

# Leveraging In-Context Learning for Language Model Agents

Shivanshu Gupta<sup>1\*</sup> Sameer Singh<sup>1</sup> Ashish Sabharwal<sup>2</sup> Tushar Khot\* Ben Bogin\*

<sup>1</sup>University of California Irvine <sup>2</sup>Allen Institute for AI

{shivag5, sameer}@uci.edu, ashishs@allenai.com

## Abstract

In-Context Learning (ICL) with dynamically selected demonstrations combines the flexibility of prompting large language models (LLMs) with the ability to leverage training data to improve performance. While ICL has been highly successful for prediction and generation tasks, leveraging it for agentic tasks that require sequential decision making is challenging—one must think not only about how to annotate long trajectories at scale and how to select demonstrations, but also what constitutes demonstrations, and when and where to show them. To address this, we first propose an algorithm that leverages an LLM with retries and demonstration selection to automatically and efficiently annotate agentic tasks with solution trajectories. We then show that set-selection of trajectories of similar tasks as demonstrations significantly improves performance, reliability, robustness, and efficiency of LLM agents. However, trajectory demonstrations have a large inference cost overhead. We show that this can be mitigated by using small trajectory snippets at every step instead of an additional trajectory. We find that demonstrations obtained from larger models (in the annotation phase) also improve smaller models, and that ICL agents can even rival costlier trained agents. Thus, our results reveal that ICL, with careful use, can be very powerful for agentic tasks as well.

## 1 Introduction

Due to advances in pretraining, instruction tuning, and scaling, Large Language Models (LLMs) are now increasingly powering autonomous agents to perform complex real-world tasks that require acting in an environment and sequential decision-making. Using LLMs to simulate such agentic behavior involves repeatedly prompting and asking them to generate the next action to be executed. However, LLM agents can be unreliable, especially

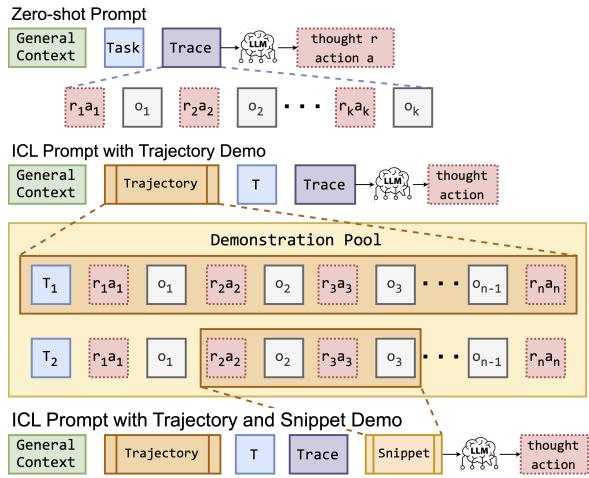


Figure 1: Different types of demonstrations for agentic tasks. **Top** Using LLMs to simulate agentic behavior involves repeatedly prompting it with the general setup, a task description, and an execution trace recording the agent’s thoughts ( $r$ ), actions ( $a$ ), and observations ( $o$ ). **Middle** Given a pool of tasks paired with solution trajectories, one way to show demonstrations is to use entire trajectories for similar tasks in the prompt. While effective, this has a large overhead. **Bottom** Another way is to use smaller trajectory snippets with similar reasoning that are post-fixed to the prompt.

for complex tasks with long trajectories. Prior work on enhancing LLM agents through structured prompting-based workflows with explicit reasoning, planning, or reflection steps (Shinn et al., 2023; Kim et al., 2023; Sun et al., 2024) used a fixed prompt for every task instance, without leveraging training data. On the other hand, approaches based on task-specific supervised finetuning or reinforcement learning (Chen et al., 2023; Mitra et al., 2024; Chen et al., 2025) are too expensive to apply to larger, more powerful LLMs and to update with the knowledge needed for new tasks after training.

In this work, we explore an alternative approach that is prompting-based yet takes advantage of training data, namely *In-Context Learning (ICL) with Demonstration Selection*, where demonstra-

\*Work done while authors were at Allen Institute for AI.

tions relevant to each instance are selected at inference time from a pool of annotations. While ICL is already effective (Brown et al., 2020), demonstration selection can dramatically boost it for traditional NLP tasks (Gupta et al., 2024, 2023; Ye et al., 2023a). However, unlike such non-sequential tasks where demonstrations can simply be input-output pairs from a train set, for agentic tasks, the training sets of tasks, even when available, are rarely annotated with *solutions* that can serve as demonstrations. Moreover, due to context length limits and recency bias of LLMs, one needs to think not just of *how* to select demonstrations, but also *what* these demonstrations comprise (e.g., entire trajectories or snippets thereof), *when* to show them (i.e., what step), and *where* to place them in the prompt.

To address these challenges, we propose an iterative annotation algorithm that leverages ICL and demonstration selection itself to automatically and efficiently annotate training tasks with solutions that can be used as demonstrations. We then use these annotations to study different demonstration granularities and placements. We begin with the simplest approach, which is to show entire trajectories of similar tasks. As shown in Fig. 1 (middle), these are placed before the test task and shown at every step. Since trajectories tend to be long and only a limited number of them can fit in the prompt, we explore how to select optimal *sets* of trajectories. Finally, as trajectory demonstrations can have a large overhead in terms of inference costs, we explore two forms of smaller demonstrations. First, we consider smaller subtask trajectories by switching to a Plan-and-Execute (PnE) solver (Yang et al., 2023; Sun et al., 2024) that decomposes tasks into a sequence of subtasks and executes them one by one. Second, we show small relevant snippets of trajectories at every step (Fig. 1 (bottom)). These are selected based on the agent’s reasoning in the latest step and shown at the end of the prompt for a single step, thus accounting for LLMs’ recency bias and also having a minimal overhead.

For our testbed, we use the AppWorld benchmark (Trivedi et al., 2024), which evaluates an LLM agent’s ability to carry out complex day-to-day user tasks involving sending emails, making payments, playing music, shopping, etc., by interacting with a variety of apps via their APIs. AppWorld’s rich code-based action space, varying observation sizes, and complex tasks make it an ideal testbed for studying various design decisions relating to demonstration selection. Using our an-

notation algorithm, we automatically annotate over 95% of training tasks with solutions to create a demonstration pool. We show that using the annotated trajectory of even a single most similar task as a demonstration boosts a zero-shot agent’s performance by 29 points, 16 points more than using a fixed, manually written trajectory. Further, when additional trajectories can be included, selecting them jointly as a set (Gupta et al., 2023) is more effective than independent ranking-based selection.

Selecting trajectory demonstrations is particularly effective at improving reliability (across multiple runs of the same task) and robustness (across variations of the same tasks), outperforming the use of a fixed trajectory by 20.8 and 23.2 points, respectively. However, while effective, trajectory demonstrations also have a large overhead in terms of inference costs—each additional trajectory adding on the order of 100K tokens on average in inference costs. Instead, as we show, providing small, relevant snippets of the trajectories as demonstrations at every step is also effective at improving performance, notably with negligible overhead. We find a combination of trajectory and snippet demonstrations to be the optimal approach, and with these, a prompted LLM agent can be made competitive with most state-of-the-art training-based approaches (Chen et al., 2025). Overall, our results show that, similar to traditional NLP tasks, demonstration selection can yield significant performance gains for LLM agents and can enable prompted LLM agents to rival even trained ones.

## 2 Related Work

**LLM Agents.** LLMs are increasingly being used to power autonomous agents for a variety of agentic tasks involving sequential decision-making. These include web navigation to answer user queries (Zhou et al., 2024b; Drouin et al., 2024) and e-commerce (Yao et al., 2022), playing games (Shridhar et al., 2021), interacting with applications and APIs to carry out user tasks (Trivedi et al., 2024), running ML experiments (Bogin et al., 2024), and more. Prior work on improving agent performance on such tasks has looked into (1) prompting based approaches to inducing structured workflows with explicit reasoning, planning, and reflection steps (Yao et al., 2023; Shinn et al., 2023; Wang et al., 2023; Kim et al., 2023; Sun et al., 2024), (2) training-based approaches including supervised finetuning on agent trajectories and reinforcement

learning (Nakano et al., 2021; Yao et al., 2022; Deng et al., 2023; Chen et al., 2023; Qin et al., 2024; Mitra et al., 2024; Chen et al., 2025).

**In-Context Learning (ICL)** (Brown et al., 2020) is the ability of LLMs to solve unseen tasks without training by merely conditioning on a few task demonstrations and without any task-specific training. However, ICL performance is highly sensitive to the choice of demonstrations (Zhao et al., 2021), and can be significantly improved by dynamically selecting demonstrations for each test input (Liu et al., 2022). There is now a large body of work on selecting better demonstrations, exploring among other things, better metrics for scoring demonstration candidates (Rubin et al., 2022; Gupta et al., 2023, 2024; Askari et al., 2025), selecting demonstrations as a set (Gupta et al., 2023; Ye et al., 2023a), selecting diverse demonstrations to reduce redundancy among them (Su et al., 2023; Levy et al., 2023; Agrawal et al., 2023; Ye et al., 2023b), etc.

**Demonstration Selection for Agentic Tasks.** Prior work on demonstration selection for ICL has primarily focused on traditional, non-sequential NLP tasks that involve mapping inputs to outputs. The two prior works that have studied demonstration selection in the context of agentic tasks are Synapse (Zheng et al., 2024) and TRAD (Zhou et al., 2024a). However, they primarily focused on web navigation tasks where the main challenge was the size of individual HTML observations rather than task complexity (in terms of number of steps). In contrast, we focus on more complex tasks which involve numerous steps with long-range dependencies, but not every step yields a large observation, allowing an entire trajectory or two can fit in the context. This setup allows us to study the impact of different granularities, selection, and placements of demonstrations. Notably, this will also become an increasingly common scenario as LLM context lengths increase, but the cost-benefit trade-offs we explore will remain.

### 3 Preliminaries

#### 3.1 LLM Agents

**ReAct.** The predominant approach to creating LLM-powered agents for agentic tasks is ReAct (Yao et al., 2023). As shown in Fig. 1 (Top), it involves repeatedly prompting the LLM with a trace of past execution and asking it to produce a *thought* (denoted  $\mathbf{r}$ ) describing its reasoning about

its progress and an *action* (denoted  $\mathbf{a}$ ), to be executed in the environment to obtain the next observation (denoted  $\mathbf{o}$ ). Formally, given (1) a context  $\mathbf{c} = \langle \mathbf{p}, \mathbf{q} \rangle$  comprising a general context  $\mathbf{p}$  which describes the setup, provides demonstrations and guidelines, etc., and a task-specific context  $\mathbf{q}$  describing the task to be carried out, and (2) a trace of past thoughts, actions, and observations  $\mathbf{h}_t = \langle \mathbf{r}_1, \mathbf{a}_1, \mathbf{o}_1, \dots, \mathbf{r}_t, \mathbf{a}_t, \mathbf{o}_t \rangle$ , the LLM is prompted to generate the next thought and action:

$$\mathbf{r}_{t+1}, \mathbf{a}_{t+1} \sim \mathcal{P}_{LM}(\cdot | \mathbf{c}, \mathbf{h}_t) \quad (1)$$

The next observation  $\mathbf{o}_{t+1}$  is then obtained by executing the action  $\mathbf{a}_{t+1}$  in the environment. This process is repeated until a terminal state is reached. We will refer to the complete execution trace  $\mathbf{h}_T$  as a trajectory  $\tau$ .

**Plan & Execute (PnE).** The ReAct approach, as described above, tries to solve the entire task in one go and retains the entire execution trace in the prompt. However, this can be expensive as the trajectories for complex agentic tasks are often very long. One way to address this is to use a Plan & Execute (PnE) approach (Yang et al., 2023; Sun et al., 2024). PnE takes advantage of the fact that a task may involve multiple simpler subtasks, and how each subtask is carried out may not be relevant to the other subtasks. It incorporates a planning step that breaks down the original task  $t$  into a sequence of subtasks  $t^1, \dots, t^m$ . Each subtask is then executed by a ReAct-styled executor agent, optionally with the plan and summaries of previous subtasks' trajectories provided in the task-specific prompt  $\mathbf{q}$ .

#### 3.2 In-Context Learning and Demonstration Selection

In-Context Learning (ICL) is the ability of LLMs to solve unseen tasks by conditioning on a few task demonstrations. Formally, for traditional NLP tasks, given demonstrations in the form of input-output pairs  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^k$  and the test input  $\mathbf{x}_{\text{test}}$ , it involves prompting the LLM with the context  $\mathbf{c} = \langle \mathbf{x}_1, \mathbf{y}_1, \dots, \mathbf{x}_k, \mathbf{y}_k, \mathbf{x}_{\text{test}} \rangle$  and letting it generate the output  $\mathbf{y}_{\text{test}}$ . Although using the same demonstrations for all test inputs allows ICL to work even for tasks lacking any training data, when a training set  $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  is available, performance can be boosted using some form of demonstration selection (Liu et al., 2022; Rubin et al., 2022; Gupta et al., 2023, 2024). Using the

training set as a pool of demonstration candidates, it involves selecting  $k \ll N$  demonstrations that, when placed in the context, increase the likelihood of the correct output being generated. Some approaches for demonstration selection proposed for traditional NLP tasks that we will experiment with include:

**Ranking-based Selection.** This involves scoring all the candidates for their relevance with respect to the test instance and using the top-K candidates as demonstrations. Formally, given the training set  $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  and the test input  $\mathbf{x}_{\text{test}}$ , the demonstrations are selected as  $\text{topk}_i \sim \text{sim}(\mathbf{x}_{\text{test}}, \mathbf{x}_i)$ , where  $\text{sim}(\cdot)$  is a similarity metric. Note that the metric operates on only the inputs that proxy as the *retrieval key* for the demonstrations. Prior work has explored Cosine Similarity (Liu et al., 2022) and BERTScore-Recall (BSR) (Zhang et al., 2020; Gupta et al., 2023) as metrics, both of which involve encoding the retrieval key using a dense encoder and using it to identify the closest candidate.

**Set Selection.** Gupta et al. (2023) showed that ranking-based selection can be sub-optimal for complex compositional tasks as it may select demonstrations that are individually relevant to the test input yet fail to provide all the relevant information needed to solve it. Instead, they argue that demonstrations should be selected as a set such that they cover all the reasoning patterns. They proposed Set-BSR, a set-extension of BSR, that is submodular and hence greedily optimizable.

## 4 Automatic Trajectory Annotation

Given the success of demonstration selection for ICL for traditional NLP tasks, in this work, we explore how to effectively and efficiently leverage it for agentic tasks. However, this requires a pool of tasks annotated with agent-style solutions<sup>1</sup> (as described in § 3.1) that can serve as demonstrations. While some agentic benchmarks provide training sets of tasks, most do not provide task solutions in a form that can be used as demonstrations for the agent; rather, they only provide a final answer or a checker that can be used to verify solution correctness.

Since manually annotating tasks with solutions is intractable at scale, we propose a simple iterative algorithm (Algorithm 1) to do this automatically. Given (1) a pool of tasks  $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^N$ , (2) a solver

---

### Algorithm 1 Iterative Annotation for Agentic Tasks

---

```

Require: Task pool  $\mathcal{T}$ ; Demonstration selector  $D$ ; Solver  $S$ ;
          Solution Checker  $C$ ; Number of Rounds  $R$ 
1:  $\mathcal{U} \leftarrow \mathcal{T}$                                  $\triangleright$  Unannotated tasks
2:  $\mathcal{T}^* \leftarrow \emptyset$                           $\triangleright$  Annotated tasks
3: for  $r = 1$  to  $R$  do
4:   for  $\mathbf{t} \in \mathcal{U}$  do
5:      $\mathcal{D} \leftarrow D(\mathbf{t}, \mathcal{T}^*)$             $\triangleright$  Select demonstrations
6:      $\mathbf{s} \leftarrow S(\mathbf{t}, \mathcal{D})$             $\triangleright$  Generate solution with
       demonstrations
7:     if  $C(\mathbf{t}, \mathbf{s})$  then
8:        $\mathcal{T}^* \leftarrow \mathcal{T}^* \cup (\mathbf{t}, \mathbf{s})$   $\triangleright$  Add to annotated tasks
9:        $\mathcal{U} \leftarrow \mathcal{U} - \{\mathbf{t}\}$             $\triangleright$  Remove from unannotated
       tasks
10:    end if
11:   end for
12:   if  $\mathcal{U} = \emptyset$  then
13:     break                                 $\triangleright$  All tasks annotated
14:   end if
15: end for
16: return  $\mathcal{T}^*$                             $\triangleright$  Annotated tasks

```

---

$S$  that is used to generate solutions  $\mathbf{s} \sim S(\mathbf{t}, \mathcal{D})$  given a task  $\mathbf{t}$  and some demonstrations  $\mathcal{D}$ , and (3) a checker that verifies solution correctness, the algorithm returns tasks annotated with solutions  $\mathcal{T}^* = \{\mathbf{t}_i, \mathbf{s}_i\}_{i=1}^N$ . Further, instead of naively retrying the solver, the algorithm also leverages currently annotated tasks as demonstrations. This not only improves efficiency in terms of the number of retries needed but also ensures that more instances are correctly annotated.

Finally, note that the algorithm, as described above, is agnostic to the kind of solver (and solutions), e.g., for the ReAct solver, the solutions would be trajectories, i.e.  $\mathbf{s}_i = \tau_i$ , while for the PnE solver, they would comprise a plan in the form of a sequence of subtasks and the corresponding trajectories, i.e.  $\mathbf{s}_i = \{\mathbf{t}_i^j, \tau_i^j\}_{j=1}^{m_i}$ . We will refer to the set of task trajectory annotations for ReAct as  $\mathcal{D}_{\text{task}}^*$ , the plan and subtask trajectory annotations for PnE as  $\mathcal{D}_{\text{plan}}^*$  and  $\mathcal{D}_{\text{subtask}}^*$ , respectively.

## 5 Demonstrations for Agents

Having annotated a pool of tasks with solutions that can serve as demonstrations, we now discuss different demonstration granularities, along with how to select them, when to show them, and where to place them in the prompt.

### 5.1 Task-level Trajectory Demonstrations

Similar to traditional ICL, a natural way to show demonstrations is in the form of trajectories for similar tasks from the pool  $\mathcal{D}_{\text{task}}^*$ . These task-trajectory pairs are selected from  $\mathcal{D}_{\text{task}}^*$  before execution and

<sup>1</sup>In real-world settings, these could also be obtained from an existing system.

are used in the prompt at every step. Specifically, they are placed in the general prompt  $\mathbf{p}$  before the description of the test task  $t$  and its execution trace  $h_t$ . To select these demonstrations, we experiment with both ranking-based and set-selection methods described in § 3.2 using the task statement as the retrieval key. Since, similarly to the tasks explored by Gupta et al. (2023), the optimal demonstrations for agentic tasks would demonstrate all necessary steps, we believe that set-selection might be more appropriate for agentic tasks as compared to ranking-based selection.

## 5.2 Fine-grained Demonstrations

As noted in § 3.1, the trajectories for complex agentic tasks can be very long. Thus, using them as demonstrations may be very expensive, if at all feasible with limited context lengths. Moreover, the trajectories for even the most similar tasks may have irrelevant steps while not being helpful for every step of the test task. Finally, LLMs have a recency bias (Liu et al., 2024), thus a *smaller* demonstration, closer to the steps it is relevant for, may be more effective than an entire trajectory early in the prompt. We explore two ways to achieve this: (1) using trajectories for similar subtasks with a PnE solver and (2) using snippets of the trajectories relevant to the current step.

### 5.2.1 Subtask-level Trajectory Demonstrations

One way to use smaller demonstrations is to use the PnE solver instead of § 3.1 that breaks the original task down to a sequence of subtasks  $t^1, \dots, t^m$  and executes each subtask separately using a React-style executor. The demonstrations for the PnE executor can be the trajectories of similar subtasks from the pool  $\mathcal{D}_{\text{subtask}}^*$ . This is similar to the trajectory demonstrations from the § 5.1 in that they are selected prior to subtask execution and included in the general prompt  $\mathbf{p}$  for every step of the subtask. However, the selection uses the subtask statement as the retrieval key, and demonstrations for different subtasks of the current task may be from different tasks. In this way, it allows for a more fine-grained demonstration to be shown for the limited scope of the subtask.

### 5.2.2 Step-level Snippet Demonstrations

A major drawback of using a PnE solver is that it introduces a planning step prior to execution, which may yield inaccurate plans. Moreover, as

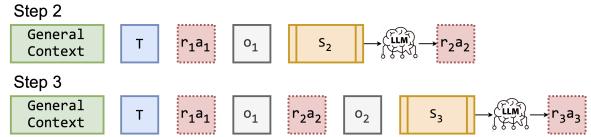


Figure 2: Snippet demonstrations are selected based on the thought at the current step (*how*) and only used to predict the next thought-action (*when*) by placing after the execution trace in the prompt (*where*). E.g., the snippets  $S_2$  are selected based on the thought  $r_1$  and used to predict  $r_2$  and  $a_2$  and so on.

they are placed early in the prompt, they still don't completely address the problem of recency bias. Thus, we experiment with showing small snippets of trajectories that are relevant to the current step as demonstrations. To identify these snippets, we use the thought produced at the current step as a retrieval key to find similar thoughts in trajectory annotations  $\mathcal{D}_{\text{task}}^*$ . A similar thought suggests a similar step, and we can use a snippet of the trajectory comprising this step as a demonstration. As shown in Fig. 3, the selected snippets are appended to the prompt after the execution trace  $h_t$ , which helps account for the recency bias. See Fig. 12 for an example of a prompt template for showing snippet demonstrations. The snippets selected for the current step are only shown for predicting a single next step, as they may not be relevant for subsequent steps, for which we select new snippets anyway. Finally, note that snippet demonstrations as described above do not require any additional annotations, as they are derived from the trajectory annotations.

## 6 Experimental Setup

### 6.1 AppWorld Benchmark

AppWorld<sup>2</sup> (Trivedi et al., 2024) is a benchmark designed to evaluate an autonomous agent's ability to carry out complex user tasks by interacting with 457 APIs associated with 9 simulated apps, viz. *Gmail*, *Venmo*, *Spotify*, *SimpleNote*, *Splitwise*, *Amazon*, *Todoist*, *Phone*, and *File System*. It provides a stateful Python interpreter that the agent can use to interact with the various APIs and a *Supervisor* app and an *ApiDoc* app it can use to obtain the information about the user and the various Apps/APIs, respectively. Note that, although due to cost constraints, we only experiment with AppWorld, its code-based action space, varying

<sup>2</sup>Our use of Appworld is permitted by its license (Apache License 2.0).

observation sizes, and complex tasks make it the ideal testbed for our experiments.

The benchmark comprises a total of 244 task templates, or scenarios, each with three variants for a total of 722 tasks. The tasks are split into a train set (90 tasks), a dev set (57 tasks), and two test sets: test-normal (Test-N, 168 tasks), which evaluates in-distribution performance, and test-challenge (Test-C, 417 tasks), containing more complex tasks involving unseen apps. Each task is associated with a suite of unit tests that check (1) whether only the requisite changes, and no extraneous changes, were made to the environment state, and (2) if required by the task, the final answer produced matches the ground truth. A task is considered solved only if all its unit tests pass.

**Annotation.** To create our demonstration pool, we use the iterative annotation algorithm (Alg. § 1) to automatically annotate 146 tasks in the combined train and dev sets. As described in § 4, we sped up the process by using already annotated instances as demonstrations for subsequent iterations. Specifically, for ReAct, we used one task-trajectory pair as a demonstration, and for PnE, we used 4 task-plan pairs and 3 subtask-trajectory pairs for the planner and executor, respectively. For annotation, all the demonstrations were selected using Cosine Similarity based on `all-mpnet-base-v2` encoder. Of the 147 tasks, we annotated 141 tasks spanning 48 scenarios with ReAct solutions. With PnE, we were able to annotate 134 tasks spanning 46 scenarios. The PnE solutions had an average of 6.2 subtasks per task for a total of 833 subtasks.

**Evaluation.** We evaluate on both the Test-N and Test-C sets. AppWorld recommends two metrics: (1) Task Goal Completion (TGC), which is the percentage of tasks solved, and (2) Scenario Goal Completion (SGC), which is the percentage of scenarios for which all three task variants passed. While TGC measures an agent’s overall performance, SGC measures its robustness across variations of a task. Additionally, to assess whether agents solve tasks reliably, rather than by chance, we also report the percentage of tasks for which multiple *runs* succeeded, called Reliable Task Goal Completion (RTGC). We also report the average Token Usage during execution of each task as a measure of efficiency and the average number of Steps taken to complete tasks as a measure of inference costs. Note that token usage would aggregate the input and output token counts across all steps.

## 6.2 Methods

As discussed in § 5, we explore the following three different types of demonstrations:

**Task Trajectory Demonstrations.** We experiment with the following selection methods to select  $k$  trajectories: (1) ranking-based selection using Cosine Similarity ( $\text{Cos}[k]$ ) and BertScore-Recall ( $\text{BSR}[k]$ ), and (2) set-selection using  $\text{SET-BSR}[k]$  (Gupta et al., 2023). Following Gupta et al. (2023), for  $\text{Cos}$ , we use `all-mpnet-base-v2` as the encoder, while for  $\text{BSR}$  and  $\text{SET-BSR}$ , we use `deberta-base-mnli-v2`. For  $\text{Cos}$  and  $\text{BSR}$ , we report results for  $k = 1$  and  $k = 2$ , while for  $\text{SET-BSR}$ , we only report results for  $k = 2$  as selection using it reduces to  $\text{BSR}$  for  $k = 1$ . As baselines, we experiment with using the agent zero-shot without any trajectory demonstrations ( $\text{ZEROSHOT}[0]$ ), and using a single fixed manually written trajectory from Trivedi et al. (2024) as demonstration for every test input ( $\text{FIXED}[1]$ ).

**Subtask Trajectory Demonstrations.** We use  $\text{BSR}$  to select subtasks whose trajectories to include in the executor’s prompt. Additionally, we use four task-plan pairs selected using  $\text{BSR}$  as demonstrations for the planner.

**Snippet demonstrations.** We use  $\text{BSR}$  for selection of up to<sup>3</sup>  $k = 2$  annotated thoughts based on the thought at every step of the current task. For each selected thought, we create the snippets using the step (thought-action-observation triple) corresponding to the selected thought and a subsequent step if there is one.

Method	TGC $\uparrow$	RTGC $\uparrow$	SGC $\uparrow$	Steps $\downarrow$
ZEROSHOT[0]	35.1	22.0	15.2	21.9
FIXED[1]	50.6	40.5	30.4	14.6
RANDOM[1]	58.9	43.5	37.5	14.8
Cos[1]	64.0	57.1	44.6	13.4
BSR[1]	61.0	52.4	44.6	13.5
RANDOM[2]	58.9	45.8	33.9	13.8
Cos[2]	63.7	57.1	44.6	12.6
BSR[2]	64.9	59.5	50.0	12.2
SETBSR[2]	<b>65.8</b>	<b>61.3</b>	<b>53.6</b>	<b>11.5</b>

Table 1: Impact of different numbers and selection of trajectory demonstrations on a GPT-4o ReAct agent on the Test-N. Even a single manually written trajectory significantly improves performance. Further, gains are obtained using actual agent trajectories, by selecting the most relevant trajectories as demonstrations, and by using set-selection when using multiple trajectories.

<sup>3</sup>To prevent spurious matches, we filter out thoughts that score less than 0.85.

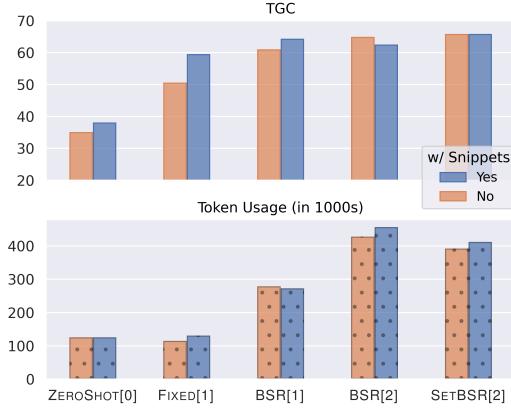


Figure 3: Effect of trajectory and snippet demonstrations (selected using BSR) on the performance in terms of TGC (Top) and inference cost in terms of token usage per task (Bottom) of GPT-4o ReAct agents on Test-N. While trajectory demonstrations are most effective at improving performance, they do so at a high cost. Snippet demonstrations are also generally effective but have a very minimal overhead.

### 6.3 Agent Implementation Details

We use OpenAI’s GPT-4o (gpt-4o-2024-08-06) as the primary LLM both for annotating our demonstration pool as well as for our ICL experiments. Detailed hyperparameters and prompt templates for the ReAct solver, PnE planner, and executor are provided in the App. A. To see if the annotation obtained from a larger LLM can benefit smaller LLMs, we also experiment with the smaller GPT-4o-mini (gpt-4o-mini-2024-07-18).

## 7 Results

**Trajectory demonstrations boost agent performance.** Table 1 shows the results on Test-N for varying numbers and selection of trajectory demonstrations. First, it is clear from the TGC numbers (task completion) that even a single manually written trajectory demonstration (FIXED) can greatly improve agent performance compared to using the agent ZEROSHOT. Moreover, the LLM agent’s own trajectory annotations are more effective as demonstrations than simplified manually written ones, even if we select them randomly (RANDOM v/s FIXED). Selecting a relevant trajectory, using any method COS or BSR, remains the most effective. Finally, when using more than one trajectory, set-selection (SETBSR) is more effective than independent ranking-based selection. Overall, SETBSR[2] beats ZEROSHOT and FIXED by 30.7 and 15.2 absolute points in TGC, respectively.

**Trajectory demonstrations also make the agent**

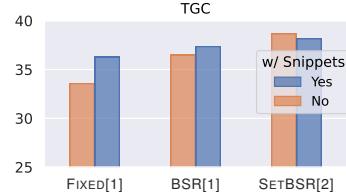


Figure 4: Effect of trajectory and snippet demonstrations (selected using BSR) on TGC of GPT-4o ReAct agent on Test-C.

**more reliable, robust, and efficient.** Trajectory demonstrations have an even greater impact on RTGC and SGC (e.g. SETBSR[2] improves on FIXED’s RTGC and SGC by 20.8 and 23.2 absolute points, respectively). This suggests that using relevant demonstrations is especially effective at making the agent more reliable (across multiple runs of the same task) and robust (across multiple variants of the task). Further, SETBSR[2] also takes 21% fewer steps than FIXED and 47% fewer steps than ZEROSHOT, implying greater efficiency and solution speed.

**Snippet demonstrations generally improve performance with minimal overhead.** Fig. 3 shows the results for the ReAct solver with and without snippet demonstrations for varying selections of trajectory demonstrations. Despite all their benefits, trajectories are very costly to use as demonstrations, and using two trajectories instead of one increases the average cost per task by 40%. On the other hand, snippet demonstrations have very minimal overhead while generally improving performance. However, since they are not as effective as trajectories, the optimal approach is to use as many trajectories as possible and then sprinkle a few snippet demonstrations.

**Demonstrations help even on out-of-domain tasks.** As shown in Fig. 4, demonstrations improve performance even on Test-C, which has more complex tasks involving unseen apps. As expected, the performance is lower than Test-N. However, although none of the annotations demonstrate the use of the unseen apps, we see that both trajectory and snippet demonstrations improve performance.

**Larger LLMs’ annotations can also improve smaller LLM agents.** As shown in Fig. 5, GPT-4o’s annotations also work well as demonstrations for the smaller GPT-4o-mini. As before, it’s clear that both trajectory and snippet demonstrations are effective at improving performance and efficiency. Using SETBSR[2] trajectories with snip-

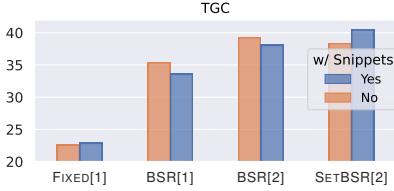


Figure 5: Effect of trajectory and snippet demonstrations (selected using BSR) on TGC of GPT-4o-mini ReAct agent on Test-N.

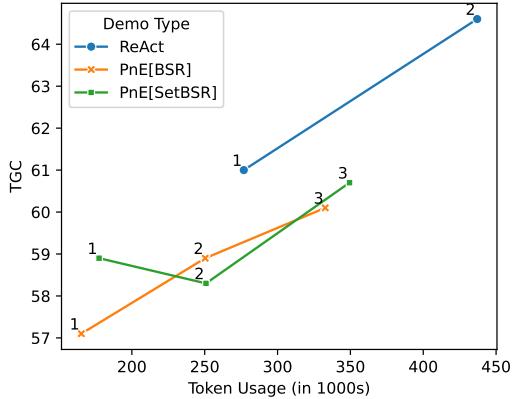


Figure 6: Comparison of GPT-4o PnE (with BSR and SETBSR planner demos) and ReAct solvers with varying number of BSR trajectory demonstrations on Test-N. As trajectories for subtasks are much shorter than for entire tasks, more of the former can be used with a PnE executor than the latter with ReAct solver. However, PnE still underperforms ReAct likely because it attempts to decompose the task prior to any execution.

pet demonstrations improves TGC rate by 14.3 absolute points and reduces the number of steps by 40% compared to using just the FIXED trajectory demonstration.

**Subtask trajectory demonstrations improve PnE solver, but it still underperforms ReAct.** Fig. 6 compares the PnE solver with BSR and SETBSR-selected planner demonstrations and the ReAct solver for a varying number of trajectory demonstrations. It is clear that subtask trajectories have much less overhead than task trajectories. Nevertheless, despite adding more trajectory demonstrations for the executor, the PnE solver cannot match the ReAct solver. This is likely because PnE plans before any execution, which may lead to inaccurate plans.

**With demonstration selection prompted agents can be competitive with trained agents.** Table 2 compares the GPT-4o ReAct agent with a variety of supervised finetuning (SFT), direct preference optimization (DPO), reinforcement learning (RL) approaches to train Qwen-2.5-32B-based agents

Approach	Test-N		Test-C		
	TGC	SGC	TGC	SGC	
SFT	SFT-GT	6.2	1.8	0.8	0.1
	RFT	47.9	26.4	26.4	11.4
	EI	58.3	36.8	32.8	17.6
DPO	DPO-MCTS	57.0	31.8	31.8	13.7
	DMPO	59.0	36.6	36.3	13.7
RL	PPO (learned critic, token)	50.8	28.9	26.4	10.5
	RLOO (traj)	57.2	35.7	36.7	17.4
	GRPO (token)	58.0	36.8	39.5	22.4
	LOOP (bandit)	53.3	33.6	27.7	13.0
	LOOP (turn)	64.1	43.5	40.8	26.5
	LOOP (token)	<b>71.3</b>	<b>53.6</b>	<b>45.7</b>	<b>26.6</b>
NFT	Traj[Fixed]	50.6	30.4	33.6	18.0
	Traj[SetBSR]	65.8	<b>53.6</b>	38.7	24.8
	Traj[SetBSR]+Snippet	65.8	<b>53.6</b>	38.2	23.4

Table 2: With selected trajectory and snippet demonstrations, a prompted GPT-4o agent performs competitively with Qwen-2.5-32B-based agents trained using a variety of approaches spanning supervised finetuning (SFT), direct preference optimization (DPO), and reinforcement learning (RL). The results for trained agents are taken from Chen et al. (2025). We refer the reader to App. B for a brief description of each approach and to Chen et al. (2025) for more details.

from Chen et al. (2025). We provide a brief description of each approach in App. B and refer the reader to Chen et al. (2025) for more details. It is clear that with selected trajectory and snippet demonstrations, GPT-4o ReAct agent can outperform all but the best-trained agents.

## 8 Conclusion

This work studied different design decisions relating to ICL with demonstration selection for LLM agents. We proposed a novel iterative annotation algorithm to automatically annotate training tasks with solutions for use as demonstrations. Using these annotations, we showed that trajectory demonstrations can effectively improve performance, reliability, robustness, and efficiency of LLM agents. Further, since trajectory demonstrations can have a large overhead in terms of inference costs, we also showed that small snippets of trajectories can be used as demonstrations at every step to boost performance with a minimal overhead. Overall, our results suggest that the optimal ICL approach is to use as many trajectory demonstrations as possible and then sprinkle a few snippets, and that this can yield prompted LLM agents that are competitive with state-of-the-art trained agents.

## Acknowledgements

This work was funded in part by the DARPA ANSR program under award FA8750-23-2-0004, in part by NSF CAREER award #IIS-2046873, and in part by NSF CCRI award #CNS-1925741. The views expressed are those of the authors and do not reflect the policy of the funding agencies.

## Limitations

In this work, we were primarily concerned with studying the right form and placement of demonstrations, but used general-purpose encoders for their selection. Further, retrieval-based approaches select demonstrations based on a retrieval key (e.g., task statements) and hence cannot take advantage of additional solution information that may only be available for the demonstration candidates but not for the test task. Future work can explore thus explore demonstration selection approaches more suitable for agentic tasks.

## References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. [In-context examples selection for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms](#).
- Thomas Anthony, Zheng Tian, and David Barber. 2017. [Thinking fast and slow with deep learning and tree search](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5360–5370.
- Hadi Askari, Shivanshu Gupta, Terry Tong, Fei Wang, Anshuman Chhabra, and Muham Chen. 2025. [Unraveling indirect in-context learning using influence functions](#).
- Ben Beglin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. [Super: Evaluating agents on setting up and executing tasks from research repositories](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. [Fireact: Toward language agent fine-tuning](#).
- Kevin Chen, Marco Cusumano-Towner, Brody Huval, Aleksei Petrenko, Jackson Hamburger, Vladlen Koltun, and Philipp Krähenbühl. 2025. [Reinforcement learning for long-horizon interactive llm agents](#).
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samual Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. [Mind2web: Towards a generalist agent for the web](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. 2024. [Workarena: How capable are web agents at solving common knowledge work tasks?](#) In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. [Coverage-based example selection for in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.
- Shivanshu Gupta, Clemens Rosenbaum, and Ethan R. Elenberg. 2024. [Gistscore: Learning better representations for in-context example selection with gist bottlenecks](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. [Language models can solve computer tasks](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Itay Levy, Ben Beglin, and Jonathan Berant. 2023. [Diverse demonstrations improve in-context compositional generalization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3?](#) In

*Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweta Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei ge Chen, Olga Vrousgos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. 2024. [Agentinstruct: Toward generative teaching with agentic flows](#).

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. [Webgpt: Browser-assisted question-answering with human feedback](#).

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. [Agent q: Advanced reasoning and learning for autonomous ai agents](#).

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. [Toollm: Facilitating large language models to master 16000+ real-world apis](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#).

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#).

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. 2021. [Alfworld: Aligning text and embodied environments for interactive learning](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. [Selective annotation makes language models better few-shot learners](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Simeng Sun, Yang Liu, Shuhang Wang, Dan Iter, Chenguang Zhu, and Mohit Iyyer. 2024. [PEARL: Prompting large language models to plan and execute actions over long documents](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 469–486, St. Julian’s, Malta. Association for Computational Linguistics.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. [AppWorld: A controllable world of apps and people for benchmarking interactive coding agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. [Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents](#).

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023. [Intercode: Standardizing and benchmarking interactive coding with execution feedback](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. [Webshop: Towards scalable real-world web interaction with grounded language agents](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. [Compositional exemplars](#)

for in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2024. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ruiwen Zhou, Yingxuan Yang, Muning Wen, Ying Wen, Wenhao Wang, Chunling Xi, Guoqiang Xu, Yong Yu, and Weinan Zhang. 2024a. TRAD: enhancing LLM agents with step-wise thought retrieval and aligned decision. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 3–13. ACM.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024b. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

## A Agent Details

Table 3 shows various hyperparameters used for a ReAct solver and PnE executor when solving a task or subtask, respectively.

### A.1 Prompt Formats

**ReAct Solver** Fig. 7 shows the format of the task context we use for the ReAct solver. It also shows the JSON format in which the agent is constrained to generate its reasoning and action. The hyperparameters used for the ReAct solver are given in Table 3.

**PnE Planner** The prompt template for the planner is given in Fig. 9. The planner is also constrained to generate the plan using a JSON format. For the planner, we use temperature 0.1 and top\_p 0.5 both during annotation and evaluation.

**PnE Executor** The PnE executor is similar to the ReAct solver. The format of the task context used for the PnE executor is given in Fig. 8. A The executor is also constrained to generate its reasoning and using the same JSON format as the ReAct solver. The hyperparameters used for the PnE executor are given in Table 3.

**Prompt Truncation** When the ReAct solver or PnE executor’s prompt exceeds the corresponding context length limit, we truncated it by first hiding the older and longer observations and then hiding any remaining older observations.

### A.2 Demonstration Templates

The templates used to show the task-trajectory, subtask-trajectory, and snippet demonstrations are given in Figs. 10, 11, and 12, respectively.

## B Trained Baselines

We compare with the following training-based approaches baselines from Chen et al. (2025):

- **Ground truth supervised fine-tuning (SFT-GT)**. SFT on ReAct-style transformation of gold solutions.
- **Rejection sampling fine-tuning (RFT)** (Yuan et al., 2023). Collects rollouts generated with the base model and finetunes on successful ones.
- **Expert iteration (EI)** (Anthony et al., 2017). Runs multiple smaller iterations of RFT using the current best model.

- **Direct Preference Optimization + MCTS (DPO-MCTS)** (Putta et al., 2024). Collects preference pairs into a replay buffer using Monte-Carlo Tree Search.
- **Proximal Policy Optimization (PPO)** (Schulman et al., 2017). PPO with a learned advantage estimate.
- **REINFORCE leave-one-out (RLOO)** (Ahmadian et al., 2024). On-policy trajectory-level REINFORCE with leave-one-out advantage estimate.
- **Group relative policy optimization (GRPO)** (Shao et al., 2024). On-policy PPO with normalized leave-one-out advantage estimate.
- **Leave-one-out PPO (LOOP)** (Chen et al., 2025). Off-policy PPO with unnormalized leave-one-out advantage estimate.

Hyperparameter	Annotation		Evaluation	
	ReAct	PnE Executor	ReAct	PnE Executor
temperature	0.1	0.3	0.1	0.1
top_p	0.5	0.5	0.5	0.5
max_context_length	40000	20000	1000000	1000000
max_steps	50	20	50	50
max_tokens	2000	2000	2000	2000

Table 3: Decoding hyperparameters for Annotation and ICL Evaluation experiments for ReAct solver and PnE Executor. `max_context_length` limits the length of the input prompt while `max_tokens` limits the length of the output. `max_steps` is the maximum number of steps for the agent to take.

I am your supervisor and you are a super-intelligent AI Assistant whose job is to assist with my day-to-day tasks involving various apps (e.g., amazon.com, gmail, calendar, etc.). To do this, we will take part in a * multi-turn conversation* with a Python REPL environment that will let you interact with the apps using their APIs.
At every step of the conversation, you will need to reason about your current progress on the task and propose the next action in the form of a Python code snippet for the environment to execute. Your response needs to be in the following JSON format {"thought": <thought>, "action": <action>} where <thought> is your reasoning and <action> the Python code snippet (without any enclosing ```).
The environment will then execute your code and respond with the output. The environment will maintain state across multiple interactions for the on-going task so that any variables defined in one step can be used in subsequent steps. Additionally, since you'll be solving a lot of tasks, it is possible that your reasoning on the current step matches with the reasoning at some step of a task you have solved before. When this happens, I will provide you with the relevant snippet of your conversation solving that task. These may be helpful for your next step.
Here are three key APIs that you can use to get more information about apps and APIs:
# To get a list of apps that are available to you: print(apis.api_docs.show_app_descriptions())
# To get the list of apis under any app listed above, e.g. amazon print(apis.api_docs.show_api_descriptions(app_name='amazon'))
# To get the full input-output specification of a particular api, e.g. amazon app's login api print(apis.api_docs.show_api_doc(app_name='amazon', api_name='show_cart'))
<task_trajectory_demonstrations>
Now, here is the actual task you need to solve using a fresh environment.
My name is Joyce Weaver. My personal email is joyce-weav@gmail.com and phone number is 3155673041. Task: Request \$13 publicly on Venmo from my friend, Stacy, with a note, "For yesterday's meal".
Here are some key guidelines that you need to follow:
(1) Make sure to produce a *single* thought and action at every step and correctly format them as {"thought": <thought>, "action": <action>}. In particular, - the JSON should be valid with any quotes, newlines, etc., properly escaped. - <action> should be just code without any enclosing backticks (```).
(2) Always look at API specifications (using <code>apis.api_docs.show_api_doc</code> ) before calling an API.
(3) Remember you can use the variables in your code in subsequent code blocks.
(4) Remember that the email addresses, access tokens and variables (e.g. <code>amazon_password</code> ) in the example above are not valid anymore. You will be provided a fresh environment to work with. Note, however, that the APIs remain the same so if it's been shown in an example above, you don't need to look at its specification again.
(5) You can use the "supervisor" app to get information about my accounts and use the "phone" app to get information about friends and family.
(6) Many APIs return items in "pages". Make sure to run through all the pages by looping over <code>page_index</code> .
(7) If your action produces output that is too long, the environment will truncate it with '... [HIDDEN FOR BREVITY] ...' replacing the middle part. E.g., this will happen if you print a large number of items returned by a search API. In such cases, you should consider using a more precise query to reduce the number of items returned.
(8) Once you have completed the task, make sure to call <code>apis.supervisor.complete_task()</code> . If the task asked for some information, return it as the answer argument, i.e., call <code>apis.supervisor.complete_task(answer=&lt;answer&gt;)</code> . Many tasks do not require an answer, so in those cases, just call <code>apis.supervisor.complete_task()</code> i.e. do not pass any argument.

Figure 7: Task context for the ReAct solver. Each box is a separate message. `<task_trajectory_demonstrations>` is the placeholder for any trajectory demonstrations (see Fig. 10 for the corresponding template).

I am your supervisor and you are a super intelligent AI Assistant whose job is to assist with my day-to-day tasks involving various apps (e.g., amazon.com, gmail, calendar, etc.). As the tasks can be complex, I will break them down into a sequence of subtasks that you will need to carry out one at a time. To do this, you will take part in a \*multi-turn conversation\* with a Python REPL environment that will let you interact with the apps using their APIs.

At every step of the conversation, you will need to reason about your current progress on the subtask and propose the next action in the form of a Python code snippet for the environment to execute. Your response needs to be in the following JSON format {"thought": <thought>, "action": <action>} where <thought> is your reasoning and <action> the Python code snippet (without any enclosing ```).

Finally, when you have completed the subtask, you will need to say FINISH as action in a separate response in the same format, i.e., {"thought": <thought>, "action": FINISH}

The environment will then execute your code and respond with the output. The environment will maintain state across multiple interactions for the on-going task so that any variables defined in one step can be used in subsequent steps. Additionally, when your reasoning for the current step matches with a task you have solved previously, I will also provide you with the relevant portion of the conversation that solved that task to help you on subsequent steps for the current task.

Here are three key APIs that you can use to get more information about apps and APIs:

```
# To get a list of apps that are available to you:  
print/apis.api_docs.show_app_descriptions()  
  
# To get the list of apis under any app listed above, e.g. amazon  
print/apis.api_docs.show_api_descriptions(app_name='amazon')  
  
# To get the full input-output specification of a particular api, e.g. amazon app's login api  
print/apis.api_docs.show_api_doc(app_name='amazon', api_name='show_cart')
```

#### <subtask\_trajectory\_demonstrations>

Now, here is the actual task you need to solve using a fresh environment.

My name is Joyce Weaver. My personal email is joyce-weav@gmail.com and phone number is 3155673041.  
Task: Request \$13 publicly on Venmo from my friend, Stacy, with a note, "For yesterday's meal".

The above task can be decomposed into the following subtasks that need to be carried out one by one

1. Login to Venmo and save access token in `venmo\_access\_token` variable.
2. Use the `venmo\_access\_token` to search for Stacy in my Venmo contacts and save her user ID in `stacy\_user\_id`.
3. Use the `venmo\_access\_token` to create a payment request to `stacy\_user\_id` for \$13 with the note 'For yesterday's meal'. Ensure the request is set to public.
4. Complete task.

Let's start by solving Subtask 1: Login to Venmo and save access token in `venmo\_access\_token` variable.

Here are some key guidelines that you need to follow:

- (1) Do not worry about the entire task. Focus ONLY on the current subtask and carry it out carefully and correctly.
- (2) Make sure to produce a \*single\* thought and action at every step and correctly format them as {"thought": <thought>, "action": <action>}. In particular,
  - the JSON should be valid with any quotes, newlines, etc., properly escaped.
  - <action> should be just code without any enclosing backticks (```).
- (3) When finished with the current subtask, make sure to say FINISH in a SEPARATE response in the format described above. Also, if it's the last subtask, make sure to call `apis.supervisor.complete\_task()` with the answer, if any, before finishing.
- (4) Always look at API specifications (using apis.api\_docs.show\_api\_doc) before calling an API.
- (5) Remember you can use the variables in your code in subsequent code blocks.
- (6) Remember that the email addresses, access tokens and variables (e.g. amazon\_password) in the example above are not valid anymore. You will be provided a fresh environment to work with. Note, however, that the APIs remain the same so if it's been shown in an example above, you don't need to look at its specification again.
- (7) You can use the "supervisor" app to get information about my accounts and use the "phone" app to get information about friends and family.
- (8) Many APIs return items in "pages". Make sure to run through all the pages by looping over `page\_index`.
- (9) If your action produces output that is too long, the environment will truncate it with '... [HIDDEN FOR BREVITY] ...' replacing the middle part. E.g., this will happen if you print a large number of items returned by a search API. In such cases, you should consider using a more precise query to reduce the number of items returned.

Figure 8: Task context for the PnE executor. Each box is a separate message. <subtask\_trajectory\_demonstrations> is the placeholder for any trajectory demonstrations (see Fig. 11 for the corresponding template).

```
I am your supervisor and you are a super intelligent AI Assistant whose job is to autonomously perform my day-to-day tasks involving various apps (e.g., spotify, simple_note, etc.).
```

```
To do this, you will need to interact with apps using their associated APIs on my behalf.  
Here are the apps:  
{ app_descriptions }
```

```
To help with this, I will provide you with an executor model that will take care of interacting with the APIs. Your job is to produce a plan on how to solve the task given access to this executor. You should respond in the following JSON format:
```

```
{
  "thought": <thought>,
  "plan": [
    "<subtask_1>",
    "<subtask_2>",
    ...
    "<subtask_n>"
  ]
}
```

Here, `<thought>` is a brief description of how you plan to solve the task and `<subtask_i>` is a description of the *i*th subtask in your plan.

**\*\*Key instructions\*\*:**

- (1) Make sure to respond in the JSON format provided above. Don't enclose your response with ` `` `.
- (2) Make sure to define all the relevant variables.
- (3) Each subtask in the plan should be clear and complete so that it can be executed independently. E.g. don't use references such as "previous list" or "repeat subtask 1".
- (4) When unsure, use more subtasks in the plan rather than fewer subtasks.
- (5) The final subtask should be "Complete task." unless the task requires an answer, in which case, it should be "Complete task with answer: <answer>" where <answer> is the answer to the task.

Figure 9: Prompt for the PnE planner. It is followed by a few demonstrative task-plan pairs and the test task.

```
Here is an example of a task and how you can interact with the environment to solve it
```

```
My name is Joyce Weaver. My personal email is joyce-weav@gmail.com and phone number is 3155673041.  
Task: How many playlists do I have in Spotify?
```

```
Lets first find which APIs are available to use in Spotify.  
```python  
print/apis.api_docs.show_api_descriptions(app_name='spotify')  
```
```

```
Output:  
```  
[  
  ...  
  {  
    "name": "login",  
    "description": "Login to your account."  
  },  
  {  
    "name": "logout",  
    "description": "Logout from your account."  
  },  
  ...  
]```
```

⋮

```
Now that the task is completed, I need to mark the task as complete and return the number of playlists found  
```python  
apis.supervisor.complete_task(answer=num_playlists)  
```
```

```
Output:  
```  
Marked the active task complete.  
```
```

Figure 10: Template used for trajectory demonstrations. Yellow boxes are user messages or environment responses, while blue boxes are agent messages (originally in JSON format).

Here is an example of a subtask and how you can interact with the environment to solve it

My name is Jennifer Powell. My personal email is jepowell@gmail.com and phone number is 4288705164.  
Task: Add a comment, "Thank you!", to all the venmo payments I received from my coworkers in the last 5 days (including today), and like those payments.

The above task can be decomposed into the following subtasks that need to be carried out one by one

1. Login to Venmo and save access token in `venmo\_access\_token` variable.
2. Use the `venmo\_access\_token` variable to retrieve a list of coworkers from the Phone app and save it in `coworker\_list`.
3. Use the `venmo\_access\_token` variable to get all payments received in the last 5 days and save them in `recent\_payments`.
4. Filter `recent\_payments` to find payments from `coworker\_list` and save them in `coworker\_payments`.
5. For each payment in `coworker\_payments`, add the comment 'Thank you!' and like the payment using the `venmo\_access\_token`.
6. Complete task.

Let's start by solving Subtask 1: Login to Venmo and save access token in `venmo\_access\_token` variable.

To login to Venmo, I need to find the appropriate API and its specifications. Let's start by checking the available APIs for the Venmo app.

```
```python
print/apis.api_docs.show_api_descriptions(app_name='venmo')
```

```

:

I have successfully logged into Venmo and obtained the access token, which is stored in the `venmo\_access\_token` variable. This completes the first subtask.

```
```python
FINISH
```

```

Figure 11: Template used for subtask trajectory demonstrations. Yellow boxes are user messages or environment responses, while blue boxes are agent messages (originally in JSON format). The main difference with the task trajectory demonstrations (Fig. 10) is that the agent is additionally provided the plan and a summary of code used to solve previous subtasks.

Your reasoning (thought) in the previous step is similar to your reasoning solving some other tasks. Providing you with the relevant snippets of your conversation solving those tasks as they may be helpful for your next step. Each snippet comprises the task statement along with the thought, the action taken, and the output for the matching step as well as the following step (if there is one).

```
# Snippet 1:
Matching Step:
Thought: {thought}
Action:
```python
{action}
```
Output:
```
{output}
```
Following Step:
Action:
```python
{action}
```
Output:
```
{output}
```
# Snippet 2:
...
```

Figure 12: Template used for showing snippet demonstrations. This is a single message appended after to all the messages.