TOWARDS UNIFIED ALIGNMENT BETWEEN AGENTS, HUMANS, AND ENVIRONMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid progress of foundation models has led to the prosperity of autonomous agents, which leverage the universal capabilities of foundation models to conduct reasoning, decision-making, and environmental interaction. However, the efficacy of agents remains limited when operating in intricate, realistic environments. In this work, we introduce the principles of **U**nified **A**lignment for **A**gents (**UA**^{2}), which advocate for the simultaneous alignment of agents with human intentions, environmental dynamics, and self-constraints such as the limitation of monetary budgets. From the perspective of **UA**^{2}, we review the current agent research and highlight the neglected factors in existing agent benchmarks and method candidates. We also conduct proof-of-concept studies by introducing realistic features to WebShop (Yao et al., 2022a), including user profiles to demonstrate intentions, personalized reranking for complex environmental dynamics, and runtime cost statistics to reflect self-constraints. We then follow the principles of **UA**^{2} to propose an initial design of our agent, and benchmark its performance with several candidate baselines in the retrofitted WebShop. The extensive experimental results further prove the importance of the principles of **UA**^{2}. Our research sheds light on the next steps of agent research with improved problem-solving abilities.

1 INTRODUCTION

Recent days have witnessed the rapid development of autonomous agents, which leverage the proficiency of Large Language Models (LLMs) or Large Multimodal Models (LMMs) (OpenAI, 2023; Touvron et al., 2023; Team et al., 2023; Jiang et al., 2024) to interact with environments for task execution. Several seminal works on foundation model agents have exhibited promising results in both digital and embodied scenarios, including but not limited to web task automation (Deng et al., 2023; Zhou et al., 2023b; Zheng et al., 2024), open-ended world exploration (Wang et al., 2023a; Zhu et al., 2023), interactive coding (Chen et al., 2023c; Qian et al., 2023; Xu et al., 2023), and robotic tasks (Ahn et al., 2022; Mirchandani et al., 2023; Huang et al., 2023b; Ma et al., 2023; Wang et al., 2023b).

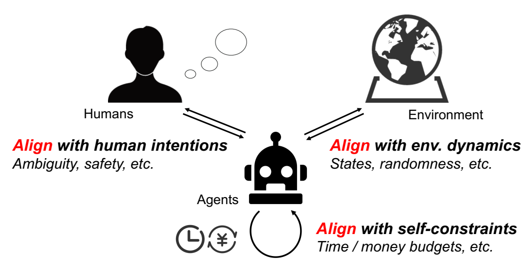

Figure 1: A working system of agents that consists of three roles: humans to be assisted, an environment to interact with, and the agents themselves. The principles of **U**nified **A**lignment for **A**gents (**UA**^{2}) suggest that the agents should align with the three roles in a unified manner by recognizing *human intentions*, adapting to *environmental dynamics*, and adhering to *self-constraints*.

Aside from existing literature, the development of foundation model agents in realistic, complex scenarios is still in its infancy. While different agent benchmarks have been proposed (Liu et al., 2023b; Mialon et al., 2023; Ma et al., 2024), the methodologies of agents are still being proposed and evaluated in synthetic, simplified settings, which results in the bottlenecked performance of agents in real-world deployment when attempting to satisfy the expectations of humans with realistic demands (Kinniment et al., 2023). This naturally leads to the question: *What are the principles that the agents should follow to improve their real-world capabilities?*

To answer the question, we first take a systematic view of agents during the operation. We recognize the working system of agents as a composition of three roles: *humans* that propose the goals to be assisted, an *environment* that provides feedback for interaction, and foundation model *agents* themselves to act in the environment to assist the human user. In complex scenarios, the intentions of humans can be ambiguous (Tamkin et al., 2022; Li et al., 2023a) or concerned with safety requirements (Ruan et al., 2023; Yuan et al., 2024). Moreover, the underlying dynamics of the environment can be complicated to identify (LeCun, 2022; Hu & Shu, 2023), and affected by temporality (Fan et al., 2022) or stochasticity (Wu et al., 2023). Last but not least, the agents themselves can also be constrained by a certain amount of budgetary limits (*e.g.*, monetary and time expenses) during operations, an aspect often overlooked in the existing agent research. While each of the aspects is noted by different aforementioned works, none of them emphasize the holistic comprehension of all the roles in the working system.

In this work, we propose the principles of **U**nified **A**lignment for **A**gents (**UA**^{2}) by drawing connections with the alignment research in the sense of both LLMs and reinforcement learning literature (Sutton & Barto, 2018; Ouyang et al., 2022; Bai et al., 2022; Ji et al., 2023; Burns et al., 2023). The goal of **UA**^{2} is to enhance the awareness of the foundation model agents to their working system, aligning with *human intentions*, *environmental dynamics*, and *self-constraints* in a unified manner. From the perspective of **UA**^{2}, we review the existing research on agents and point out the neglected factors in the design of existing benchmarks and candidate methodologies of agents.

To further demonstrate the essence of **UA**^{2}, we conduct proof-of-concept studies by constructing an upgraded version of WebShop (Yao et al., 2022a). In the retrofitted WebShop, we add the design of the *human intentions* of shoppers for agents to track and infer, the *environmental dynamics* with personalized re-ranking algorithms that evolve with agent actions, and the *self-constraints* by implementing a counter of monetary and temporal costs. On top of the retrofitted environment, we initiate an agent method guided by the principles of **UA**^{2}, and benchmark its performance as well as several other candidate agent baselines. The results reveal the suboptimality of the agent baselines that violate the principles of **UA**^{2}. The results further support our advocacy that the agents should achieve a unified alignment with humans, the environment, and the agents themselves. Our research sheds light on the future steps of autonomous agents, including synergizing agents with alignment techniques, constructing agent benchmarks and methods that follow the principles of **UA**^{2}, and envisioning self-evolving agents through lifelong interaction and continual alignment.

2 Principles of Unified Alignment for Agents

2.1 Roles in a Working System of Agents

A working system of agents consists of three roles (Figure 1): agents, humans, and the environment.

Agents are the core component of the entire system. Agents are responsible for understanding human intentions and generating appropriate responses or actions to interact with the environment. Proficient agents should provide accurate, informative, and engaging interactions during task execution.

Humans are the main role to be assisted in the system. The tasks assigned by humans can be viewed as the initial inputs to the working system, which reflects the underlying goals and human intentions.

Environment refers to the situation where the agents operate. It encompasses the external factors and conditions that can influence the agents' behavior, performance, and interactions. The feedback from the environment affects the reasoning of the agents, as well as their following actions.

Realistic working systems of agents are composed of diverse, ambiguous human intentions, changing environments with complex dynamics, as well as self-constraints over the agents themselves. This leads to the necessity of agents to operate towards the unified alignment with all the roles.

2.2 Unified Alignment with All the Roles

While three distinct roles exist in a working system of agents, we argue that the agents should align with all the roles in a unified manner. To promote the orchestration of agents, humans, and the environment, the agents should work in the direction of eliminating the gap between agents and

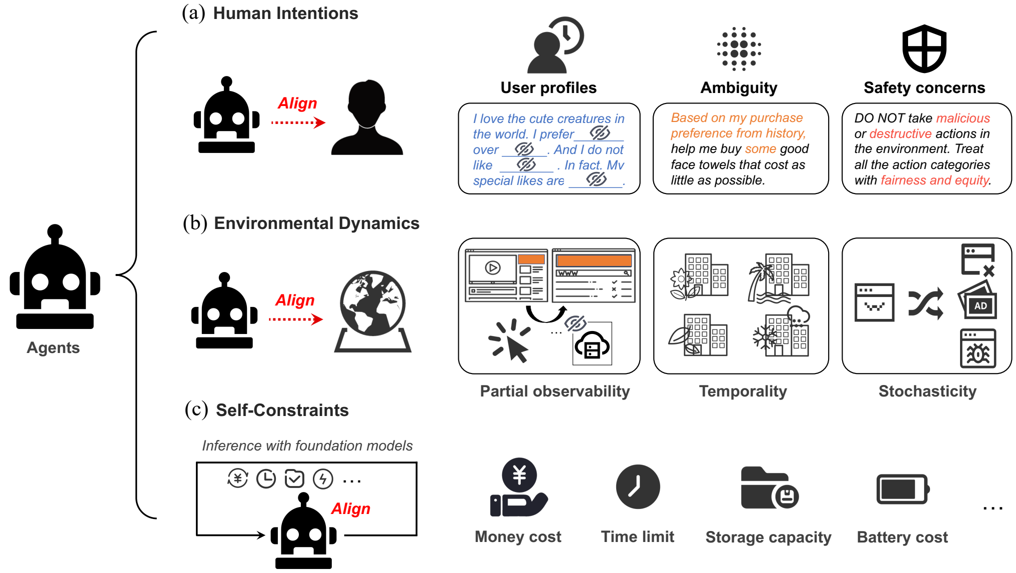

Figure 2: Illustrations of the principles of unified alignment with (a) *human intentions*, (b) *environmental dynamics*, and (c) *self-constraints*. The principles of unified alignment for agents emerge from all the roles in an agent working system: agents, humans, and environment.

humans, agents and the environment, as well as adapting to the constraints imposed on the agents themselves. Based on this, we propose the principles of **U**nified **A**lignment for **A**gents (**UA**^{2}):

1. Alignment with *human intentions*. The agents need to correctly recognize the intentions of the human users. While the goal is usually specified with a textual sentence, the ambiguity of language expression can affect the understanding and decision-making of agents.
2. Alignment with *environmental dynamics*. The agents need to interact with the environment to achieve the goal required by human users. To succeed, the agents should raise their awareness of the operation laws of the environment. This is also advocated in (LeCun, 2022; Hu & Shu, 2023) that proposes to incorporate a world model into an agent system.
3. Alignment with *self-constraints*. The underscored factor of current agent research comes from the constraints imposed on the agents themselves, including time/money budget limits. For foundation model agents, the underlying models (*e.g.*, proprietary LLMs/LMMs) are costly for inference, which hurdles the performance in realistic scenarios.

2.3 Challenges from the Principles of UA^2

Figure 2 illustrates the principles of **UA**^{2}. In this section, we pose the challenges raised from **UA**^{2}.

Challenges in the alignment with *human intentions*. When the interaction between humans and the agent is a single-turn process, it is equivalent to LLM alignment (Ouyang et al., 2022) in the form of a prompt-response pair. However, in realistic settings, human intentions are often not perfectly covered in a single prompt, but rather reflected by preferences not directly visible from instructions (*e.g.*, personal preferences and safety concerns). Challenges arise for the agents to infer authentic human intentions with multiple turns of interactions by either eliciting human preferences (Li et al., 2023a), or learning to self-correct from environmental feedback (Huang et al., 2023a), or both.

Challenges in the alignment with *environmental dynamics*. The interactive environments for agents in realistic scenarios can be highly complicated, which requires the agent to recognize the hidden state from the history of observations. Considering the dynamics function $\mathbf{s}_{n+1} \sim \pi(\mathbf{s}_n, \mathbf{a}_n)$ where $\mathbf{s}_n$ and $\mathbf{a}_n$ stand for the n-th step state and action, respectively, its complexity includes:

- Partial observability. This is reflected by the complexity of the function that transforms the historical observations $\{\mathbf{o}_{\leq n}\}$ into the authentic state of the current step $\mathbf{s}_n$.

Table 1: Existing agent benchmarks from the perspective of alignment with *human intentions*, *environmental dynamics*, and *self-constraints*. "Temp." stands for temporality, and "Stoch." stands for stochasticity. "#Actions" means that the step count in the environment will be reported as a metric. † WebShop is fully observable as long as the URL is covered in each observation.

Type	Benchmarks	Human Intentions	Environmental Dynamics	Self-Constraints
Digital	Androidenv (Toyama et al., 2021)	None	Partial Obs.	None
	WebShop (Yao et al., 2022a)	None	Full Obs.†	None
	Mind2Web (Deng et al., 2023)	None	Partial Obs.	None
	ToolBench (Qin et al., 2023)	None	Full Obs. & Temp. & Stoch.	None
	WebArena (Zhou et al., 2023b)	Fixed and Given	Partial Obs.	None
Embodied	VirtualHome (Puig et al., 2018)	None	Partial Obs.	None
	BabyAI (Chevalier-Boisvert et al., 2019)	None	Partial Obs.	None
	ALFWorld (Shridhar et al., 2020)	None	Partial Obs.	None
	MineDojo (Fan et al., 2022)	None	Partial Obs. & Stoch.	None
	ScienceWorld (Wang et al., 2022a)	None	Partial Obs.	None
	Interactive Gibson (Xia et al., 2020)	None	Partial Obs.	#Actions
	AGENT (Shu et al., 2021)	None	Partial Obs.	#Actions
	RFUniverse (Fu et al., 2022)	Fixed and Given	Partial Obs.	#Actions
	BEHAVIOR-1K (Li et al., 2023b)	None	Full Obs.	#Actions
	HAZARD (Zhou et al., 2024)	None	Partial Obs. & Temp.	#Actions
Mixed	SmartPlay (Wu et al., 2023)	None	Partial Obs. & Stoch.	None
	AgentBench (Liu et al., 2023b)	None	Partial Obs.	None
	AgentBoard (Ma et al., 2024)	None	Partial Obs. & Temp. & Stoch.	None

- Time-variant property. This is reflected by the temporal effect in the dynamics function, where the evolution of time t leads to the variation of $\mathbf{s}_{n+1} \sim \pi(\mathbf{s}_n, \mathbf{a}_n, t)$.
- Stochasticity. The $\pi(\mathbf{s}_n, \mathbf{a}_n)$ can be interlaced with (nearly) independent random events.

In this way, constructing a precise world model for an agent system requires delicate techniques beyond ad-hoc exploration, coarse-grained memory, or ungrounded planning.

Challenges in the alignment with *self-constraints*. The self-constraints of agents are the often-overlooked desiderata in the design of existing agent methodologies. Taking the budgetary limits into account, the agent system should re-use the accumulated experiences during the lifelong learning process (Majumder et al., 2023), and balance the resources invested in the different learning modules. Furthermore, in scenarios where the self-constraints change with different episodes, additional challenges emerge for the agents to adapt to the constraints autonomously.

3 Literature Review from the Lens of UA^2

3.1 Benchmarks

In this section, we begin with a comprehensive review of current benchmarks in agent research, from the perspective of **UA**^{2}. Representative benchmarks in both digital (Toyama et al., 2021; Yao et al., 2022a) and embodied (Puig et al., 2018; Chevalier-Boisvert et al., 2019) environments are summarized in Table 1. By rendering realistic simulations (Puig et al., 2023; Szot et al., 2021) and carefully configured tasks (Li et al., 2023b), current benchmarks offer diverse environments for both language-based and embodied agents (Xi et al., 2023) to operate and interact within (Maes, 1995). Instead of focusing on environmental authenticity (Fu et al., 2022) or general task complexity, we assess the benchmarks prioritizing the alignment principles of **UA**^{2}. In practice, we consider the following three aspects:

1. *Human intentions*: Whether the authentic goals need to be inferred during task execution, or the intentions of humans are precisely conveyed in the descriptions.
2. *Environmental dynamics*: Whether the state transitions of the environment are intrinsically endowed with partial observability, temporality, or stochasticity.
3. *Self-constraints*: Whether the status of budgetary resources is reflected, including time consumption, the maximum number of actions or reasoning steps, etc.

In terms of human intentions, most benchmarks (Qin et al., 2023; Liu et al., 2023b) provide explicit task instructions for more effective evaluation, rather than considering human intentions as hidden

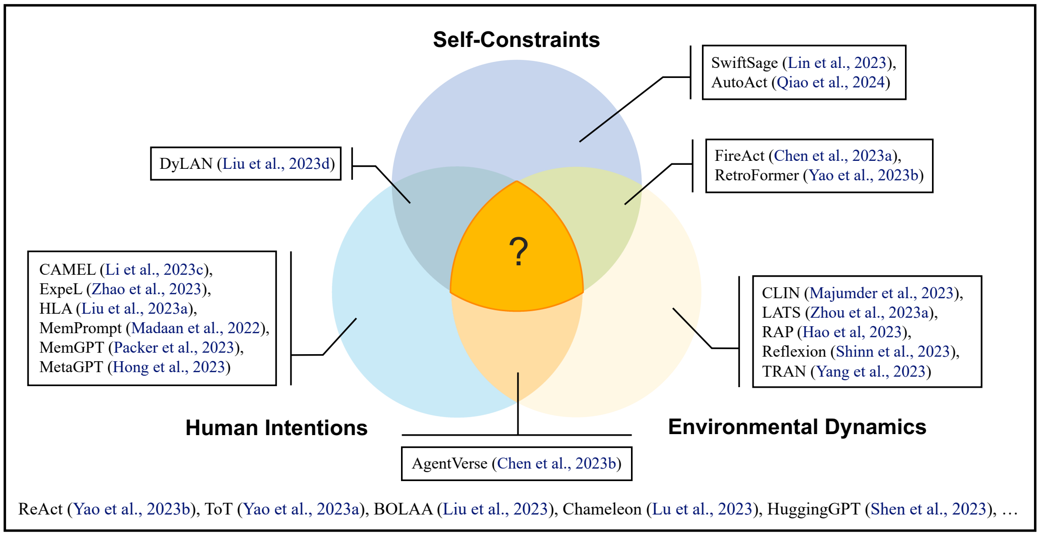

Figure 3: The dissection of alignment endeavors for different representative agent techniques. Generally speaking, the methods that actively coordinate with humans excel at aligning with *human intentions*. The methods that are grounded with external feedback from the environment align well *environmental dynamics*. The methods that adopt adaptive strategies or fine-tuning demonstrate better alignment with *self-constraints*. While the advanced techniques mostly align with one role or two in the working system of agents, much room lies in the quest for $\mathbf{UA}^2$.

attributes for agents to discover. By incorporating human interactions, several embodied simulators (Puig et al., 2023; Xia et al., 2019) facilitate tasks with vague goal descriptions (Paul et al., 2022; Liu et al., 2023c), necessitating agents to engage with humans to gather sufficient information for task completion. In contrast, digital benchmarks hardly account for this aspect. The most relevant digital environment in this aspect is WebArena (Zhou et al., 2023b), which deliberately defines consistent human intentions across episodes. However, the intentions are also explicitly stated in the instructions, which bypasses the intention elicitation process of agents with humans.

The benchmarks for agents are designed to mirror the complexities of the real-world dynamics (Puig et al., 2023). Most benchmarks assume the environment is partially observable where agents are required to accomplish tasks through exploration and interaction (Xia et al., 2020). Some benchmarks also include stochastic factors (Wu et al., 2023; Zhou et al., 2024) or evolve with time (Qin et al., 2023). Nevertheless, the synthesis of fine-grained realistic dynamics remains underdeveloped in benchmark design, resulting in the lack of evaluations of agent methodologies therein.

As for *self-constraints*, embodied benchmarks (Xia et al., 2020; Li et al., 2023b) use the number of actions as a metric to reflect the operational cost in real-world deployments, such as the path length in navigation tasks (Anderson et al., 2018). In this context, AGENT (Shu et al., 2021) further explicitly evaluates the trade-offs between cost and reward. However, existing digital benchmarks overlook cost and time constraints in the assessments, which should be equally important.

In essence, existing agent benchmarks are still inadequate from the lens of $\mathbf{UA}^2$. In general, the development of digital benchmarks lags behind that of embodied benchmarks. This underscores the need for more realistic environments to enhance the development of agent techniques.

3.2 Methods

In this section, we review the representative agent methods. For each method, we investigate whether it actively seeks alignment with *human intentions*, *environmental dynamics*, or *self-constraints*.

To align with *human intentions*, the agent methods should coordinate with humans through reasoning or experience summarization. HLA (Liu et al., 2023a) and MemPrompt (Madaan et al., 2022) interact with humans for multiple rounds to solicit authentic human intentions. Multi-agent frameworks like CAMEL (Li et al., 2023c), AgentVerse (Chen et al., 2023b), and DyLAN (Liu et al., 2023d) leverage a group of agents for role-playing and inter-discussion to improve the understanding of human instructions. ExpeL (Zhao et al., 2023) and MemGPT (Packer et al., 2023) also align with human intentions through the analysis of human goals in an iterative manner.

To align with *environmental dynamics*, the agents should ground themselves with external information from the environment. Reflexion (Shinn et al., 2023), LATS (Zhou et al., 2023a), and AgentVerse (Chen et al., 2023b) use external reward feedback as conditions to rectify their actions and improve the alignment with the environment. RetroFormer (Yao et al., 2023b), TRAN (Yang et al., 2023), CLIN (Majumder et al., 2023), and FireAct (Chen et al., 2023a) integrate the (abstracted) trajectories accumulated through the interaction with the environment into the prompts or data for fine-tuning. This results in an in-context or parametrized world model, which narrows the gap of alignment with the environment. RAP (Hao et al., 2023) can also be categorized as aligning with the environment through simulation of the underlying foundation models.

To align with *self-constraints*, the agents should adopt an adaptive strategy in the process of task execution and/or group construction. The representative works in this vein include SwiftSage (Lin et al., 2023), Retroformer (Yao et al., 2023b), and DyLAN (Liu et al., 2023d). Finetuning a small-sized foundation model is also beneficial to the obedience of self-constraints (Chen et al., 2023a; Qiao et al., 2024), which eliminates the need to call the costly APIs of proprietary models.

In addition to the aforementioned frameworks, there are also basic techniques for agents, such as ReAct (Yao et al., 2022b) and Tree-of-Thoughts (Yao et al., 2023a), that serve as the foundational elements in most of the advanced agents. An overview of the analysis is illustrated in Figure 3.

Despite the emergence of diverse agent methodologies, plenty of room still exists for the unified alignment of agents with *human intentions*, *environmental dynamics*, and *self-constraints* simultaneously. Challenges lie in the construction of the agent framework (Sumers et al., 2023), which requires an elaborate design to strike a good balance of alignment with all three roles. Counterexamples in this sense are Reflexion (Shinn et al., 2023) and LATS (Zhou et al., 2023a), which leverage multiple rounds of sampling to achieve better alignment with the environment, but the self-constraints are significantly violated at the same time due to the high cost. Moreover, the capability of the underlying foundation model dominates the potential of the sophisticated alignment endeavors of an agent. Therefore, it is essential to promote the synergy between the development of foundation models (such as alignment techniques) and the research of agents.

4 Proof-of-Concept Studies

In this section, we conduct proof-of-concept studies to validate the importance of **UA**^{2} in the design of both benchmarks and methods for agents. Section 4.1 covers several realistic features we introduced into WebShop (Yao et al., 2022a), which are selected according to the principles of **UA**^{2}. In Section 4.2, we introduce our agent method design following the principles of **UA**^{2}. Section 4.3 covers the experiments of several agent candidate baselines and our method in the retrofitted environment, and Section 4.4 reports the results as well as our discussions and findings.

4.1 Environment Construction

We conduct the case studies by first upgrading the WebShop environment. WebShop is a simulated online shopping environment with 1.18M real-world shopping items gathered from Amazon, and 12,087 textual shopping instructions collected from human annotators. While serving as a high-quality testbed for the instruction-following and planning abilities of foundation model agents, we further improve the complexity of WebShop by introducing the realistic factors around the three roles in the agent working system: *human intentions*, *environmental dynamics*, and *self-constraints*.

Human intentions. In reality, different human users own unique, potentially invisible preferences about the properties and categories of shopping items. Given this, we configure 10 different users for testing, each possessing a basic preference (in text) that corresponds with a certain hidden attribute of items. We equip each user with a group of 50 consecutive artificially constructed instructions with user profiles, ambiguous descriptions, and preferences to be inferred by tracking the purchase history. The rules of reward computation for each instruction follow those of the original WebShop. For details of the task construction and the instructions with user profiles, see Appendix A.1.

Environmental dynamics. To narrow the gap with realistic online shopping scenarios, we implement fine-grained personalized reranking algorithms on top of the original search engine in WebShop. The algorithms include collaborative filtering (Sarwar et al., 2001) and a Determinantal Point Process (DPP) based method (Chen et al., 2018). With personalized reranking schemes, the website is constantly evolving with user actions, which better reflects the complexity of realistic environmental dynamics. The details of the implementation are listed in Appendix A.2.

Self-constraints. To measure the expenses of the agents themselves during the operating process, we implement the runtime environment to count for the temporal and monetary expenditures for the agent working system. The monetary cost consists of the API calls of the proprietary foundation models, and the time consumption indicates the normalized endurance of interaction between the agents and the interactive environment (detailed in Appendix A.3).

4.2 Agent Design with the Principles of UA^2

Following the principles of **UA**^{2}, we initiate our agent by introducing the structured memory module on top of ReAct (Yao et al., 2022b). Shown in Figure 4, the introduced module is formed by two components: *low-level* action insights and *high-level* intra-task experience.

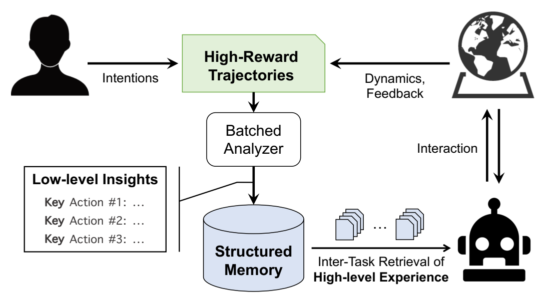

Figure 4: Overview of our agent design that follows the principles of **UA**^{2}. By continually analyzing and retrieving structured memory from similar tasks of the same user, the agent extrapolates experience across different tasks.

Low-level action insights are a list of key actions exploited from different runs in the environment under the same task instruction. The key actions are extracted from the high-reward trajectories with an analyzer, with which the contributions of actions are computed in task-solving. The analyzer adopts a batched inference (Cheng et al., 2023) to tag all actions at the same time. The structured memory is then maintained with the key actions paired with their corresponding human instructions.

High-level intra-task experience is formed by the retrieval of the *low-level* action insights accumulated in the structured memory. According to the similarity of the current human instruction with the previous ones stored in the memory, the key actions are gathered to form an initial plan for the current task. The re-use of high-level experience throughout the stream of tasks promotes efficient intra-task generalization.

We design the agent to differ from previous works, which rely on LLM summarization of unstructured insights (Majumder et al., 2023; Zhao et al., 2023) or multiple-round LLM reflections within a single task (Shinn et al., 2023). Our method aligns with *human intentions*, *environmental dynamics*, as well as *self-constraints*: (i) The maintenance of the structured memory contributes to the lifelong profiling of a human user. (ii) The storage and retrieval of key actions analyzed from different trajectories improves the awareness of the agent to the environment. (iii) The reuse of structured records saves the agents from planning from scratch for each task, which aligns with *self-constraints* by cost minimization. Appendix B covers the formal descriptions and implementation details.

4.3 Experiments

Baselines. We compare the performance of our method with several widely-used agent techniques on the retrofitted WebShop in Section 4.1, including (1) ReAct (Yao et al., 2022b), which harmonizes internal reasoning and external actions, (2) ReAct-SC (ReAct with Self-Consistency), which equips ReAct with sampling and marginalization (Wang et al., 2022b), (3) Reflexion (Shinn et al., 2023), which conducts self-correction by reflecting on past actions and observations, and (4) LATS (Zhou et al., 2023a), which leverages a combination of techniques including ReAct, self-reflection, and Monte Carlo Tree Search (MCTS). Note that we leave the implementation of techniques categorized

Table 2: The performance of averaged reward, success rate (SR) (%), alignment gap (%) with human intentions ($\mathbf{G}_{\text{HI}}$) and environment dynamics ($\mathbf{G}_{\text{ED}}$), time (s) and money ($) cost of all methods tested in our retrofitted WebShop environment. The best result for each metric is in **bold**. The better performance under each metric is indicated by the darker green shades. *LATS is tested on 1/10 subset of the entire task instructions due to the significant cost.

Method	Reward ↑	SR (%) ↑	$\mathbf{G}_{\text{HI}}$ (%) ↓	$\mathbf{G}_{\text{ED}}$ (%) ↓	Time (s) ↓	Money ($) ↓
ReAct	50.3	8.0	11.7	14.9	**1.716**	**0.013**
ReAct-SC	49.9	7.4	14.4	14.6	1.720	0.039
Reflexion	44.4	**13.8**	22.5	25.7	5.539	0.045
LATS*	**52.4**	10.0	18.5	**14.3**	125.935	5.508
Ours	51.9	9.6	**6.7**	14.8	1.779	0.014

as aligning with human intentions in Section 3.2 as future work, since great effort should be taken by involving humans in the interaction loop and adapting to our settings.

Evaluation Metrics. Following the settings of Yao et al. (2022a), we measure the performance of task completion with the average reward and success rate incurred per task. To quantitatively investigate the alignment of different methods under the principles of $\mathbf{UA}^2$, we introduce three extra metrics. We report the averaged monetary and time cost to reflect the alignment of each method with *self-constraints*. For *human intentions* and *environmental dynamics*, we build ablated versions of the retrofitted WebShop that exclude the introduced feature, respectively. We then test each agent technique on the pair of fully-retrofitted / ablated environments, and finally investigate the difference between the pair of the evaluated rewards. More specifically:

To evaluate the alignment with *human intentions*, we construct an ablated version of the environment in Section 4.1, where the hidden attributes corresponding with user profiles or preferences are excluded from the reward computation. In this ablated environment, the performance of each method should be better than that in the fully-retrofitted environment. We define the alignment gap with human intentions $\mathbf{G}_{\text{HI}}$ as the relative difference between the two performances:

$$\mathbf{G}_{\text{HI}} = (R_{\text{full}} - R_{\text{HI}})/R_{\text{full}} \times 100\%, \tag{1}$$

where R_{full} and R_{HI} stand for the reward of an agent in the fully-retrofitted environment and the environment excluding the computation of human intentions, respectively.

Similarly, to evaluate the alignment with *environmental dynamics*, we build an ablated environment without the implementation of the personalized reranking algorithms, and define the alignment gap with environmental dynamics $\mathbf{G}_{\text{ED}}$:

$$\mathbf{G}_{\text{ED}} = (R_{\text{full}} - R_{\text{ED}})/R_{\text{full}} \times 100\%, \tag{2}$$

with R_{ED} as the reward of an agent in the ablated environment that excludes the personalized reranking algorithms.

4.4 Results and Discussions

The performances of different methods in all the metrics are shown in Table 2. According to the results, Our framework achieves the top unified performance among all the methods, with the best balance between task completion performance and measures of different alignment sources.

LATS achieves the highest average reward, and Reflexion obtains the top success rate. This is because they both employ trial-and-error approaches with multiple rounds of interactions. However, the money and time costs of the two methods are significantly higher than other methods, suggesting their weaknesses in aligning with the *self-constraints* of agents. To be specific, Reflexion incurs a cost over 5× in time and 3× in money compared to other methods, while LATS, in contrast with other methods, entails a cost exceeding 100× in time and nearly 200× in money.

ReAct-SC achieves a comparable average reward and success rate (SR) with ReAct. This might be attributed to the complexity of our retrofitted environment, where even more runs of sampling are required in ReAct-SC to vote for better actions. In addition, The incorporation of self-consistency in ReAct-SC requires more calls of the API of the proprietary foundation model, resulting in approximately three times the cost of money compared to ReAct. The time cost of ReAct and ReAct-SC is nearly identical. This is because we only document the endurance within the interactive environment (the time of API requests is neglected), and at the same time, ReAct might exhibit similar planning abilities as ReAct-SC. Finally, Our framework achieves the top performance in averaged rewards and success rates, which underscores the significance of the principles of **UA**^{2}.

As for the alignment gap, the results of $\mathbf{G}_{\mathrm{HI}}$ and $\mathbf{G}_{\mathrm{ED}}$ in Table 2 indicate that almost all baselines possess the gap above 10% in terms of aligning with humans or the complex environment. Notably, our method demonstrates a significantly lower $\mathbf{G}_{\mathrm{HI}}$ than other methods, which might benefit from its capacity to adapt to diverse human intentions by intra-task experience generalization with the structured memory. In contrast, LATS demonstrates a relatively low $\mathbf{G}_{\mathrm{ED}}$ of 14.3%. This is because of the accumulation of trials from the exhaustive sampling in the environment, which meanwhile limits its practical applicability. For comparison, neither $\mathbf{G}_{\mathrm{HI}}$ nor $\mathbf{G}_{\mathrm{ED}}$ of Reflexion is satisfactory, which might indicate that the mechanism of the self-reflection is inferior to other techniques in this setting. These results highlight the need for agent techniques following the principles of **UA**^{2}.

5 ACTIONABLE INSIGHTS

Envisioning the future of autonomous agents powered by foundation models in real-world applications, in this section, we provide insights on the next steps of research from **UA**^{2}.

Synergizing agents with alignment research. Alignment research aims to steer a model to follow instructions faithfully. To achieve unified alignment in an agent system, techniques in the field of alignment research can be helpful to the foundation model agents in following the principles of **UA**^{2}. For instance, humans can leverage ideas like Constitutional AI (Bai et al., 2022) to integrate the principles of unified alignment into the objectives of the agents.

Constructing realistic agent benchmarks. While appreciating the existing efforts on the benchmark construction for agents, we advocate for more realistic simulation and sandbox design reflecting the intricate scenarios with nuanced logistics and details. As shown in our proof-of-concept studies, the principles of **UA**^{2} are also helpful in the design of the benchmark. Taking **UA**^{2} into account, the gap between agents and realistic human demands and interactive environment can be revealed more faithfully, laying the foundation for the next breakthrough of agent techniques.

Developing holistic evaluations for agents. Existing research on agents mainly uses the final success of task completion as the evaluation metric. In our work, we propose the principles of unified alignment for agents, suggesting the proficiency of agents can be reflected by the quality of alignment with *human intentions*, *environmental dynamics*, and *self-constraints*. Given this, the dissection of the performance of agents is necessary for the development of agent techniques, since an analysis of alignment gaps with different roles indicates the direction of improvement for agents. This suggests the importance of holistic evaluations for the development of autonomous agents.

Toward self-evolving agents through continual alignment. While the sources of alignment have been categorized by **UA**^{2}, it requires the elaborated design of agent methods that carefully balance the different alignment sources in a unified manner. Envisioning agents with next-level autonomy, we expect the agents to self-evolve through lifelong interaction with humans and the environment with continual alignment. In this vein, agents improve themselves with better use and efficiency, leading to general problem-solving abilities in complex, real-world scenarios.

6 CONCLUSION

In this work, we propose the principles of unified alignment for agents with *human intentions*, *environmental dynamics*, and *self-constraints*. We start by recognizing the three components in a working system of agents: agents, humans, and environment, then state the necessity of agents to align with the three roles in a unified manner and propose the principles of **UA**^{2}. We demonstrate the significance of **UA**^{2} by literature review and proof-of-concept studies. Eventually, we shed light on the impact of **UA**^{2} on the future of agent research with enhanced general problem-solving abilities.

BROADER IMPACT

The prosperity of autonomous agents with foundation models has posed exciting avenues for future research toward the automatic execution of daily tasks for humans. In our work, we advocate for the unified alignment of agents ($\mathbf{UA}^2$) with humans, the environment, and the agents themselves simultaneously. To align with humans means to improve the understanding of *human intentions*, and especially safety concerns, to provide better assistance. By doing so, the agents also need to align with the environment to enhance the awareness of *environmental dynamics*, so that the agents can be cautious about whether the next actions could be malicious or destructive. The agents should also align with themselves in terms of *self-constraints*, adhering to the running cost of money, time, battery, etc. In our work, we conduct proof-of-concept studies by introducing realistic features, such as human profiles, personalized reranking algorithms, and runtime cost counters into the original WebShop. While the results have proved the essence of $\mathbf{UA}^2$, we plan to experiment with extra alignment factors in the future, including safety concerns from human intentions, temporal variation and random events from the environment, as well as other types of self-constraints.

Our work covers the principles for agents to follow, and we expect the future of agents with narrowed alignment gaps in a unified manner. We also expect the construction of more realistic sandboxes or simulators as the testbeds for agents, where both the capability and safety of agents can be better studied and improved under realistic settings. Eventually, our principles of unified alignment for agents lay the foundation for the next-level agents more intelligent and more responsible.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023a.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/dbbf603ff0e99629dda5d75b6f75f966-Paper.pdf`.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2023b.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023c.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. Batch prompting: Efficient inference with large language model APIs. In Mingxuan Wang and Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 792–810, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-industry.74. URL `https://aclanthology.org/2023.emnlp-industry.74`.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning, 2019.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web, 2023.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18343–18362. Curran Associates, Inc., 2022.

Haoyuan Fu, Wenqiang Xu, Han Xue, Huinan Yang, Ruolin Ye, Yongxi Huang, Zhendong Xue, Yanfeng Wang, and Cewu Lu. Rfuniverse: A physics-based action-centric interactive environment for everyday household tasks. *arXiv preprint arXiv:2202.00199*, 2022.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

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

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023b.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*, 2023.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023a.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023b.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023c.

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

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pp. 2356–2362, 2021.

Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. Llm-powered hierarchical language agent for real-time human-ai coordination. *arXiv preprint arXiv:2312.15224*, 2023a.

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

Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. Active sparse conversations for improved audio-visual embodied navigation. *arXiv preprint arXiv:2306.04047*, 2023c.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023d.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv preprint arXiv:2401.13178*, 2024.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv: Arxiv-2310.12931*, 2023.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve GPT-3 after deployment. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2833–2861, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.183. URL `https://aclanthology.org/2022.emnlp-main.183`.

Pattie Maes. Artificial life meets entertainment: lifelike autonomous agents. *Communications of the ACM*, 38(11):108–114, 1995.

Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. Clin: A continually learning language agent for rapid task adaptation and generalization. *arXiv preprint arXiv:2310.10134*, 2023.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*, 2023.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL `https://doi.org/10.48550/arXiv.2303.08774`.

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

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. *Advances in Neural Information Processing Systems*, 35:6236–6249, 2022.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development, 2023.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. Autoact: Automatic agent learning from scratch via self-planning. *arXiv preprint arXiv:2401.05268*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. *arXiv preprint arXiv:2309.15817*, 2023.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pp. 285–295, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133480. doi: 10.1145/371920.372071. URL `https://doi.org/10.1145/371920.372071`.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pp. 9614–9625. PMLR, 2021.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. Task ambiguity in humans and language models. *arXiv preprint arXiv:2212.10711*, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. Androidenv: A reinforcement learning platform for android. *arXiv preprint arXiv:2105.13231*, 2021.

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

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader?, 2022a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023b.

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

Yue Wu, Xuan Tang, Tom M Mitchell, and Yuanzhi Li. Smartplay: A benchmark for llms as intelligent agents. *arXiv preprint arXiv:2310.01557*, 2023.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Fei Xia, Chengshu Li, Kevin Chen, William B Shen, Roberto Martın-Martın, Noriaki Hirose, Amir R Zamir, Li Fei-Fei, and Silvio Savarese. Gibson env v2: Embodied simulation environments for interactive navigation. *Stanford University, Tech. Rep.*, 2019.

Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2): 713–720, 2020.

Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*, 2023.

Zeyuan Yang, Peng Li, and Yang Liu. Failures pave the way: Enhancing large language models through tuning-free rule accumulation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1751–1777, 2023.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL `https://openreview.net/forum?id=5Xc1ecxO1h`.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*, 2023b.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, et al. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*, 2024.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. *arXiv preprint arXiv:2308.10144*, 2023.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024.

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

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, Hongxin Zhang, Yilun Du, Joshua B Tenenbaum, and Chuang Gan. Hazard challenge: Embodied decision making in dynamically changing environments. *arXiv preprint arXiv:2401.12975*, 2024.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023b.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

A Environment Construction in Section 4.1

In this section, We introduce the realistic features we introduce into the original WebShop in detail. Note that as we aim to conduct *proof-of-concept* studies, the features are implemented for the purpose of *reflecting the three lines of alignment only*. We also anticipate realistic configurations and more nuanced logistics in a dedicated benchmark in the future.

A.1 Task Design in the Retrofitted WebShop

Different from the precise human instructions in the original WebShop environment, we design tasks to reflect the necessity of agents to align with *human intentions*. In reality, different human users own unique preferences about the properties and categories of shopping items. Such preferences form the profile of a user, which dominates their authentic intentions in the stream of shopping instructions. As it is not always easy for human users to explicitly write down the precise instructions for all their shopping intentions, the agent should assist humans by continually inferring human intentions.

Given this, we configure 10 different users, each possessing a basic preference (described in text) that corresponds with a certain hidden attribute of items. For example, for the hidden attribute `cruelty-free`, we design the corresponding basic human profile sentence to be `I cannot care enough for the cute creatures in this world.` We follow the reward computation rules in the original WebShop, therefore the match of hidden attributes is essential to the final reward. We equip each user with a group of 50 consecutive artificially constructed instructions. In all the instructions of the group, the aforementioned profile sentence always appears.

The specific instruction for the purchase of this round falls into three cases:

- The basic preference of the user should be considered. An example in this category reads: `i am interested in a 60 count of toner that is suitable for sensitive skin, and price lower than 50.00 dollars.` For this instruction, the expected item to be found contains two hidden attributes: `sensitive skin` according to the task instruction, and `cruelty-free` according to the user profile.
- The basic preference of the user does not need to be considered. An example in this category reads: `i am looking for a nightstand that is easy to install.` Our motivation for this case is that the users with whatever preferences always need to buy some specific items, which can be irrelevant to their profiles. In this example, the underlying hidden attributes contain just the task-related `easy install`.
- The basic preference of the user should be considered, *and* an extra preference should be recognized according to the recent instruction histories. An example in this category reads: `Based on my purchase preference from history, help me to buy eco friendly face towels.` And the corresponding instruction history for this example is listed in the following Table 3. It should be inferred according to the history that the hidden attribute (as the invisible intention) that frequently appears in recent instructions are `sensitive skin`. We design the order of the instruction group so that such to-be-inferred attributes appear at least five times more than other attribute candidates in a sliding window. As a result, in this example, the set of the hidden attributes consists of the `cruelty-free` according to the user profile, the `eco friendly` according to the description in the textual instruction, and also the `sensitive skin` that is to be tracked and inferred from history.

Most of the task instructions in the former two groups are selected from the crowdsourced instructions, whose corresponding ground-truth items are labeled with the hidden attributes of both the basic user preference and the task-related instruction. We artificially rewrite the instructions in the third category by introducing indicating words like `Based on my purchase from the history`. In the 10 groups of 50 consecutive instructions, the statistics of the three categories is 298/97/105 for the first/second/third category, approximately 3:1:1. The task completion performances of all agent techniques should be tested on the 10 groups of 50 instructions each, with the overall averaged reward and success rate reported. In the ablated version of the retrofitted environment for the calculation of R_{HI} and $\mathbf{G}_{\mathrm{HI}}$ in Section 4.3, all the hidden attributes about the basic user profiles (*e.g.*, the `cruelty free`) and the preferences to be inferred (*e.g.*, the `sensitive skin`) are excluded from the reward computation.

Table 3: The instruction history for the instruction example: *Based on my purchase preference from history, help me to buy eco friendly face towels.*

The following instruction history is listed in reverse chronological order:
i am interested in a 60 count of toner that is suitable for sensitive skin.
i am looking for a sulfate & paraben free shampoo that is also suitable for my sensitive skin.
i want some hand cream for dry and sensitive hands in a grapefruit scent.
i need men's non toxic deororizing body spray for sensitive skin. get the 2pack 3.4 ounce size.
get me a body scrub to remove dead skin. pick a 3.4 fl oz pack that is meant for sensitive skin.
i need low rise boot cut pants in grey color.
buy me some paraben free coconut oil lip balm for my sensitive skin.
i'm looking for some juicy watermelon lip gloss that is paraben and oil free and suitable for sensitive skin.
i need soaps for dry and sensitive skin that are made with argan oil.
i'm looking to buy a body wash that has tea tree oil as an ingredient that would work well for sensitive skin.

A.2 Personalized Reranking Algorithms

A.2.1 Overview

To narrow the gap with realistic online shopping scenarios, we implement personalized reranking algorithms in WebShop. With the reranking algorithms, the search results of the shopping item list are re-ordered according to the click histories of the users. Specifically, on top of the Pyserini (Lin et al., 2021) search engine used in the original Webshop, we integrate a Collaborative Filtering (CF) (Sarwar et al., 2001) algorithm and a Determinantal Point Process (DPP) based algorithm (Chen et al., 2018) for fine-grained personalized re-ranking. The DPP-based algorithm provides the reranking weights based on the historical actions of the agent itself, while the CF-based algorithm provides the reranking weights based on the similarity with other users. The two sets of weights by the two algorithms are eventually linearly averaged with the coefficient of 0.2 for DPP-based weights and 0.8 for CF-based weights. Driven by the reranking algorithms, the environment is constantly evolving with user actions, which could better reflect the complexity of realistic environmental dynamics. In the ablated version of the retrofitted environment for the calculation of R_{ED} and $\mathbf{G}_{\mathrm{ED}}$ in Section 4.3, the two algorithms are disabled.

Algorithms 1 and 2 brief the collaborative filtering and DPP-based reranking, respectively.

Algorithm 1 User-Based Collaborative Filtering

Input: Prime user-item rating matrix $\mathbf{R}$, User Click Through Rate $\mathbf{U}$, Top-n items $\mathbf{I}$
Output: CF reranking score of top-n items $\mathbf{Y}$

```
for each prime user i do
    for j ∈ I do
        M_{i,j} = R_{i,j}
    end for
end for
for each prime user i do
    /* Calculate the intersection of items contained in the current agent and prime users */
    P = U ∩ R_i
    S_i = Σ_{p∈P}(R_{i,p} − R̄_i)(U_p − Ū) / ( sqrt(Σ_{p∈P}(R_{i,p} − R̄_i)^2) sqrt(Σ_{p∈P}(U_p − Ū)^2) )
end for
Y = SM / Σ S_i
```

$$\mathbf{M}_{i,j} = \mathbf{R}_{i,j}$$

$$P = \mathbf{U} \cap \mathbf{R}_i$$

$$\mathbf{S}_i = \frac{\sum_{p\in P}(\mathbf{R}_{i,p}-\overline{\mathbf{R}}_i)(\mathbf{U}_p-\overline{\mathbf{U}})}{\sqrt{\sum_{p\in P}(\mathbf{R}_{i,p}-\overline{\mathbf{R}}_i)^2}\sqrt{\sum_{p\in P}(\mathbf{U}_p-\overline{\mathbf{U}})^2}}$$

$$\mathbf{Y} = \mathbf{S}\mathbf{M}/\sum \mathbf{S}_i$$

In the implementation of the CF-based algorithm, we first employed ChatGPT (`gpt-3.5-turbo-1106`) as an assistant to simulate 30 different users and gather their preference data for collaborative filtering (detailed in Appendix A.2.2). During the shopping process, we record the click-through rates (CTR) of the agent on every item. We then re-rank the item list of the search results according to the agent's CTR and its Pearson correlation between the simulated user preferences.

Algorithm 2 Deterministic Point Process Based Reranking (Chen et al., 2018)

Input: Item score vector $\mathbf{I}$, Item similarity matrix $\mathbf{S}$, Top K
Output: DPP-based reranking score of top-n items Y_g

$\mathbf{c}_i = [], d_i^2 = \mathbf{L}_{ii}, j = \arg\max_{i \in Z} \log(d_i^2), Y_g = \{j\}$
$\mathbf{L} = \mathrm{diag}(\mathbf{I})\ \mathbf{S}\ \mathrm{diag}(\mathbf{I})$
while $|Y_g| < K$ **do**
  for $i \in Z \setminus Y_g$ **do**
    $e_i = (\mathbf{L}_{ji} - \langle \mathbf{c}_j, \mathbf{c}_i \rangle)/d_j$
    $\mathbf{c}_i = \mathbf{c}_i \| e_i$
    $d_i^2 = d_i^2 - e_i^2$
  end for
  $j = \arg\max_{i \in Z \setminus Y_g} \log(d_i^2), Y_g = Y_g \cup \{j\}$
end while

A.2.2 User Behavior Simulation

We employed ChatGPT to design 30 different roles, simulating the process of shopping and gathering the ranking results for 50 products for each role, the prompts are shown in Table 4. The collected data will be used to simulate user behavior for simulating the dynamic environment similar to recommendation systems with changeable displayed items for different users and behaviors. The information of 30 roles is shown in Table 5 and Table 6. Note that these roles are completely generated by ChatGPT, including their genders and other information. In this work, we construct the 30 roles to obtain the set of preference weights for only the purpose of introducing the reranking mechanisms into the environment. We plan to refine the construction of the profiles with a broader coverage of demographic groups in the future.

Table 4: The prompt for user behavior simulation using ChatGPT. We define roles (colored in green) and the query (colored in blue), asking for reranking items (colored in orange) given by the original Webshop. The generated results (colored in red) can serve as a simulation of user click behavior.

User Behavior Simulation

/* Prompt */
Your name is [NAME] and here is your profile:
Gender: [Gender]
Age: [Age]
Occupation: [Occupation]
Shopping Habits: [Shopping Habits]

You are searching for [Query] on a shopping website and obtain 50 results:

Id: [Id_1]; Description: [Desc_1]; Price: [Price_1]
Id: [Id_2]; Description: [Desc_2]; Price: [Price_2]
...
Id: [Id_{50}]; Description: [Desc_{50}]; Price: [Price_{50}]

Please sort all 50 products according to your preferences using the format of “Id-Ranking”.

/* Response */
[Id_1]-[Rank_1]; [Id_2]-[Rank_2]; ...; [Id_{50}]-[Rank_{50}];

Human Evaluation for the Ranking Results by ChatGPT. Furthermore, we conducted a human evaluation on the ranking results of ChatGPT, and the results in Table 7 show that the NDCG score of ChatGPT is 0.871, indicating that the simulation results are close to human ranking preferences.

Table 5: The generated 30 different simulated users by using ChatGPT (part 1).

Profile	**Role #1**	**Role #2**	**Role #3**	**Role #4**	**Role #5**
Name	Sarah	Juan	Lisa	Michael	Emma
Gender	Female	Male	Female	Male	Female
Age	32	21	40	45	55
Occupation	Software Engineer	College Student	Stay-at-home mom	Construction Manager	Retired Teacher
Habits	Sarah loves online shopping for the latest gadgets and tech accessories. She researches extensively, reads reviews, and compares prices before making a purchase. She's always on the lookout for the newest tech trends.	Juan is passionate about fashion and enjoys shopping for trendy clothing and accessories. He follows fashion influencers on social media, visits local boutiques, and regularly updates his wardrobe to stay stylish on campus.	Lisa prioritizes her family's health and wellness. She shops for organic groceries, supplements, and natural skincare products. She also invests in fitness equipment and enjoys trying new workout routines.	Michael is passionate about home improvement projects. He frequently visits hardware stores, researches tools and materials, and enjoys renovating and enhancing his living space. He seeks quality products for long-term durability.	Emma is an avid reader and loves hosting book club meetings. She enjoys browsing bookstores, collecting literary classics, and exploring various genres. She values recommendations from fellow book lovers.
Profile	**Role #6**	**Role #7**	**Role #8**	**Role #9**	**Role #10**
Name	Alex	Ryan	Maya	Daniel	Olivia
Gender	Non-binary	Male	Female	Male	Female
Age	27	35	28	50	42
Occupation	Etsy Shop Owner	Chef	Environmental Scientist	Pet Store Owner	Financial Analyst
Habits	Alex loves creating unique handmade items and runs an online store. They actively seek out specialty craft supplies, materials, and tools to produce high-quality products. They also enjoy attending craft fairs and networking with other artisans.	Ryan is passionate about cooking and constantly seeks out new ingredients and culinary tools. He enjoys shopping at local markets, specialty food stores, and online platforms for gourmet products. He values quality and freshness.	Maya loves hiking, camping, and exploring nature. She invests in high-quality outdoor gear, such as tents, hiking boots, and backpacks. She actively researches and reads reviews to ensure durability and functionality.	Daniel owns a pet store and constantly seeks out pet-related products for his business. He actively sources pet food, toys, grooming supplies, and accessories to cater to various pet owners' needs.	Olivia loves finding the best deals and discounts. She enjoys using coupons, comparing prices, and exploring online platforms to save money on her purchases. She values both quality and affordability.
Profile	**Role #11**	**Role #12**	**Role #13**	**Role #14**	**Role #15**
Name	Ahmed	Sophia	Carlos	Emily	Javier
Gender	Male	Female	Male	Female	Male
Age	38	25	30	27	34
Occupation	Physical Education Teacher	Social Media Influencer	Automotive Engineer	Environmental Activist	Travel Blogger
Habits	Ahmed is passionate about sports and fitness. He shops for athletic apparel, sports equipment, and supplements. He enjoys exploring local sports stores and stays updated on the latest fitness trends.	Sophia is a beauty enthusiast and creates content about cosmetics, skincare, and makeup tutorials. She actively seeks out new beauty products, follows trends, and shares her recommendations with her followers.	Carlos has a deep interest in cars and enjoys shopping for automotive accessories, performance parts, and maintenance tools. He actively researches and stays updated on the latest automobile technology.	Emily is focused on sustainable living and seeks out eco-friendly products. She shops for ethically sourced clothing, reusable items, and environmentally friendly household products.	Javier loves traveling and exploring new destinations. He shops for travel gear, luggage, and outdoor accessories. He values lightweight and durable products for his adventures.

Table 6: The generated 30 different simulated users by using ChatGPT (part 2).

Profile	**Role #16**	**Role #17**	**Role #18**	**Role #19**	**Role #20**
Name	Lily	Oliver	Emma	Noah	Ava
Gender	Female	Male	Female	Male	Female
Age	29	19	52	65	45
Occupation	Marketing Manager	College Student	Antique Store Owner	Retired Engineer	Art Gallery Owner
Habits	Lily recently became a new parent and actively shops for baby products, including clothing, toys, and nursery essentials. She seeks out trusted brands and prioritizes safety and quality.	Oliver is passionate about music and loves shopping for musical instruments, equipment, and vinyl records. He actively explores local music stores and online platforms for unique finds.	Emma has a keen interest in vintage items and actively seeks out antique furniture, clothing, and collectibles. She enjoys visiting flea markets, estate sales, and auctions to expand her collection.	Noah embraces technology and enjoys shopping for the latest gadgets, smartphones, and smart home devices. He actively seeks user-friendly products and keeps up with technological advancements.	Ava is passionate about art and actively collects paintings, sculptures, and other fine art pieces. She frequents art fairs, galleries, and auctions to discover new artists and expand her collection.
Profile	**Role #21**	**Role #22**	**Role #23**	**Role #24**	**Role #25**
Name	Max	Olivia	Liam	Sophia	Ethan
Gender	Male	Female	Male	Female	Male
Age	20	35	32	26	50
Occupation	College Student	Interior Designer	Personal Trainer	Graphic Designer	Landscape Architect
Habits	Max is an avid gamer and actively shops for the latest gaming consoles, accessories, and video games. He stays updated on gaming news, follows esports tournaments, and seeks out merchandise from his favorite games.	Olivia specializes in creating beautiful spaces and frequently shops for furniture, decor, and lighting fixtures. She stays updated on design trends, visits trade shows, and sources unique pieces for her clients.	Liam is dedicated to fitness and actively shops for workout apparel, equipment, and supplements. He seeks out high-quality gear that withstands intense training sessions and recommends products to his clients.	Sophia has a passion for stationery and actively shops for notebooks, pens, art supplies, and planners. She values aesthetically pleasing and functional products that inspire her creativity.	Ethan enjoys gardening and frequently shops for plants, seeds, gardening tools, and outdoor decor. He seeks out sustainable and eco-friendly products that enhance his garden.
Profile	**Role #26**	**Role #27**	**Role #28**	**Role #29**	**Role #30**
Name	Mia	Noah	Isabella	James	Harper
Gender	Female	Male	Female	Male	Female
Age	38	75	30	35	23
Occupation	CEO	Retired Teacher	Environmental Scientist	Stay-at-home dad	Vintage Clothing Store Owner
Habits	Mia appreciates luxury and actively shops for high-end fashion, accessories, and designer items. She seeks out exclusive brands, attends fashion events, and values premium craftsmanship.	Noah prefers simplicity when it comes to technology and shops for user-friendly devices, such as easy-to-use smartphones, tablets, and assistive technology. He values products with clear instructions and reliable customer support.	Isabella is committed to sustainable living and actively shops for eco-friendly products, including reusable bags, zero-waste toiletries, and environmentally friendly cleaning supplies. She values products with minimal environmental impact.	James is a hands-on parent and frequently shops for baby gear, including strollers, baby carriers, and childproofing items. He seeks out functional and safe products that make parenting easier.	Harper has a passion for vintage fashion and actively shops for retro clothing, accessories, and antique jewelry. She enjoys visiting thrift stores, vintage markets, and online platforms for unique finds.

Table 7: NDCG@10 scores of the ranking results according to the results by ChatGPT and eight human evaluators.

Annotator	Query #1	Query #2	Query #3	Query #4	Query #5	Average.
#1	0.807	0.830	0.826	0.857	0.819	0.828
#2	0.823	0.908	0.871	0.798	0.814	0.843
#3	0.795	0.956	0.999	0.962	0.964	0.935
#4	0.927	0.896	0.974	0.931	0.763	0.898
#5	0.843	0.900	0.910	0.961	0.876	0.898
#7	0.836	0.934	0.883	0.860	0.915	0.885
#6	0.851	0.905	0.872	0.749	0.841	0.844
#8	0.798	0.824	0.920	0.764	0.877	0.837
Average.	0.835	0.894	0. 907	0. 860	0.859	0.871

A.3 RUNTIME ENVIRONMENT

To measure the expenses of the agents themselves we implement the runtime environment for the agent working system that tracks the temporal and monetary expenditures. We compute the monetary cost of each API call of the proprietary foundation models based on their official pricing. We also track the time consumption of the interaction between the agents and the environment. As the response of the website can be affected by networking issues, we leverage the following benchmarking measures for simplicity: In our environment construction, we documented all kinds of actions taking place in the environment and pre-calculated a static list of their estimated response delays. We also disregard the duration of API calls in the runtime environment as the tracked monetary cost also reflects the expense of API calls. With the runtime environment as a wrapper of the working system, humans can monitor the obedience of the agents to their self-constraints.

Specifically, we estimate the response delay of each action in the interactive environment by artificially sampling actions, trying them out, and then recording the delays. After gathering the delay statistics for each action, we fit the data with a uniform distribution, and use the expected value to reflect the estimated time for the action. Such estimated time is static, and is leveraged to benchmark the time cost in the experiments. The estimated time for all the actions is listed in the Table 8.

Table 8: Estimated time for different actions.

Action	**Time (s)**
`reset`	0.1874
`search`	0.5966
`click[Instruction History]`	0.2645
`click[Back to Search]`	0.1197
`click[Next >]`	0.2693
`click[< Prev]`	0.2545
`click[Descriptions]`	0.2401
`click[Features]`	0.2167
`click[Reviews]`	0.1275
`click[Buy Now]`	0.1920
`click[other valid tag]`	0.2896
`think`	0.0000
Invalid Action	0.3234
Average	**0.2370**

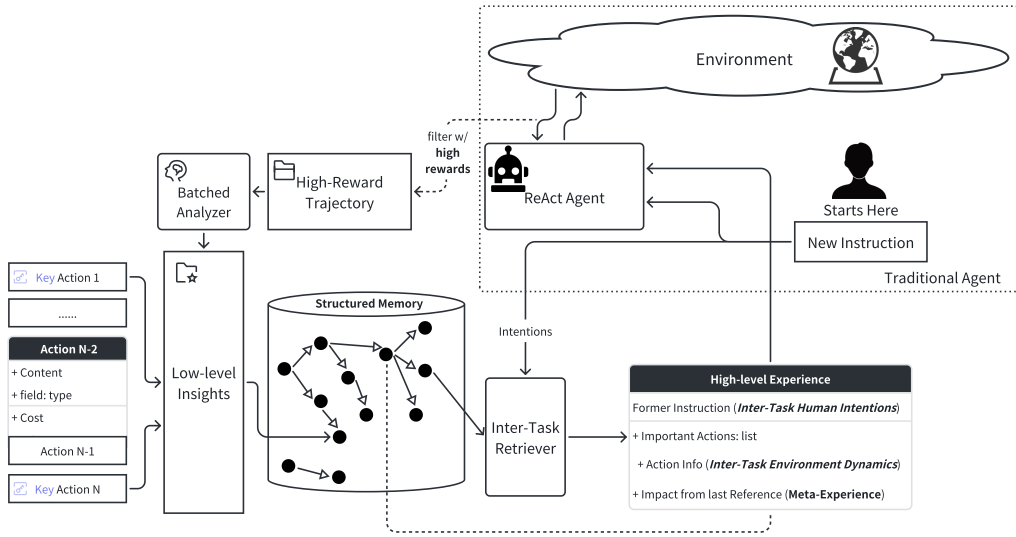

Figure 5: The details of our agent design that follows the principles of **UA**^{2}. Compared to traditional `ReACT` agents, we append structured experience as the long-term memory: By filtering and analyzing raw trajectories, we extracted key actions from prior successes as low-level insights in reasoning/action paths. By retrieving reference low-level insights under the same user, we can find the high-level experience under most similar user instructions, expressing similar human intentions. Agents are able to understand human intentions and environment dynamics by extrapolating key actions from a similar, prior task.

B Experiments

B.1 Descriptions on Agent Design with the Principles of UA^2

From Section 3.2, we identify three core capacities essential to agents and how they are related to the alignment principles. However, the trade-off between stronger capacities and fulfilling these principles makes it challenging to design a universally satisfactory method. Instead, to make the very first attempt, we consider improving well-known techniques to satisfy most principles.

We introduce structured memory for unified alignment principles for agents (**UA**^{2}), as depicted in Figure 5, on top of the original ReAct (Yao et al., 2022b) agent.

Definition B.1. Trajectory is a list of actions (a_i) and observations (o_i) that agents have observed from environment ($\mathcal{E}$) after each action: $\mathcal{T} = \{a_i, o_i\}_{i=1}^n, o_i = \mathcal{E}(\{a_k\}_{k=1}^i)$.

After filtering high-reward trajectories (Definition B.1) under human intentions from the given instruction q and temporal environment conditions ($\mathcal{E}$), we utilize a batched analyzer (Cheng et al., 2023) that tags key actions from the whole trajectory within one API call. Thus we integrate insights from environment dynamics with low costs compared to Reflexion (Shinn et al., 2023) and LATS (Zhou et al., 2023a).

Definition B.2. Key actions (a_i^*) are those having a positive impact on the efficiency or efficacy of task completion. We obtain the key actions $\mathcal{T}_q^* = \{a_i^*\}_{i=1}^m$ each time the agent completes a task.

Definition B.3. Low-level insights is a list of key actions (a_i^*) under a given instruction (q): $\mathcal{T}_q^* = \{a_i^*\}_{i=1}^m$. On top of this, structured experiences at t-th query are formed by a set of paired previous instructions and key actions: $\mathcal{S} = \{(q, \mathcal{T}_q^*) : q \in \mathcal{Q}_{t-1}\}$, where $\mathcal{Q}_{t-1}$ denotes the set of previous instructions before t-th query.

Thus, we enhance agents' memory by mapping low-level insights with corresponding instructions of previous tasks, which formulates a structured memory ($\mathcal{S}$) (Definition B.3). It is worth noting that we construct the structured memory under each user, which allows agents to comprehend human intention from prior instructions. When a new instruction is given, denoted as q_{given}, a reference containing high-level experience under the instruction q_r could be retrieved by simply calculating the most similar instructions from its memory using *BM25* (Robertson et al., 2009) scoring: $\mathcal{T}^*_{q_r} \in \mathcal{S}, q_r = \mathrm{argmax}_{q \in \mathcal{Q}}\{\mathrm{BM25}(q, q_{given})\}$.

The retrieved high-level experience acts as a plan across tasks which guides the agent towards the goal accurately and rapidly, also reducing costs of time and money. After completing a task under the environment, the agent analyzes the new trajectory as well as the experience learned from the former reference, termed "meta-experience". We then record the new insights adjacent to the former reference. Also, we copy the "meta-experience" in the attachment to the former reference, which could be retrieved for upcoming tasks.

B.2 Implementation Details

We evaluate our method and baseline methods across all 10 users on our retrofitted Webshop, each comprising 50 tasks, except for LATS which is evaluated with only one user due to its high cost. All methods utilize `gpt-3.5-turbo-instruct-0914` as the underlying model for their agents except for LATS where we utilize `gpt-3.5-turbo-1106` to keep the same setting as the original paper. In executing each task, we limited the interaction with the web to a maximum of 15 steps per task, inclusive of any invalid actions.

B.3 Details of Experimental Setups

For ReAct, Reflexion, and our method, we set the temperature as 0.0. For ReAct-SC, we set the number of samples k to be 3 and the temperature to be 0.05. We also experiment with the family of chain-of-thought methods: CoT (Wei et al., 2022), CoT-L2M (Zhou et al., 2022), CoT-SC (Wang et al., 2022b). Since their performances in task completion are significantly less competitive than other methods (see the next section), we exclude them from the main experiments in Section 4.3, but include them here for reference. For CoT and CoT-L2M, we set the temperature as 0.0; and for CoT-SC, we set $k = 3$ and the temperature to be 0.05. To adhere to the same settings with (Zhou et al., 2023a), we set the temperature as 1.0, k as 5, the number of iterations n as 30 for LATS.

All methods were tested in each of the following three environments respectively:

- The fully retrofitted environment: configured exactly as described in Section 4.1.
- The ablated environment that excludes *human intentions*: Based on the fully retrofitted environment, the hidden attributes from the user profiles and to be inferred from purchase histories are not excluded from reward computation. The alignment gap with *human intentions* ($\mathbf{G}_{\mathrm{HI}}$) can be identified by comparing the test performances therein with those in the fully retrofitted environment.
- The ablated environment that excludes *environmental dynamics*: The fully retrofitted environment with the re-ranking algorithms in Appendix A.2 disabled. The alignment gap with *environmental dynamics* ($\mathbf{G}_{\mathrm{ED}}$) can be identified by comparing the test performances therein with those in the fully retrofitted environment.

B.4 Results

We present the comparative results on our retrofitted WebShop in Figure 6, Figure 7, and Table 9. Note that due to the significantly lower reward and success rate of CoT-related methods compared to others, the relative differences $\mathbf{G}_{\mathrm{HI}}$ and $\mathbf{G}_{\mathrm{ED}}$ can be dominantly affected by stochastic issues, and are therefore not of reference and comparison value.

In each figure, the X-axis represents the alignment gap with *self-constraints*, and the Y-axis represents the performance. In terms of success rate, our proposed agent demonstrates comparable performance to Reflexion. To be specific, our approach places a greater emphasis on minimizing costs, whereas Reflexion prioritizes performance improvement. Our proposed agent, guided by the principles of $\mathbf{UA}^2$, achieves a good balance between reward and cost considerations, while there remain a substantial gap between our agent and the ultimate goal of an oracle agent.

Table 9: Reward, success rate (SR), alignment gap with human intentions and environment dynamics, time and money cost result on our retrofitted WebShop. *LATS is tested on 1/10 subset of the entire task collection due to the significant cost.

Method	Reward ↑	SR (%) ↑	$\mathbf{G}_{\mathrm{HI}}$ (%) ↓	$\mathbf{G}_{\mathrm{ED}}$ (%) ↓	Time (s) ↓	Money ($) ↓
CoT	8.5	1.2	22.4	67.1	1.858	0.011
CoT-L2M	9.8	0.8	6.1	8.2	1.939	0.037
CoT-SC	11.8	1.4	32.2	-61.9	1.883	0.032
ReAct	50.3	8.0	11.7	14.9	1.716	0.013
ReAct-SC	49.9	7.4	14.4	14.6	1.720	0.039
Reflexion	44.4	13.8	22.5	25.7	5.539	0.045
LATS*	52.4	10.0	18.5	14.3	125.935	5.508
Ours	51.9	9.6	6.7	14.8	1.779	0.014

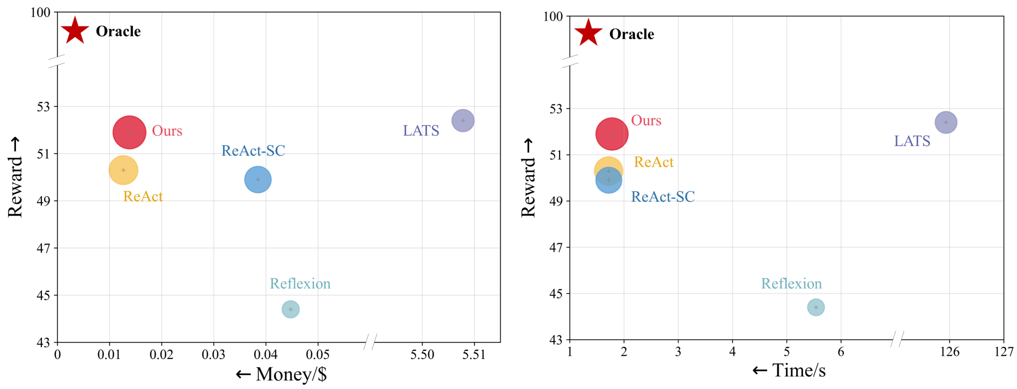

Figure 6: Agent's performance against the alignment gap with self-constraints tested on the retrofitted WebShop. The size of each circle represents the alignment gap with *human intentions* ($\mathbf{G}_{\mathrm{HI}}$). The red star symbolizes our ultimate goal of developing an oracle agent capable of flawlessly completing complex tasks with minimal cost.

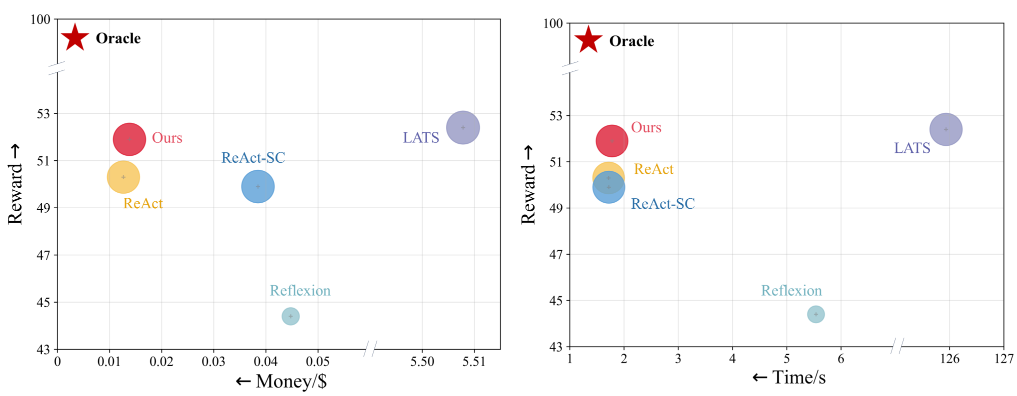

Figure 7: Agent's performance against the alignment gap with self-constraints tested on the retrofitted WebShop. The size of each circle represents the alignment gap with *environmental dynamics* ($\mathbf{G}_{\mathrm{ED}}$). The red star symbolizes our ultimate goal of developing an oracle agent capable of flawlessly completing complex tasks with minimal cost.