TRAD: Enhancing LLM Agents with Step-Wise Thought Retrieval and Aligned Decision

Ruiwen Zhou
skyriver@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yingxuan Yang
zoeyyx@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Muning Wen
muningwen@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Ying Wen
ying.wen@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Wenhao Wang
wangwenhao-009@cpic.com.cn
China Pacific Insurance
Shanghai, China

Chunling Xi
xichunling@cpic.com.cn
China Pacific Insurance
Shanghai, China

Guoqiang Xu
xuguoqiang-009@cpic.com.cn
China Pacific Insurance
Shanghai, China

Yong Yu
yyu@apex.sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Weinan Zhang*
wnzhang@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

ABSTRACT

Several large language model (LLM) agents have been constructed for diverse purposes such as web navigation and online shopping, leveraging the broad knowledge and text comprehension capabilities of LLMs. Many of these works rely on in-context examples to achieve generalization without requiring fine-tuning. However, few have addressed the challenge of selecting and effectively utilizing these examples. Recent approaches have introduced trajectory-level retrieval with task meta-data and the use of trajectories as in-context examples to enhance overall performance in some sequential decision making tasks like computer control. Nevertheless, these methods face issues like plausible examples retrieved without task-specific state transition dynamics and long input with plenty of irrelevant context due to using complete trajectories. In this paper, we propose a novel framework (*TRAD*) to tackle these problems. *TRAD* first employs ***T**hought **R**etrieval* for step-level demonstration selection through thought matching, enhancing the quality of demonstrations and reducing irrelevant input noise. Then, ***A**ligned **D**ecision* is introduced to complement retrieved demonstration steps with their preceding or subsequent steps, providing tolerance for imperfect thought and offering a balance between more context and less noise. Extensive experiments on ALFWorld and Mind2Web benchmarks demonstrate that TRAD not only surpasses state-of-the-art models but also effectively reduces noise and promotes generalization. Furthermore, TRAD has been deployed in real-world scenarios of a global business insurance company and yields an improved success rate of robotic process automation. Our codes are available at: https://github.com/skyriver-2000/TRAD-Official.

*Corresponding author.

SIGIR '24, July 14-18, 2024, Washington D.C., USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657788

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Large Language Model, LLM Agent, Sequential Decision Making, LLM Reasoning, Information Retrieval

ACM Reference Format:
Ruiwen Zhou, Yingxuan Yang, Muning Wen, Ying Wen, Wenhao Wang, Chunling Xi, Guoqiang Xu, Yong Yu, and Weinan Zhang. 2024. 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 '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657788

1 INTRODUCTION

Large Language Models (LLMs) [3, 30] have achieved remarkable success on various tasks like question answering [42], chatbot [18], and code synthesis [22] due to their wide knowledge and excellent ability of text understanding and generation. Recently, a series of works have attempted to build powerful agents based on LLMs for various sequential decision-making tasks, including text-based games [39], web navigation [4], and information retrieval [45].

Among existing LLM agents, some are trained with large-scale expert data by supervised fine-tuning (SFT) [7, 8, 16], while some are tuning-free and utilize in-context learning (ICL) with few expert demonstration examples [12, 32, 40, 43]. In this paper, we focus the scope on tuning-free ICL methods, as they are highly cost-effective and can seamlessly generalize to different tasks using only a small amount of expert samples. Most existing ICL-based agents are prompted with expert trajectories carefully selected by human [26, 36, 40], which work well when few expert trajectories are available. However, when we have access to a large dataset of expert trajectories or an expert policy, the automatic and personalized selection of expert trajectories for each task instruction becomes necessary, and can have an essential influence on task performance.

Recently, Zheng et al. [43] study the problem of demonstration selection and propose *Synapse*, which retrieves relevant expert trajectories by task meta-data, and then prompts LLMs with these retrieved trajectories. *Synapse* performs well on computer control

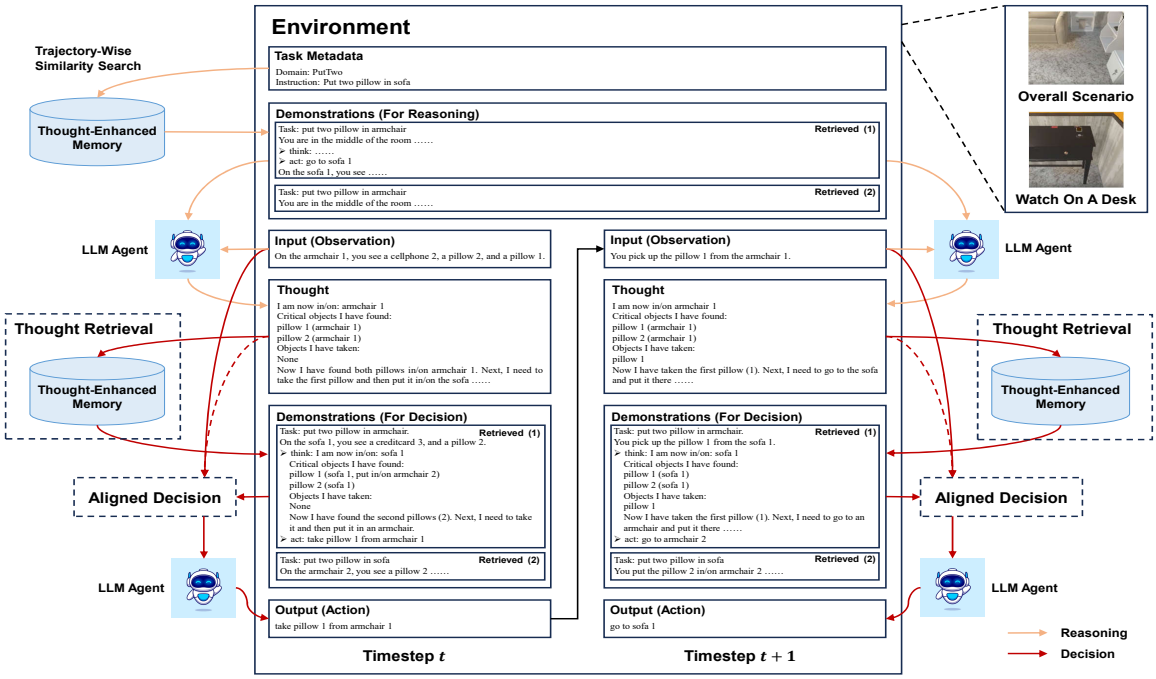


Figure 1: An overall illustration of *TRAD* agent (on ALFWorld [28] enviroment). *TRAD* first pre-processes expert trajectories, labeling each step with high-quality thoughts. At inference time, *TRAD* first conducts *thought retrieval*, which generates thought with trajectory-wise retrieved demonstrations as the query and keys for a more precise step-wise demonstration retrieval. Given the retrieved steps, TRAD employs *aligned decision* module to complement their temporally neighboring steps and corresponding position information (Fig. 2). Finally, the next action is generated according to the enhanced demonstration.

tasks (MiniWob++ [25]) and web navigation tasks (Mind2Web [4]). Nevertheless, retrieving and prompting with complete trajectories can be problematic in the following three aspects.

Plausible examples. Sometimes generalization to data from various domains can be critical. For example, in cross-website and cross-domain subsets of Mind2Web, agents operate on websites unseen in the training set, i.e., memory. In this case, retrieving trajectories with only task meta-data is very likely to provide plausible examples, which share similar task instructions to the current one but require totally different solutions. As shown by experiments in [43], plausible examples provide no more information than random examples and can usually mislead LLM agents to wrong decisions.

Context limit of LLMs. When facing tasks with long horizons and complex observations, prompting with complete trajectories will result in input sequences longer than the allowed length of LLMs. *Synapse* thus has to reduce the number of trajectory examples or even fail to complete the task directly. Though some long-context LLMs can receive very long prompts, the performance can be harmed due to the issue of long-term forgetting [29].

Irrelevant information in prompts. LLMs are found sensitive to their prompts, and can easily copy their recent input [10, 20]. The decision at the current timestep can be related to very few steps in a retrieved trajectory, while other steps do not provide any helpful information. Therefore, irrelevant steps will have unpredictable effects on the decision of LLM agents. As shown by our experiments, they negatively impact the performance most of the time.

To address the problems of trajectory-wise retrieval and prompting, we delve into step-wise demonstration retrieval and prompting. We discover that, via demonstrating with relevant steps, the input context of the LLM agent can be significantly reduced. Thus, the issue of context limit and irrelevant information can be alleviated. Therefore, the critical part is to retrieve step demonstrations that are truly relevant and helpful. To achieve this, we utilize step-by-step reasoning, i.e. *Chain-of-Thought* technique [36], to abstract the state at each timestep as retrieval queries and keys. The generated *thoughts* can involve historical information or future plans, which is more specific with state transitions and helpful in reducing plausible examples.

In this paper, we propose ***T**hought **R**etrieval* and ***A**ligned **D**ecision* (*TRAD*), a novel framework that achieves step-wise demonstration retrieval via thought matching and enhances the context for action prediction with temporally neighboring steps and their order information. Our contribution can be summarized in four-folds:

- We propose a *thought retrieval* method, where we label thoughts for expert demonstration steps in advance with an LLM, prompt LLM agents to reason at inference time, and achieve step-wise retrieval by a similarity search on thought. To the best of our knowledge, this is the first work that enables the LLM agent with thought retrieval techniques for sequential decision-making.
- Based on the thought retrieval operation, we further propose an *aligned decision* method, where we supply the retrieved steps with their temporal neighbors to overcome imperfect thoughts and enhance task-relevant information.
- We conduct extensive experiments and analysis on Mind2Web [4] tasks and ALFWorld [28], showing that TRAD achieves state-of-the-art (SoTA) performance compared to existing works. *TRAD* brings a 2.99% improvement over the strongest baseline (93.78% → 96.77%) to the success rate (SR) on ALFWorld. On Mind2Web, *TRAD* improves element accuracy, step SR, and SR remarkably over the powerful *Synapse* agent [43] by 2.1%, 1.4%, and 0.5%.

- We have deployed TRAD to the real-world robotic process automation scenarios of a global business insurance company, where *TRAD* enables the LLM agent to significantly improve the success rate in a bunch of practical tasks. In average, *TRAD* raises step SR from 90.2% to 98.1% and SR from 65.0% to 92.5%.

2 RELATED WORK

2.1 LLM Agents

In recent years, there has been a rapidly growing trend to utilize pre-trained LLMs as the central controller to obtain human-level decision-making capabilities [33]. Among these works: Nakano et al. [16] fine-tune the GPT-3 [3] model for question answering in a text-based web browsing environment. Yao et al. [38] develop WebShop, a simulated e-commerce website environment, and fine-tune a BERT [5] model with imitation learning and reinforcement learning. Yao et al. [40] insert a reasoning section between observation input and action output, significantly improving the performance on ALFWorld [28] and WebShop [38] tasks. Shinn et al. [26] further improve over [40] via reflection on task feedback. Schick et al. [24] teach LLMs to use external tools via simple APIs in a self-supervised learning way. Park et al. [19] introduce *Generative Agents*, extending LLMs with natural language memories and retrieving them dynamically to plan behavior. Wang et al. [35] propose *DEPS*, an interactive planning approach, which facilitates better error correction by integrating a description of the plan execution process and an explanation of failure feedback. Wang et al. [32] employ an exploration curriculum, a growing skill library, and a novel iterative prompting mechanism, leading to better proficiency in playing Minecraft. Deng et al. [4] construct the Mind2Web dataset from real-world webpages, which consists of three subsets requiring different degrees of generalization, and compare the performance of imitation learning and few-shot inference.

As can be seen above, most existing LLM agents focus on: 1) improving task performance by direct fine-tuning [4, 16, 38]; 2) enhancing planning or reasoning by explicitly prompting [26, 35, 40]; 3) extending the application with an external memory or tool library [19, 24, 32]. However, providing more relevant information in prompts, as a fundamental way to elicit better task understanding, does not receive sufficient attention. When near-optimal demonstrations are accessible, selecting few-shot demonstrations properly can be a simple yet very effective way to improve task performance, which is investigated in our work.

2.2 In-Context Example Selection

LLMs have been shown excellence of few-shot learning [3], and the selection of in-context examples can yield a significant improvement on the overall performance. Liu et al. [15] first propose to retrieve the k-nearest neighbors (k-NN) of the input as in-context examples, and achieve improvement over random retrieval baselines. Rubin et al. [23] select relevant samples with an encoder trained with label similarity, and obtain better performance over BM25 and pre-trained encoder baselines. Zhang et al. [41] consider selecting and labeling unlabeled examples as demonstrations to achieve the best performance, and view this problem as a sequential decision making task to solve by reinforcement learning. Wu et al. [37] further select examples in a subset recalled from k-NN search via minimizing the entropy of output.

IRCoT [31] should be the most relevant work to ours, which retrieves relevant documents with reasoning steps on question-answering tasks. However, their method retrieves with a complete historical trajectory and accumulates retrieved documents over time, which are not transferable to complex sequential decision-making tasks, and we propose a method different from theirs in that: (i) Our method focuses on both providing more relevant demonstrations and reducing irrelevant context for decision-making tasks, while theirs is limited to question-answering tasks and only addresses the first issue. (ii) Our method retrieves different steps across timesteps and complements the retrieval results with temporal information, while theirs only accumulates relevant documents at every reasoning step and heuristically cuts off the earliest ones to fit in the context limit of LLMs. (iii) Our method prepares pseudo-golden thoughts for expert trajectories in the memory to enable retrieval with trajectories without thoughts, and utilizes single-step thoughts as both queries and keys for precise retrieval, while theirs uses thoughts only as queries with raw documents as keys.

The selection of in-context examples has been studied thoroughly for non-sequential tasks like question answering and sentiment analysis. However, for sequential decision-making tasks, how to select the examples to improve the overall performance remains unclear. Zheng et al. [43] propose a trajectory-wise retrieval solution, while a more precise step-wise solution is still desired as discussed in Section 1, which motivates our work.

2.3 LLM Planning and Reasoning

Our work proposes to use thought, which can be viewed as a general abstraction of the current state, as queries and keys for retrieval. Nevertheless, plans, code comments, and any other text that extracts comprehensive information about the current state can serve as an alternative. Therefore, we particularly review some remarkable reasoning and planning works based on LLMs, and most of them are complementary to our work.

Wei et al. [36] first introduce the concept of *Chain-of-Thought* (CoT) by providing with explicit step-by-step reasoning process in example outputs improving performance on arithmetic, commonsense, and symbolic reasoning tasks. Wang et al. [34] further find that a single reasoning path can be sub-optimal, and propose *self-consistency* to address this problem by sampling multiple reasoning paths. For efficient yet flexible search of reasoning paths, Yao et al. [39] apply tree search with self-evaluation to find globally excellent thoughts. Besta et al. [2] later extend the tree-search structure to a graph search for even better flexibility and overall performance.

The works mentioned above consider problems that are non-sequential or solvable by a single complete reasoning path after receiving the input. For harder sequential decision-making problems: Zhou et al. [44] introduce *least-to-most* prompting to solve hard problems by decomposing the problem and solving sub-problems sequentially. *ReAct* proposed by Yao et al. [40] interacts with the environment in a reason-then-act style, which enriches the context for action prediction. *Code-as-Policies* [13] writes executable codes for embodied control by hierarchically expanding undefined programs, which can be viewed as implicit reasoning or CoT process. Liu et al. [14] propose to incorporate the strength of classical planners by translating the original problem into a PDDL [1] problem to solve by classical planners. Hao et al. [9] and Ding et al. [6] share a similar insight that reasoning can be implemented indeed by planning, where [9] use LLMs as world models and [6] conduct MCTS for thought generation with a light-weight extra network.

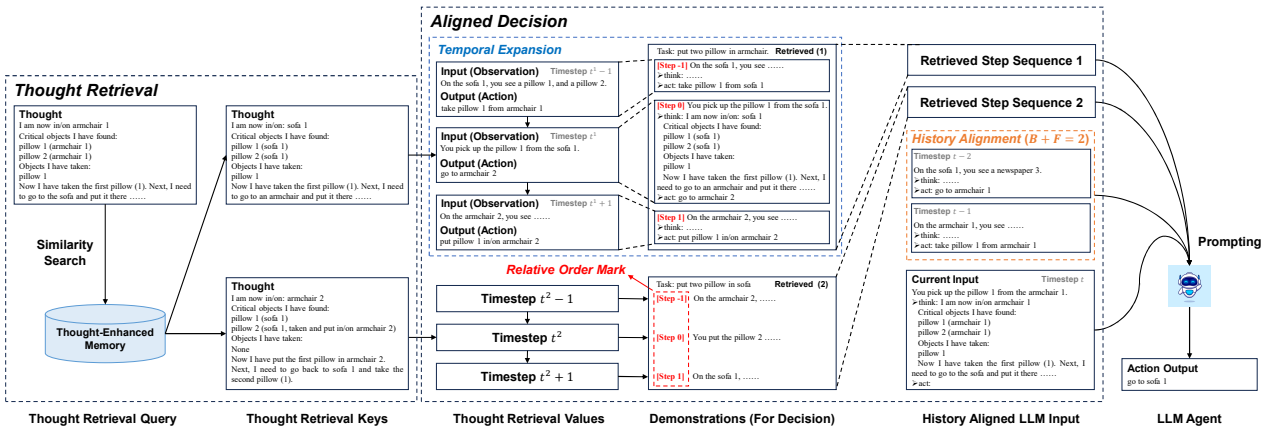


Figure 2: An illustration of our *aligned decision* method, where $B = F = 1$ and the i-th retrieved step is at time t^i in its trajectory. The aligned decision method consists of three sub-processes to the retrieved step demonstrations and prompting: 1) Temporal Expansion: Collect at most B previous steps and F subsequent steps for each retrieved step, and transform each step into a sequence of length $B + F + 1$ from $t^i - B$ to $t^i + F$; 2) Relative Order Mark: For each step in one demonstration step sequence, we label its relative position to the retrieved step in this sequence, i.e., the previous one ($t^i - 1$) with `[Step -1]` and the next one ($t^i + 1$) with `[Step 1]`; 3) History Alignment: For the current episode, we complement current observation (and thought, optional) with $B + F$ previous steps to enrich information and align with demonstrations.

To summarize, LLM planning and reasoning have continuously received huge attention from researchers in recent years. This makes our work flexible and improvable with more powerful planning and reasoning methods in the future.

3 THE TRAD FRAMEWORK

As discussed in Section 1, trajectory-wise retrieving and prompting lead to issues of plausible examples, LLM context limits, and irrelevant information. To resolve these issues, we propose a novel method called ***T**hought **R**etrieval* and ***A**ligned **D**ecision* (*TRAD*), as illustrated in Fig. 1. Our *TRAD* agent utilizes thought, which is obtained by reasoning about its current state, to retrieve similar steps from expert trajectories, and is then complemented with steps temporally correlated to the retrieved ones and their temporal position information to predict the action. Formally, our *TRAD* agent can be summarized in one equation:

$$\pi_{TRAD}(a_t|\xi, o_{0:t}, a_{0:t-1}) = \text{LLM}(\text{AD}(\text{TR}(\tau_t, \mathcal{M}), \xi, o_{0:t}, a_{0:t-1})) \,,$$

where ξ is the current task, $o_{0:t}$ and $a_{0:t-1}$ are historical observations and actions, τ_t is the thought generated by LLM about the current state, TR and AD denote our *thought retrieval* and *aligned decision* modules, and $\mathcal{M}$ refers to the thought-enhanced memory. We will present each module of *TRAD* in the following subsections.

3.1 Thought Preparation

Most expert trajectories, collected by either human or other expert agents, do not contain their reasoning process. Therefore, before we utilize thoughts for retrieval, we should prepare thoughts for each demonstration step in the memory. Specifically, we start from a small subset of expert demonstrations and provide thoughts written by human experts for each step in it. Given this small subset as few-shot examples in prompts, we can query LLMs to label thoughts for a large memory. Although ground-truth actions are not accessible at inference time, we can prompt LLMs with them to generate thoughts of higher quality. In this way, LLMs produce pseudo-golden thoughts consistent with expert actions, and we obtain a *thought-enhanced memory* $\mathcal{M}$ supporting both trajectory-wise retrieval with task meta-data and step-wise retrieval with thoughts.

3.2 Thought Retrieval

Given pseudo-golden thoughts for all steps in the memory, which can serve as keys for step-wise similarity search, we now present our *thought retrieval* method to select relevant demonstrations at inference time. To be specific, we first conduct trajectory-wise demonstration retrieval as in [43] for thought generation. With these trajectory demonstrations, at each timestep t we prompt the LLM to generate a thought τ_t for step-wise retrieval. Note that this process does not directly effects decision-making, hence it can be further simplified if necessary and the issues mentioned in Section 1 will not impact the agent severely.

With the thought τ_t, which can be viewed as an abstraction, about current state, we conduct dense retrieval to find relevant steps in the *thought-enhance memory* $\mathcal{M}$. Here any encoder pre-trained on a large corpus for retrieval, e.g., Sentence-BERT [21] and DPR [11], can be utilized to encode the query thought and key thoughts into dense vectors. Using a cosine similarity between the query and keys, we then collect top-K relevant steps that belong to mutually different trajectories and their corresponding task instructions.

3.3 Aligned Decision

Now we have relevant demonstration steps from *thought retrieval*. However, the query thought can be imperfect due to the lack of expert action information at inference time. As we will show by ablation experiments in Section 4.4, directly using these steps to form single-step demonstrations does not provide satisfactory performance, which is similar to the plausible example issue of trajectory-wise retrieval. Therefore, we propose an *aligned decision* method to incorporate more information during the decision-making process. *Aligned decision* complements LLM agents with steps temporally correlated to the retrieved ones and their temporal position information. As illustrated in Fig. 2, the *aligned decision* method can be decomposed into following three sub-processes.

Temporal expansion. For each retrieved step, we first expand it into a step sequence involving B previous steps and F subsequent steps. When the number of previous or subsequent steps is smaller than B or F, we simply take all previous or subsequent steps. This

Table 1: Success Rate of Different Methods on 6 Types of ALFWorld Tasks. We compare *TRAD* with *ReAct* [40], *Synapse* [43], and their strong combination. *TRAD* significantly outperforms all baselines in terms of overall performance, achieves the best performance in 5 out of 6 types of task, and shows a decent performance on Heat task. The improvement of *TRAD* over all baselines on overall performance is statistically significant (measured by student's t-test at $p < 0.05$).

Method	Put	Examine	Clean	Heat	Cool	PutTwo	All
ReAct (Random)	0.8472±0.0393	0.8333±0.0454	0.9570±0.0304	0.8841±0.0205	0.9841±0.0224	0.8431±0.0277	0.8980±0.0093
ReAct (Fixed)	0.7778±0.0708	**0.9630±0.0262**	0.9032±0.0263	**0.9275±0.0205**	**1.0000±0.0000**	0.8824±0.0480	0.9055±0.0186
Synapse	0.9444±0.0196	0.7037±0.0262	0.9355±0.0000	0.9130±0.0615	**1.0000±0.0000**	0.8039±0.0555	0.8955±0.0106
Synapse + ReAct	0.9167±0.0340	0.9444±0.0454	**1.0000±0.0000**	0.9130±0.0000	0.9524±0.0000	0.8627±0.0555	0.9378±0.0035
TRAD (Ours)	**0.9583±0.0000**	**0.9630±0.0524**	**1.0000±0.0000**	0.8986±0.0205	**1.0000±0.0000**	**0.9804±0.0277**	**0.9677±0.0141**

transforms each retrieved step into at most $(B + 1 + F)$ temporally successive steps, allowing LLM agents to correct their imperfect thoughts by looking at more related steps at decision-making time.

Relative order mark. Given K expanded step sequences by *temporal expansion*, we insert a mark for each step (including the retrieved ones) indicating the relative position w.r.t. its corresponding retrieved step, and incorporate this rule of mark in the prompt for decision. For example, the last step before the retrieved one will be marked as `[Step -1]`, the retrieved step as `[Step 0]`, and the first step after the retrieved one as `[Step 1]`. This provides temporal information about the $(B + 1 + F) \times K$ demonstration steps, and promotes more accurate demonstration following.

History alignment. Sometimes the optimal policy to a task, like ALFWorld, can be history-dependent, hence using single-step input for action prediction is unreasonable. Since we aim to reduce input content for less forgetting and noise, we should neither use all historical observations and actions. Moreover, even if we include previous actions as auxiliary information, there exists a mismatch where expert demonstrations are given as sequences of length $B + 1 + F$ while current input is a single step. We thus propose to insert at most $B + F$ previous input-output pairs (i.e. $o_{t-(B+F):t-1}$, $a_{t-(B+F):t-1}$) before current input o_t, transforming current input into a similar sequence to demonstrations.

4 EXPERIMENTS

In this section, we aim to study the following research questions:

RQ1 How does *TRAD* perform against existing SoTA methods?
RQ2 Does *thought retrieval* help to reduce irrelevant context and improve the overall performance?
RQ3 Does *aligned decision* help to supply information when generalization is important?
RQ4 Diving into *aligned decision*, are all *temporal expansion* (TE), *relative order mark* (ROM), and *history alignment* (HA) necessary for improvement?
RQ5 How will the performance and advantage of *TRAD* be effected by critical hyper-parameters?

4.1 Experiment Setup

To answer the above research questions, we conduct extensive experiments on ALFWorld [28] and Mind2Web [4] tasks. For each task, we introduce the details of evaluation as follows.

ALFWorld [28] is a text-based game aligned with ALFRED [27] benchmark. It involves 6 types of tasks where an agent must take a series of actions (e.g. *go to shelf 1*, *take vase 2 from shelf 1*, *put vase 2 in/on cabinet 5*) to achieve a high-level goal given by a natural language instruction (e.g. *put some vase on a cabinet*). This environment is challenging in three aspects: 1) Agent should determine likely places of a householding object and explore them one by one to find such object; 2) Agent should understand the usage of some objects like microwaves, fridges, and desklamps; 3) Some tasks can take an agent more than 30 steps to solve, requiring substantial long-term memorization.

Following Shridhar et al. [28], we evaluate on the subset of 134 out-of-distribution tasks, comparing the task success rates of *TRAD* to *ReAct* [40] and *Synapse* [43] (without state abstraction as observations are short). As *ReAct* and *Synapse* has provided sufficiently strong performances, we do not include more complex reasoning and planning baselines and corresponding variants of *TRAD* due to our API cost limit. Note that the original *ReAct* uses fixed but not retrieved trajectories as demonstrations, hence we test two *ReAct* baselines to eliminate such an effect:

- *ReAct* (Fixed) uses fixed human-written trajectories as demonstrations;
- *ReAct* (Random) randomly samples trajectories from the memory as demonstrations.

For fair comparison, *TRAD* uses thoughts in exactly the same format as *ReAct*, and shares a consistent memory of expert trajectories with *Synapse*. We also add a strong baseline (*Synapse+ReAct*) combining the trajectory-level retrieval in *Synapse* and the reasoning in *ReAct*. On ALFWorld, all methods are built with GPT-4 [17] and 2 in-context examples.

Mind2Web [4] is an HTML-based web navigation benchmark collected from real-world webpages, involving various tasks such as searching, trip booking, social network subscription, etc. It contains 3 subsets, i.e., cross-task, cross-website, cross-domain. This environment is challenging in two aspects: 1) Existing LLM agents can hardly understand HTML input well; 2) Unseen tasks and websites can require substantial generalization. Deng et al. [4] find that the cross-website and cross-domain subsets are significantly harder due to the need for generalization to unseen websites.

Since Mind2Web was introduced only about half a year ago, there is a lack of suitable baseline algorithms, and thus we compare our *TRAD* agent to *Synapse* [43] and *ReAct* [40]. Following Zheng et al. [43], we evaluate on all 3 subsets, comparing the element accuracy (Ele. Acc), step success rate (Step SR), and trajectory success rate (SR). For fair comparison, we follow [43] and summarize observations into 5 web elements with the pre-trained element ranker provided by [4] for all methods. Since the observations are still very complex on Mind2Web, including thoughts for every step in trajectories is not available, hence: 1) we do not include a *Synapse + ReAct* baseline; 2) *TRAD* generates thoughts and predicts actions by a single-step prompt with the current observation and previous actions (without previous observations). To eliminate the effect of prompting style and reasoning, we build two *ReAct* baselines using the same format of prompt as *TRAD*:

- *ReAct* (Random), for which we prompt *ReAct* with completely random demonstration steps.

Table 2: Results (%) of all methods on Mind2Web benchmark. *TRAD* achieves the best overall performances and the most improvement on the two harder subsets, especially the most out-of-distribution Cross-Domain subset. The improvement of *TRAD* over all baselines on three overall metrics is statistically significant (measured by student's t-test with $p < 0.01$).

Method	Cross-Task			Cross-Website			Cross-Domain			All		
	Ele. Acc	Step SR	SR	Ele. Acc	Step SR	SR	Ele. Acc	Step SR	SR	Ele. Acc	Step SR	SR
MindAct	20.3	17.4	0.8	19.3	16.2	0.6	21.0	18.6	1.0	20.6	18.0	0.9
ReAct (Random)	31.0	24.7	1.6	25.7	19.1	0.6	27.9	22.9	1.8	28.3	22.7	1.6
ReAct (Relevant)	31.3	26.0	1.2	26.7	20.5	0.6	28.0	23.1	1.6	28.5	23.4	1.4
Synapse w/o Retrieval	33.1	28.9	3.2	27.8	22.1	**1.1**	30.0	26.5	1.4	30.4	26.4	1.7
Synapse	34.4	30.6	2.0	28.8	23.4	**1.1**	29.4	25.9	1.6	30.4	26.6	1.6
TRAD (Ours)	**35.2**	**30.8**	**3.6**	**30.4**	**24.0**	0.6	**32.0**	**28.0**	**2.0**	**32.5**	**28.0**	**2.1**

- *ReAct* (Relevant), for which we prompt *ReAct* with demonstrate steps randomly chosen from trajectories retrieved by *Synapse*.

We do not include the *ReAct* (Fixed) baseline as it is hard to write or pick demonstrations commonly helpful for such diverse test sets. We also provide the results of the simplest MindAct [4] baseline without reasoning and retrieval for completeness. On Mind2Web, all methods are built with GPT-3.5-turbo and 3 in-context examples.

4.2 Evaluation on ALFWorld

The success rate of each method tested on ALFWorld is shown in Tab. 1. Generally, our *TRAD* agent achieves an average success rate of 96.77%, significantly outperforming *ReAct* (∼90%), *Synapse* (89.55%), and even their strong combination (93.78%). It is also worth noting that the worst trial of *TRAD* among 3 random seeds achieves a success rate of 94.8%, outperforming the best trial produced by any other method (94.0%).

Down to the success rate on each type of task, we observe that the success rate of each method varies more on the simplest *Put* task and the hardest *PutTwo* task. We discuss the results of these two tasks respectively as follows:

- On the simplest *Put* task, *ReAct* performs even more poorly than other harder tasks. We find that the two vital reasons for *ReAct*'s failure on *Put* task are incorrect location and usage of objects, e.g. trying to put an object in a closed safe. As this issue can be alleviated through a combination with *Synapse*, the necessity of retrieving relevant demonstrations thus justified.
- *TRAD* achieves the largest improvement on the hardest *PutTwo* task. *PutTwo* requires to correct the locations of two objects and a comprehensive understanding of its task process. Since *TRAD*'s outstanding performance on this hardest task is obtained from a reduced input context at decision-making time, we can conclude that step-wise *thought retrieval* is helpful by reducing the noise of irrelevant steps and finding relevant examples more precisely.

4.3 Evaluation on Mind2Web

To verify the capability of *TRAD* under more realistic scenarios, we compare *TRAD* to *ReAct* and the current SoTA method, *Synapse*, on the Mind2Web benchmark, and the results are shown in Tab. 2. We also include the results of *Synapse* without retrieval here to better illustrate the effect of different retrieval methods.

Generally, *TRAD* achieves the highest performance in terms of all 3 metrics averaged on 3 subsets. Considering that the trajectory-level retrieval of *Synapse* only brings marginal boosts on Cross-Task and Cross-Website subsets, and even slightly impacts the performance on the Cross-Domain subset, our *TRAD* method can be thus justified in two aspects:

- By reducing input context and utilizing step-wise relevant demonstrations, our step-wise *thought retrieval* helps more than the trajectory-wise retrieval with task meta-data in *Synapse* to improve on the simplest Cross-Task subset.
- By eliminating plausible examples and complementing temporal correlated steps, *aligned decision* helps to improve on the two harder subsets, especially the most out-of-distribution Cross-Domain subset.

Furthermore, we observe that the two *ReAct* baselines perform poorly on this task, which indicates that:

- The thoughts generated by GPT-3.5-turbo on Mind2Web tasks are not sufficient for LLM agents to infer the correct action.
- The single-step prompting style which removes previous observations does not benefit overall performance.

On the contrary, *TRAD* utilizes these imperfect thoughts for retrieval rather than direct decision-making, and is complemented with temporally correlated steps via *aligned decision*. Therefore, *TRAD* is not negatively impacted by the imperfect thoughts, but transforms them into helpful information.

Before we start the study on detailed design and hyper-parameter choices of *TRAD*, we can summarize our performance evaluation on ALFWorld and Mind2Web benchmarks and answer the first three research questions as follows.

Answer to RQ1: On both householding (ALFWorld) and web navigation (Mind2Web) tasks, *TRAD* significantly outperforms curernt SoTA methods and becomes the new SoTA method.

Answer to RQ2: On ALFWorld benchmark, *Synapse* + *ReAct* generates thoughts in exactly the same way with our *TRAD*, and uses entire relevant trajectories (more information than *TRAD*) as demonstrations for action prediction. However, *TRAD* shows obvious advantage over this baseline. Therefore, we can conclude that *TRAD* benefits from more relevant demonstrations and less irrelevant input context brought by *thought retrieval*.

Answer to RQ3: On Mind2Web benchmark, *TRAD* achieves the most improvement over *Synapse* on the Cross-Domain subset which requires the most generalization. Therefore, we can tell that the *aligned decision* method complements critical information for decision-making on unseen input.

4.4 Ablation Studies

We have verified the effectiveness of *TRAD* on two different scenarios, i.e., automatic householding and web navigation. Next, we are to examine the effect of each module in *TRAD*. Due to our limited budget for API usage, all ablation studies are conducted on the Mind2Web benchmark with GPT-3.5-turbo.

Table 3: Results (%) of ablation studies on Mind2Web benchmark. TE builds the basic structure of *aligned decision* and is thus critical for performance boost on all three subsets. HA and ROM work well to promote generalization on the two harder Cross-Website and Cross-Domain subsets but provide little help on the Cross-Task subset. The improvement of *TRAD* over all ablation baselines on Ele. Acc and Step SR is statistically significant (measured by student's t-test with $p < 0.05$).

Method	Cross-Task			Cross-Website			Cross-Domain			All		
	Ele. Acc	Step SR	SR	Ele. Acc	Step SR	SR	Ele. Acc	Step SR	SR	Ele. Acc	Step SR	SR
TRAD w/o TE	34.2	28.4	1.2	27.4	20.4	0.6	29.1	24.0	1.4	30.0	24.5	1.3
TRAD w/o HA	**36.2**	**31.1**	**4.0**	28.3	22.2	0.6	29.4	24.9	1.8	30.8	25.9	**2.1**
TRAD w/o ROM	35.7	30.5	3.6	28.9	22.3	0.6	31.5	27.2	1.9	32.1	27.2	2.0
TRAD (Ours)	35.2	30.8	3.6	**30.4**	**24.0**	0.6	**32.0**	**28.0**	**2.0**	**32.5**	**28.0**	**2.1**

4.4.1 The Effect of Aligned Decision. First, we study the effect of macro building blocks of *TRAD*. Since eliminating *thought retrieval* will disable *aligned decision* at the same time and break the framework fundamentally, we do not remove the *thought retrieval* module, but ablate each component of *aligned decision*, i.e., *temporal expansion* (TE), *relative order mark* (ROM), and *history alignment* (HA), and compare the corresponding performances. The results are shown in Tab. 3.

From Tab. 3, we observe that the performance without each component varies differently on the simplest Cross-Task subset and the two harder subsets:

- On the harder Cross-Website and Cross-Domain subsets, the elimination of all three modules in *aligned decision* results in a significant performance drop, and the effect of TE is the most significant. This is intuitive, since only retrieved steps are provided to the agent without TE, and thus the agent becomes more vulnerable to imperfect thoughts.
- On the simplest Cross-Task subset, however, HA and ROM are not that helpful and even cause performance drop. As discussed earlier (Section 1 and Section 3.3), when the issue of plausible examples is not severe, reducing context and prompting with the most relevant demonstration becomes the dominant factor of performance boost. Therefore, only TE remains beneficial for recovering from imperfect thoughts, while the other two components lead to sub-optimal performance.

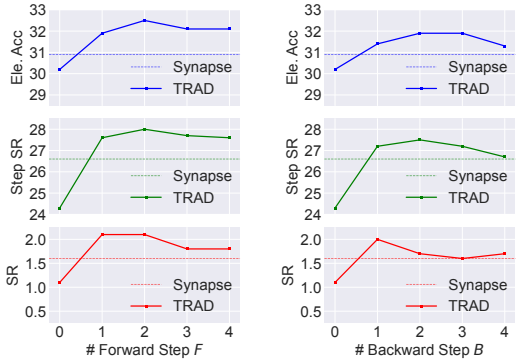
Generally, the *aligned decision* method provides more information about the source trajectories of retrieved steps and the current trajectory, and helps especially for scenarios where generalization is essential. We can now summarize these observations and answer the fourth research question.

Answer to RQ4: Among the sub-processes in *aligned decision*, 1) *temporal expansion* provides tolerance for imperfect thoughts and improves the overall performance of *TRAD* consistently; 2) *relative order mark* and *history alignment* complement *TRAD* with temporal information about the trajectories of retrieved steps and the current trajectory, which serve as useful context for out-of-distribution decision-making but may become less useful for in-distribution decision-making.

4.4.2 The Effect of Expansion Steps B and F. Next we vary a critical hyper-parameter, the number of temporal expansion steps, and investigate how the overall performance will change accordingly. To avoid an expensive grid search on B and F, we consider only one-side expansion by varying B or F from 0 to 4 with the other set to 0. The results are shown in Fig. 3.

From Fig. 3, we can have the following observations:

- Both forward expansion ($F > 0$) and backward expansion ($B > 0$) achieve improvement compared to no expansion ($F = B = 0$). This justifies our design of *aligned decision*.



(a) Varying F **(b) Varying B**

Figure 3: The effect of varying subsequent steps F and previous steps B on Mind2Web benchmark. Solid lines correspond to the performance metrics of *TRAD* given different F and B, and the dashed lines correspond to the *Synapse* baseline. Forward expansion ($F > 0$) generally provides more improvement than backward expansion ($B > 0$) over no expansion ($F = B = 0$) and the *Synapse* baseline. F or B does not help more when they are sufficiently large.

- Either forward expansion or backward expansion does not benefit from increasing a large enough F or B further. This proves our hypothesis that irrelevant context too far from the current state is of little value and even noisy.
- Generally, forward expansion performs better than backward expansion when varying F and B. The reason for this phenomenon might be that historical information has been incorporated in thoughts and thus future information helps more.
- *TRAD* achieves its best performance when $F = 2$ and $B = 0$, and consistently outperforms *Synapse* with forward expansion.

4.4.3 The Effect of Demonstration Amount K. Finally, we look into a common yet important hyper-parameter, the number of retrieved demonstrations K, and see how the advantage of *TRAD* over the baseline (*Synapse*) will change given different $K \in \{1, 2, 3, 4, 5\}$. We show the results in Fig. 4. Note that the trajectory-wise prompting in *Synapse* frequently exceeds the context limit when $K = 5$, and thus we omit this result.

From Fig. 4, we see that K has a mild effect on the performance of *TRAD* and *Synapse*, and that the advantage of *TRAD* over *Synapse* consistently remains for all $K \in \{1, 2, 3, 4\}$.

With results in Section 4.4.2 and Section 4.4.3, we now respond to our last research question.

Answer to RQ5: The performance and advantage of *TRAD* generally remains stable with different hyper-parameter choices, i.e., temporal expansion steps, number of retrieved demonstrations. Its

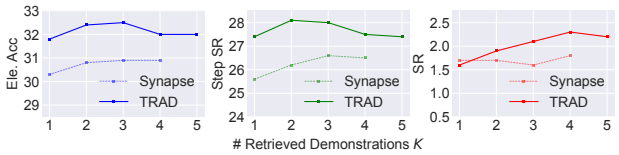


Figure 4: The effect of varying the number of retrieved demonstrations K on Mind2Web benchmark. Solid lines correspond to the performance metrics of *TRAD* given different K, and the dashed lines correspond to the *Synapse* baseline. K has a mild effect on the performance of *TRAD* and *Synapse*, and the advantage of *TRAD* over *Synapse* remains stable when K varies.

performance and advantage only degrade when using long backward extension, which is possibly due to the fact that historical information has already been incorporated in thoughts and does not provide further help for decision-making.

4.5 Case Studies

At the end of this section, we present some representative trajectories or steps, where we can intuitively learn the advantages of *TRAD*. We show two cases produced by *Synapse* and our *TRAD* agent on the cross-domain subset of Mind2Web in Fig. 5, to demonstrate: 1) the difference between task meta-data retrieval and *thought retrieval*; 2) the reason for retrieval rather than direct prediction with thought and the tolerance for imperfect thoughts.

In Fig. 5a, the trajectory-wise retrieval of *Synapse* is obviously problematic, which only considers **"search"** in task instructions and the retrieved trajectories are completely irrelevant to the current one. However, when we use these irrelevant demonstrations for thought production and conduct *thought retrieval* afterwards, the retrieved demonstrations become much more relevant as they all relate to **baby (toddler)** and reflect the process of interacting with **navigation** links or buttons to unfold invisible web pages during web browsing. With the demonstrations from *thought retrieval*, *TRAD* is capable of making the correct decision.

In Fig. 5b, both *Synapse* and *TRAD* seem to retrieve relevant examples trying to find something in **New York**, but if we examine the trajectories retrieved by task meta-data, 2/3 of them fulfill the condition "New York" by clicking some link or button rather than **typing** in a text box. Unfortunately, the correct action under the current state is typing, not clicking, and thus *Synapse* fails to type the correct content. On the contrary, *TRAD* learns to type the correct content "New York" into the text box, even if its thought is incorrect. This also validates our hypothesis that using thought for retrieval instead of prediction helps to correct imperfect thoughts.

5 REAL-WORLD DEPLOYMENT OF TRAD

Since Dec. 2023, we have deployed our *TRAD* agent to automate some real-world office tasks in a mainstream insurance company, which owns a global business with approximately 170 million customers worldwide. We select 4 different websites and collect 100 expert trajectories for some representative tasks on each website as our memory. For evaluation, we collect 20 unseen tasks on each website, using step success rate (Step SR) and trajectory success rate (SR) as evaluation metrics. Tasks involve filling in insurance inquiry forms, implementing advanced information retrieval, etc. Since the websites are complex and contain thousands of web elements, prompting with complete trajectories is not available, hence we only consider single-step prompting with historical actions as auxiliary information.

To verify the effectiveness of *TRAD*, we use two different *ReAct* agents that the company has attempted as our baseline:

- *ReAct*-RD: randomly selects expert steps **in random trajectories** as demonstrations.
- *ReAct*-RV: randomly selects expert steps **in relevant trajectories** retrieved by task instruction as demonstrations.

To be specific, the difference between *TRAD* and *ReAct*-RV is using thought for a second-time step retrieval and the aligned decision module. To further investigate the effect of *thought retrieval* and *aligned decision*, we also deploy a TR agent which removes our *aligned decision* method, namely the *TRAD* w/o TE baseline in Tab. 3. We list the results in Tab. 4.

Table 4: Evaluation results on real-world websites from a mainstream global business insurance company.

Method		ReAct-RD	ReAct-RV	TR	TRAD (Ours)
Website 1 (form filling)	**Step SR**	0.843	0.826	0.941	**0.950**
	SR	0.500	0.450	**0.800**	**0.800**
Website 2 (advanced IR)	**Step SR**	0.941	0.937	0.958	**0.974**
	SR	**0.900**	0.850	0.850	**0.900**
Website 3 (advanced IR)	**Step SR**	0.962	0.987	**1.000**	**1.000**
	SR	0.850	0.800	0.850	**1.000**
Website 4 (form filling)	**Step SR**	0.820	0.860	0.845	**1.000**
	SR	0.350	0.350	0.400	**1.000**
Average	**Step SR**	0.891	0.902	0.936	**0.981**
	SR	0.650	0.613	0.725	**0.925**

As can be seen in Tab. 4, *TRAD* achieves the best performance on all 4 websites, showing its advantage can remain when deployed to real-world scenarios. Moreover, we observe that *TRAD* w/o TE baseline also outperforms both *ReAct* agents, but exhibits noticeable disadvantages compared to the complete *TRAD* agents. This justifies our design of both *thought retrieval* and *aligned decision*.

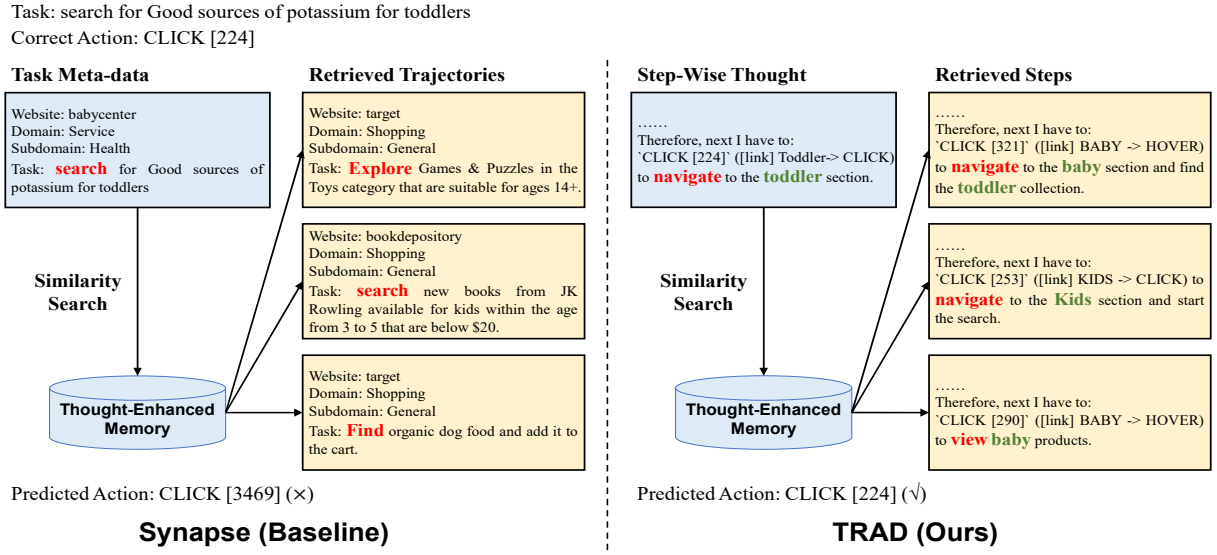
Inference efficiency of *TRAD*. At inference time, our *TRAD* agent only introduces little extra time consumption in *thought retrieval* compared to *ReAct*. We profile the inference process of *TRAD* and *ReAct* on all websites and tasks, and in average *TRAD* takes only 11.7% more time than *ReAct*-RD, which indicates that our method achieves improvement without much sacrifice on efficiency.

6 DISCUSSIONS

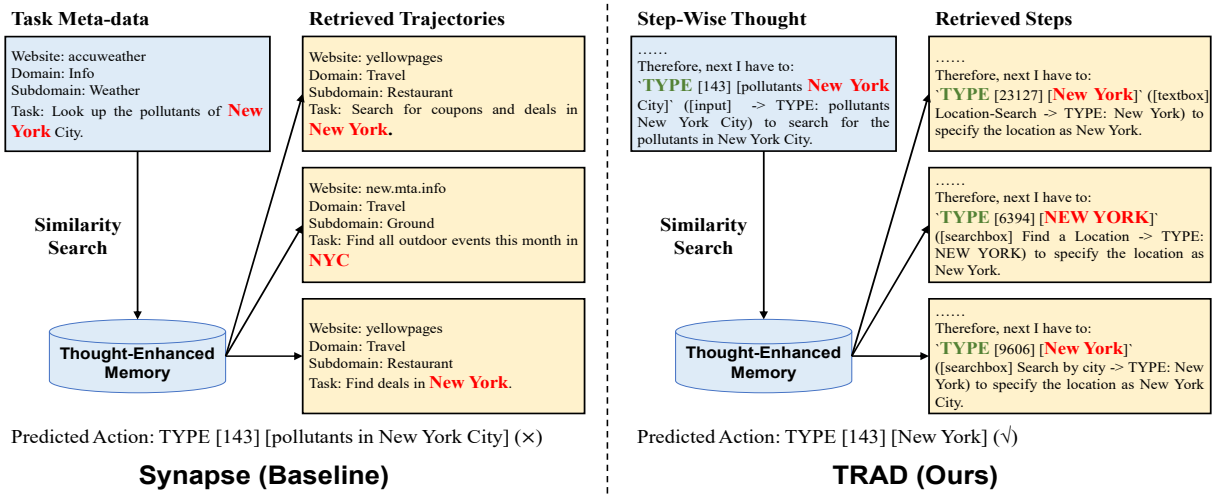
Although *TRAD* exhibits excellent performances over a diverse set of tasks, it still has limitations like dependence on high-quality thought and trade-off between information and noise in *temporal expansion*, and we briefly discuss about them here.

Dependence on high-quality thought. *TRAD* alleviates the issue of imperfect thoughts by its *aligned decision* module, but its capability still depends heavily on the quality of thoughts. To make *TRAD* work well, the abstraction of current state is critical since it serves as the query and key for retrieval, hence the LLM used in *TRAD* should at least have a decent understanding of the task.

Trade-off in temporal expansion. *TRAD* expects to keep relevant information but reduce irrelevant input context by step-wise *thought retrieval*, while preserving some chance for correcting imperfect thoughts by *temporal expansion*. Here exists a trade-off: a longer *temporal expansion* brings not only more tolerance to imperfect thoughts, but also more irrelevant noise in demonstrations. This trade-off requires careful consideration for different tasks.



(a) Representative Case 1



(b) Representative Case 2

Figure 5: Comparison between Synapse trajectory-wise retrieval with task meta-data and TRAD step-wise retrieval with thought. (a) The trajectory-wise retrieval of *Synapse* only considers "search" in task instructions and the retrieved trajectories are completely irrelevant. However, by generating thoughts with these irrelevant trajectories, *thought retrieval* finds more relevant step-wise demonstrations related to baby (toddler) and navigation. (b) The trajectory-wise retrieval of *Synapse* retrieves plausible examples which do not type in a text box with task meta-data. Although thoughts are imperfect, *thought retrieval* finds more relevant demonstrations and *TRAD* learns to input "New York".

7 CONCLUSIONS

In this work, we propose a novel LLM agent augmented by step-wise demonstration retrieval (*TRAD*) for sequential decision-making tasks. *TRAD* first retrieves relevant step demonstrations by its thought about current state, and then complements temporally correlated steps for more informative action prediction. Extensive experiments are conducted on two different sequential decision-making tasks to validate the effectiveness of our solution, and thorough ablation studies justify the design choice and stability of our method. We further present the results from real-world deployment of our method, showing its superior performance and decent efficiency in real-world applications. In the future, we plan to further improve our work by enhancing the thought retrieval process with more powerful reasoning or planning methods and learned dense representations of states.

ACKNOWLEDGMENTS

The SJTU team is partially supported by National Key R&D Program of China (2022ZD0114804), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and National Natural Science Foundation of China (62322603, 62076161).

REFERENCES

[1] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. 1998. Pddl| the planning domain definition language. *Technical Report* (1998).

[2] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *arXiv preprint arXiv:2308.09687* (2023).

[3] 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a Generalist Agent for the Web. In *Proceedings of the 37th Advances in Neural Information Processing Systems (NeurIPS)*.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. Everything of thoughts: Defying the law of penrose triangle for thought generation. *arXiv preprint arXiv:2311.04254* (2023).

[7] Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2024. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis. In *Proceedings of The 12th International Conference on Learning Representations (ICLR)*.

[8] Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2023. Understanding HTML with Large Language Models. In *Findings of the Association for Computational Linguistics (EMNLP)*. 2803–2821.

[9] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8154–8173.

[10] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

[11] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[12] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language Models can Solve Computer Tasks. In *Proceedings of the 37th Advances in Neural Information Processing Systems (NeurIPS)*.

[13] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as Policies: Language Model Programs for Embodied Control. In *Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*. 9493–9500.

[14] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477* (2023).

[15] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804* (2021).

[16] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[17] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[18] 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. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS)*. 27730–27744.

[19] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. 1–22.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* (2019).

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3980–3990.

[22] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).

[23] 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 (NAACL-HLT)*. 2655–2671.

[24] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Proceedings of the 37th Advances in Neural Information Processing Systems (NeurIPS)*.

[25] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of Bits: An Open-Domain Platform for Web-Based Agents. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Vol. 70. 3135–3144.

[26] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 37th Advances in Neural Information Processing Systems (NeurIPS)*.

[27] 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10737–10746.

[28] 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 *Proceedings of 9th International Conference on Learning Representations (ICLR)*.

[29] The LongChat Team. 2023. *How Long Can Open-Source LLMs Truly Promise on Context Length?* https://lmsys.org/blog/2023-06-29-longchat/

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[31] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 10014–10037.

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

[33] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023).

[34] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The 11th International Conference on Learning Representations, (ICLR)*.

[35] Zihao Wang, Shaofei Cai, 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. In *Proceedings of the 37th Advances in Neural Information Processing Systems (NeurIPS)*.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS)*.

[37] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 1423–1436.

[38] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Proceedings of 36th Conference on Neural Information Processing Systems (NeurIPS)*.

[39] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Proceedings of 37th Conference on Neural Information Processing Systems (NeurIPS)*.

[40] 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 *Proceedings of The 11th International Conference on Learning Representations (ICLR)*.

[41] Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9134–9148.

[42] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Step-Back Prompting Enables Reasoning Via Abstraction in Large Language Models. In *Proceedings of The 12th International Conference on Learning Representations (ICLR)*.

[43] Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2024. Synapse: Trajectory-as-Exemplar Prompting with Memory for Computer Control. In *Proceedings of 12th International Conference on Learning Representations (ICLR)*.

[44] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The 11th International Conference on Learning Representations (ICLR)*.

[45] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).