

# ADAPT: As-Needed Decomposition and Planning with Language Models

Archiki Prasad<sup>\*</sup> Alexander Koller<sup>♡</sup> Mareike Hartmann<sup>♡</sup>  
Peter Clark<sup>♣</sup> Ashish Sabharwal<sup>♣</sup> Mohit Bansal<sup>\*</sup> Tushar Khot<sup>♣</sup>

<sup>♣</sup> UNC Chapel Hill <sup>♣</sup> Allen Institute for AI <sup>♡</sup> Saarland University

## Abstract

Large Language Models (LLMs) are increasingly being used for interactive decision-making tasks requiring planning and adapting to the environment. Recent works employ LLMs-as-agents in broadly two ways: iteratively determining the next action (iterative executors) or generating plans and executing sub-tasks using LLMs (plan-and-execute). However, these methods struggle with task complexity, as the inability to execute any sub-task may lead to task failure. To address these shortcomings, we introduce **As-Needed Decomposition and Planning** for complex Tasks (ADAPT), an approach that explicitly plans and decomposes complex sub-tasks *as-needed*, i.e., when the LLM is unable to execute them. ADAPT recursively decomposes sub-tasks to adapt to both task complexity and LLM capability. Our results demonstrate that ADAPT substantially outperforms established strong baselines, achieving success rates up to 28.3% higher in ALFWorld, 27% in WebShop, and 33% in TextCraft – a novel compositional dataset that we introduce. Through extensive analysis, we illustrate the importance of multi-level decomposition and establish that ADAPT dynamically adjusts to the capabilities of the executor LLM as well as to task complexity.<sup>1</sup>

## 1 Introduction

Recent advances in Large Language Models (LLMs) have expanded their application beyond conventional NLP tasks to more complex tasks involving mathematical, symbolic, and common-sense reasoning (Wei et al., 2022; Huang and Chang, 2023). Recent models have even been applied to *decision-making* tasks, such as performing household chores, navigating a webpage, etc., that require interactions with external environments or tools (Yao et al., 2023b; Qin et al., 2023).

<sup>1</sup>Project: <https://allenai.github.io/adaptllm>

Prior works on using LLMs for decision-making, such as ReAct (Yao et al., 2023b), iteratively generate the next action to be executed in the environment given the history of actions and observations (see Fig. 1; top-left). However, as the tasks become more complex, LLMs struggle due to their limited composition ability (Dziri et al., 2023) and inability to deal with the distractors (Shi et al., 2023) in a long action-observation trajectory.

To mitigate this, modular approaches (Khot et al., 2023; Yang et al., 2023; Sun et al., 2023) incorporate a separate planner module that utilizes an LLM to create a high-level plan.<sup>2</sup> The planner then delegates simpler sub-tasks to an executor LLM module thereby reducing the compositional complexity and length of action trajectory required by the executor. We refer to this category broadly as *plan-and-execute* approaches (see Fig. 1; top-right). While the plans enable these methods to guide the execution and track progress (Wang et al., 2023b), their non-adaptive nature poses a limitation when confronting unachievable sub-tasks. These approaches inherently lack the flexibility to adapt to task complexity and manage execution failures, as shown in Fig. 1 (top-right), where just one sub-task that is too complex results in overall task failure.

To address such failures, we propose **As-Needed Decomposition and Planning** for complex Tasks (ADAPT), a recursive algorithm that further decomposes sub-tasks *when necessary*, to dynamically accommodate to task complexity. We utilize separate *planner* and *executor* LLM modules within our framework but *only* decompose a task using the planner, if the executor LLM detects a failure. As shown in Fig. 1, the overall task of putting a clean mug on a desk in an unfamiliar

<sup>2</sup>By “planning”, we refer to the colloquial concept of designing a list of sub-tasks to accomplish a complex task rather than its usage in classical AI-planning literature. E.g., a “plan” for preparing a lasagna could be to cook the pasta, prepare the sauce, layer the ingredients, and then bake it.

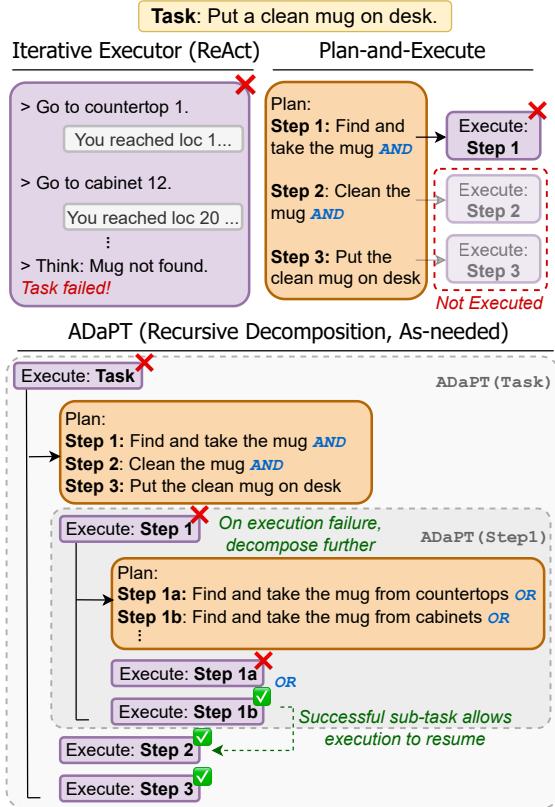


Figure 1: **Top-Left:** Iterative executors such as ReAct (Yao et al., 2023b) interact directly with the environment, performing planning implicitly. **Top-Right:** Plan-and-Execute, e.g., Yang et al. (2023), creates a fixed plan for the task, without accounting for complexity in executing step 1. **Bottom:** ADaPT dynamically decomposes based on success of the executor.

household is too complex for the model, leading to failure of the iterative executor. While a plan-and-execute-style approach initially breaks down the task into three sub-tasks, it falls short in accounting for the complexity in finding a mug. Moreover, it is challenging to anticipate the difficulty of such a sub-task in advance, as the executor could find a mug in the first attempt or in an obscure location. Therefore, ADaPT employs its recursive structure to *dynamically adapt* to execution failures (assessed by LLMs), by *further decomposing* the complex sub-task of *finding a mug* via the planner.

Empirically, we demonstrate the effectiveness of ADaPT on three datasets involving interactive environments: ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022), and a new compositional text game for crafting Minecraft recipes called *TextCraft* (Sec. 4.1). Using GPT-3.5 as the underlying LLM, ADaPT outperforms strong baselines (discussed in Sec. 4.2) such as ReAct (Yao et al., 2023b), and Plan-and-Solve (Wang et al.,

2023b) by up to 28.3%, 27%, and 33% absolute points on ALFWorld, WebShop, and TextCraft respectively (Sec. 5). Compared to Reflexion (Shinn et al., 2023), an adaptive approach that addresses *failures in the full task trajectory*, ADaPT yields higher success rates by 14.1%, 9%, and 20% on ALFWorld, WebShop, and TextCraft respectively. Through extensive analysis of ADaPT, we establish the importance of recursive decomposition (Sec. 6.1) and showcase dynamic adaptation to the capabilities of the executor LLM including open-source models such LLaMA-2 (Touvron et al., 2023) and Lemur (Xu et al., 2023) in Sec. 6.2. Lastly, we demonstrate that ADaPT incorporates task complexity (Sec. 6.3), where the extent of recursive decomposition aligns with the inherent task complexity. To summarize, our contributions are:

1. We present ADaPT, a recursive algorithm that dynamically decomposes complex sub-tasks on an as-needed basis, i.e., *intervening only if the task is too complex for the executor*.
2. On three diverse datasets, ALFWorld, WebShop, and TextCraft, ADaPT improves success rate of GPT-3.5 over previous approaches by up to 28.3%, 27%, and 33% points respectively.
3. Analysis of ADaPT underscores the significance of recursive decomposition and the ability to adapt dynamically to varying LLM execution capabilities and task complexities.

## 2 Related Work

**LLMs for Decision-Making.** LLMs have been successfully used as agents to perform a wide variety of decision-making tasks such as robotic navigation (Ahn et al., 2022; Huang et al., 2023b; Singh et al., 2023), complex multi-modal games like Minecraft (Fan et al., 2022; Wang et al., 2023a), text-based environments (Shridhar et al., 2021; Liu et al., 2023). While most of these works focus on learning from trajectories, ReAct (Yao et al., 2023b) uses few-shot prompting to build an agent that reasons about the current state (thoughts) and generates the next action in the environment, given prior actions and observations. Their iterative approach (shown in Fig. 1; top-left) can handle failures, but they have to keep track of the entire plan *implicitly* while deciding every local action (c.f. ADaPT in Fig. 9 of Appendix A). By incorporating planning and execution into separate modules and enabling dynamic adaptation we are able to achieve higher success rates (refer to Sec. 5).

Several follow-up works improve upon the ReAct framework by incorporating feedback in future trials (Madaan et al., 2023; Shinn et al., 2023), or using LLMs to develop heuristics for search (Yao et al., 2023a; Zhou et al., 2023). In contrast to ADAPT, they do not employ task decomposition, leading to unnecessary computation as they explore multiple trajectories or trials for the whole task, even though the LLM struggles with just one sub-task. Such works are complementary to ADAPT as they can be incorporated within the planner or executor modules to strengthen LLM performance (just like they are incorporated in ReAct).

**Decomposition and Modularity.** Our work follows extensive literature in NLP on decomposing tasks into neural modules (Andreas et al., 2016; Gupta et al., 2019; Jiang and Bansal, 2019) or seq2seq models (Min et al., 2019; Talmor and Beirant, 2018; Khot et al., 2021; Perez et al., 2020; Saha et al., 2023b). With the advent of few-shot prompted black-box LLMs, this paradigm of programmatic decomposition into LLMs has become more popular (Yao et al., 2023b; Khot et al., 2023; Wang et al., 2023b, *inter alia*), referred to as LLM Programs (Schlag et al., 2023; Dohan et al., 2022). Additionally, past works in program synthesis (Murali et al., 2018; Nye et al., 2019; Zheng et al., 2023) also employ task decomposition via generating a “program sketch” prior to program generation.

ADAPT not only decomposes tasks via the planner module and delegates them to the executor module but also *automatically* adapts to executor failures by further decomposing complex tasks *as-needed*. This dynamic capability distinguishes ADAPT from prior works with a non-adaptive structure. ADAPT extends the recursive and hierarchical decomposition in Khot et al. (2023), enabling inter-module communications, and robust strategies for execution failures, excelling in real-world textual environments like online shopping.

**Hierarchical Problem Solving.** In AI problem-solving, there is a longstanding tradition of hierarchical task decomposition employed in planning (Ghallab et al., 2004; Georgievski and Aiello, 2014; Höller et al., 2020), reinforcement learning (Sutton et al., 1999; Barto and Mahadevan, 2003; Nachum et al., 2018; Zhang et al., 2021), and navigation (She et al., 2014; Sharma et al., 2022; Blukis et al., 2022; Min et al., 2022; Song et al., 2023). These approaches, such as Hierarchical Task Networks (Erol et al., 1994), leverage domain knowl-

edge, e.g., hand-specified library of plans, to break complex problems into simpler tasks. Our work embraces this tradition but distinguishes itself by exploring how LLMs can autonomously decompose tasks by leveraging their extensive world knowledge, without predefined plan libraries. Lastly, ADAPT performs dynamic hierarchical planning by employing its recursive structure.

### 3 Methodology

We introduce **As-Needed Decomposition and Planning for complex Tasks** (ADAPT), a modular approach for decision-making that integrates an LLM as an *executor* and a *planner* (Secs. 3.1 and 3.2) within an LLM program called the controller (Sec. 3.3). In Fig. 1, when ADAPT is given a complex task, it first attempts to accomplish the entire task by running the executor iteratively, and resorting to the LLM planner for further decomposition into sub-tasks if the executor fails. Subsequently, ADAPT is recursively called for each sub-task to ensure their successful completion, ultimately leading to overall task success.

#### 3.1 LLM as an Executor

**Overview.** In a given environment, the executor is provided with a concise natural language task specification, as shown in Fig. 2 (left). Following Yao et al. (2023b), the executor iteratively interacts with the environment via actions generated by the LLM. This interaction continues until the task is either completed or a preset maximum iteration limit is reached. Consistent with Ahn et al. (2022), we provide the LLM with in-context demonstrations of low-level “atomic” skills specific to the environment (listed in Table 5 of Appendix A), such as knowing how to correctly heat objects in ALFWorld. This approach offers two advantages: (i) it allows us to employ the same executor with environment-specific knowledge for all baselines (Sec. 4.2); and (ii) it enables the planner (discussed in Sec. 3.2) to work at a higher level of abstraction, leveraging the LLM’s general world knowledge.

**Execution Capabilities of an LLM.** At a minimum, the LLM executor should reliably execute atomic skills. While we provide demonstrations for successful execution of atomic skills, LLMs can adapt to failures by combining multiple skills to perform complex tasks, as discussed in Sec. 6.2. For instance, in Fig. 2 (left), we show the LLM successfully cleaning a mug it’s carrying (an atomic

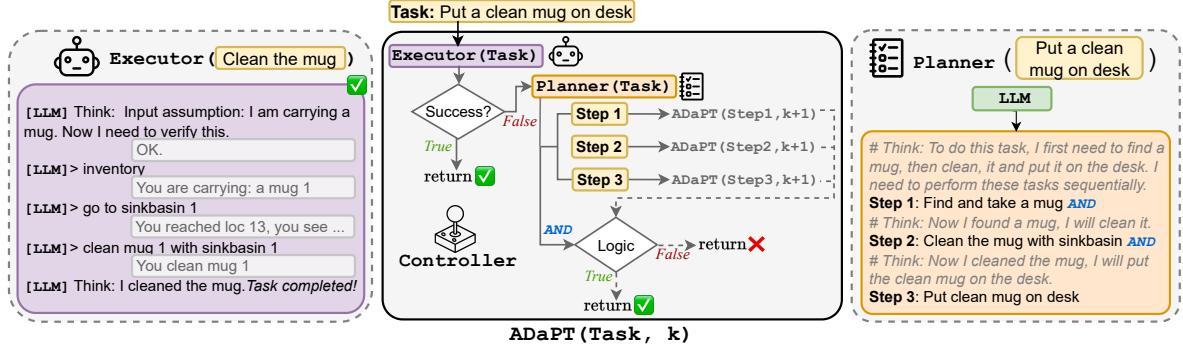


Figure 2: Block diagram of the ADAPT pipeline with an example from ALFWorld. **Left:** Use of LLM as an executor to interact iteratively with the environment along with an example execution trajectory. **Middle:** Overall recursive algorithm (depth  $k \leq d_{\max}$ ) that embeds the executor and planner, refer to Algorithm 1 for details. **Right:** Outline of using LLM as a planner to generate sub-tasks (steps) and logical operators combining them.

skill). An advanced executor could combine “finding a mug” with the “cleaning” skill to accomplish “find a clean mug” without an explicit planner.

**Self-generated Success Heuristic.** In order to decompose based on the abilities of the executor, we need to determine whether the executor is capable of finishing the given (sub-)task independently or if further decomposition is required. To this end, we employ the executor LLM to determine the completion of the (sub-)task *without relying on the environment* for obtaining gold rewards for (sub-)tasks. We include a simple instruction in the executor prompt to output “*task completed*” if it determines it has succeeded, otherwise output “*task failed*” in case it cannot proceed. Refer to example in Fig. 2 (left). Our success heuristic aligns with binary classification models employed in Shinn et al. (2023), providing a way to simulate intermediate rewards, which complements end-of-task environment rewards (Rengarajan et al., 2022). We study this LLM-generated heuristic in Appendix F and show that it closely matches the gold reward.

### 3.2 LLM as a Planner

**Overview.** The objective of the planner is to break down complex tasks into smaller sub-tasks. To achieve this, we instruct the LLM to generate a concise yet comprehensive plan consisting of a few steps, typically 3-5, as shown in Fig. 2 (right). We opt for shorter, more abstract plans because expecting a detailed, fine-grained plan upfront can be impractical, especially in unexplored environments. E.g., devising a 10-step plan to put a clean mug on a desk without prior knowledge of the mug’s location can lead to cascading errors due to incor-

rect assumptions. Therefore, we task the LLM to generate short plans, with the *flexibility to decompose further* in subsequent iterations, based on the executor’s capabilities.

**Composition Logic for Sub-tasks.** Along with the sub-tasks, we prompt the planner to generate logical operators to combine various sub-tasks in the plan to accomplish the task. We allow for two logical operators: “AND” and “OR”. Sub-tasks are linked using AND when they must be executed sequentially for the task to succeed. However, in cases requiring exploration, such as finding an item in an unknown room, we employ the OR operator to simulate conditional checks. Here, the task succeeds if any of the sub-tasks are successful. For instance, in Fig. 1, the plan to “*find a mug*” would be to “*find a mug on the countertop*” OR “*find a mug in the cabinet*”. We execute the latter only if the agent has not found the mug yet. While examples in Figs. 1 and 2 show homogeneous logic, ADAPT can handle complex logical expressions as described in Appendix B.

### 3.3 Controller – LLM Program

**Overall Pipeline.** Thus far, we describe two LLM-based modules that can perform the roles of low-level execution and high-level planning. We incorporate these modules into ADAPT via the controller which is a pre-determined and recursive algorithm – making the overall pipeline of ADAPT an LLM program (Schlag et al., 2023; Dohan et al., 2022), shown in Algorithm 1. The overall flow of the controller program is as follows: (i) given an input task, the controller calls the executor to check if it can succeed in performing the task directly; (ii)

if the executor does not succeed, the controller delegates decomposing the complex task to the planner and recursively calls ADAPT for each sub-task until we hit a termination criterion, i.e., if a maximum depth  $d_{\max}$  ( $\geq 1$ ) is reached.

Fig. 2 (mid) shows the control flow of ADAPT. A complex task such as “put a clean mug on the desk” is first assigned to the executor. If the executor does not succeed, then ADAPT calls the planner to decompose the task into sub-tasks along with a logical operator (AND or OR) indicating how to compose them. Each sub-task (referred to as ‘step’ in Fig. 2) is then assigned recursively to ADAPT and is combined using the logical operator. In the end, the success of sub-tasks after recursive decomposition ensures overall task success (unrolled calls to planner and executor are shown in Fig. 1).

## 4 Experimental Setup

We describe the datasets used in our experiments and baselines used for comparison with ADAPT.

### 4.1 Datasets

We employ LLMs-as-agents to perform tasks in the following three environments and use task **success rate** as our evaluation metric in Secs. 5 and 6.

**ALFWorld.** ALFWorld (Shridhar et al., 2021) is a text-based game version of the embodied ALFRED benchmark (Shridhar et al., 2020) implemented in the TextWorld environment (Côté et al., 2019). It encompasses 6 distinct task types, where an agent is required to accomplish high-level tasks through navigation and interaction via text-based actions in a simulated household that gives textual feedback to an agent (e.g., *put a clean mug on desk* discussed earlier in Fig. 2). Following Shridhar et al. (2021), we present results on 134 unseen evaluation games (test set) with a separate dev set of 10 games per task from the seen evaluation games split. Along with atomic skills, we add example gold trajectories, following Yao et al. (2023b), for two tasks: heat and look in the executor prompt.<sup>3</sup>

**WebShop.** WebShop (Yao et al., 2022) is an online shopping website environment featuring 1.18 million real-world products containing 500 user queries in the test set. It serves as a complex

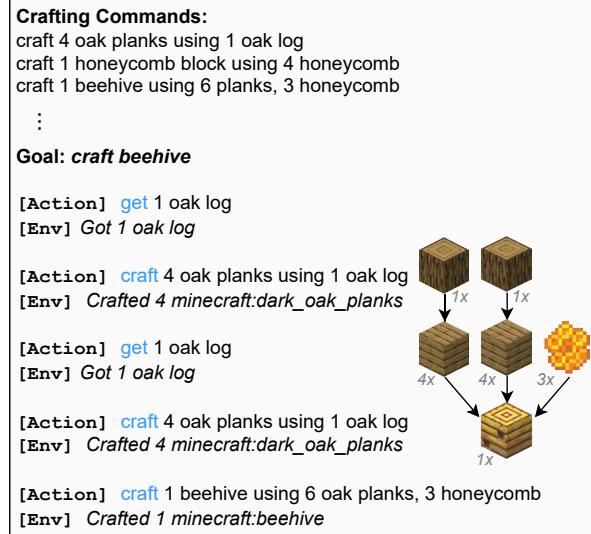


Figure 3: Example gold trajectory in TextCraft for a task with recipe depth of 2.

decision-making environment with practical applications wherein an agent must navigate a website through a variety of commands to purchase an item matching a user specification (e.g., *grey sectional sofa priced less than \$300 with fast delivery*). Following Shinn et al. (2023), we report performance on 100 user instructions and use a different subset of 40 queries as the dev set.

**TextCraft.** We create a new text-only environment for crafting Minecraft<sup>4</sup> items similar to WordCraft (Coenen et al., 2021). Unlike existing agent-based environments, tasks in TextCraft exhibit a natural compositional structure, resembling cooking recipes with steps of varying complexity, where some sub-tasks are more intricate, such as layering a lasagna, while others are simpler, like baking it.

Tasks in TextCraft are inherently decomposable. In Fig. 3, crafting a beehive necessitates crafting its ingredients, like planks and honeycomb, which may require further decomposition. The agent thus needs to identify and adapt to varying task complexity, e.g., crafting a plank is *easier* than crafting a beehive. Moreover, some recipes allow using any item from a particular category. For instance, crafting a beehive uses planks (a category), requiring the agent to use linguistic knowledge for proper item selection (e.g., select oak planks, a specific item in the category planks). We evaluate our approach on a test set of 200 tasks where the target items have recipe trees of depth 2, 3, and 4 (example tree of depth 2 is shown in Fig. 3). We use the

<sup>3</sup>Unlike Yao et al. (2023b), we use a standardized executor prompt for all ALFWorld tasks, avoiding the agent to know the task-type *a priori*. Table 6 in Appendix C further demonstrates that ADAPT still improves over task-specific executors.

<sup>4</sup><https://www.minecraft.net>

Method ( $d_{\max} = 3$ )	Pick	Clean	Heat	Cool	Look	Pick2	All
ReAct	33.3	<b>67.7</b>	43.5	33.3	55.6	<u>11.8</u>	43.3
Plan-and-Execute	29.2	61.3	47.8	38.1	<b>61.1</b>	<u>11.8</u>	43.3
Try Again with ReAct	50.0	51.6	<u>60.8</u>	47.6	<b>61.1</b>	5.9	47.8
Reflexion	70.8	61.3	<b>61.0</b>	<u>66.7</u>	<b>61.1</b>	5.9	<u>57.5</u>
ADAPT (Ours)	<b>87.5</b>	<b>80.6</b>	<u>60.8</u>	<b>76.2</b>	<b>61.1</b>	<b>52.9</b>	<b>71.6</b>

Table 1: ADAPT yields the highest the overall success rates (%) compared to baselines from prior work (discussed in Sec. 4.2) on ALFWorld (test split). Best (highest) success rates are highlighted in bold and second-highest rates are underlined.

items with recipe tree depth of 3 (123 tasks), depth of 4 (11 tasks) and depth of 2 (77 out of 297) in our test set, and the rest of depth 2 tasks constitute the dev set. Additional details about creating the environment are present in Appendix E.

## 4.2 Baseline Approaches

We compare ADAPT with four classes of baseline approaches described below.

**Iterative Executor-Only (ReAct).** In this setting, we employ the executor to interact iteratively with the environment, adopting the think-act-observe prompting style from ReAct (Yao et al., 2023b). All methods discussed below, including ADAPT, share the *same* executor, ensuring a standardized impact of the executor’s strength and design choices when comparing relative performance in Sec. 5. When  $d_{\max} = 1$ , ADAPT solely relies on this executor.

**Plan-and-Execute.** As shown in Fig. 1, in this setting, we generate a plan first and then assign each sub-task to the executor. This approach only plans once and as a result has a non-adaptive structure (consistent with Wang et al. (2023b); Yang et al. (2023); Sun et al. (2023)). To ensure each plan step is executable without further decomposition, we design new prompts with more detailed plans. Note that ADAPT with  $d_{\max} = 2$  differs from plan-and-execute as it is adaptive, i.e., decomposes only when executor fails and generates relatively shorter plans (refer to Appendix B).

**Try Again with ReAct.** By design, ADAPT makes multiple calls to the executor module, albeit with different (sub-)tasks. Like Yang et al. (2023), we design a simple controller that requests the executor to retry the task in a total of  $d_{\max}$  separate trials and then uses the trial with the best performance for each task instance.

Method	WebShop	TextCraft
ReAct	32.0	19.0
Plan-and-Execute	17.0	27.0
Try Again with ReAct	30.0	15.0
Reflexion	35.0 <sup>†</sup>	<u>32.0</u>
LATS (Zhou et al., 2023)	<u>38.0<sup>†</sup></u>	—
ADAPT (Ours)	<b>44.0</b>	<b>52.0</b>

Table 2: ADAPT yields the highest success rate on WebShop and TextCraft (test split) with  $d_{\max} = 3$  and 4 respectively. <sup>†</sup>Performance reported by Zhou et al. (2023)

**Reflexion.** Shinn et al. (2023) execute the entire task first, and if unsuccessful, reflect and store feedback in memory for subsequent  $d_{\max} - 1$  trials. While adaptive, this approach repeats the entire trial even if a single sub-task fails, redundantly re-executing previously successful sub-tasks.

## ADAPT and Shared Implementation Details.

Following (Yao et al., 2023b; Shinn et al., 2023; Zhou et al., 2023), by default, we use the GPT-3.5 (Ouyang et al., 2022) LLM for both planning and execution in ADAPT and other baselines. We use the completion-based models for ALFWorld and TextCraft and the chat-based model for WebShop.<sup>5</sup> Further, we use ADAPT (and other baselines) with  $d_{\max} = 3$  for ALFWorld, and WebShop and increase to  $d_{\max} = 4$  for TextCraft to accommodate recipes with a depth of 4 (Sec. 4.1). For additional details, refer to Appendix A. We increase the maximum number of iterations for the ReAct baseline by a factor of  $d_{\max}$  and ensure all baselines use a comparable number of LLM calls (Sec. 6.5).

## 5 Main Results

Using GPT-3.5 as the underlying LLM, in this section, we show that ADAPT yields the highest success rate compared to baselines from prior work on ALFWorld, WebShop, and TextCraft datasets.

**ALFWorld.** In Table 1, we observe that ADAPT achieves the *highest overall success rate*, while using ReAct alone results in the lowest overall performance. By leveraging adaptive decomposition, ADAPT improves over ReAct’s performance by 28.3% points (absolute) as well as over Plan-and-Execute and Try Again by 28.3% and 23.8% points, respectively. Lastly, we find that ADAPT

<sup>5</sup>We use the completion model as chat variants of GPT-3.5 consistently underperform their completion counterparts (Liu et al., 2023; Yang et al., 2023). We discuss the effectiveness of ADAPT different LLMs in Sec. 6.2.

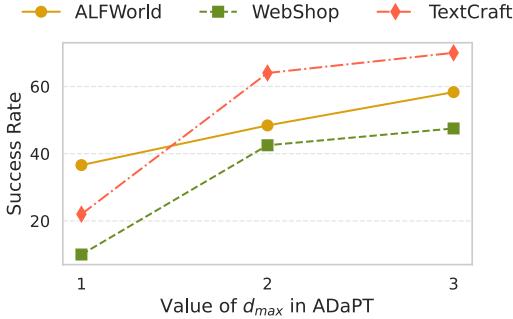


Figure 4: Success rate of ADaPT increases with the maximum depth  $d_{max}$  for all datasets (dev splits).

yields 14.1% points higher overall success rate than Reflexion, despite the latter having access to dedicated memory and natural language feedback. Specifically, we find baselines yield poor results on ‘pick2’ tasks (<12% success rate) as they require the agent to compose two ‘pick’-style tasks involving a longer action history. However, ADaPT yields significant improvements (by over a factor of 4×) for this type of tasks.

**WebShop.** Table 2 shows a similar trend with ADaPT *surpassing all baselines* and achieving the highest success rate. ADaPT outperforms ReAct, Plan-and-Execute, and Try-Again baselines by up to 27% points. We corroborate the findings of Shinn et al. (2023) and observe that natural language feedback offers limited gains in performance, as compared to ADaPT (which surpasses Reflexion by 9% points). Additionally, we compare with a recent search-based baseline LATS (Zhou et al., 2023) and find that ADaPT outperforms the success rate of LATS by 6% points.

**TextCraft.** Our results on TextCraft are summarized in Table 2. First, we observe that ADaPT *achieves an improvement of 33%* compared to the ReAct executor. In contrast to Plan-and-Execute, i.e., starting with a fixed plan, having the dynamic ability to adapt to complex sub-tasks (in this case, crafting complex ingredients) in ADaPT improves performance by 25% points. Lastly, ADaPT outperforms Reflexion by 20% points, highlighting the importance of adaptive and as-needed planning. We hypothesize that ADaPT consistently outperforms Reflexion across datasets as the latter relies on generating feedback based on errors in the entire trajectory. In contrast, due to its design, ADaPT often handles failures of small sub-tasks and redirects more resources in the form of calling the planner

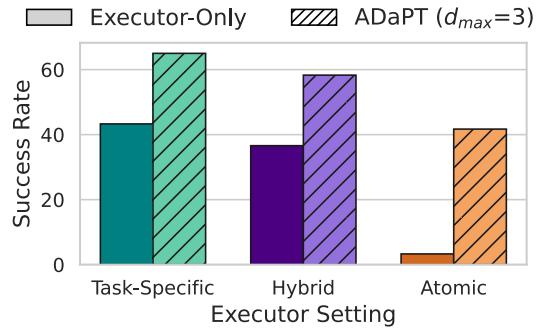


Figure 5: ADaPT improves success rates across varying settings capturing different executor capabilities (i.e., executor-only performance) on ALFWorld (dev).

and decomposition to the challenging sub-tasks.

## 6 Analysis and Discussion

We analyze ADaPT in detail by addressing the following research questions on dev data splits.

### 6.1 How does performance of ADaPT scale with the depth of decomposition?

**Setup.** To assess the impact of adaptive decomposition, we study ADaPT under three settings with increasing maximum depth  $d_{max} \in \{1, 2, 3\}$  for ALFWorld, WebShop, and TextCraft. Note that  $d_{max} = 1$  setting corresponds to the iterative executor-only baseline (ReAct).

**Results.** Fig. 4 shows that across all datasets, performance of ADaPT scales with increasing the maximum depth  $d_{max}$ . Consistently, we find a significant improvement in success rates as we move from  $d_{max} = 1$  to  $d_{max} = 2$ , i.e., adding the planner to decompose a complex task when executor fails proves to be effective. Finally, the performance increase from  $d_{max} = 2$  to  $d_{max} = 3$  validates our hypothesis that some sub-tasks are difficult for the LLM to directly execute successfully, and decomposing these further boosts overall performance.

### 6.2 Does ADaPT cater to different execution capabilities of LLMs?

**Same LLM, different execution capabilities.** We run ADaPT on three different executor prompts on ALFWorld: (i) task-specific gold trajectories, (ii) atomic skills and common gold-trajectories for 2 tasks used in Sec. 5 (hybrid), and (iii) only atomic skills. Using gold trajectories aligns closely with the task at inference-time and thus, should exhibit high performance. In contrast,

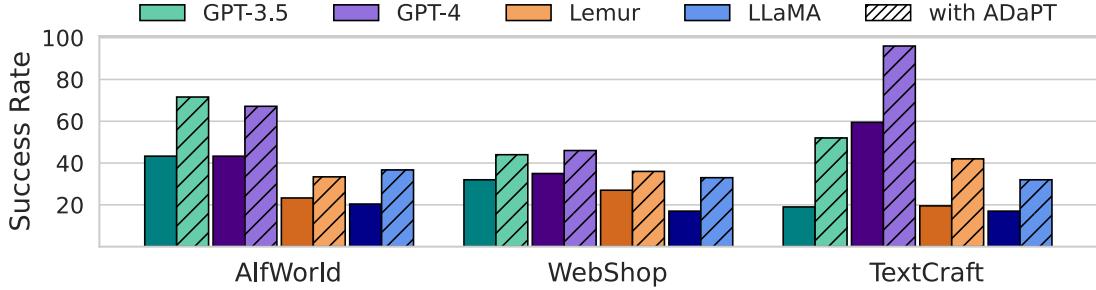


Figure 6: ADaPT improves (test) performance of GPT-3.5, GPT-4, LLaMA, and Lemur LLMs across datasets.

executor using only atomic skills relies on the inherent composition abilities of the LLM, yielding weaker performance. Here we examine if ADaPT can improve success rates for all three settings.

**Results.** In Fig. 5, we observe that ADaPT consistently improves over the executor-only baseline for *all diverse executor settings*. As expected, the executor prompted with task-specific trajectories performs the best (left), while the executor with only atomic skills performs the worst (right). Notably, ADaPT substantially improves performance of the relatively weak executor, improving success rate from 3.3% to 41.7%.

**ADaPT with different LLMs.** We study the ability of ADaPT to improve performance across different LLMs (as planners and executors): (i) GPT-3.5, (ii) GPT-4 (OpenAI, 2023), (iii) LLaMA-2 70B (Touvron et al., 2023), and (iv) Lemur 70B (Xu et al., 2023) on test splits of all datasets.

**Results.** Fig. 6 shows that ADaPT consistently improves downstream performance for *all* models across *all* three datasets. Consistent with Liu et al. (2023), we find that the gated GPT models outperform the open-source models based on absolute success rates. Nevertheless, ADaPT is effective across LLMs and improves performance of GPT-4, the strongest LLM, by up to 37%, as well as LLaMA, the least performant LLM, by up to 15% on the TextCraft dataset.

### 6.3 Does ADaPT handle task complexity?

**Setup.** By the compositional design of TextCraft, complexity of each task in the dataset can be defined with respect to the depth of the crafting recipe, i.e., recipes with higher depth would be more complex to craft. We evaluate efficacy of ADaPT and the ReAct baseline on the test set of TextCraft

Method	Recipe Depth	$k_{\max}$	Success Rate
ReAct	2	1.0	26.9
ADaPT ( $d_{\max} = 4$ )	2	1.9	<b>78.2</b>
ReAct	3	1.0	1.8
ADaPT ( $d_{\max} = 4$ )	3	2.8	<b>38.7</b>

Table 3: ADaPT improves TextCraft (test) performance even as recipe depth increases. The maximum decomposition depth used by ADaPT to succeed at the task ( $k_{\max}$ ) also scales with the recipe depth.

with increasing recipe depth.<sup>6</sup> Furthermore, while we provide ADaPT with a maximum budget of  $d_{\max} = 4$ , we study how the maximum decomposition depth utilized by ADaPT to succeed ( $k_{\max}$ ) varies with task complexity.

**Results.** In Table 3 we observe that ADaPT improves success rates for games with recipe depth of 2 from 26.9% to 78.2%, and of depth 3 from 1.8% to 38.7% as compared to the ReAct baseline. As expected, the executor alone is unable to handle complex recipes with depth  $\geq 3$ , but with the help of ADaPT the performance improves significantly. Additionally, given the same budget  $d_{\max} = 4$ , as the recipe depth (complexity) increases from 2 to 3, ADaPT’s level of decomposition ( $k_{\max}$ ) also increases from 1.9 to 2.8. This showcases that ADaPT leverages as-needed decomposition in order to handle task complexity.

### 6.4 Can we use different planner and executor LLMs within ADaPT?

**Setup.** The planner and executor modules of ADaPT do not need to necessarily use the same underlying model. Following, Lin et al. (2023) we explore if a relatively smaller LLM can be used to perform local actions in the executor and a more

<sup>6</sup>As we have only 11 tasks with recipe depth of 4, we exclude them from this analysis.

Executor LM	Planner LM	Success Rate
GPT-3.5	—	38.4
GPT-3.5	GPT-3.5	<b>58.3</b>
LLaMA-2-70B	—	20.4
LLaMA-2-70B	GPT-3.5	<b>43.3</b>

Table 4: ADAPT improves performance on ALFWorld (dev) when using different planner and executor LLMs.

advanced LLM be used to devise plans. To this end, we explore different combinations of planner and executor LLM, with the latter using both gated and open-source models on ALFWorld.

**Results.** Table 4 shows that ADAPT can successfully be used to generate plans from one LLM that are useful to a different, possibly smaller, executor LLM, improving success rates by up to 19.9% compared to the executor-only (ReAct) setting. Interestingly, using an open-source model, such as LLaMA-2-70B-chat (Touvron et al., 2023) can be used as an executor with a more advanced LLMs such as GPT-3.5 to improve success rates by 22.9% points. Since the planner LLM is used sparingly, open-source executors can dramatically decrease the monetary or computational costs of using ADAPT. We defer combining knowledge from stronger and weaker LMs within ADAPT to future work, as examined in the context of mathematical reasoning (Fu et al., 2023; Saha et al., 2023a).

## 6.5 How does ADAPT compare to baselines in terms of LLM calls?

**Setup.** Performance of decision-making agents can be enhanced by increasing the number of calls allowed to an LLM, e.g., number of retrials in Reflexion. To verify that the gains in ADAPT are not simply due to higher number of LLM calls, we compare the average of number of LLM calls made by ADAPT to the baselines.

**Results.** Fig. 7 shows that a ADAPT employs a comparable number of LLM calls w.r.t. Try-Again and Reflexion baselines in order to yield performance improvements discussed in Sec. 5 (Tables 1 and 2). Note that while all methods including ReAct and Plan-and-Execute baselines are offered a comparable computational budget, the actual number of LLM calls used by the latter is often lower due to their inability to handle intermediate execution failures. This strengthens the argument for effectiveness of ADAPT as the improvements do

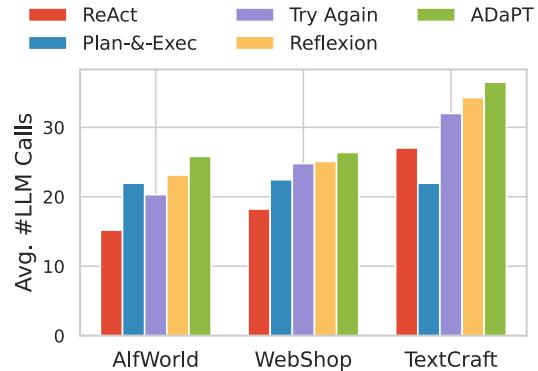


Figure 7: Average number of LLM calls for each approach including ADAPT and baselines discussed in Sec. 4.2 with GPT-3.5 LLM across datasets.

not simply stem from using substantially higher number of calls to the LLM.

## 7 Conclusion

We introduce ADAPT, a recursive algorithm designed to harness the planning capabilities of LLMs, dynamically decomposing complex tasks when the LLM acting as an executor encounters challenges. Our evaluation across three diverse decision-making tasks, ALFWorld, WebShop, and TextCraft, reveals impressive performance of ADAPT, surpassing existing baselines by substantial margins of up to 28.3%, 27%, and 33% points, respectively. This not only underscores the effectiveness of ADAPT but also highlights the significance of as-needed decomposition in enhancing task performance. Moreover, our findings demonstrate that ADAPT not only adapts to the capabilities of the underlying executor LLM but also takes into account the complexity of individual task instances, showcasing its versatility and effectiveness.

## Acknowledgements

Part of this work was done during internship at AI2 and was partially supported at UNC by NSF-CAREER Award 1846185, NSF-AI Engage Institute DRL-2112635, DARPA Machine Commonsense (MCS) Grant N66001-19-2-4031,. We sincerely thank Bodhisattwa Prasad Majumder, Chris Callison-Burch, Shashank Gupta, Peter Jansen, Bill Yuchen Lin and the Aristo team for their valuable feedback. We also thank Swarnadeep Saha, Elias Stengel-Eskin, and Peter Hase for their feedback.

## Limitations

ADAPT relies on the success heuristic generated by the executor LLM to determine if the model is capable of performing a complex task. For decision-making tasks studied in this work, we find that LLMs can reliably determine task success based on past action trajectories and textual feedback from the environment (see Appendix F). However, Huang et al. (2023a); Stechly et al. (2023) discuss the limits of LLM’s ability to self-evaluate and self-refine. In such situations, future works may additionally employ external verifiers (Lightman et al., 2023; Shridhar et al., 2023), theory-of-mind strategies among multiple LMs (Saha et al., 2023a), and other calibration and self-evaluation techniques (Kadavath et al., 2022). These improved self-evaluation techniques could be useful to extend our framework to non-decision making tasks such as question answering.

## References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Andrew G Barto and Sridhar Mahadevan. 2003. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1-2):41–77.

Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. 2022. A persistent spatial semantic representation for high-level natural language instruction execution. In *Conference on Robot Learning*, pages 706–717. PMLR.

Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-Dickstein, et al. 2022. Language model cascades. *arXiv preprint arXiv:2207.10342*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Kutluhan Erol, James Hendler, and Dana S Nau. 1994. Htn planning: Complexity and expressivity. In *AAAI*, volume 94, pages 1123–1128.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Ilche Georgievski and Marco Aiello. 2014. An overview of hierarchical task network planning. *arXiv preprint arXiv:1403.7426*.

Malik Ghallab, Dana Nau, and Paolo Traverso. 2004. *Automated Planning: theory and practice*. Elsevier.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2019. Neural module networks for reasoning over text. In *International Conference on Learning Representations*.

Daniel Höller, Gregor Behnke, Pascal Bercher, Susanne Biundo, Humbert Fiorino, Damien Pellier, and Ron Alford. 2020. Hddl: An extension to pdll for expressing hierarchical planning problems. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9883–9891.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2023b. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR.

Yichen Jiang and Mohit Bansal. 2019. [Self-assembling modular networks for interpretable multi-hop reasoning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4474–4484, Hong Kong, China. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. [Text modular networks: Learning to decompose tasks in the language of existing models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1264–1279, Online. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. [Decomposed prompting: A modular approach for solving complex tasks](#). In *The Eleventh International Conference on Learning Representations*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.

Bill Yuchen Lin, Yicheng Fu, Karina Yang, Prithviraj Ammanabrolu, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2023. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *arXiv preprint arXiv:2305.17390*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hanneh Hajishirzi. 2019. [Multi-hop reading comprehension through question decomposition and rescoring](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

So Yeon Min, Devendra Singh Chaplot, Pradeep Kumar Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. 2022. Film: Following instructions in language with modular methods. In *International Conference on Learning Representations*.

Vijayaraghavan Murali, Letao Qi, Swarat Chaudhuri, and Chris Jermaine. 2018. [Neural sketch learning for conditional program generation](#). In *International Conference on Learning Representations*.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. 2018. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31.

Maxwell Nye, Luke Hewitt, Joshua Tenenbaum, and Armando Solar-Lezama. 2019. Learning to infer program sketches. In *International Conference on Machine Learning*, pages 4861–4870. PMLR.

OpenAI. 2023. [Gpt-4 technical report](#).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. [Unsupervised question decomposition for question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online. Association for Computational Linguistics.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Desik Rengarajan, Gargi Vaidya, Akshay Sarvesh, Dileep Kalathil, and Srinivas Shakkottai. 2022. [Reinforcement learning with sparse rewards using guidance from offline demonstration](#). In *International Conference on Learning Representations*.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023a. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *arXiv preprint arXiv:2306.09299*.

Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2023b. [Summarization programs: Interpretable abstractive summarization with neural modular trees](#). In *The Eleventh International Conference on Learning Representations*.

Imanol Schlag, Sainbayar Sukhbaatar, Asli Celikyilmaz, Wen-tau Yih, Jason Weston, Jürgen Schmidhuber, and Xian Li. 2023. Large language model programs. *arXiv preprint arXiv:2305.05364*.

Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. 2022. [Skill induction and planning with latent language](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1713–1726, Dublin, Ireland. Association for Computational Linguistics.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 89–97.

Freida Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Huai hsin Chi, Nathanael Scharli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *International Conference on Machine Learning*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#). *arXiv preprint arXiv:2303.11366*, 14.

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ram Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. 2023. [The art of Ilm refinement: Ask, refine, and trust](#). *arXiv preprint arXiv:2311.07961*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. [Alfred: A benchmark for interpreting grounded instructions for everyday tasks](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. [ALFWorld: Aligning Text and Embodied Environments for Interactive Learning](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. [Prog-prompt: Generating situated robot task plans using large language models](#). In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. [Llm-planner: Few-shot grounded planning for embodied agents with large language models](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. [Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems](#). *arXiv preprint arXiv:2310.12397*.

Simeng Sun, Y. Liu, Shuo Wang, Chenguang Zhu, and Mohit Iyyer. 2023. [Pearl: Prompting large language models to plan and execute actions over long documents](#). *ArXiv*, abs/2305.14564.

Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

Alon Talmor and Jonathan Berant. 2018. [The web as a knowledge-base for answering complex questions](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. [Voyager: An open-ended embodied agent with large language models](#). *arXiv preprint arXiv:2305.16291*.

Lei Wang, Wanyu Xu, Yihui Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pieric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. [Huggingface's transformers: State-of-the-art natural language processing](#). *arXiv preprint arXiv:1910.03771*.

Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, Zhoujun Cheng, Siheng Zhao, Lingpeng Kong, Bailin Wang, Caiming Xiong, and Tao Yu. 2023. [Lemur: Harmonizing natural language and code for language agents](#).

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023. [Intercode: Standardizing and benchmarking interactive coding with execution feedback](#). *arXiv preprint arXiv:2306.14898*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. **React: Synergizing reasoning and acting in language models**. In *The Eleventh International Conference on Learning Representations*.

Jesse Zhang, Haonan Yu, and Wei Xu. 2021. Hierarchical reinforcement learning by discovering intrinsic options. In *International Conference on Learning Representations*.

Wenqing Zheng, SP Sharan, Ajay Kumar Jaiswal, Kevin Wang, Yihan Xi, Dejia Xu, and Zhangyang Wang. 2023. Outline, then details: Syntactically guided coarse-to-fine code generation. In *International Conference on Machine Learning*, pages 42403–42419. PMLR.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2023. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*.

## A ADAPT Implementation Details

**Executor.** We use a common ReAct executor for each dataset. To this end, we provide the LLM in the executor with in-context example trajectories for each atomic skill (refer to Table 5 for an exhaustive list). Atomic skills are inherently task dependent, and thus, vary with the underlying environment. For ALFWorld, in which the agent needs to navigate and perform tasks in the household, the atomic skills include: taking an object, putting it down at a location, cleaning, heating, etc. On the other hand, the goal in WebShop is to buy a product based on user queries, thus, atomic skills include: searching a specified query, shortlisting products based on search page, matching if a product satisfies a criteria, and buying a product. Lastly, the atomic skills in TextCraft are fetching objects from the environment, and crafting them given the recipe and the ingredients. Following Yao et al. (2023b), we add gold trajectories for two tasks: heat and look in the executor prompt for ALFWorld, and one full gold trajectory for TextCraft.

	Atomic Skill	Description
ALFWorld	put	Assuming that the robot is carrying an object, put it on a given receptacle.
	take	Take a specified object from a specified receptacle.
	clean/heat/cool	Assuming that the robot is carrying an object, clean/heat/cool the object.
	examine	Assuming the robot is at a desk with a desk lamp, use it to look at an object.
WebShop	search	Put a given query in the search box, results in a page with list of products.
	shortlist	Based on the search page and query, get list of any matching products.
	match	Given a product ID and query, navigate to the product page and verify it matches the query.
	buy	Given a product ID and query, buy product by selecting relevant options.
TextCraft	craft	Assuming the agent has all the ingredients in the inventory, craft a target object by picking an appropriate command from the list of crafting recipes.
	fetch	Look for a given object in the inventory or get it directly from the game.
	inventory	Look-up the game inventory.

Table 5: Overview of atomic skills used in Sec. 3.1.

**Planner.** We provide the LLM with a brief description of atomic skills and in-context demonstrations of few task decompositions for each dataset.

- **ALFWorld:** The planner includes 6 demonstrations of task decompositions for one household configuration. Specifically, “*find*” is not an atomic skill for the executor, and therefore, needs to be handled by the planner (refer to Fig. 2).
- **WebShop:** The planner breaks down a given task in terms of the atomic skills described in Table 5 via 2 in-context demonstrations.
- **TextCraft:** The planner determines the necessary ingredients for each item and creates a plan to obtain them and then craft the item, illustrated via 2 examples with different crafting commands.

**Controller.** The controller performs two crucial roles in the overall functioning of ADAPT. First, it serves as the *communication bridge* between planner and executor, propagating salient information across the two depending on the task. Second, since ADAPT is a recursive algorithm, the controller determines the *termination criterion* using the logical expression from the planner and success heuristic from the executor or if a maximum depth  $d_{\max} (\geq 1)$  is reached. The controller propagates task-dependent salient information described below:

- **ALFWorld:** In the controller, we propagate the last successful action from a previous execution

Method	Pick	Clean	Heat	Cool	Look	Pick2	All
ReAct	66.7	41.9	47.8	80.9	83.3	23.5	56.7
Plan-and-Execute	87.5	58.1	73.9	52.4	83.3	17.6	63.4
Try Again with ReAct	75.0	38.7	60.9	76.2	66.7	23.5	56.7
Reflexion	83.3	61.3	73.9	85.7	61.1	29.4	67.2
ADAPT (Ours)	<b>91.7</b>	<b>67.7</b>	<b>78.3</b>	<u>81.0</u>	<b>100</b>	<b>64.7</b>	<b>79.8</b>

Table 6: Comparison of success rates (%) achieved by ADAPT and other baselines from prior work on ALFWorld (test split) with executor used by Yao et al. (2023b)

### Algorithm 1 Algorithm for ADAPT

```

1: function ADAPT(Task  $T$ , Current depth  $k$ )
2:   // ADAPT(.) Generates success heuristic value completed for the task  $T$ . Initialized with  $k = 1$ .
3:   // Base case: terminate on reaching maximum depth
4:   if  $k > d_{\max}$  then return False
5:   // Execute the task/sub-task to assess if the LLM can directly perform it using LLM-generated success.
6:    $completed \leftarrow \text{executor}_{\text{LLM}}(T)$ 
7:   // Plan only when the executor fails.
8:   if  $completed$  is False then
9:     // Using the LLM, decompose the task into a set of sub-tasks,  $\mathcal{P}$ , and a Boolean function, logic(.), that combines output of the sub-tasks.
10:     $\mathcal{P}, logic \leftarrow \text{planner}_{\text{LLM}}(T)$ 
11:    // Get the outputs for individual sub tasks
12:     $\mathcal{O} = \{\text{ADAPT}(T_{\text{sub}}, k+1) | T_{\text{sub}} \in \mathcal{P}\}$ 
13:    // Combine the outputs of the sub tasks
14:     $completed \leftarrow logic(\mathcal{O})$ 
15:   return  $completed$ 

```

run to subsequent calls of the executor. Note that information is only propagated from successful sub-tasks. For sub-tasks connected via “OR”, each receives the same information from the controller. Unlike Shinn et al. (2023), executor does not get text feedback from prior failures.

- **WebShop:** We propagate the current page visible to the agent along with past unsuccessful executor tasks to the planner (without any rationales). Once we find a matching product, we also propagate the product ID in future executor calls.
- **TextCraft:** We propagate the current inventory of the agent to the executor. This is akin to executors starting with the inventory command as the first step to keep stock of which items are missing and need to be fetched or crafted.

For partial rolled-out trajectories with ADAPT refer to Figs. 9 to 11. Communication between planner and executor is highlighted in gray box(es).

**LLM-related Hyperparameters.** Following previous works (Shinn et al., 2023; Liu et al., 2023) we use text-davinci-003 from the OpenAI API for ALFWorld. For WebShop, we use the gpt-3.5-turbo models, and for TextCraft we use

Method	Score	Success Rate
Iterative Executor-Only	42.1	29.0
Static Decomposition	27.7	17.0
Retry Execution	45.4	30.0
Naive	58.3	24.0
Reflexion*	64.2	35.0
LATS (Zhou et al., 2023)*	75.9	38.0
ADAPT (Ours)	60.0	<b>44.0</b>

Table 7: Performance comparison of different methods on WebShop.

the gpt-3.5-turbo-instruct models. All executors have a maximum budget of iterations to interact with the environment and execute the task. We set this budget to 20, 15, and 20 respectively for ALFWorld, WebShop, and TextCraft respectively. For try again with ReAct, we sample additional trajectories with a temperature of 0.7. As discussed in Sec. 4.2, we run the iterative executor-only baseline for 60, 45, 60 iterations for ALFWorld, WebShop, and TextCraft respectively. In Sec. 6.2, we use publicly available checkpoints for LLaMA 70B<sup>7</sup> and Lemur 70B<sup>8</sup> available on Huggingface (Wolf et al., 2019). For both planner and executor modules, we use a fixed prompt consisting of few in-context examples (as described above) for each dataset. We show all executor and planner prompts to the LLM in Appendix G. Due to cost constraints, we report success rates for a single run of each LLM in Secs. 5 and 6.

## B Handling Complex Logic in Plans

While the examples in Figs. 1 and 2 show homogeneous logic across sub-tasks in the plan, our controller can handle complex logical expressions including both “AND” and “OR” operators. Specifically, we provide instructions to the planner to output this logical expressing at the end of the plan with a fixed prefix: Execution Order. We then build a deterministic parser that can parse complex logical expressions that the controller can process. We do so by splitting the logical expression into a series of homogeneous expression each passed to ADAPT. Whenever the task given to ADAPT comprises of multiple sub-tasks connected via (one) logical operator, we automatically decompose this task as per the logical expression. For example, in Fig. 8, a detailed plans used by the plan-and-execute baseline (discussed in Sec. 4.2) comprised

<sup>7</sup><https://huggingface.co/meta-llama/Llama-2-70b-hf>

<sup>8</sup><https://huggingface.co/OpenLemur/lemon-70b-chat-v1>

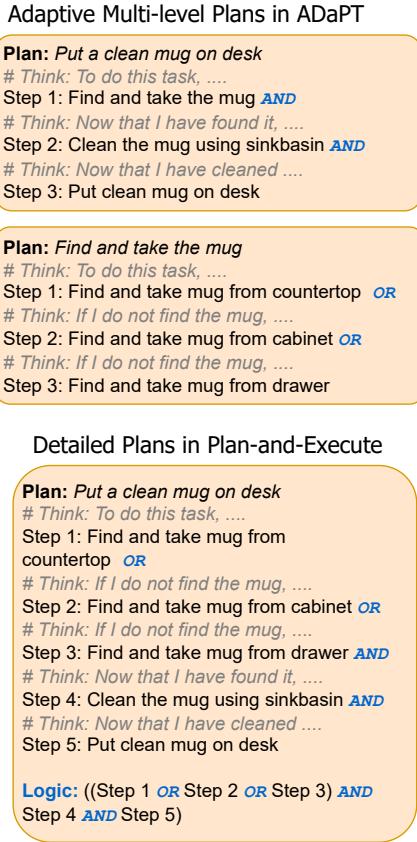


Figure 8: Illustration of how multiple levels of plans from ADAPT, can be collapsed into one detailed plan in non-adaptive settings as used in the plan-and-execute baseline (Sec. 4.2). Our controller can handle complex (non-homogeneous) logical expressions.

of logical expressions using both AND, and OR operators. Therefore, the parser will break automatically break this into multiple levels, i.e., Step 6 = Step 1 OR Step 2 OR Step 3, followed by Step 6 AND Step 4 AND Step 5. While such complex logical expressions are mostly associated with the plan-and-execute baseline, they can be easily used within the ADAPT framework. Furthermore, this allows the plan-and-execute baseline to simulate a multi-level planning structure via detailed plans without being adaptive to the executor.

## C Task-specific Executors in ALFWorld

In Table 1, we use a standardized executor with in-context demonstrations of atomic skills and two gold trajectories. While this allows for a common executor across different sub-tasks, task-specific executors yield higher performance on the specific sub-tasks. We now show ADAPT can also be used on top of task-specific executors used by Yao et al. (2023b). The results are shown in Table 6. First,

Method	#Products	Success Rate
ReAct	3	27.5
ADAPT ( $d_{\max} = 3$ )	3	<b>47.5</b>
ReAct	10	20.0
ADAPT ( $d_{\max} = 3$ )	10	<b>42.5</b>

Table 8: ADAPT improves WebShop (dev) performance irrespective of how many products (3 or 10) are chosen from the search page.

we observe that ADAPT yields the overall success rate by up to 23.1% points and also surpasses baselines on all but 1 task types. Interestingly, we find strong performance of the plan-and-execute baseline when using a stronger executor (as compared to Table 1) possibly as such an executor can handle complex sub-tasks better. Consistent with Table 1, ADAPT outperforms Reflexion by 12.6% points despite lack of dedicated memory and natural language feedback.

## D Additional WebShop Experiments

**Evaluation Metrics.** We focus on success rate and not the (soft) score as the primary metric for this task because it is possible to get a non-zero score by naively buying a product. To this effect, we construct a naive executor that inputs the user query in the search bar and buys the first available product. Table 7 shows that while this baseline yields the lowest success rate, it surprisingly yields a high success rate of 58.3. In contrast, our executors often do not buy products especially when the previous sub-goals fail which can adversely impact scores even though the success rate remains unaffected. Therefore, we argue for optimizing the success rate instead of the score as opposed to prior works (Zhou et al., 2023).

**ADAPT accommodating task complexity.** By default, Yao et al. (2023b) use a search page with only the top-3 search results displayed. Intuitively, increasing the number of products on the search page requires the model to choose from a wider array of products and track all their information to determine the best fit to the user query, making the overall task harder. Therefore, we apply ADAPT on Webshop in two settings with 3, and 10 products per search page.

**Results.** From Table 8, we observe that ADAPT effectively improves success rate by 20.0% and 22.5% for 3 and 10 products respectively over the

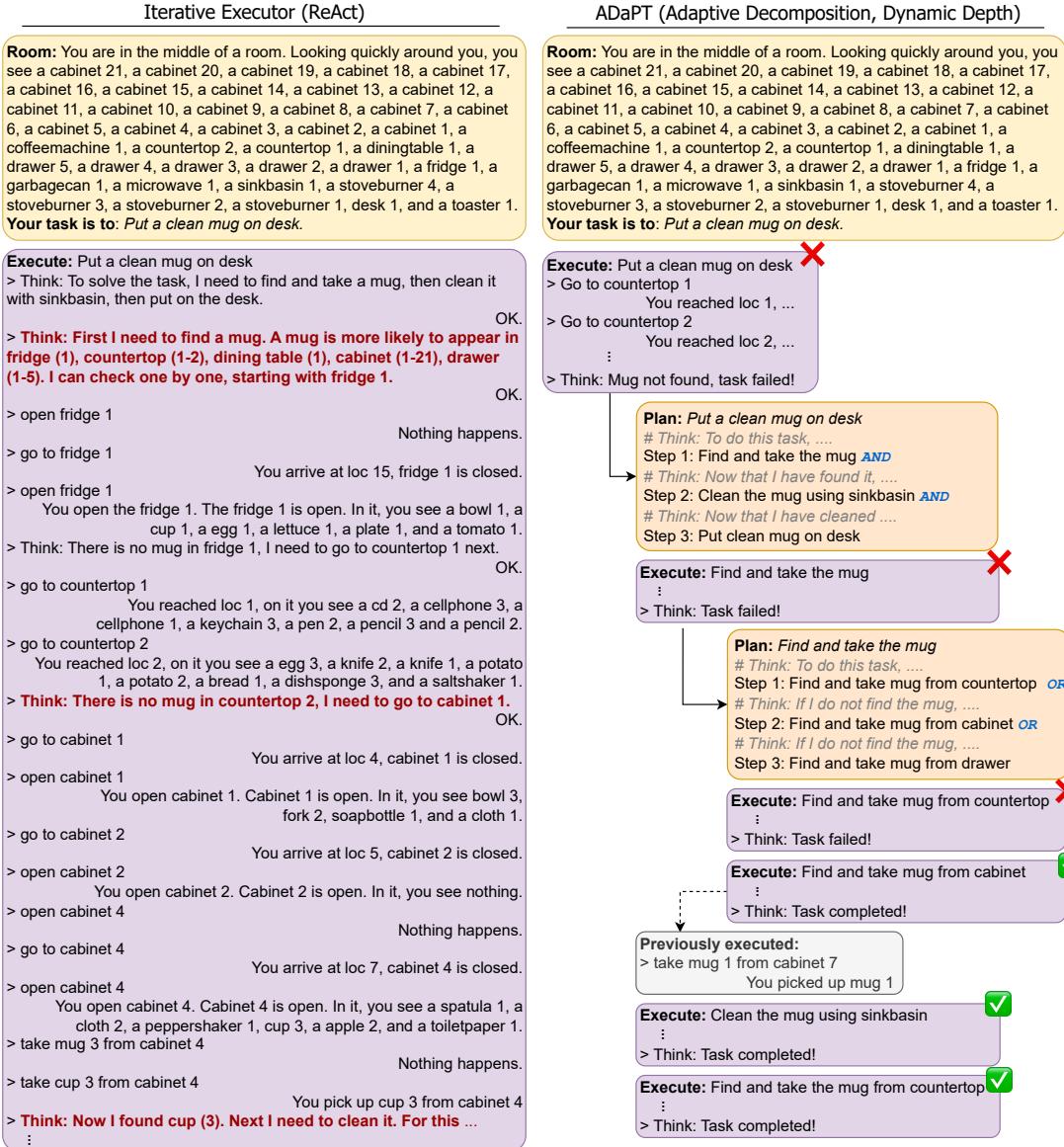


Figure 9: Comparison of iterative executors such as ReAct with ADAPT. On left, ReAct uses interleaved “thought” statements to set milestones and track their progress. However, due to a large action history, it struggles to follow the plan exactly and hallucinates the wrong object (highlighted in red). ADAPT, on the right, decomposes complex tasks into smaller sub-tasks whenever the executor fails, leading to shorter action trajectories for easy execution.

ReAct baseline. The difference in ReAct performance for both settings corroborates our hypothesis that increasing number of products on the search page increases task complexity, all else equal. Notably, we show that ADAPT yields *higher* improvement for *more complex* task settings.

## E TextCraft

**TextCraft: Environment Details.** In TextCraft, the objective is to obtain target Minecraft items by crafting them from available items in the environment. We define an environment with three ac-

tions: `craft <item> using <ingredients>`, `get <item>`, and `inventory`. We utilize Minecraft’s crafting recipes to specify craftable items and their ingredients, assuming that all other items are obtainable from the environment. Similar to Alf-World, our agent can directly execute these operations in the embodied game. The game begins with a list of crafting commands provided to the agent that detail recipes that can be used to craft the final target, its ingredients along with some distractors (details in Appendix E). A reward of 1 is generated when the target item gets added to the

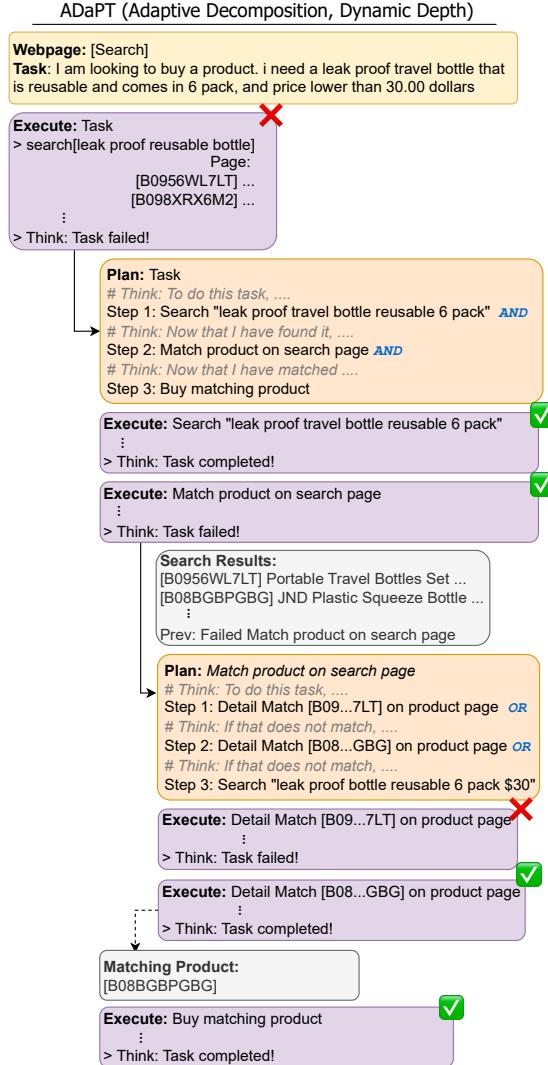


Figure 10: Partial rolled out trajectories for WebShop with ADaPT. In the gray box we communicate to the planner the current (search) page that is visible to the agent, and once a matching product is found, we propagate it to future executor runs. Note “match on search page” corresponds to shortlist skill in Table 5, and “detail match on product page” corresponds to match skill.

agent’s inventory. An illustrative gold trajectory from TextCraft is shown in Fig. 3.

We create the TextCraft environment using Minecraft v1.16.5 recipes. We only consider the recipes craftable using a crafting table. We consider both shapeless (only count matters) and shaped (position of ingredients matters) recipes and convert them into crafting commands (e.g. craft 4 sticks using 2 planks). Items that do not have any recipe are considered obtainable via the get command, e.g. get 4 diamond.

Since the entire set of crafting commands would not fit in the context of modern LLMs, we create

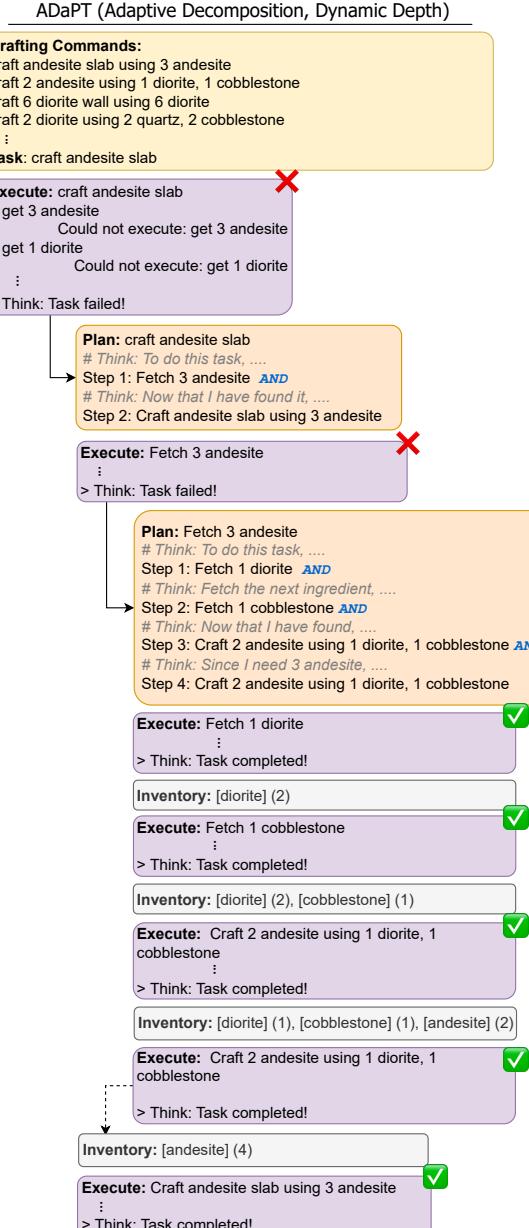


Figure 11: Partial rolled out trajectories for TextCraft using ADaPT. In the gray box, we propagate the inventory of the agent to subsequent executor calls. Note that while “diorite” is not directly present in the environment, i.e., it needs to be crafted. The executor LLM is able to inherently compose skills to fetch it without further decomposition.

a set of relevant crafting commands for every task. Apart from the set of gold crafting commands (i.e, crafting commands for all the items in the recipe tree), we also add up to 10 distractor commands. To create this distractor set, we sub-sample up to 10 recipes for every ingredient in the recipes of our gold recipe tree. We finally sub-sample up to 10 distractors from this entire set to ensure a reason-



Figure 12: Comparison of LLM-generated success heuristic with gold environment rewards to compute success rates for all datasets.

able context size. Note that we do not provide the list of valid `get` commands as that can be inferred from the `craft` commands.

## F Evaluation of Success Heuristic

In Sec. 3.1, we describe the executor module used in ADAPT. For tasks assigned to the executor, we prompt the LLM to generate a binary success heuristic. We use this heuristic repeatedly to evaluate if the (sub-)task needs to be decomposed further. We now study the ability of LLMs to generate this success heuristic on all our datasets. To this end, we run ADAPT and in the end compare the success rate when using the LLM’s self-assessed task success with the gold reward from the environment in Fig. 12. On ALFWorld and TextCraft, we find the LLM slightly over-estimates its overall task success. This is to be expected as the underlying tasks involve minimal subjectivity (e.g., the agent either has an item on its inventory or not). However, on WebShop, where a product can match the user criteria to different degrees (partially or fully), we find that the LLM’s assessment is significantly inflated compared to the environment reward ( $> 30$  points). This imperfect feedback affects downstream performance of ADAPT, as the algorithm terminates even though further decomposition is needed. We leave it to future work to address the shortcomings of self-evaluation with LLMs (Huang et al., 2023a; Stechly et al., 2023).

## G Prompts

We provide all the prompts used in our planner and executor modules for ALFWorld, WebShop, and TextCraft datasets in the following pages.

## ALFWorld Hybrid Executor Prompt

Here is a demo of actions you can perform.

You are in the middle of a room. Looking quickly around you, you see a desk 1, microwave 1, a cabinet 3, a cabinet 9, a drawer 2, a coffeemachine 1, a stoveburner 4, a drawer 5, a cabinet 11, a drawer 3, a stoveburner 1, a drawer 1, a toaster 1, a fridge 1, a stoveburner 2, a cabinet 6, a cabinet 10, a countertop 1, a cabinet 13, a cabinet 7, a garbagecan 1, a cabinet 2, a cabinet 8, a cabinet 12, a drawer 4, a cabinet 1, a sinkbasin 1, a cabinet 5, a stoveburner 3, and a cabinet 4.

Your task is to: Put a tomato in cabinet 1.

> think: Input assumption: I am carrying a tomato. Now I need to verify this.

OK.

> inventory

You are carrying: a tomato 1.

> examine tomato 1

There's nothing special about tomato 1.

> think: Based on previous context assumption is true, now I can proceed.

> go to cabinet 1

The cabinet 1 is closed.

> open cabinet 1

You open the cabinet 1. The cabinet 1 is open. In it, you see a candle 1, and a spraybottle 2.

> put tomato 1 in/on cabinet 1

You put the tomato 1 in/on the cabinet 1.

> think: End state: Now I put tomato (1) in cabinet (1). Task completed!

Your task is to: Take a mug from cabinet 5

> think: Input assumption: None

OK.

> go to cabinet 5

The cabinet 5 is closed.

> open cabinet 5

You open the cabinet 5. The cabinet 5 is open. In it, you see a mug 3.

> take mug 3 from cabinet 5

You pick up the mug 3 from the cabinet 5.

> think: End state: Now I take a mug (3) from cabinet (5). Task completed!

Your task is to: Heat an egg.

> think: Input assumption: I am carrying an egg. Now I need to verify this.

OK.

> inventory

You are carrying: a egg 2.

> examine egg 2

There's nothing special about egg 2.

> think: Based on previous context assumption is true, now I can proceed.

OK.

> go to microwave 1

The microwave 1 is closed.

> heat egg 2 with microwave 1

You heat the egg 2 using the microwave 1.

> think: End state: Now I heat a egg (2). Task completed!

Your task is to: Cool a mug.

> think: Input assumption: I am carrying a mug. Now I need to verify this.

OK.

> inventory

You are carrying: a mug 3.

> examine mug 3

There's nothing special about mug 3.

> think: Based on previous context assumption is true, now I can proceed.

OK.

> go to fridge 1

The fridge 1 is closed.

> cool mug 3 with fridge 1

You cool the mug 3 using the fridge 1.

> think: End state: Now I cool a mug (3). Task completed!

Your task is to: Clean a lettuce.

> think: Input assumption: I am carrying a lettuce. Now I need to verify this.

OK.

> inventory

You are carrying: a lettuce 1.

> examine lettuce 1

There's nothing special about lettuce 1.

> think: Based on previous context assumption is true, now I can proceed.

OK.

> go to sinkbasin 1

On the sinkbasin 1, you see a apple 2, a ladle 2, a spoon 1, and a tomato 3.

> clean lettuce 1 with sinkbasin 1

You clean the lettuce 1 using the sinkbasin 1.

> think: End state: Now I clean a lettuce (1). Task completed!

### ALFWorld Hybrid Executor Prompt (cont.)

Your task is to: Use a desklamp to look at tomato 1.

> think: Input assumption: I am at the location that contains desklamp (1). Now I need to verify this.

OK.

> inventory

You are carrying: a tomato 1.

> look

You are facing desk 1. Next to it, you see nothing.

> examine desk 1

On desk 1, you see a spoon 1, a desklamp 1, and a pencil 2.

> think: Based on previous context assumption is true, now I can proceed.

OK.

> use desklamp 1

You turn on the desklamp 1.

> think: Now I use desklamp (1). Task completed!

Here are two example tasks.

You are in the middle of a room. Looking quickly around you, you see a cabinet 13, a cabinet 12, a cabinet 11, a cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeeemachine 1, a countertop 1, a diningtable 1, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a shelf 3, a shelf 2, a shelf 1, a sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.

Your task is to: put a hot apple in fridge.

> think: To solve the task, I need to find and take an apple, then heat it with microwave, then put it in fridge.

OK.

> think: First I need to find an apple. An apple is more likely to appear in fridge (1), diningtable (1), coffeetable (1), drawer (1), cabinet (1-13), garbagecan (1). I can check one by one, starting with fridge 1.

OK.

> go to fridge 1

The fridge 1 is closed.

> open fridge 1

You open the fridge 1. The fridge 1 is open. In it, you see a cup 1, and a egg 1.

> go to diningtable 1

On the diningtable 1, you see a apple 1, a bread 1, a fork 3, a fork 2, a fork 1, a mug 2, a peppershaker 3, a plate 2, a pot 1, a soapbottle 3, a spatula 1, a spoon 3, a spoon 2, a spoon 1, a winebottle 3, and a winebottle 2.

> think: Now I find an apple (1). Next, I need to take it.

OK.

> take apple 1 from diningtable 1

You pick up the apple 1 from the diningtable 1.

> think: Now I take an apple (1). Next, I need to go to a microwave (1) and heat it.

OK.

> go to microwave 1

The microwave 1 is closed.

> heat apple 1 with microwave 1

You heat the apple 1 using the microwave 1.

> think: Now I heat an apple (1). Next, I need to put it in/on fridge 1.

OK.

> go to fridge 1

The fridge 1 is open. In it, you see a cup 1, and a egg 1.

> put apple 1 in/on fridge 1

You put the apple 1 in/on the fridge 1.

> think: Task completed!

You are in the middle of a room. Looking quickly around you, you see a bed 1, a desk 1, a drawer 3, a drawer 2, a drawer 1, a garbagecan 1, a safe 1, a shelf 5, a shelf 4, a shelf 3, a shelf 2, a shelf 1, a sidetable 2, and a sidetable 1.

Your task is to: look at bowl under the desklamp.

> think: To solve the task, I need to find and take a bowl, then find and use a desklamp.

OK.

> think: First I need to find a bowl. A bowl is more likely to appear in drawer shelf (1-5), drawer (1-3). I can check one by one, starting with drawer 1.

OK.

> go to shelf 1

On the shelf 1, you see nothing.

> go to shelf 2

On the shelf 2, you see a bowl 1.

> think: Now I find a bowl (1). Next, I need to take it.

OK.

> take bowl 1 from shelf 2

You pick up the bowl 1 from the shelf 2.

> think: Now I take a bowl (1). Next, I need to find a desklamp. A desklamp is more likely to appear in desk (1), sidetable (1-2), and drawer (1-3). I can check one by one, starting with desk 1.

OK.

> go to desk 1

On the desk 1, you see a alarmclock 2, a book 1, a cellphone 1, a keychain 1, a laptop 2, a laptop 1, and a pen 3.

> go to sidetable 1

On the sidetable 1, you see a cd 1, a pen 1, and a pencil 1.

> go to sidetable 2

On the sidetable 2, you see a alarmclock 1, a desklamp 1, and a pen 2.

> think: Now I find a desklamp (1). Next, I need to use it.

OK.

> use desklamp 1

You turn on the desklamp 1.

> think: Task completed!

## ALFWORLD Planner Prompt

Here are some examples.

You are in the middle of a room. Looking quickly around you, you see a desk 1, microwave 1, a cabinet 3, a cabinet 9, a drawer 2, a coffeemachine 1, a stoveburner 4, a drawer 5, a cabinet 11, a drawer 3, a stoveburner 1, a drawer 1, a toaster 1, a fridge 1, a stoveburner 2, a cabinet 6, a cabinet 10, a countertop 1, a cabinet 13, a cabinet 7, a garbagecan 1, a cabinet 2, a cabinet 8, a cabinet 12, a drawer 4, a cabinet 1, a sinkbasin 1, a cabinet 5, a stoveburner 3, and a cabinet 4.

Goal: Put a mug in/on desk.

Come up with an abstract plan to perform this task in a couple of steps.

# Think: To perform this task, I need to find and take mug and then put it on desk. First, I will focus on finding mug.

Step 1: Find and take mug

# Think: Now that I am carrying mug, I will focus on putting it in/on desk.

Step 2: Put mug in/on desk

Execution Order: (Step 1 AND Step 2)

Goal: Clean mug and put it in/on desk.

Come up with an abstract plan to perform this task in a couple of steps.

# Think: To perform this task, I need to find and take mug, clean it, and then put it on desk. First, I will focus on finding mug.

Step 1: Find and take mug

# Think: Now that I am carrying mug, I will focus on cleaning it.

Step 2: Clean mug with sinkbasin

# Think: Now that I have cleaned mug, I will focus on putting it in/on desk.

Step 3: Put cleaned mug in/on desk

Execution Order: (Step 1 AND Step 2 AND Step 3)

Goal: Cool mug and put it in/on desk.

Come up with an abstract plan to perform this task in a couple of steps.

# Think: To perform this task, I need to find and take mug, cool it, and then put it on desk. First, I will focus on finding mug.

Step 1: Find and take mug

# Think: Now that I am carrying mug, I will focus on cooling it.

Step 2: Cool mug with fridge

# Think: Now that I have cooled mug, I will focus on putting it in/on desk.

Step 3: Put cooled mug in/on desk

Execution Order: (Step 1 AND Step 2 AND Step 3)

Goal: Heat mug and put it in/on desk.

Come up with an abstract plan to perform this task in a couple of steps.

# Think: To perform this task, I need to find and take mug, heat it, and then put it on desk. First, I will focus on finding mug.

Step 1: Find and take mug

# Think: Now that I am carrying mug, I will focus on heating it.

Step 2: Heat mug with microwave

# Think: Now that I have heated mug, I will focus on putting it in/on desk.

Step 3: Put heated mug in/on desk

Execution Order: (Step 1 AND Step 2 AND Step 3)

Goal: Look at mug under desklamp.

Come up with an abstract plan to perform this task in a couple of steps.

# Think: To perform this task, I need to find and take mug, and then go to the desklamp and use it. First, I will focus on finding mug.

Step 1: Find and take mug

# Think: Now that I have found and taken mug, I will focus on using the desklamp.

Step 2: Use the desklamp

Execution Order: (Step 1 AND Step 2)

Goal: Find and take mug

Come up with an abstract plan to perform this task in a couple of steps.

# Think: To perform this task I need to find mug in the room. mug is likely to be in desk, cabinet, countertop, or drawer. Now I will focus on finding mug in each of these locations one by one.

Step 1: Find and take mug from desk

# Think: If mug not found so far, I will next look in the cabinet.

Step 2: Find and take mug from cabinet

# Think: If mug not found so far, I will next look in the countertop.

Step 3: Find and take mug from countertop

# Think: If mug not found so far, I will next look in the drawer.

Step 4: Find and take mug from drawer

Execution Order: (Step 1 OR Step 2 OR Step 3 OR Step 4)

Here is the goal.

<room>

Goal: <task>.

Come up with an abstract plan to perform this task in a couple of steps. Constraints: The robot can hold/take/put only one object at a time to a location.

Ensure each step can be understood independently and mentions the name of object.

When stating the execution order, ensure that 'AND'/'OR' statements are properly nested using brackets '()'.

### WebShop Executor Prompt: Buy

Instruction: Buy product [B078GWRC1J] that matches criteria: 3 ounce bottle of bright citrus deodorant for sensitive skin, and price lower than 50.00 dollars

Action: think['I should be on the search page where I can see a list of products, I should verify that.']  
Observation: OK.

Action: load

Observation:

[Back to Search]

Page 1 (Total results: 50)

[Next >]

[B078GWRC1J]

Bright Citrus Deodorant by Earth Mama | Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce

\$10.99

[B08KBVJ4XN]

Barrel and Oak - Aluminum-Free Deodorant, Deodorant for Men, Essential Oil-Based Scent, 24-Hour Odor Protection, Cedar & Patchouli Blend, Gentle on Sensitive Skin (Mountain Sage, 2.7 oz, 2-Pack)

\$35.95

[B078GTVKXY]

Ginger Fresh Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce

\$10.99

Action: think['I can spot product [B078GWRC1J] and it is a clickable link since its enclosed in square brackets. I should click on it.']

Observation: OK.

Action: click[B078GWRC1J]

Observation:

[Back to Search]

[< Prev]

scent [assorted scents][bright citrus][calming lavender][ginger fresh][simply non-scents]

size [travel set (4-pack)][3 ounce (pack of 1)][3-ounce (2-pack)]

Bright Citrus Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce

Price: \$10.99

Rating: N.A.

[Description]

[Features]

[Reviews]

[Buy Now]

Action: think[I'm already on the product page. I see multiple possible options enclosed in '[]' for scent and size for this product. To match the input criteria, I should click on options '[bright citrus]' for scent and '[3 ounce (pack of 1)]' for size one by one and then buy in the end.]

Observation: OK.

Action: click[bright citrus]

Observation: You have clicked bright citrus.

Action: click[3 ounce (pack of 1)]

Observation: You have clicked 3 ounce (pack of 1).

Action: think[My task is to buy the product, for it should to click 'buy now']

Observation: OK.

Action: click[Buy Now]

Observation: You have clicked buy now.

Action: think[I finished buying the product. Task completed!]

Here is another task in which you need to buy a product. When you finish buying the product with the most relevant choices, use 'think[Task completed]'. If you cannot find the matching options or proceed, think['Task failed']. Note that you can only click on text enclosed in '[]' on the webpage. Everything else is only a description, not valid with the "click" action.

Instruction: Buy product [{} that matches the criteria: {}

### WebShop Executor Prompt: Match (cont.)

You are given a webpage of an item on an online shopping website and a criteria. Your task is to answer if the product on the page exactly matches the criteria. Not the criteria could have multiple requirements that should be checked one by one and all must satisfy for an exact match.

Here are a few examples:

Criteria: 3 ounce bottle of citrus deodorant for sensitive skin that is priced lower than \$30 and natural.

Item Page:

[Back to Search]

[< Prev]

scent [assorted scents][bright citrus][calming lavender][ginger fresh][simply non-scents]  
size [travel set (4-pack)][3 ounce (pack of 1)][3-ounce (2-pack)]

Bright Citrus Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce

Price: \$10.99

Rating: N.A.

[Description]

Features:

NEW from Earth Mama (formerly Earth Mama Angel Baby), formulated especially for pregnancy, breastfeeding and sensitive skin

Contains organic grapefruit, tangerine and calendula

NO propylene glycol, artificial fragrance, parabens or aluminum

Dermatologist tested and clinically tested for irritation

Better than natural organic! NSF/ANSI 305 Certified by Oregon Tilth

[Reviews]

[Attributes]

[Buy Now]

Answer: The product is available in 3 ounce size, is citrus and suitable for sensitive skin. It is also organic or natural.

Its price is \$10.99 which is less than \$30.

Thus, the answer is True (exact match).

Criteria: 3 ounce bottle of citrus deodorant for sensitive skin that is priced lower than \$30 and natural.

Item Page:

[Back to Search]

[< Prev]

size [3 ounce][3 ounce (pack of 1)]

unit count [2.0][3.0]

Barrel and Oak - Aluminum-Free Deodorant, Deodorant for Men, Essential Oil-Based Scent, 24-Hour Odor Protection, Cedar & Patchouli Blend, Gentle on Sensitive Skin (Mountain Sage, 2.7 oz, 2-Pack)

Price: \$15.95

Rating: N.A.

[Description]

Features:

About this item WHY ALUMINUM-FREE DEODORANT? Aluminum-free deodorants use more natural ingredients unlike antiperspirants, which use chemicals to block sweat. Safely fight odor for 24 hours with Barrel & Oak's deodorants our gentle formula is easy on sensitive skin. START SMELLING LIKE THE MAN YOU WANT TO BE: Our mountain sage aluminum-free men's deodorant is naturally fragranced with an outdoorsy scent of crisp conifer, sage, & citrus. Think sweet notes of citrus with earthy tones of cedar & patchouli. PREMIUM INGREDIENTS FOR NATURAL FRAGRANCES: Our deodorants for men are composed of natural, essential oil-based scents. These natural fragrance deodorants are more subtle than their synthetic counterparts, but they're better for you & the planet. DESIGNED FOR THE MODERN MAN: Barrel & Oak has a full spectrum of grooming & body care products that are designed with function, fragrance, & effective ingredients for the health-conscious & practical modern man. Give your body what it deserves. EARTH-FRIENDLY, YOU-FRIENDLY, WALLET-FRIENDLY: Our premium products for men are scented with natural fragrances & essential oils, free of parabens, phthalates, & SLS, packaged in recyclable materials, cruelty-free, & vegan or vegetarian.

[Reviews]

[Attributes]

[Buy Now]

Answer: The product is not citrus in nature. It does not match the criteria. It's price is \$15.95 which is less than \$30.

Thus, the answer is False (not an exact match).

Now here is the criteria and item page for the another task. Try you best to determine exact match, otherwise, respond with "False", i.e., no exact match. Generate an explanation before the answer to justify your decision.

Criteria: {}

Item Page:

{}

Answer:

### WebShop Executor Prompt: Shortlist (cont.)

You are given a search page on an online shopping site with a list of products along with name and price. Based on this information, your task is return a list of product IDs (enclosed in []) of all products that exactly match all requirements in the criteria. If the information provided is not enough to make a determination, return an empty list.

Here are a few examples.

Search Page:

[Back to Search]

Page 1 (Total results: 50)

[Next >]

[B078GWRC1J]

Bright Citrus Deodorant by Earth Mama | Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce

\$10.99

[B08KBVJ4XN]

Barrel and Oak - Aluminum-Free Deodorant, Deodorant for Men, Essential Oil-Based Scent, 24-Hour Odor Protection, Cedar & Patchouli Blend, Gentle on Sensitive Skin (Mountain Sage, 2.7 oz, 2-Pack)

\$35.95

[B078GTVKXY]

Ginger Fresh Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce

\$10.99

[B08SMG4WB9]

Each & Every 2-Pack Natural Aluminum-Free Deodorant for Sensitive Skin with Essential Oils, Plant-Based Packaging (Citrus & Vetiver, 2.5 Ounce (Pack of 2))

\$25.0

[B08KVCCSD6]

Each & Every 3-Pack, Natural Aluminum-Free Deodorant for Sensitive Skin Made with Essential Oils, 2.5 Oz. (Lavender & Lemon, Citrus & Vetiver, and Coconut & Lime)

\$35.0

Criteria: less than 5 ounce citrus deodorant sensitive skin, price less than \$30.

Answer: My requirements are 5 ounce, citrus deodorant, suitable for sensitive skin, and price less than \$30. Looks like this information is available on the search page, so I can proceed.

Products B078GWRC1J, B08SMG4WB9 look suitable as they are less than 5 ounce, citrus and have price 10.99 and \$25 less than \$30. Thus, shortlisted IDs are shortlisted=['B078GWRC1J', 'B08SMG4WB9']

Criteria: less than 5 ounce citrus deodorant sensitive skin, cruelty free.

Answer: My requirements are 5 ounce, citrus deodorant, suitable for sensitive skin, and cruelty-free. Since there is no information about cruelty free on the search page, I cannot proceed. Task failed!

Here is another task with a different search page and criteria. List all the product ids (enclosed in []) from the search page that match ALL the requirements in the criteria. Name this list shortlisted. If you cannot make the determination about even 1 sub-criteria, do not make a guess, output "task failed!". Generate an explanation before the answer to justify your decision.

Search Page:

{}

Criteria: {}

Answer:

## WebShop Planner Prompt

Write an abstract plan to successfully complete the goal. In each step of the plan mention which module (including arguments) that need to be called. Learn from and incorporate information from previous runs, e.g. do not repeat previously successful or unsuccessful commands. Here are some examples: Information from previous run: -

Goal: Buy 3 ounce bottle of citrus deodorant for sensitive skin, that is natural and priced less than 50.00 dollars.

# Think: Based on the criteria and the search bar, I should query 3 ounce citrus deodorant sensitive skin. I have the following constraints: natural and price lower than \$30 which I can use to narrow down search results.

Step 1: Search[3 ounce citrus deodorant sensitive skin]

# Think: Now I will need to narrow down the search results for price lower than \$30 and natural

Step 2: SimpleMatch[3 ounce citrus deodorant sensitive skin with price lower than \$50 and natural]

# Think: Since it returns a list of up to 3 products, I will pick the first suitable product. For now, I'll denote its id as prod\_id for placeholder.

Step 3: Buy[prod\_id, "3 ounce bottle of citrus deodorant for sensitive skin, that is natural and priced less than 30.00 dollars"]

# Think: My plan requires all these steps to succeed sequentially, so I will use the "AND" operator.

Execution Order: (Step 1 AND Step 2 AND Step 3)

Information from previous run:

- Unable to get matching product using: SimpleMatch[3 ounce citrus deodorant sensitive skin with price lower than \$30 and natural]
- Search results page:  
[Back to Search]  
Page 1 (Total results: 50)  
[Next >]  
[B078GWRC1J]  
Bright Citrus Deodorant by Earth Mama | Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce  
\$10.99  
[B08KBVJ4XN]  
Barrel and Oak - Aluminum-Free Deodorant, Deodorant for Men, Essential Oil-Based Scent, 24-Hour Odor Protection, Cedar & Patchouli Blend, Gentle on Sensitive Skin (Mountain Sage, 2.7 oz, 2-Pack)  
\$35.95  
[B078GTVKXY]  
Ginger Fresh Deodorant by Earth Mama | Natural and Safe for Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula 3-Ounce  
\$10.99  
[B08SMG4WB9]  
Each & Every 2-Pack Natural Aluminum-Free Deodorant for Sensitive Skin with Essential Oils, Plant-Based Packaging (Citrus & Vetiver, 2.5 Ounce (Pack of 2))  
\$25.00  
[B08KVCCSD6]  
Each & Every 3-Pack, Natural Aluminum-Free Deodorant for Sensitive Skin Made with Essential Oils, 2.5 Oz. (Lavender & Lemon, Citrus & Vetiver, and Coconut & Lime)  
\$35.00  
[B087WKS2G]

Goal: Narrow down search results for 3 ounce bottle of citrus deodorant for sensitive skin that is priced lower than \$30 and natural. You cannot search again.

# Think: Based on the search results and previous information, SimpleMatch failed because my criteria was too complex. Price constraint is easy to verify, I will narrow down based on that first then examine in detail for natural constraint

# Think: Based on price, I narrow down my search to B078GWRC1J, B08SMG4WB9 as they look suitable. These are on my shortlist to examine the natural constraint in detail one by one.

Step 1: DetailMatch[B078GWRC1J, 3 ounce bottle of for sensitive skin, that is natural and priced less than 30.00 dollars]

Step 2: DetailMatch[B08SMG4WB9, 3 ounce bottle of citrus deodorant for sensitive skin, that is natural and priced less than 30.00 dollars]

# Think: If none of the products exactly match my criteria, I will search again with a new query that includes the natural criteria too. This ensures my plan is complete.

Step 3: Search[3 ounce citrus deodorant natural and sensitive skin]

# Think: Since these steps are linked by an if condition, I only need one of them to succeed. I will connect them using the "OR" operator.

Execution Order: (Step 1 OR Step 2 OR Step 3)

Here is a new goal. Write an abstract plan to successfully complete the goal. In each step of the plan mention which module (including arguments) that need to be called. Learn from and incorporate information from previous runs, e.g. do not repeat previously successful or unsuccessful commands. In the end, output the intended execution order.

Information from previous run: {}

Goal: {}

## TextCraft Executor Prompt

You are given few useful crafting recipes to craft items in Minecraft. Crafting commands are of the format "craft [target object] using [input ingredients]". You can either "fetch" an object (ingredients) from the inventory or the environment or "craft" (target) using any of the crafting commands. You can use ONLY these crafting commands provided, do not use your own crafting commands. However, if the crafting command uses a generic ingredient like "planks", you can use special types of the same ingredient e.g. "dark oak planks" in the command instead. For any other natural language or thoughts, use prefix 'think: '.

Here is a demo of how to fetch and craft objects.

Crafting commands:

```
craft 3 dark oak sign using 6 dark oak planks, 1 stick
craft 4 dark oak planks using 1 dark oak log
craft 1 stick using 1 planks
craft 4 stick using 2 bamboo
craft 4 oak planks using 1 oak log
craft 1 dark oak fence using 2 stick, 4 dark oak planks
craft 1 warped stairs using 6 warped planks
craft 3 oak sign using 6 oak planks, 1 stick
```

Goal: craft dark oak sign

> think: I should check if I can fetch dark oak sign directly from the environment or the inventory.

OK.

> inventory:

Inventory: [stick] (1) [dark oak planks] (8)

> get dark oak sign

Could not find dark oak sign

> think: I cannot get dark oak sign directly, I need to craft it. From the crafting commands, I can use: craft dark oak sign using 6 dark oak planks, 1 stick. Ingredients needed: 6 dark oak planks, 1 stick. Input assumption: I have all the necessary ingredients in my inventory. Let me verify this first.

OK.

> inventory

Inventory: [stick] (1) [dark oak planks] (8)

> think: I found my ingredients: 6 dark oak planks, 1 stick in my inventory. My assumption is true, I can proceed. I will use the crafting command: craft dark oak sign using 6 dark oak planks

OK.

> craft 1 dark oak sign using 6 dark oak planks, 1 stick

Crafted 1 minecraft:dark\_oak\_sign

> inventory

Inventory: [dark oak sign] (1)

> think: I now have dark oak sign in my inventory. Task Completed!

OK.

Goal: fetch 2 dark oak logs.

> think: I should check my inventory first, to see if I already have dark oak sign. Otherwise, I will directly try to get it from the environment.

OK.

> inventory

Inventory: [stick] (1)

> get 2 dark oak logs.

Got 2 dark oak logs

> inventory

Inventory: [dark oak log] (2) [stick] (1)

> think: I have 2 dark oak logs in my inventory. Task Completed!

OK.

## TextCraft Executor Prompt (cont.)

Here is an example of a complex goal.

Goal: craft 2 oak planks

> think: I should check if I can fetch dark oak sign directly from the environment or the inventory.

OK.

> inventory:

Inventory: [stick] (1)

> get 2 oak planks

Could not find 2 oak planks

> think: I cannot get oak planks directly, I need to craft it. From the crafting commands, I can use: craft 4 oak planks using 1 oak log. Ingredients needed: 1 dark oak log. Input assumption: I have the ingredients needed in my inventory. Let me verify this first.

OK.

> inventory

Inventory: [stick] (1)

> think: My assumption is false. I need to get the ingredients: 1 oak log first.

OK.

> get 1 oak log

Got 1 oak log

> craft 4 oak planks using 1 oak log

Crafted 4 minecraft:oak\_planks

> inventory

Inventory: [stick] (1) [oak\_planks] (4)

> think: I now have 2 oak planks in my inventory. Task Completed!

OK.

Now here is a different goal. You can use these crafting commands to accomplish the goal. When you the desired item in your inventory, think: Task Completed! If you have tried your best but cannot proceed, think: task failed!

## TextCraft Planner Prompt

Your task is to come up with a short plan to help me accomplish my goal in a couple of steps using at most ONE of the provided crafting commands. You can take the help of crafting commands below to create new objects. Craft command can be understood as follows: craft [target] using [ingredients], where target is item/object generated by the craft command as output and ingredient are the inputs. You are given an agent that can "craft" or "fetch" objects.

Here are some examples.

Crafting commands:

```
craft 3 dark oak sign using 6 dark oak planks, 1 stick
craft 4 dark oak planks using 1 dark oak log
craft 1 stick using 1 planks
craft 4 stick using 2 bamboo
craft 4 oak planks using 1 oak log
craft 1 dark oak fence using 2 stick, 4 dark oak planks
craft 1 warped stairs using 6 warped planks
craft 3 oak sign using 6 oak planks, 1 stick
```

Goal: craft dark oak sign.

```
# Think: My target is a dark oak sign. From the list of crafting commands, only 1 command generates my target: craft 3 dark oak sign using 6 oak planks, 1 stick. I will use this command to devise a plan. My ingredients are: 6 dark oak planks, 1 stick. I should first get all the ingredients and then use the crafting command.
```

Step 1: fetch 6 dark oak planks

Step 2: fetch 1 stick

```
# Think: Now that I have collected the input ingredients, I can craft the dark oak sign using given command.
```

Step 3: craft dark oak sign using 6 dark oak planks, 1 stick

```
# Think: To succeed, I need to perform all these steps, one after the other. So I need to use the "AND" operator.
```

Execution Order: (Step 1 AND Step 2 AND Step 3)

Goal: fetch 6 dark oak planks.

```
# Think: My target is 6 dark oak planks. From the list of crafting commands, only 1 command generates my target: craft 4 dark oak planks using 1 dark oak log. My ingredients are: 1 dark oak log. To successfully accomplish the goal, I should first get all the ingredients and then use the crafting command.
```

Step 1: fetch 1 dark oak log

```
# Think: Now that I have collected the input ingredients, I can craft dark oak planks using given command. I know that I cannot use a partial recipe.
```

Step 2: craft 4 dark oak planks using 1 dark oak log

```
# Think: This gives me 4 dark oak planks which is less than my desired 6 dark oak planks. I know that I cannot use a partial recipe. So my goal is not satisfied, I need to craft more dark oak planks by repeating Step 2 one more time.
```

Step 3: craft 4 dark oak planks using 1 dark oak log

```
# Think: To succeed, I need to perform all these steps, one after the other. So I need to use the "AND" operator.
```

Execution Order: (Step 1 AND Step 2 AND Step 3)

Here is a different goal with different craft commands. Your task is to come up with a short plan to help me accomplish my goal in a couple of steps using at most ONE of the provided crafting commands. You can take the help of crafting commands below to create new objects. Keep in mind that:

- It is okay to generate more target objects than your goal.
- Be very careful with the count of objects, SAME object counts mentioned in the input crafting command.
- You cannot use a partial crafting command recipe, i.e. if the recipe generates 2 objects you CANNOT alter it to produce just 1.
- Also, you can use ONLY 1 crafting command in your plan.