rather than generating plans greedily, we instead generate step-by-step by sampling several candidate next steps and searching for those with a high log-probability under both the distilled student and a verifier. The verifier is tasked to check for sequential ordering and semantic completeness. In an embodied setting, the verifier could be taken over by any affordance or safety module [1] that determines the executability of an action in a given environment.

**Step Verifier.** We introduce an independent verifier, which is trained to check the validity of plan steps and encourage PLASMA to produce more temporally and causally valid plans. The verifier takes as input a goal, the plan-so-far and a candidate next step and outputs a continuous validity score  $p_{\text{verifier}}(s_t|g, s_{<t}) \in [0, 1]$ .

We implement the verifier by fine-tuning a RoBERTa model [18] to classify whether a candidate step is valid or invalid. For training data, we use steps from available human-written plans<sup>3</sup> as positive

<sup>3</sup>Note that only a small-scale set of ground-truth plans is needed to train a verifier.

examples (valid steps). However, since no negative examples are readily available, we automatically create a set of invalid steps as pseudo-negative examples. Inspired by the common errors made by models, we design perturbations over ground-truth plans to target sequential ordering, semantic completeness, topicality, and fluency. See Appendix B.3 for details.

**Verifier-guided Step-wise Beam Search.** We illustrate our *verifier-guided decoding* in Figure 2. The procedure generates a plan  $\mathbf{y} = (s_1, \dots, s_T)$  by sequentially sampling and pruning the next step candidate  $s_t$ . Concretely, at each iteration<sup>4</sup>, it selects and expands a size- $K$  beam of plan-so-far,  $Y_{t-1} = \{s_{\leq t}^k\}_{k=1}^K$ , and generates  $N$  next-step candidates,

$$Y_t = \cup_{s_{\leq t} \in Y_{t-1}} \{(s_{\leq t} || s_t^n) \mid s_t^n \sim q(\cdot | \mathcal{T}(x, s_{\leq t})\}_{n=1}^N \quad (3)$$

where  $||$  is concatenation,  $x$  is a task-specific input, and  $q$  is a decoding algorithm. We encourage exploration at each step, by generating candidates using multiple decoding methods such as beam search, and nucleus sampling with temperature 1.0.

To select the top- $K$  scoring next-step candidates  $S_t^*$ , we use a value function  $v(s_{\leq t}) \rightarrow \mathbb{R}$  which returns the weighted sum of normalized sequence log-likelihood from the student model and the verifier validity score,

$$S_t^* = \arg \text{top-K}_{s_{\leq t} \in Y_t} v(s_{\leq t}) \quad (4)$$

$$v(s_{\leq t}) = \alpha \log p_\theta(s_{\leq t}) + (1 - \alpha) \log p_{\text{verifier}}(s_t | g, s_{\leq t}) \quad (5)$$

with  $\alpha$  controlling the impact of the distilled student and the verifier. The search ends when the beam contains  $K$  completed plans. We return the highest-scored plan as the final output. Our step-wise beam search strategy maintains a diverse set of candidate plans during the decoding process, allowing the model to explore multiple plausible paths before converging on a most promising one.

### 3 Experiments

**Implementation Details.** While any model with few-shot capabilities could be used, we choose our teacher model  $\mathcal{M}$  to be GPT-3 `text-curie-001` [4] for collecting the goals and initial plans, and GPT-3 `text-davinci-003` for collecting conditions and counterfactual plans.<sup>5</sup> We sample data points from GPT-3 using nucleus sampling ( $p = 0.98$ ) and temperature of  $T = 0.9$ . For our student models, we try a range of model sizes in T5 family [33], such as T5-large, T5-3B, and T5-11B. Student models are trained using Huggingface Transformers [42]. Main experiments can be done on 2 GPUs with 48GB of memory.

During inference, we use a beam of size  $K = 5$  for regular beam search, and  $N = 10$  (next-step candidates), beam  $K = 5$  and  $p = 0.9$  for our verifier-guided step-wise decoding (see §2.3).

**Baselines.** For each task, we compare our distilled students with their corresponding teacher, zero-shot and few-shot variants of GPT-3 [4], CoCoGEN [23] and human performance (when available). CoCoGEN frames the planning task as a code generation task and use a pre-trained code LM (`code-davinci-002`) in a few-shot setting.

Next, we present the experimental setup for each task, along with their results.

#### 3.1 Goal-based Planning

In this section, we aim to study two key research questions through our experiments. Firstly, we seek to investigate the extent to which scale impacts the distillation of procedural knowledge. Secondly, we aim to examine whether the scale gap can be bridged through the use of multitasking and/or a novel decoding algorithm. In essence, we seek to determine whether small language models can perform procedural planning tasks with the same level of proficiency as large language models.

**Evaluation Set.** For the original planning task, we use human-written plans from the test set of ProScript [34] dataset as our evaluation data.

---

<sup>4</sup>Iteration refers to a full step in a plan.

<sup>5</sup>In our preliminary experiment, we found `text-davinci-003` (the strongest GPT-3 version at the time) to be helpful for the more challenging counterfactual data collection.

**Setup.** We compare several student models of varying scales (770M-11B) with the teacher model, `text-curie-001`, and extremely large scale models (175B). For all student models, we decode using both regular beam search (PLASMA) and our verifier-guided step-wise beam search (PLASMA+).

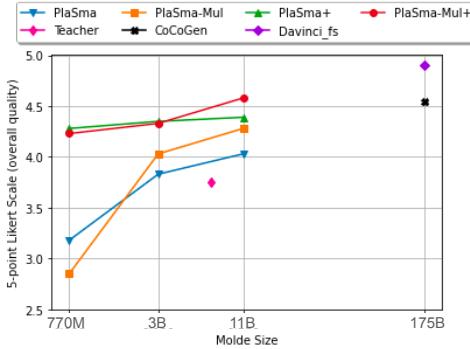


Figure 3: Visualization of bridging the scale gap in goal-based planning task. Smaller models are able to achieve comparable performance and sometimes surpass larger models via multi-task distillation and step-wise guided decoding.

**Does scale matter?** Larger models perform relatively better across all aspects.

**Does multi-task distillation help bridge the scale gap?** As we observe, multi-task distillation almost always wins over its task-specific counterpart with the exception of the smallest student, PLASMA (770M). We posit that very small student models might not have enough capacity to leverage the related tasks efficiently during multi-tasking.

**Does verifier-guided decoding help bridge the scale gap?** Pairing models with our proposed verifier-guided step-wise decoding substantially improves performance across students of varying sizes over all aspects. Specifically, compared with regular beam search, our proposed decoding results in 7%-48% relative improvements in overall quality across different student sizes. The improvements achieved by the verifier-guided decoding is larger for smaller students. We showcase the comparisons with qualitative examples in Appendix Table 8.

The best distilled students with 770M, 3B, and 11B parameters achieved respectively 14.13%, 16%, and 22.59% relative improvements over their teacher model (`text-curie-001`). Finally, our best distilled model (11B PLASMA-Mul+) performs equally well as human and is competitive with orders-of-magnitude larger models (175B).<sup>6</sup> Figure 3 visualizes how we bridge the scale gap using our multi-task distillation and verifier-guided step-wise decoding.

**Metrics.** Since there may exist many equally valid plans to a goal, we conduct human evaluations for the main results and report automatic metrics such as BLEU [25], ROUGE [16] and BERTScore [47] in Appendix Table 7.

We ask human annotators on the Amazon Mechanical Turk (AMT) platform to rate the generated plans for 250 randomly sampled goals on three aspects: 1) Order: how well-ordered the plan is (captures sequential correctness), 2) Completeness: how well the plan covers the necessary steps to accomplish the goal (captures semantic completeness), and 3) Overall quality: overall quality and correctness of the plan. Details of the human evaluation can be found in Appendix D.2 Figure 9.

Table 1 and Figure 3 summarize the human evaluation results for the original planning task.

| Model_size      |               | Coverage    | Order       | Overall Quality |
|-----------------|---------------|-------------|-------------|-----------------|
| Distilled 770M  | PLASMA        | 3.18        | 3.64        | 3.17            |
|                 | PLASMA+       | 4.25        | 4.55        | 4.28            |
|                 | PLASMA-Mul    | 2.84        | 3.36        | 2.85            |
|                 | PLASMA-Mul+   | 4.16        | 4.48        | 4.23            |
| Distilled 3B    | PLASMA        | 3.78        | 4.07        | 3.83            |
|                 | PLASMA+       | 4.38        | 4.60        | 4.35            |
|                 | PLASMA-Mul    | 3.96        | 4.35        | 4.03            |
|                 | PLASMA-Mul+   | 4.29        | 4.62        | 4.33            |
| Distilled 11B   | PLASMA        | 4.01        | 4.33        | 4.03            |
|                 | PLASMA+       | 4.33        | 4.60        | 4.39            |
|                 | PLASMA-Mul    | 4.24        | 4.59        | 4.28            |
|                 | PLASMA-Mul+   | 4.53        | 4.77        | 4.58            |
| Curie (Teacher) | few-shot (5)  | 3.75        | 4.27        | 3.75            |
| Davinci (175B)  | zero-shot     | 4.83        | 4.87        | 4.84            |
|                 | few-shot (5)  | <b>4.88</b> | <b>4.90</b> | <b>4.90</b>     |
| CoCoGEN (175B)  | few-shot (16) | 4.48        | 4.70        | 4.55            |
| Human           |               | 4.56        | 4.61        | 4.57            |

Table 1: Averaged 5-point Likert scale human evaluation for the original planning task. Small students paired with our decoding algorithm consistently outperform their teacher model (`text-curie-001`) and are competitive with order of magnitude larger models in zero/few-shot settings. \*CoCoGen [23] is a 16-shot baseline using code LLM.

<sup>6</sup>Pairwise annotator agreements (i.e., how often do two annotators agree on the answer) are 0.78, 0.84, and 0.80 for coverage, order and overall quality, respectively.

**Effect of symbolic distillation.** In this experiment, we compare models trained/tested on human-written pairs of (goal, plan) from ProScript dataset [34], our model-generated dataset CoPLAN, and the mix of both.

Models are initialized with T5-11B. We generate plans using our proposed verifier-guided decoding for randomly sampled 50 and 150 goals from ProScript and CoPLAN, respectively. We use the same human evaluation setup as before. Table 2 shows that training on our LLM-generated CoPLAN dataset, consistently transfers better to human-written dataset, ProScript. Training on the mix of both datasets, however, achieves the best performance. Intuitively, we observe that models are in general better at tackling LLM-generated data.

| Test on → | ProScript   |             |             | CoPLAN          |             |             |                 |
|-----------|-------------|-------------|-------------|-----------------|-------------|-------------|-----------------|
|           | Train on ↓  | Coverage    | Order       | Overall Quality | Coverage    | Order       | Overall Quality |
| ProScript | 4.38        | 4.54        | 4.35        | 4.51            | 4.81        | 4.58        |                 |
| CoPLAN    | 4.55        | 4.74        | 4.63        | 4.72            | 4.86        | 4.73        |                 |
| Mix       | <b>4.77</b> | <b>4.88</b> | <b>4.65</b> | <b>4.77</b>     | <b>4.88</b> | <b>4.78</b> |                 |

Table 2: Effect of symbolic knowledge distillation. The model trained on our CoPLAN dataset transfers better to other dataset, ProScript.

### 3.2 Counterfactual Planning and Revision

Here, we seek to benchmark language models’ planning abilities under constrained (contextually grounded) situations. This task goes beyond the original planning task, requiring models to produce novel linguistic alternatives to unseen situations.

**Evaluation Set.** To create an evaluation set, we generate conditions and counterfactual plans for the test set of ProScript following Step 1 in §2.1. We then only use human-verified tuples of (goal, plan, condition, counterfactual plan) as our test set for counterfactual planning and revision tasks.

**Setup.** We compare 3B and 11B student models with GPT-3 Curie and the 175B teacher model, text-davinci-003 in zero/few-shot settings. During inference, we use our proposed verifier-guided step-wise beam search with  $\alpha = 0.75$  to outweigh student model’s probability over the verifier validity score.<sup>7</sup>

**Metric.** We conduct human evaluation on the AMT platform. We generate (counterfactual) plans for 300 randomly sampled examples using each model. We ask 3 human annotators to rate each generated plan based on whether it contains the necessary steps to make the goal achievable *while satisfying the condition*. We provide 3 options for the annotators to pick from: **A**: The plan contains all the necessary steps to meet the requirements of the condition on the goal, **B**: The plan addresses the condition, but it is trivial and lacks thoughtfulness<sup>8</sup>, and **C**: The plan does NOT address the condition or does so very poorly. We take the majority vote for the final results. Details on crowd-sourcing human evaluation can be found in Appendix Figure 11.

**Results.** Figure 4 depicts the results. Large students perform better on both tasks. In counterfactual planning, our 11B PLASMA-Mul+ demonstrates a 93.33% success rate in producing high-quality plans while adhering to the given condition, which is comparable to the performance of the 175B parameter Davinci model in a zero-shot setting. Furthermore, our model generates slightly fewer low-quality plans, only 7 as opposed to 12 by Davinci. While multi-tasking seems to be helpful in (counterfactual) planning, this is not always the case for counterfactual revision. We hypothesize that the reason for this could be that the original and counterfactual planning tasks, which do not involve modifying an existing plan, may negatively impact the revision task. The best performance for the counterfactual plan revision is achieved by Davinci (90%) followed by PLASMA+ (86.33%).<sup>9</sup> We also collect additional feedback from annotators on the errors made by models. Results are reported in Appendix Table 11, showing “missing necessary steps” is the most prevalent mistakes.

We provide qualitative examples of model generations across all three tasks in Table 4. More examples of (good and bad) generations according to human annotators are provided in Appendix Tables 9, 10.

<sup>7</sup>We performed a hyperparameter search over  $\alpha = \{0.5, 0.75, 0.8\}$ .

<sup>8</sup>An example of trivial modification is addressing the condition “you have no money” with adding an step “find money” in the plan.

<sup>9</sup>Pairwise annotator agreements are 0.96 and 0.94 for counterfactual planning and revision, respectively.

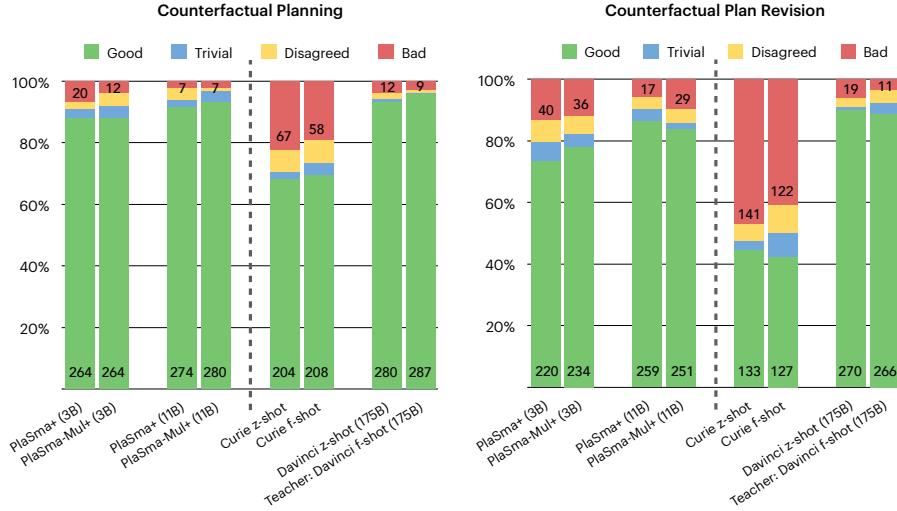


Figure 4: Human evaluation results of 300 generations for counterfactual planning and revision tasks. Left: in counterfactual planning, our best student model PLASMA-Mul+ (11B) with  $\times 16$  fewer parameters is on par with GPT-3 Davinci model. Right: in counterfactual revision, our best student model PLASMA+ (11B) is able to generate good counterfactual plans 86.33% of the time.

### 3.3 Application to Embodied Agents

An important application enabled by PLASMA is that of enabling an agent to plan according to a given high-level goal. We evaluate PLASMA on the task of planning in the VirtualHome [29] environment. In this environment, agents can perform household activities, e.g. “paint ceiling”, through programs, in the form of supported actions (42 in total) and arguments. For evaluation, we use their test set consisting of 88 goals (and corresponding gold programs).

We compare our best student PLASMA-Mul+ (11B) with Planner [11], a 1-shot GPT-3 (175B) model with several inference-time strategies to ensure executability in embodied environments. We follow their procedure to translate generated steps from natural language to steps executable in the environment. To apply our model to VirtualHome, we finetune PLASMA-Mul on  $\sim 4K$  human labeled examples and also finetune the step verifier on the same data using the method described in Section 2.3. We show, in Table 3, that our model generates steps that are significantly more executable (according to automatic metric) and also more complete (according to human judges). More experimental details can be found in Appendix E.

| model                           | Executability | LCS          | Correctness  |
|---------------------------------|---------------|--------------|--------------|
|                                 | (%)           | (%)          | (%)          |
| Planner (175B) [11]             | 77.17         | 19.10        | 18.33        |
| PLASMA-Mul <sup>FT</sup> (11B)  | 76.38         | 28.36        | 41.38        |
| PLASMA-Mul+ <sup>FT</sup> (11B) | <b>94.18</b>  | <b>31.93</b> | <b>43.68</b> |
| Human                           | 100           | N/A          | 66.66        |

Table 3: Human-evaluated correctness along with (automatic) executability and LCS scores on VirtualHome environment [29]. Steps generated by our model is more executable and correct for accomplishing the task.

## 4 Related Works

**Procedural Planning** The problem of planning to accomplish a goal via sub-steps is widely studied in two contexts. One is script knowledge generation, which is a long-standing NLP problem [36]. Collecting script knowledge requires either human annotation [39], unsupervised feature-based extraction [5], or, more recently, methods that utilize task-specific fine-tuned LLMs [34] and pipeline-based approaches [35]. In addition, there is a line of procedural planning that involves planning with executable actions that can be executed by robots in real-life environments [11, 1, 43, 12]. Recent approaches view planning as a conditional text generation problem and use LLMs in a zero/few-shot prompting mode to tackle the task [23, 11, 1, 22]. Despite showing strong performance, their success

| Goal                                      | Condition                               | Initial Plan  | Generated (Counterfactual) Plan  |
|---|---|---|--|
| Empty lint filter                         |   |   | Find lint filter > Remove lint filter > Clean lint filter > Replace lint filter  |
|   | If the lint trap is too hot to touch    |   | Unplug dryer > Open dryer > Locate lint trap > Use heat-resistant glove or mitt to pick up lint trap > Pull lint trap out of dryer > Empty lint trap   |
|   |   | Load clothes into dryer > Locate lint trap > Move hand towards lint trap > Grip lint trap with fingers > Pull lint trap out of dryer > Use other hand to lift lint from lint trap > Empty lint filter                                     | Load clothes into dryer > Locate lint trap > Use gloved hand to move hand toward lint trap > Grip lint trap with fingers > Use other hand to lift lint from lint trap > Remove lint trap from dryer  |
| Eat some ice cream from an ice cream shop |   |   | Find ice cream shop > Enter ice cream shop > Find ice cream > Choose ice cream > Eat ice cream   |
|   | If you have allergies to dairy products |   | Decide to eat some ice cream from an ice cream shop > Find an ice cream shop > Walk into ice cream shop > Ask for a dairy-free ice cream > Pick out ice cream > Pay for ice cream > Eat some ice cream from an ice cream shop  |
|   |   | Decided to eat some ice cream from an ice cream shop > Decide to eat ice cream > Gather the family > Drive to ice cream parlor > Get family out and lock car > Walk in ice cream shop and sit > Eat some ice cream from an ice cream shop | Decide to eat some ice cream from an ice cream shop > Research ice cream shops in your area that offer dairy-free options > Gather the family > Drive to ice cream parlor > Get family out and lock car > Walk in ice cream shop and sit > Ask the staff about their dairy-free options > Select a dairy-free option > Eat some ice cream from an ice cream shop |

Table 4: PLASMA generations for (*counterfactual planning*) and *revision* tasks.

heavily relies on scale. However, in this paper, we seek to achieve comparable performance while using more parameter-efficient and accessible models.

**Symbolic Knowledge Distillation** Crowd-sourcing human-written datasets at scale is both challenging and costly. Therefore, there has been a growing interest in using LLM-generated data to train smaller models. This approach which falls under the conceptual framework of symbolic knowledge distillation [41] has been applied to simpler classification tasks [37], reasoning [38, 10, 46, 7], as well as commonsense and general knowledge base construction [41, 3]. This approach not only achieves promising performance on smaller models but is also cost-efficient compared to pre-training smaller models from scratch [13]. In a concurrent work, Yuan et al. [45] proposed a similar approach to distill script knowledge from LLMs for constrained planning task. However, unlike our “conditions” which can take free-form format, their constraints are limited to specific types by extending an original goal with a modifier, intent or method.

**Decoding-time Algorithm** Decoding-time algorithm is an emerging approach for adapting language models’ output for task-specific characteristics. Works in this line often focus on incorporating explicit lexical constraints at inference time so that the model is bounded with certain generation words [20, 19, 9, 26]. In addition to discrete lexical constraints, applying continuous optimization functions such as KL loss has also been found to be effective [30, 31, 15, 8]. Perhaps our approach is most similar to function-guided decoding methods. Krause et al. [14] and Yang et al. [44] fuse next-token probability with desired attributes’ probabilities at inference using a discriminator model. These and related token-level beam search variants assume access to per-token logits and gradient updates. Our decoding method however only relies on model log-probabilities and a verifier to facilitate semantic and temporal constraints at a step level.

## 5 Conclusions and Future Work

In this paper, we focus on procedural planning, a challenging task that involves decomposing high-level goals into ordered steps. We introduce PLASMA as an effective approach that uses smaller and more accessible models. By leveraging symbolic procedural knowledge distillation and an inference-time algorithm, we have endowed smaller models with enhanced procedural knowledge and planning capabilities. Furthermore, we introduced the task of Counterfactual Planning, which involves generating/revising plans to accommodate realistic counterfactual scenarios. Our results demonstrate that significantly smaller models can effectively compete with and often outperform

their larger teacher models in both original and counterfactual settings. We hope our work sheds light on new directions towards developing smaller yet powerful multi-modal models for (counterfactual) procedural planning and reasoning.

## 6 Acknowledgements

This work was funded in part by the DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI. We also thank the Beaker Team at the Allen Institute for AI for helping with the compute infrastructure and OpenAI for providing access to the GPT-3 API.

## References

- [1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.
- [2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.
- [3] Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. I2d2: Inductive knowledge distillation with neurologic and self-imitation, 2023.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [5] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June 2008. Association for Computational Linguistics.
- [6] Katherine M Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Josh Tenenbaum. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.
- [7] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *ArXiv*, abs/2212.10071, 2022.
- [8] Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. Towards decoding as continuous optimisation in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 146–156, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [9] Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics.

- [10] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023.
- [11] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR, 2022.
- [12] Peter Jansen. Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4412–4417, Online, November 2020. Association for Computational Linguistics.
- [13] Junmo Kang, Wei Xu, and Alan Ritter. Distill or annotate? cost-efficient fine-tuning of compact models, 2023.
- [14] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [15] Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. Controlled text generation as continuous optimization with multiple constraints. In *Neural Information Processing Systems*, 2021.
- [16] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
- [18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [19] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic a\*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, July 2022. Association for Computational Linguistics.
- [20] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online, June 2021. Association for Computational Linguistics.
- [21] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [22] Yujie Lu, Weixi Feng, Wanrong Zhu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Neuro-symbolic procedural planning with commonsense prompting. In *The Eleventh International Conference on Learning Representations*, 2023.

- [23] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [24] OpenAI. Openai api pricing. 2023. Accessed: 2023-05-15.
- [25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [26] Damian Pascual, Béni Egressy, Florian Bolli, and Roger Wattenhofer. Directed beam search: Plug-and-play lexically constrained language generation. *ArXiv*, abs/2012.15416, 2020.
- [27] Douglas Pearson and John Laird. Incremental learning of procedural planning knowledge in challenging environments. *Computational Intelligence*, 21:414–439, 11 2005.
- [28] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *NeurIPS*, 2021.
- [29] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018.
- [30] Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805, Online, November 2020. Association for Computational Linguistics.
- [31] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 2022.
- [32] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.
- [33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [34] Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. proScript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2138–2149, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [35] Abhilasha Sancheti and Rachel Rudinger. What do large language models learn about scripts? In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 1–11, Seattle, Washington, July 2022. Association for Computational Linguistics.
- [36] Roger C. Schank and Robert P. Abelson. Scripts, plans and knowledge. In *International Joint Conference on Artificial Intelligence*, 1975.
- [37] Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [38] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions, 2022.

- [39] Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. A crowdsourced database of event sequence descriptions for the acquisition of high-quality script knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3494–3501, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [40] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.
- [41] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics.
- [42] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
- [43] Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [44] Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.
- [45] Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. Distilling script knowledge from large language models for constrained language planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4303–4325, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [46] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [47] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

## Supplementary Material

### A COPLAN Analysis Details

#### A.1 Goal diversity

In this section, we demonstrate that the goals in our COPLAN dataset broadly cover a diverse set of everyday, real-world human activities.

For this analysis, we first define seven goal-relevant categories based on categories defined by the US Bureau of Labor Statistics<sup>10</sup>: (1) **career** and work related activities; (2) **education** and professional

<sup>10</sup><https://www.bls.gov/news.release/atus.t12.htm> defines 11 categories to cover common everyday civilian activities. We cluster these categories into five.

growth; (3) **financial** and commercial activities; (4) **fitness** and health; (5) **service** and civic activities; (6) **social** activities and relationships; and (7) **self-improvement** and leisure.

Next, using the seven categories, we manually annotate 200 most frequent verb unigrams, 300 most frequent noun unigrams, and 300 most frequent nominal (nouns + adjectives) bigrams extracted from the goals statement. Only when the unigram (e.g. “make”) or the bigram (e.g. “new word”) indicate one of the seven categories (e.g., “close friend” for relationship or “college university” for education) the instance is annotated with the category. Otherwise, it is annotated with an eight category, **other**. For each goal in COPLAN, each (verb, noun) unigram or (nominal) bigram casts a category as a vote if found in the annotated data. If not found, then it casts other as vote. Majority vote is taken as the category of the larger goal statement.

Figure 5 shows the distribution of the activity types in COPLAN. Education is the largest category (“join an online course to learn a new language”) followed by self-improvement (“develop my creative writing skills”). Service (“cooking meals for a homeless shelter”), career (“get interview for a new job”), and financial (“upgrade to a new car”) are the next largest categories. The other category includes miscellaneous activities like chores and events like “vacuum the livingroom floor”.

## A.2 Condition diversity

We assess the diversity of the conditions in CoPLAN by analyzing the verbal use and nominal trigrams employed in the statements.

We manually analyze 20 most frequent verbs and phrasal verbs (e.g., “have access”) appearing in the condition statements. The verbs are grouped into 5 semantic categories: (1) **want** (to want, to desire, etc); (2) **possess** (to have, to possess, etc); (3) **access** (to obtain, to get, to procure etc); (4) **able** (to be able to, be capable of, etc); and (5) **trust** (to be safe, to rely, etc). Note that each of these categories include conditions of both polarity; for example, for **possess**, it includes both the condition imposed by having (“have enough money”) and by lacking (“not have enough money”). A sixth category, **other**, was included for the verbs not included in the above categories. For each condition in COPLAN, the first trigram made up of verbs, adjectives, and nouns appearing after the main verb (e.g., “If you *want* to [apply to an online program]” → main verb: *want*, trigram: *apply online program*) were extracted. Trigrams were then associated with each of the 5 semantic categories based on the main verb.

Figure 6 shows the most frequent unique trigrams in each category. The graph includes the 20 most frequent trigrams for each category. The displayed trigrams were manually clustered when appropriate for readability purposes (e.g., “take course online” clustered with “take online course”).

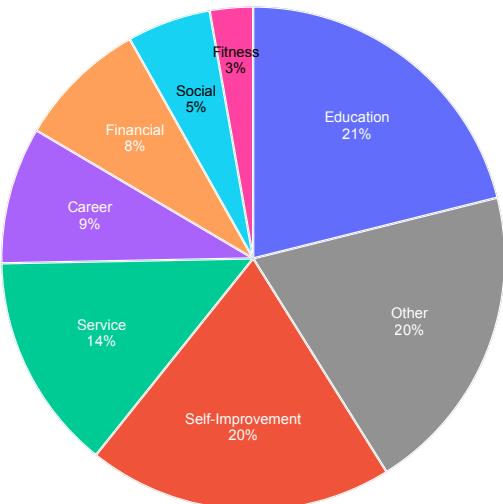


Figure 5: Goal diversity in CoPLAN

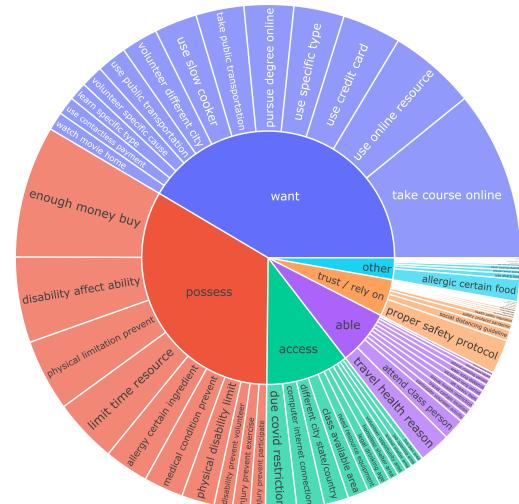


Figure 6: Condition diversity in CoPLAN

We find a wide variety of real-world constraints that pose varying levels of restriction such as preference and desire (“want to take an online course”) and hindrances posed by the state of having or not having something (“not having enough money” or “having a disability”).

## B Additional Experimental Details

### B.1 Critic Models: Collecting Human Annotations

We gather human annotations of *valid* vs. *invalid* teacher generations. Annotations are crowdsourced through the Amazon Mechanical Turk (AMT) platform. We qualify 263 best performing workers through a paid qualification round. Additionally, we chose annotators among those who were located in US, GB and CA, and had 98% approval rate for at least 10,000 previous annotations. Crowdworker compensation for qualification and annotation HITs is maintained at an average of \$15 per hour.

**Plans.** For plans, the crowdworkers were presented with randomly-sampled 13K generated (goal, plan) pairs, and were asked to evaluate the plans along three dimensions: *topicality*—the topic of the plan is relevant and appropriate for the goal, *ordering*—the steps in the plan are appropriately ordered, and *completeness*—the plan provides complete and informative steps to achieve the goal. We asked the workers to evaluate the goal’s *achievability* as a separate (fourth) dimension. Each dimension was rated on a 5-point likert scale with three *valid* labels (Definitely, Mostly, and Somewhat; numeric value 1) and two *invalid* labels (Hardly, Not at all; numeric value 0). Each (goal, plan) pairs were annotated by three crowdworkers. The template used is shown in Figure 9.

We determine the validity of a (goal, plan) pair in the following manner. We then calculate the mean score (over the three annotator responses) for each of the dimensions. A (goal, plan) pair is considered *valid* only if: (1) it receives a score greater than 0.25 for each of the *achievability*, *topicality*, or *ordering* dimensions, and (2) receives a scores greater or equal to 0.65 on the *completeness* dimension. Failing that, a pair is considered *invalid*.

**Conditions.** For conditions, we collect human judgements on whether the condition makes the goal more specific or harder to achieve (but not impossible) on a randomly-sampled set of 6100 generated tuples of (goal, plan, condition). We include screenshot of our annotation template in Figure 10.

**Counterfactual Plans.** And finally, for counterfactual plans, we collect 10.5K human judgements on whether the modified plan contain all the necessary steps to make the goal achievable while adhering to the condition. We include screenshot of our annotation template in Figure 11.

### B.2 Critic Models: Training Details

We train 3 binary classifiers (critics) for filtering out low quality teacher generations in §2.1 using pre-trained RoBERTa-Large [17]. We conduct a small grid search on validation loss for batch size  $bs = \{16, 32, 64\}$  and learning rate  $lr = \{1e-4, 1e-5, 1e-6, 5e-6\}$ . We report the effective hyper-parameters for each critic in Table 5. We use early stopping on validation loss.

### B.3 Training the Verifier

**Constructing Pseudo-negative Examples.** For training the step verifier, we use the human-written plans [34] to construct positive examples of (plan-so-far, next-step) pairs and devise three main perturbation strategies to automatically construct negative examples as explained below:

- **Reordered Steps:** Conflicting logical order results from inaccurate causal or temporal dependencies in a plan. Thus, we apply both *near* and *distant* reordering by randomly reordering two consecutive and two distant steps.
- **Repetitive Steps:** Degeneration i.e., generating repetitive text is commonly observed in language models. Similarly, we include both *near* and *distant* repetition by repeating the immediate previous step and distant previous step as a pseudo-negative next-step.

|                       | batch size | learning rate |
|-----------------------|------------|---------------|
| Plan Critic           | 16         | 1e-6          |
| Condition Critic      | 32         | 1e-5          |
| Counterfactual Critic | 32         | 1e-6          |

Table 5: Hyper-parameter values for training different critic models.

- **Missing Steps:** Another common mistake made by language models is missing necessary steps, leading to incoherent plans. To simulate this behaviour, we randomly select a non-immediate step as a pseudo-negative next-step.

We collect a training set of 47k positive and negative pairs of (plan-so-far, next-step) using only 3k human-written plans.

**Training Details.** We fine-tune RoBERTa Large [17] as a binary classifier identifying the validity of a candidate next-step. We train for 10 epochs with early stopping on validation accuracy using batch size of 32 and learning rate of  $1e - 5$ .

| Category                               | Goal                            | Condition   |
|--|---------------------------------|---|
| <b>Location</b>                        | Purchase gardening supplies     | there are no local gardening stores nearby        |
|  | Sing the lyrics                 | you want to sing the lyrics in a recording studio |
| <b>Equipment</b>                       | Studying for the exam           | you want to use a laptop or computer              |
|  | Practice pottery techniques     | you don't have the right tools or clay            |
| <b>Safety</b>                          | Take out several plates         | the plates are too heavy or fragile               |
|  | Transport materials home        | the car breaks down or runs out of gas            |
| <b>User's condition/ specification</b> | Practice playing the instrument | you are unable to read sheet music                |
|  | Rent rock climbing equipment    | you need size specific equipment                  |

Table 6: Examples for different categories of conditions in COPLAN dataset.

| model_size             |                | BLEU  | ROUGE-2 | ROUGE-L | BERT-f1 |
|------------------------|----------------|-------|---------|---------|---------|
| <b>Distilled 770M</b>  | PLASMA         | 12.97 | 14.02   | 28.23   | 84.31   |
|                        | PLASMA +       | 14.26 | 16.31   | 31.02   | 85.30   |
|                        | PLASMA-Mul     | 14.47 | 14.43   | 27.99   | 84.02   |
|                        | PLASMA-Mul+    | 14.49 | 16.70   | 31.49   | 85.35   |
| <b>Distilled 3B</b>    | PLASMA         | 12.89 | 14.39   | 28.57   | 84.70   |
|                        | PLASMA +       | 13.92 | 15.56   | 30.83   | 85.19   |
|                        | PLASMA-Mul     | 13.62 | 15.42   | 29.31   | 84.80   |
|                        | PLASMA-Mul+    | 14.96 | 16.80   | 31.97   | 85.28   |
| <b>Distilled 11B</b>   | PLASMA         | 12.64 | 13.93   | 28.14   | 84.56   |
|                        | PLASMA +       | 14.65 | 15.84   | 31.04   | 85.33   |
|                        | PLASMA-Mul     | 13.61 | 15.67   | 29.99   | 85.10   |
|                        | PLASMA-Mul+    | 15.54 | 16.76   | 31.98   | 85.37   |
| <b>Curie (Teacher)</b> | few-shot (3-5) | 7.13  | 9.24    | 22.78   | 83.08   |
| <b>Davinci (175B)</b>  | zero-shot      | 4.98  | 7.81    | 21.38   | 81.20   |
|                        | few-shot (3-5) | 10.27 | 10.27   | 24.57   | 83.48   |
| <b>CoCoGen (175B)</b>  | few-shot (16)  | 18.07 | 21.01   | 33.99   | 86.41   |

Table 7: Automatic Evaluation Results for the Planning task. Note that the human-written plans are not the only possible plans, hence these automatic metrics may not provide an informative measure of performance.

## C Out-of-domain Evaluation

[6] proposed two out-of-distribution reasoning one of which involved constrained planning. For a given goal and one or more conditions, the task is to generate a plan. We evaluate PLASMA on the 28 constrained goals provided by the paper. We compare our generations to the GPT-generated plans provided by the paper and `text-davinci-002` prompted in a zero shot manner. To evaluate the generations we perform a human evaluation, as described in §D.2.

| Model            | % good    |
|------------------|-----------|
| PLASMA           | <b>71</b> |
| GPT-3 (from [6]) | 36        |
| GPT-3 zero shot  | 64        |

Table 12: Percent of generated counterfactual plans which have been rated as *good* by annotators.

| Model  | Plan   |
|--|--|
| <b>Goal: put coffee in coffee machine</b>      |  |
| Curie (Teacher)                                | Step 1: find coffee machine Step 2: unlock coffee machine Step 3: put coffee in coffee machine Step 4: close coffee machine Step 5: wait a few minutes Step 6: coffee should be ready  |
| PLASMA-Mul (11B)                               | Step 1: get coffee Step 2: put coffee in coffee machine Step 3: turn on coffee machine Step 4: wait for coffee to brew Step 5: drink coffee  |
| PLASMA-Mul+ (11B)                              | Step 1: buy coffee beans Step 2: grind the coffee beans Step 3: put the ground coffee into the coffee filter Step 4: put the filter into the machine Step 5: turn the machine on Step 6: wait for the coffee to brew Step 7: enjoy your coffee           |
| <b>Goal: put on baseball jersey</b>            |  |
| Curie (Teacher)                                | Step 1: get baseball jersey Step 2: put baseball jersey on Step 3: take baseball jersey off  |
| PLASMA (11B)                                   | Step 1: find a baseball jersey Step 2: put on baseball jersey Step 3: take off baseball jersey Step 4: put on another baseball jersey Step 5: take off second baseball jersey Step 6: put on third baseball jersey Step 7: put on fourth baseball jersey |
| PLASMA+ (11B)                                  | Step 1: walk to closet Step 2: open closet Step 3: look for baseball jersey Step 4: find baseball jersey Step 5: take baseball jersey out of closet Step 6: put on baseball jersey   |
| <b>Goal: sit at the table with dinner date</b> |  |
| Curie (Teacher)                                | Step 1: text or call the person to let them know that you will be late Step 2: arrive at the restaurant on time Step 3: take a seat at the table Step 4: wait for the other person to arrive Step 5: eat dinner Step 6: say goodbye and thank you        |
| PLASMA (3B)                                    | Step 1: find dinner date Step 2: go to dinner date's house Step 3: sit at the table Step 4: talk to dinner date Step 5: eat dinner Step 6: leave dinner date's house   |
| PLASMA+ (3B)                                   | Step 1: choose a restaurant Step 2: make a reservation Step 3: drive to the restaurant Step 4: park the car Step 5: enter the restaurant Step 6: find a seat Step 7: sit at the table with dinner date   |
| <b>Goal: brush all teeth</b>                   |  |
| Curie (Teacher)                                | Step 1: get in the car Step 2: drive to the dental office Step 3: park the car Step 4: go inside the office Step 5: pay the bill Step 6: get a receipt Step 7: brush all teeth Step 8: leave office  |
| PLASMA (770M)                                  | Step 1: Get toothbrush Step 2: Get toothpaste Step 3: Get toothbrush Step 4: Get toothbrush Step 5: Get toothbrush Step 6: Get toothbrush Step 7: Brush all teeth  |
| PLASMA+ (770M)                                 | Step 1: Get toothbrush Step 2: Get toothpaste Step 3: Put toothpaste on toothbrush Step 4: Grab toothbrush Step 5: Hold toothbrush up to mouth Step 6: Brush all teeth   |

Table 8: Qualitative examples generated by different models for the original planning task.

The human evaluation results in Table 12 show that PLASMA outperforms the other baselines in this out-of-domain subset of counterfactual planning task.

## D Evaluation Details

### D.1 Automatic Evaluation

We report automatic evaluation of models for the original planning task in Table 7. Note that the human-written plans are not the only possible plans, hence these automatic metrics may not provide an informative measure of performance.

### D.2 Human Evaluation on AMT

All human evaluations were conducted on the Amazon Mechanical Turk (AMT). We sourced our annotators from the same pool of qualified workers (see B.1). Throughout the entirety project, we maintained an average of \$15/hour pay rate based on our estimation of time needed to complete the task. Each examples were annotated by 3 workers and majority vote was taken for the reported

| Type | Conditional Plan   |
|------|--|
| good | <p><b>Goal:</b><br/> <i>eat ice cream if you have an ice cream maker</i></p> <p>Step 1: Decide to eat ice cream<br/>     Step 2: Locate the ice cream maker<br/>     Step 3: Turn on the ice cream maker<br/>     Step 4: Add ingredients to the ice cream maker<br/>     Step 5: Follow the instructions to churn the ice cream<br/>     Step 6: Remove the ice cream from the ice cream maker<br/>     Step 7: Put the ice cream in a bowl<br/>     Step 8: Eat ice cream</p>                |
| good | <p><b>Goal:</b><br/> <i>brush all teeth if you are out of toothpaste</i></p> <p>Step 1: Decided to brush all teeth<br/>     Step 2: Locate a store that sells toothpaste<br/>     Step 3: Purchase toothpaste<br/>     Step 4: Drive to the store<br/>     Step 5: Purchase toothpaste<br/>     Step 6: Drive back home<br/>     Step 7: Unpack the toothpaste<br/>     Step 8: Open the toothpaste<br/>     Step 9: Brush all teeth</p>   |
| bad  | <p><b>Goal:</b><br/> <i>write names on the cards if you have a shaky hand</i></p> <p>Step 1: find a pen or pencil with a large tip<br/>     Step 2: hold the pen or pencil in one hand<br/>     Step 3: hold the pen or pencil in the other hand<br/>     Step 4: use a pen or pencil to write the names on the cards<br/>     Step 5: once the names are on the cards, place them in a drawer or file<br/>     Step 6: close the drawer or file<br/>     Step 7: write names on the cards</p> |
| bad  | <p><b>Goal:</b><br/> <i>walk into town if you have a disability that limits your mobility</i></p> <p>Step 1: Decide to walk into town<br/>     Step 2: Put on shoes<br/>     Step 3: Get a cane or other mobility aid if necessary<br/>     Step 4: Put the cane or other mobility aid in a backpack<br/>     Step 5: Put the backpack on<br/>     Step 6: Leave the house<br/>     Step 7: Walk to the bus stop<br/>     Step 8: Take the bus into town<br/>     Step 9: Walk into town</p>   |

Table 9: Qualitative examples of the *counterfactual planning* task.

| Type | Original Plan   | Counterfactual Plan   |
|------|---|---|
| good | <p><b>Goal:</b> empty lint filter</p> <p>Step 1: load clothes into dryer</p> <p>Step 2: locate lint trap</p> <p>Step 3: move hand towards lint trap</p> <p>Step 4: grip lint trap with fingers</p> <p>Step 5: pull lint trap out of dryer</p> <p>Step 6: use other hand to lift lint from lint trap</p> <p>Step 7: empty lint filter</p>  | <p><b>Goal:</b> empty lint filter <i>if the lint trap is too hot to touch</i></p> <p>Step 1: Load clothes into dryer</p> <p>Step 2: Locate lint trap</p> <p>Step 3: Move hand towards lint trap</p> <p>Step 4: Use a heat-resistant glove to grip lint trap with fingers</p> <p>Step 5: Pull lint trap out of dryer</p> <p>Step 6: Use other hand to lift lint from lint trap</p> <p>Step 7: Empty lint filter</p>  |
| good | <p><b>Goal:</b> take a shower after work</p> <p>Step 1: decided to take a shower after work</p> <p>Step 2: drive home from work</p> <p>Step 3: go to bath room</p> <p>Step 4: take off clothes</p> <p>Step 5: start water to get heat</p> <p>Step 6: take hot shower</p> <p>Step 7: take a shower after work</p>  | <p><b>Goal:</b> take a shower after work <i>if you want to take a shower at the gym</i></p> <p>Step 1: Decide to take a shower after work</p> <p>Step 2: Drive to gym</p> <p>Step 3: Go to locker room</p> <p>Step 4: Remove clothes</p> <p>Step 5: Enter shower room</p> <p>Step 6: Start water to get heat</p> <p>Step 7: Take hot shower</p> <p>Step 8: Take a shower after work</p>   |
| bad  | <p><b>Goal:</b> empty old left overs from the fridge</p> <p>Step 1: go through the house</p> <p>Step 2: walk into the kitchen</p> <p>Step 3: locate the refridgerator</p> <p>Step 4: walk to the refridgerator</p> <p>Step 5: grab the door handle</p> <p>Step 6: open the refrigerator door</p> <p>Step 7: empty old left overs from the fridge</p>  | <p><b>Goal:</b> empty old left overs from the fridge <i>if you have difficulty bending down</i></p> <p>Step 1: go through the house</p> <p>Step 2: walk into the kitchen</p> <p>Step 3: locate the refridgerator</p> <p>Step 4: walk to the refridgerator</p> <p>Step 5: grab a chair or stool</p> <p>Step 6: place the chair or stool in front of the refrigerator</p> <p>Step 7: use the chair or stool to open the refrigerator door</p> <p>Step 8: empty old left overs from the fridge</p> |
| bad  | <p><b>Goal:</b> buy a new watch</p> <p>Step 1: decided to buy a new watch</p> <p>Step 2: search the internet for watch retailers</p> <p>Step 3: pick a reliable retailer</p> <p>Step 4: search the retailer site for watches</p> <p>Step 5: add watch to cart</p> <p>Step 6: click check out</p> <p>Step 7: add payment information</p> <p>Step 8: add address information</p> <p>Step 9: buy a new watch</p> | <p><b>Goal:</b> buy a new watch <i>if your payment information is compromised</i></p> <p>Step 1: decide to buy a new watch</p> <p>Step 2: search the internet for watch retailers</p> <p>Step 3: pick a reliable retailer</p> <p>Step 4: search the retailer site for watches</p> <p>Step 5: add watch to cart</p> <p>Step 6: click check out</p> <p>Step 7: add payment information</p> <p>Step 8: verify payment information</p> <p>Step 9: buy a new watch</p>                               |

Table 10: Qualitative examples of the *counterfactual plan revision task*.

| Error Type            | Counterfactual Planning |         |             | Counterfactual Revision |         |             |
|-----------------------|-------------------------|---------|-------------|-------------------------|---------|-------------|
|                       | Edits                   | Missing | Unnecessary | Edits                   | Missing | Unnecessary |
|                       | Required                | steps   | steps       | Required                | steps   | steps       |
| Plasma+ (3B)          | 4.66                    | 8.33    | 3.66        | 13.33                   | 19.33   | 6.00        |
| Plasma-Mul+ (3B)      | 4.33                    | 7.66    | 3.66        | 10.66                   | 14.66   | 4.33        |
| Plasma+ (11B)         | 3.66                    | 5.00    | 3.33        | 4.66                    | 10.00   | 3.33        |
| Plasma-Mul+ (11B)     | 3.00                    | 3.33    | 3.66        | 6.00                    | 11.66   | 4.66        |
| curie-001 zero-shot   | 7.00                    | 27.00   | 6.66        | 26.00                   | 49.33   | 13.66       |
| curie-001 few-shot    | 6.00                    | 25.33   | 5.00        | 30.00                   | 48.00   | 13.33       |
| davinci-003 zero-shot | 1.33                    | 6.33    | 0.66        | 5.33                    | 7.33    | 2.66        |
| davinci-003 few-shot  | 1.33                    | 3.00    | 0.66        | 4.33                    | 8.66    | 2.66        |

Table 11: Percent of generated (counterfactual) plans with each error type. “Missing Steps” is the most common error types across all models.

results. The layout templates for evaluating plans and counterfactual plans are shown in Figures 9 and 11, respectively.

## E Experimental Details of VirtualHome Evaluation

We follow the same experimental setup and metrics for evaluation as Planner [11]. The test set consists of 88 high-level goals. To translate a generated natural language step into an executable step, we follow [11] and find an executable action closest in embedding space to the generated step. To compute these embeddings, we use the `stsbert-large` model. Executability and LCS are computed identical to [11]. Due to challenges with reproducibility of GPT-3 outputs, evaluation results of GPT-3 do not exactly match between our works.

## F Additional Checklist Support

### F.1 IRB and Annotation Ethics

We obtained IRB exemption for our data collection and evaluation from our institution’s internal review board. In full compliance to the exemption clauses as published in the code of federal regulations (45 CFR 46.104(d)(2,3)), we did not collect any deanonymizing information, and we do not publish our dataset with worker specific information such as the MTurk’s worker id. Based on our exempted status, according to our internal regulations, does not require for us to use consent forms with our crowdsourcing.

Additionally, our data collection and evaluation efforts only involve human judgments about world knowledge relating to general real-world goals and plans. We have no reason to believe that our crowdsourcing posed harm or discomfort beyond the minimal risk as defined by 45 CFR 46.102(i).

### F.2 Limitations

One potential limitation of our work is that the verbalization component of our framework involves open text generation from large-scale language models (GPTs). Works such as Bender et al. [2] have argued that generations from LLMs can be prone to harmful biases stemming from the massive language data they are trained on. In the process of constructing the dataset, we have not directly observed levels of biases to cause us alarm. We believe harmful and discriminatory generations are largely mitigated by the very nature of the goals and scripts we obtain: our data is primarily composed of low-level everyday situations such as education, self-care, and mundane chores like vacuuming the floor or cooking a meal (see §A.1,A.2). This said, we acknowledge that prejudices like gender roles, for example, do also surface in the most mundane scenarios.

A related limitation is that LLMs have been trained on primarily English pretraining data, likely sourced from texts that reflect North American or European culture or norms. Consequently, we note that the goals in COPLAN may reflect the goals that are most culturally expected or appropriate to the cultures of English-speaking countries. This is also expected of the plans that may include culturally limited processes and procedures. This should be a consideration that any follow-up studies using our data and model should attend to. Extending our study to include more socio-culturally inclusive goals and plans is a compelling direction for our future research.

### **F.3 Broader Impacts**

Related to the concerns discussed in the Limitations section above, it is important for any downstream application to be aware that our data may have a limited representation of the goals and procedures of dominant cultures of the English-speaking countries.

**Example Template:**

Given a goal write down a list of steps to achieve the goal:

Goal: take a nap on the bed  
Step 1: sit on the bed for a little  
Step 2: pull back the blanket  
Step 3: pull back the sheet  
Step 4: fluff up the pillow  
Step 5: lay down on the bed  
Step 6: fall asleep on the bed  
Step 7: take a nap on the bed  
...

Goal: hire a dog walker  
Step 1:

**Prompt Prefix Generator:**

```
def generate_prompt_prefix():
    w1_list = ["For a given goal", "Given a goal"]
    w2_list = ["write down", "break down into", "put down" "jot
               down"]
    w3_list = ["steps", "subgoals", "a list of steps", "several
               steps", "several subgoals",
               "some steps", "some small
               steps"]
    w4_list = ["to achieve the goal", "for achieving the goal",
               "to attain the goal"]

    w1 = random.sample(w1_list, 1)[0]
    w2 = random.sample(w2_list, 1)[0]
    w3 = random.sample(w3_list, 1)[0]
    w4 = random.sample(w4_list, 1)[0]

    prompt_prefix = f"{w1}, {w2} {w3} {w4}.\n\n"
    return prompt_prefix
```

Figure 7: Randomize prompt template for eliciting plans.

| Prompt Template (Conditions)  | Prompt Template (Counterfactual Plan)  |
|---|--|
| <p>You want to use social media. How can you do this in 7 steps?<br/>     step 1: decided to use social media; step 2: Grab the phone;<br/>     step 3: Open, Start phone; step 4: Go to app store; step 5: Download Facebook from store; step5: Open and use facebook; step6: use social media<br/> <u>What is the hindrance that might affect the plan above?</u><br/> <u>If your phone screen is cracked.</u></p> <p>You want to plant a tomato plant. How can you do this in 7 steps?<br/>     step 1: decided to plant a tomato plant; step 2 : Go to nursery; step 3: Purchase tomato seedling.; step 4: Purchase potting soil and a pot.; step 5: Return to home.; step 6: Plant seedling in soil and pot.; step 7: plant a tomato plant<br/> <u>What is a specification that might affect the plan above?</u><br/> <u>If you want to use compost for soil.</u></p> <p>... x 3</p> <p>You want to print the report. How do you do this in 7 steps?<br/>     step 1: type the edited draft; step 2: save the edited draft; step 3: open the file menu in the word processor; step 4: select print from the file menu; step 5: select printer settings; step 6: send document to the printer; step 7: print the report<br/> <u>What is the hindrance that might affect the plan above?</u></p> | <p>You want to learn how to swim. How can you do this in 7 Steps?<br/> <u>Step 1: Decided to learn how to swim; Step 2: Find swimming instructor; Step 3: Travel to pool; Step4: Meet swimming teacher; Step 5: Practice swimming during classes; Step 6: Review mistakes with teacher until right; Step 7: Learn how to swim.</u><br/> <p>You want to learn how to swim. How can you do this in several steps <u>if you forget your swimsuit?</u></p> <p><u>Step 1: Decided to learn how to swim; Step 2: Find swimming instructor; Step 3: Travel to pool; Step4: Meet swimming teacher; Step 5: If you have forgotten your swimsuit, ask the instructor if it is possible to borrow one or if there is a place where you can purchase one; Step 6: Practice swimming during classes; Step 7: Review mistakes with teacher until right; Step 8: Learn how to swim</u></p> <p>... x 3</p> <p>You want to pick up pen. How can you do this in 6 steps?<br/>     step 1: look for a pen; step 2: find a pen; step 3: walk over to pen; step 4: extend hand out to pen; step 5: reach for pen; step 6: pick up pen<br/> <p>You want to pick up pen. How can you do this in several steps <u>if you want to pick up the pen from a high shelf?</u></p> </p></p> |

Figure 8: Prompt templates for acquiring Conditions and Counterfactual Plans.

Instructions (click to expand/collapse)

**WARNING:** This HIT may contain adult content. Worker discretion is advised.

Thanks for participating in this HIT!

In this HIT, imagine you are in the business of teaching people how to go about achieving everyday or life-long goals. You are handed a **goal** and a **plan** you can use to teach.

Your task is to evaluate plans based on several criteria. A good plan **doesn't contain repetitive or unnecessary steps**, is **on-topic**, **well-ordered**, and **complete**.

**Goal**    A life goal to achieve. Goal can be as simple as "cooking a dinner" to more elaborate "visiting Hawaii". It can also be fairly ambitious like "travelling to every country in the world" or time-consuming like "becoming the best sushi chef in the country".

**Plan**    The proposed plan given goal. A list of typical subgoals or steps to achieve the main goal. (e.g., for "visit Hawaii": buy ticket to Hawaii, decide what you want to see, book lodging, pack, leave for the airport)

**Grade with the following rubric.** We define the **Definitely**, **Somewhat**, and **Not at all**. Use middle values as needed.

1. **Is the plan on-topic?**
  - **Definitely**: Topic in the plan is relevant and appropriate to the goal.
  - **Somewhat**: Topic in the plan is wandering a bit from goal, but it is okay overall.
  - **Not at all**: Topic in the plan is overall irrelevant to the goal.
2. **Is the plan well-ordered?**
  - **Definitely**: The ordering is just fine as is.
  - **Somewhat**: I could see reordering some of these, but it would be more of a stylistic change.
  - **Not at all**: Ordering is bad or nonsensical.
3. **Is the plan complete and informative?**
  - **Definitely**: The plan provides a complete and informative picture of what's needed to achieve the goal.
  - **Somewhat**: The steps are somewhat general, but you overall get what you need. You might need a few more minor details.
  - **Not at all**: The plan is really bland and is dominated by unnecessary, irrelevant, and/or repetitive steps, -or- key steps are missing.
4. **Is the plan overall good?**
  - **Definitely**: The plan is overall good. A good plan should be well-ordered, complete and contains no repetitive steps.
  - **Somewhat**: The steps are somewhat general, but overall you get what you need.
  - **Not at all**: The plan is really bland and not good with repetitive steps.

NOTES:

- Steps are allowed to be general so long as the key information is there. Think: is the plan enough to give students solid grounding to start of asking relevant questions and taking relevant steps to achieve the goal?
- Note that an overall good plan should be topically relevant, ordered correctly, almost complete and contains no repetitive or unnecessary steps.
- Please do not hover too much over fine-grained differences. When in doubt, choose go with your gut instinct.
- If the goal is an incomplete thought or is nonsensical, then please choose **Not at all**

Examples (click to expand/collapse)

Goal: \${goal}  
Steps:  
1. \${steps\_html}

|                               | Definitely            | Mostly                | Somewhat              | Hardly                | Not at all            |
|-------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Is plan on-topic?             | <input type="radio"/> |
| Is plan well-ordered?         | <input type="radio"/> |
| Is plan complete/informative? | <input type="radio"/> |
| Is plan overall good?         | <input type="radio"/> |

(Optional) Please let us know if anything was unclear, if you experienced any issues, or if you have any other feedback for us.

**Submit**

Figure 9: AMT human evaluation template for the original planning task. For validation round we substituted goal *achievability* (is the goal achievable with appropriate steps?) for *overall* question (is the plan overall good?).

**Instructions (click to expand/collapse)**

Thanks for participating in this HIT!

In this HIT you are shown a pair of goal and plan, followed by 2 sets of conditions. Each condition might make achieving the goal more complex or harder which would require a change of the plan. You will evaluate the quality of each of the condition for the given pair of goal and plan. Additionally, you will evaluate if the condition has an association to a given tag.

Each question consists of a goal, plan, a set of conditions, and corresponding tags:

|                  |   |
|------------------|---|
| <i>Goal</i>      | A goal/desire that can be <b>achievable</b> through a plan (e.g., go to Hawaii, spend my Sundays at the beach, and so on).  |
| <i>plan</i>      | A <b>step-by-step proposed actions</b> to achieve the goal which consists of a list of typical subgoals or steps to achieve the main goal (e.g., for "visit Hawaii": buy ticket to Hawaii, decide what you want to see, book lodging, pack, leave for the airport). |
| <i>condition</i> | A constraint that may impede achieving the goal, require some changes in the plan, or make it more specific, but does not make it impossible (e.g., for the above goals, cannot find flight ticket, not having a car, and so on.).                                  |
| <i>tag</i>       | A tag is a general term that describes the type of the condition and how it might add constraints to the given plan (e.g., for the above conditions, not having a car is associated with equipment tag, etc.)   |

For a condition to be of good quality the following should be satisfied:

- The goal itself should make sense and can be achievable.
- The condition should be relevant to and realistic for the given goal.
- The condition should require a change in the plan that does not makes achieving the goal impossible. In other words, there should be alternative ways to achieve the goal. E.g., it should not negate an existing steps of the plan (or the goal) in a way that makes it impossible to achieve the goal.
- One should be able to come up with a list of steps to achieve the goal given the condition.

For each set of (goal, plan, condition), you will need to answer whether the condition alters the plan in achieving the goal, or make it impossible to achieve or has some other issues.

**Examples (click to expand/collapse)**

| Goal                 | Plan                       | Condition                  | Tag |
|----------------------|----------------------------|----------------------------|-----|
| <code>#{goal}</code> | <code>#{plan_part1}</code> | <code>#{plan_part2}</code> |     |

**Condition: \${condition\_1}**

Condition makes the goal more specific or harder to achieve (but not impossible)  
 Condition will make the goal impossible to achieve  
 Others (condition and/or goal do(es)n't make sense, condition is irrelevant to the goal, condition is simply negating a step, etc.)

Does the condition associate with the tag `#{tag_1}`?

Yes  No

**Condition: \${condition\_2}**

Condition makes the goal more specific or harder to achieve (but not impossible)  
 Condition will make the goal impossible to achieve  
 Others (condition and/or goal do(es)n't make sense, condition is irrelevant to the goal, condition is simply negating a step, etc.)

Does the condition associate with the tag `#{tag_2}`?

Yes  No

(Optional) Please let us know if anything was unclear, if you experienced any issues, or if you have any other feedback for us.

**Submit**

Figure 10: AMT template for assessing validity of conditions for critic model training.

**Instructions (click to expand/collapse)**

Thanks for participating in this HIT!

**Please read carefully:**

- Trivial modification is a little tricky so please be sure to look at the examples carefully.
- As you are working through these hits, you may see repeats in goal/plan. In such cases, conditions are different.

In this HIT, you are shown a **goal**, a corresponding **plan** of action, and a **condition** to the goal. You will also be provided with a **modified plan** that is meant to overcome the hindrance to make the goal achievable. Following are the definitions.

|                      |   |
|----------------------|---|
| <b>Goal</b>          | A <b>goal</b> /desire that can be achievable through a <b>plan</b> .<br>(e.g., go to Hawaii, spend my Sundays at the beach, and so on)  |
| <b>Plan</b>          | A <b>step-by-step proposed actions to achieve the goal</b> which consists of a list of typical subgoals or steps to achieve the main goal.<br>(e.g., for "visit Hawaii": 1. buy ticket to Hawaii, 2. decide what you want to see, 3. book lodging, 4. pack, 5. leave for the airport)   |
| <b>Condition</b>     | A <b>hindrance/blocker</b> or a <b>particular preference</b> that requires <b>modifications or updates</b> to the <b>plan</b> .<br>(e.g., for the above goals, "not having a car", "visit during a Aloha Week", and so on.)   |
| <b>Modified Plan</b> | A <b>new plan</b> for achieving the <b>goal</b> that takes <b>condition</b> into account. The <b>modified plan</b> should be a <b>modification of the original plan</b> .<br>(e.g., for the condition "not having a car" when the goal is to "visit Hawaii": a valid modification would require an additional step like "look for shared cabs", etc.) |

**YOUR TASK** is to evaluate the **quality** of the **modified plan** by answering two questions.

**Q1:** Does the **modified plan** contain all the necessary steps to make the **goal** achievable even with the **condition**?

- Yes**: The **modified plan** contains **all the necessary steps** to meet the requirements of the **condition** on the **goal**.
- Yes, but**: The **modified plan** addresses the **condition**, but the modification is **trivial and lacks thoughtfulness**.
- No**: The **modified plan** does **NOT** address the **condition** or does so very poorly. One or more **important steps are missing**.

**Q2:** What are the problems with the **modified plan**?

- Are the steps wrong (are edits required)?
- Are there missing steps (additional steps are required)?
- Are there extra unnecessary steps (are deletes required)?
- If everything is okay, there's a choice for that.

**Example of a bad modification:**

|  |   |  |
|--|---|--|
| Your goal is to...<br><b>Buy a new car</b>   | But you are hindered by...<br><b>Don't have enough money for a new car</b>  | Q1: Does the <b>modified plan</b> contain all the necessary steps to make the <b>goal</b> achievable even with the <b>condition</b> ?  |
| Your initial plan was...<br>1. Research vehicle and features<br>2. Go to dealership<br>3. Test-drive a car of your choice<br>4. Check sale price and warranties<br>5. Negotiate a best price<br>6. Make the purchase | Now, your modified plan is...<br>1. Research vehicle and features<br>2. Go to a used dealership<br>3. Go to a dealership<br>4. Check sale price and warranties<br>5. Negotiate a best price | <b>No</b> : one or more important steps are missing.<br><br><b>Explanation:</b> The modified plan doesn't address the condition. Also there are random repetitions. And where did the "purchase" step go?                    |
| →  |   | <b>Q2: What are the problems with the <b>modified plan</b>? (select all that apply)</b>  |
|  |   | <input type="checkbox"/> Some step(s) is/are <b>wrong</b> (edits are necessary).<br><input type="checkbox"/> Some step(s) is/are <b>missing</b> .<br><input type="checkbox"/> There is/are <b>extra unnecessary</b> step(s). |

|   |   |
|---|---|
| Your goal is to...<br><b>\${goal}</b>         | But you are hindered by...<br><b>\${condition}</b>          |
| Your initial plan was...<br>1. \${steps_html} | Now, your modified plan is...<br>1. \${modified_steps_html} |
| →   |   |

**Q1: Does the **modified plan** contain all the necessary steps to make the **goal** achievable even with the **condition**?**

- Yes**, all of the important steps are present.
- Yes, but**, the changes are trivial (e.g. just negates the hindrance)
- No**, one or more important steps are missing.

**Q2: What are the problems with the **modified plan**? (select all that apply)**

- No problems**: All important steps present. No missing or extra steps.  
Note this choice is mutually exclusive with other choices.
- Some step(s) is/are wrong** (edits are necessary).
- Important step(s) is/are missing**.
- There is/are extra unnecessary** step(s).

Figure 11: AMT human evaluation template for counterfactual plan revision. We use a similar layout for counterfactual planning task only removing the initial plan.

| Goal                | Condition   | Initial Plan   | Generated (Counterfactual) Plan  |
|---------------------|---|--|--|
| Buy a new videogame |   |  | Find a store that sells videogames > Go to the store > Find the videogame you want to buy > Buy the videogame > Take the videogame home > Play the videogame   |
|                     | If the game store is too far away                 |  | Decide to buy a new videogame > Research online game stores > Compare prices and reviews > Select the game to buy > Purchase the game online > Wait for the game to be delivered > Buy a new videogame   |
|                     |   | Decided to buy a new videogame > Decide on game > Get into car > Drive to game store > Walk into game store > Find the desired game > Buy a new videogame  | Decided to buy a new videogame > Decide on game > Search for the game online > Compare prices online > Purchase the game online > Wait for the game to be delivered > Unbox the game > Play the game > Enjoy the game  |
| get extra exercise  |   |  | Set a goal for how many hours a week you want to exercise > Research different types of exercise > Choose one type of exercise to start with > Make a plan to follow the exercise plan > Start exercising > Keep exercising  |
|                     | If the fitness activity presents a risk of injury |  | Decide what type of exercise you would like to do > Research the activity to make sure you understand the risks and benefits > Consult with a medical professional to ensure the activity is safe for you > Find a place to do the activity > Set a goal for how often you would like to do the activity > Start doing the activity > Monitor your progress and adjust the activity as needed > Get extra exercise |
|                     |   | Decided to get extra exercise > Open the calendar > Identify open time slots during the week > Schedule a fitness activity during free time > Set reminder for activity on phone > Attend weekly fitness activity > Get extra exercise | Decide to get extra exercise > Open the calendar > Identify open time slots during the week > Research and select a fitness activity that presents a low risk of injury > Schedule the fitness activity during free time > Set reminder for activity on phone > Attend weekly fitness activity > Get extra exercise  |
| chop the wood       |   |  | Get a chopping block > Get a chopping knife > Chop the wood > Clean the chopping block > Put the chopping block away   |
|                     | If you have a chainsaw                            |  | Pick up the wood > Place the wood on the ground > Set up the chainsaw > Sharpen the chainsaw > Start the chainsaw > Cut the wood > Chop the wood   |
|                     |   | Pick up axe > Pick up a piece of wood > Check the condition of wood > Make sure wood is dry enough > Set the wood upright > Swing the axe at wood > Chop the wood  | Pick up the chainsaw > Pick up a piece of wood > Check the condition of wood > Make sure wood is dry enough > Set the wood upright > Start the chainsaw > Guide the chainsaw along the wood > Chop the wood  |

Table 13: Additional PLASMA generations for (*counterfactual*) planning and *revision* tasks.