

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 VESX: A FRAMEWORK FEATURED BY VERIFICATION, SELF-CORRECTION AND IN-CONTEXT LEARNING FOR WEB AUTOMATION TASKS

Anonymous authors

Paper under double-blind review

ABSTRACT

While large language models have achieved remarkable success in tasks such as reasoning and question answering, applying LLMs to interactive tasks like web automation remains challenging. In web automation, existing planning-execution workflow often faces limitations due to the infeasible subtasks. We propose VeSX, a framework designed to enhance subtask feasibility through verification, self-correction, and in-context learning. VeSX introduces three key improvements: (1) subgoal-guided verification, which verifies the execution results of subtasks based on the preset subgoals; (2) hierarchical self-correction, which combines reflection and replanning, targeting to self-correct mistakes in both planning and execution phases; (3) exemplar bank, which improves in-context learning by partitioning execution trajectories and heuristically generating metadata for exemplars. We evaluate VeSX on WebArena benchmark and achieve the state-of-the-art average success rate of 0.34, which significantly outperforms existing methods without human guidance on all five scenarios.

1 INTRODUCTION

Large language models (LLMs) have achieved notable success in tasks such as logical reasoning, commonsense understanding, and question answering (Achiam et al., 2023; Touvron et al., 2023; Brown et al., 2020; Wei et al., 2022; Yao et al., 2023). Building agents with LLM as backbones to handle tasks in interactive scenarios has been identified as a feasible and promising approach but applying these agents in the real world remains highly challenging (Wang et al., 2023b; Kim et al., 2024; Wang et al., 2024). Web automation is one such scenario where LLM agents autonomously control web browsers to perform online activities such as shopping and posting by executing actions like clicking or typing.

When handling web automation tasks, two predominant agent workflow paradigms are empirically employed by current LLM agents. The first involves sequentially generating a series of actions (Koh et al., 2024; Deng et al., 2024; Zheng et al., 2023), as illustrated in Figure 1(a). The second commences with a planning phase, breaking down the task into a series of subtasks which are to be tackled one by one(Kim et al., 2024; Sodhi et al., 2024). For complex tasks that require multi-step reasoning, the second workflow is beneficial for that it breaks actions down into easier subtasks like chain-of-thought instead of a hard final task. However, such workflows are often constrained by the lack of guarantees regarding subtask feasibility (Wang et al., 2023b). An effective and commonly-used remedy is to rely on expertly designed subtask candidates with manually crafted prompts to decompose the entire task as illustrated in Figure 1(b). However, this method requires human annotation for new tasks, which limits the ability of LLM agents to autonomously adapt to more advanced and diverse web environments Sodhi et al. (2024); Zhou et al. (2022). Additionally, there is a significant shortage of data available for learning within these workflows, which collectively restricts the capabilities of LLMs during the planning and execution phases Lai et al. (2024). To overcome these limitations, we propose VeSX, a framework for LLM agents in web automation that seeks to improve subtask feasibility and empower LLM both in planning and execution through Verification, Self-correction, and in-context learning.

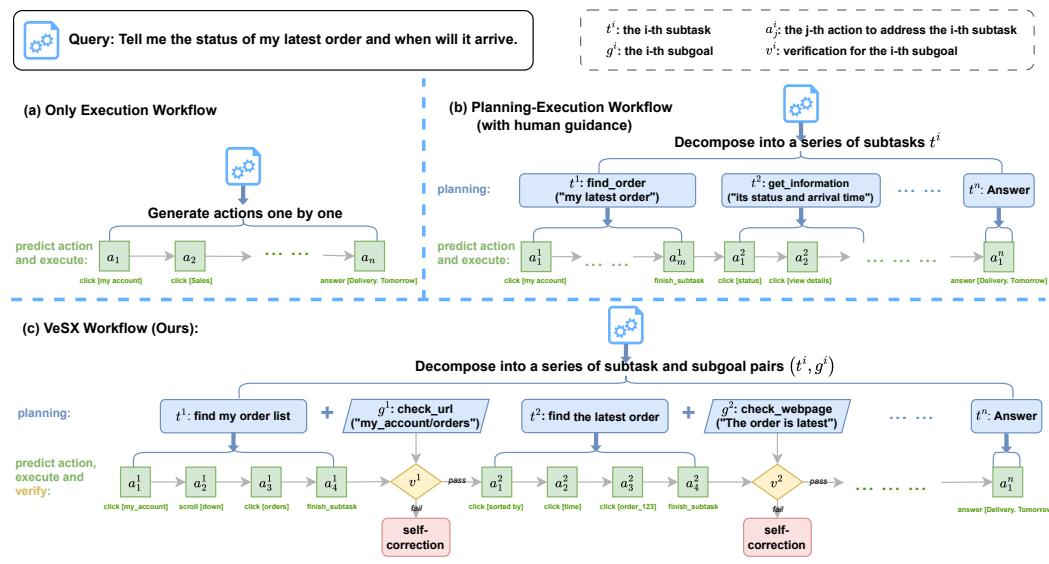


Figure 1: Comparison among different agent workflows for web automation. (a) is the workflow only engages the execution phase that predicts actions one by one. (b) is the planning-execution workflow with human guidance. (c) is the proposed VeSX workflow, featured with verification and self-correction.

VeSX introduces two significant modules to the planning-execution workflow for web automation tasks, as illustrated in Figure 1(c). The first involves adopting **subgoal-guided verification** to ascertain the successful completion of subtasks. VeSX generates subgoals in the planning phase and conducts verification to check whether the execution result satisfies the subgoal once execution is finished. We devised two modes for verification: self-verification which makes LLMs do verification by themselves, and external verification which calls for external functions to conduct verification. Verification ensures the completion of the subtasks and avoids the issue of subsequent subtasks becoming infeasible due to errors in the execution of previous subtasks.

The second improvement is **hierarchical self-correction**, which aims to self-correct both execution and planning. When verification fails, VeSX first tries to identify and correct errors in the current subtask’s execution trajectory, a process referred to as “reflection”. If reflection is unsuccessful, the subtask is likely infeasible, prompting VeSX to regenerate the remaining incomplete subtasks in a process called “replanning”. This hierarchical self-correction mechanism facilitates error correction during both the execution and planning phases, enabling the framework to effectively assess the feasibility of subtasks and generate new plans as replacements.

Furthermore, VeSX develops an **exemplar bank** for in-context learning within the planning and execution phase. Existing research on in-context learning related to web automation primarily focuses on the execution-only workflow and is limited to simple simulated tasks, leading to a dependence on complete execution trajectories as exemplars (Kim et al., 2024; Zheng et al., 2023). However, complex web automation tasks in the real world involve long action sequences, and using entire trajectories as exemplars includes a lot of irrelevant information, which will consume excessive content space. To address these issues, VeSX samples actions and planning strategy from the entire trajectory as exemplars for execution and planning, and heuristically generates metadata for retrieval. This method creates a larger and more diverse dataset for the LLM to learn from, and therefore improve its performance in both the planning and execution phases.

We evaluate VeSX on 5 scenarios under the web automation benchmark, WebArena(Zhou et al., 2023b). VeSX outperforms all publicly reported methods without human guidance by a huge margin and pushes the state-of-the-art success rate to 34.0, which even exceeds the performance of the method using human guidance.

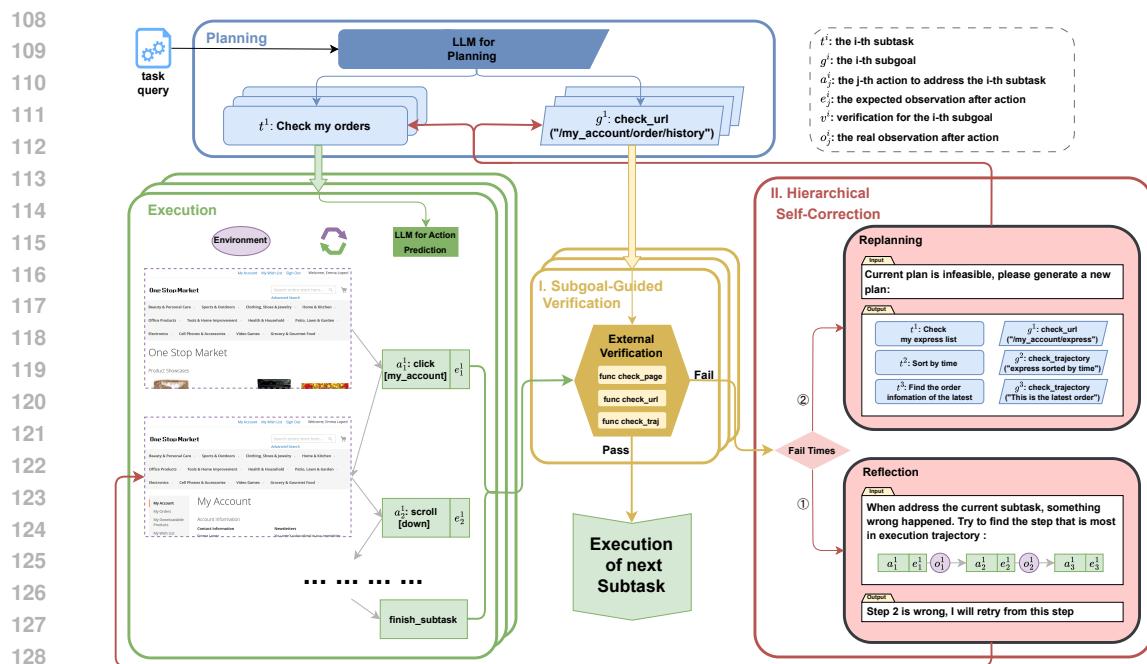


Figure 2: The framework of VeSX, which is made up of four phases and featured with subgoal-guided verification and hierarchical self-correction.

2 VESX

2.1 OVERVIEW

The VeSX involves innovative design in both process and data aspects. In terms of process, VeSX implements a highly accurate verification scheme based on subgoals generated in the planning phase, which we refer to as subgoal-guided verification (Section 2.2). Additionally, when verification fails, we design a hierarchical self-correction scheme that involves local reflection to tackle the problems within one certain subtask and global replanning to address the errors happening in the planning (Section 2.3). In terms of data, to address the low efficiency of in-context-learning in the previous web automation methods, VeSX uses a single step instead of the entire trajectory for in-context-learning, and improves the metadata construction method to efficiently and accurately retrieve effective exemplars (Section 2.4).

2.2 SUBGOAL-GUIDED VERIFICATION

Generate subgoals alongside with subtasks In order to verify whether execution results align with the intentions, VeSX instructs the LLM to generate not only subtasks but also corresponding subgoal pairs during the planning phase, as illustrated in Figure 2. The requirements for tasks in real-world scenarios are often highly diverse, leading to a wide array of corresponding solutions. However, the methods for assessing whether these requirements can be met are typically simplistic and fixed. For example, in web automation, although there may be hundreds of distinct task intents to handle, the evaluation methods are relatively limited. For navigation tasks, the evaluation may be based on whether a specific webpage has been reached. For information-gathering tasks, the presence of the required information on the final webpage is often the criterion. We observe that evaluating the execution result of a subtask is much simpler and more standardized compared to addressing the subtask itself. From this perspective, we propose the subgoal-guided verification, which evaluates the execution results of the subtasks according to the subgoals. Specifically, in VeSX, we generate pairs of subtasks and subgoals during the planning phase. Subtasks are articulated in natural language to capture the diverse task requirements, which in turn guide the generation of actions during the execution phase. Meanwhile, subgoals, which are formulated in either natural

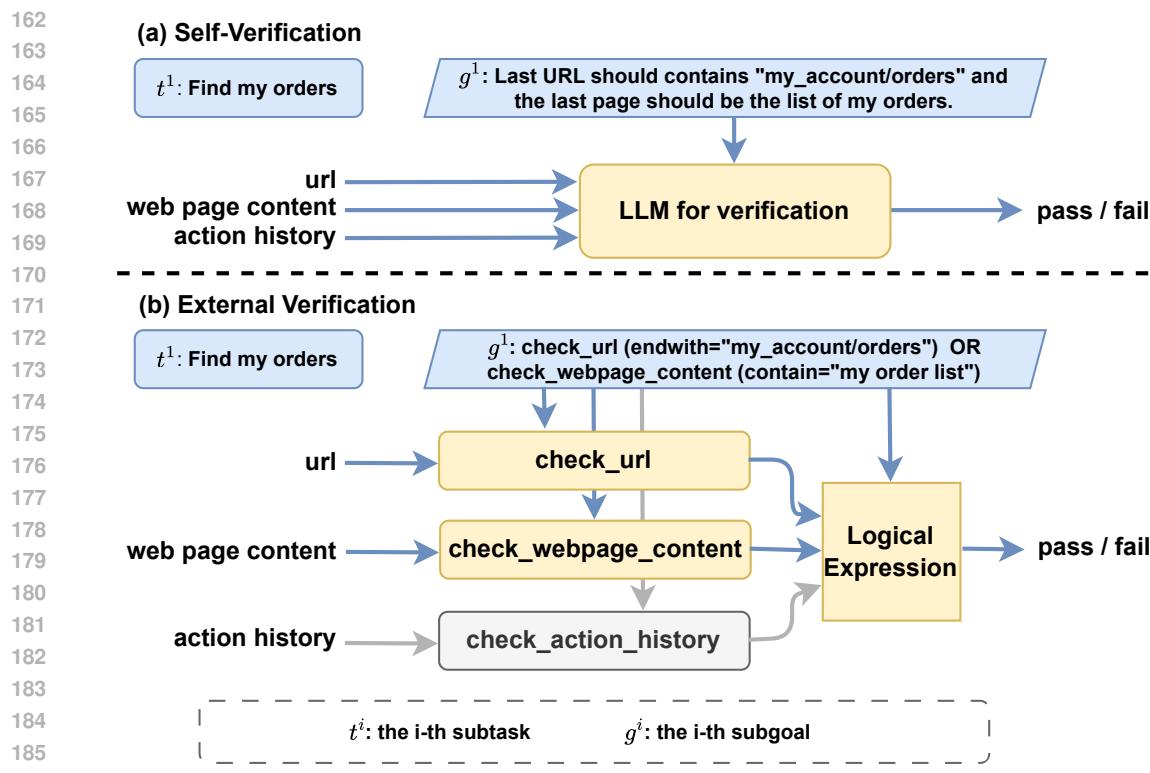


Figure 3: Two methods for verification. (a) is the self-verification using the LLM itself to check the subgoals are accomplished or not. (b) is the external verification using generally predefined verification functions to check the accomplishment, which is more stable and reliable.

or structured language—corresponding to the two verification methods we discuss later—are used to invoke external models or pre-designed tools to verify whether the subtasks have been successfully completed.

Verification Method We employed two methods to implement verification as illustrated in Figure 3. The first method is self-verification, which involves using natural language for subgoals. Self-verification incorporates all external environmental information and execution processes, making the LLM itself verify whether the execution results align with the previous intent. Self-verification has the advantage of flexibility, which means it can give the LLM to make decisions according to the current environment rather than totally depending on the predefined scheme.

The second method is external verification, which involves presenting subgoals in a structured language akin to function calls. In this approach, subgoals parameterize a set of predefined verification functions, creating a logical expression that defines the pass/fail criteria. During the verification process, external functions will be invoked according to the specifications of the subgoals. The results of these invocations will undergo logical operations to produce the final verification outcome, as illustrated in Figure 3(b). The advantage of this method is that it relies on external functions for verification, which is more stable, controllable, allows for swift adjustments, and reduces dependence on the LLM’s capabilities.

2.3 HIERARCHICAL SELF-CORRECTION

Due to the inherent unpredictability in web environments, errors during an agent's execution are often unavoidable, making self-correction essential. However, since planning can lead to infeasible subtasks, merely reflecting on and retrying execution actions is insufficient for effective error correction. From this perspective, we propose a hierarchical self-correction approach. Specifically, the hierarchical self-correction phase will be engaged when the verification fails. First, VeSX will em-

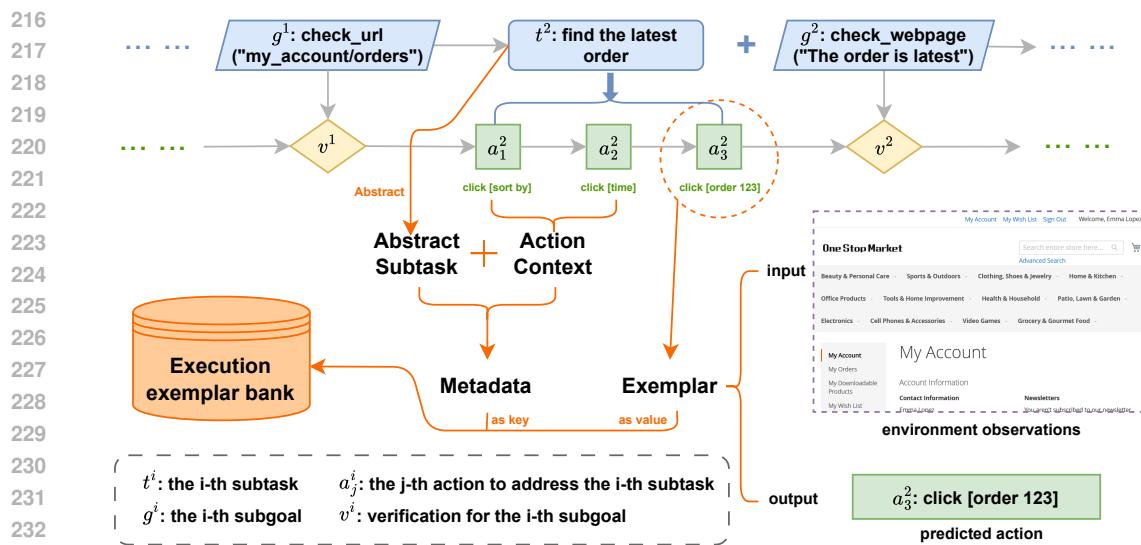


Figure 4: Construction of execution exemplar bank. We use the abstract subtask and the action context as the metadata for the exemplar. The key-value pairs of metadata’s embedding are stored to construct exemplar bank.

ploy reflection to identify mistakes in the execution trajectories associated with the current subtask. If reflection fails, it indicates that the generated subtasks may be infeasible or too challenging to solve. In that case, replanning will adjust the subtasks and subgoals dynamically.

Reflection In reflection, the LLM takes the execution trajectory and verification results as inputs to identify the step where the error occurred and subsequently re-execute from that point. Previous work on making LLMs reflect primarily focused on errors that occur when actions are generated by the environment and task queries. However, in real interactive scenarios, errors are likely to occur when changes in the environment after action execution do not align with the expected outcomes of those actions. Adjustments to previous actions are also needed in such cases. To address this issue, we enhanced the outputs during the execution phase by having the LLM additionally output the expected results e_j^i following an action. This allows the model to reflect on whether there is a discrepancy between expectations and environment changes during reflection thereby adjusting the action generation accordingly.

Replanning To enable dynamic adjustment capabilities in planning, VeSX employs replanning when reflection fails to correct errors. Specifically, when execution can not pass verification and reflection cannot rectify the issue, we consider the current subtask to potentially be infeasible. The previous error information and environment feedback are then fed into the LLM, prompting the LLM to regenerate subsequent subtasks and subgoals. Compared to the common framework that relies solely on reflection, the hybrid self-correction approach allows for adaptive adjustment, avoiding infeasible subtasks and thereby enhancing the long-term planning capabilities within the framework.

2.4 EXEMPLAR BANK

Previous research has typically relied on recalling the entire correct execution trajectories to perform in-context learning for execution. However, real-world tasks often require multi-step reasoning and operations, with the underlying logic of these steps varying significantly. This makes it challenging to use the full trajectory as an exemplar to guide the prediction of single-step actions, as it incorporates a significant amount of extraneous information and consumes substantial context. We decompose trajectories into individual actions as exemplars and have meticulously designed a heuristic approach to generate their metadata, which will serve as the key for recalling these exemplars later. There is a lack of research on collecting and recalling planning exemplars.

270 **Metadata of execution exemplars** VeSX utilizes single-step action instead of the entire trajectory
 271 as an exemplar to avoid the redundancy of the whole trajectory and enables LLM to learn from
 272 more diverse action examples. This involves designing metadata to describe the exemplars effec-
 273 tively. Specifically, for each action, VeSX will abstract the tackling subtask and concatenate with
 274 the action context, which means the sequence of prior actions within that subtask to get the meta-
 275 data for the exemplar, as shown in Figure 4. For instance, in the subtask “Find the latest order” the
 276 correct execution trajectory constitutes three actions `click [sort by]`, `click [time]`, and
 277 `click [order 123]`. We treat each action as an individual exemplar and generate its metadata.
 278 The metadata for the third action would be “Abstract Subtask: find conditional order; Action Con-
 279 text: click sort by, click time”. For each piece of exemplar, we store its embedding of metadata as the
 280 key and the exemplar as the value to construct the execution exemplar bank. Subsequently, during
 281 execution, VeSX recalls the most relevant examples from the exemplar bank through similarity-
 282 matching embeddings for in-context learning.

283 **Collecting planning exemplars** Due to the planning phase playing a significant role in the
 284 planning-execution framework, VeSX also collects the planning exemplars from the correct tra-
 285 jectory except for the execution exemplars. The metadata of planning exemplars is generated from
 286 its corresponding task query, abstracting the specific concepts in it, and then the planning exemplar
 287 bank is constructed with the key-value pairs of embedding of metadata and exemplars. Construct
 288 planning exemplar bank enhance models’ reasoning ability in the planning phase and facilitate the
 289 generation of more feasible subtasks.

290 3 EXPERIMENTS

291 3.1 EXPERIMENTAL SETUP

292 We evaluate VeSX on 5 different scenarios within the benchmark WebArena. WebArena is a bench-
 293 mark featured with complex web tasks across multiple scenarios including Shop (Amazon Shop-
 294 ping site), CMS (Amazon Commercial Management System), Red (Reddit), Git (Gitlab), and Map
 295 (Open Street Map). It builds several virtual websites from the snapshots of real-world websites like
 296 Amazon Shopping, GitLab, and Reddit, and allows agents to take actions freely like humans do in
 297 browsers, making it well-suited for evaluating what LLM web agents will do in the real world.

298 To ensure fair comparisons, we choose `gpt-4o` the same as previous works without finetuning
 299 the LLMs. We use `text-embedding-3-small` as the embedding model and retrieve the top 3
 300 relevant exemplars for in-context learning. We follow the browser system and action space built by
 301 Zhou et al. (2023b). We also set the temperature of the LLM to zero to minimize the uncertainty
 302 to generate reproducible results. For the data collection for the exemplar, we sampled 60 tasks
 303 randomly for each scenario and let the LLM try to solve them but not evaluate them with the correct
 304 answer. We heuristically filter out the bad trajectories like repeating actions or reaching the max
 305 actions, then make LLM self-evaluate the result by itself. If LLM considers the whole execution
 306 trajectory to be reasonable and the final result may be correct, the actions in this trajectory will be
 307 collected as exemplars. We use external verification in default unless a specific verification method
 308 is explicitly specified.

309 3.2 MAIN RESULTS

310 **Baselines** We selected five works from different perspectives as the baseline for the experiment.
 311 Zhou et al. (2023b) and Lai et al. (2024) focus on improving the single model’s reasoning ability
 312 by prompt or reinforcement learning: Zhou et al. (2023b) tries to improve the model performance
 313 using various combinations of prompt strategies. Lai et al. (2024) uses reinforcement learning to
 314 fine-tune an LLM with more knowledge of the web environment. Drouin et al. (2024) tries to refine
 315 this problem from the perspective of the environment, which builds a new browser system with more
 316 versatile actions and tries to resolve the tasks from the perspective of the environment. Koh et al.
 317 (2024) and Sodhi et al. (2024) improve from the perspective of the LLM agent workflow, using
 318 the strategy like reflection and planning: Koh et al. (2024) enhances the workflow of generating
 319 actions sequentially by tree search method. And Sodhi et al. (2024) improves the planning-execution
 320 workflow with the planning in a stack of subtasks and expert-designed examples.

Method	Model	Human-Feedback	Shop.	CMS	Red.	Git.	Map	Avg
Zhou et al. (2023b)	gpt-4o	×	-	-	-	-	-	0.131
Lai et al. (2024)	AutoWebGLM	×	-	-	-	-	-	0.182
Drouin et al. (2024)	gpt-4o	×	0.172	0.148	0.202	0.190	0.255	0.235
Koh et al. (2024)	gpt-4o	×	0.281	0.165	0.105	0.133	0.255	0.192
Sodhi et al. (2024)	gpt-4o	✓	0.369	0.242	0.594	0.317	0.303	0.335
VeSX	gpt-4o	×	0.412	0.249	<i>0.545</i>	<i>0.283</i>	<i>0.291</i>	0.340

Table 1: Task success rate on five scenarios of WebArena. The best performance is displayed in bold, and the second performance is displayed in italics. It constitutes of five scenarios: Shop (Amazon Shopping site), CMS (Amazon Commercial Management System), Red (Reddit), Git (Gitlab), and Map (Open Street Map)

Do Verification?	Methods	SR	ΔSR
✗	-	0.288	0.0
✓	self	0.346	+0.058
✓	external	0.412	+0.134

Table 2: Ablation study on subgoal-guided verification. SR represents the success rate.

Reflection	Replan	SR	ΔSR
✗	✓	0.261	0.0
✓ (w/o expect)	✗	0.314	+0.052
✓	✗	0.340	+0.078
✓	✓	0.412	+0.150

Table 3: Ablation study on hierarchical self-correction. SR represents the success rate.

Methods	SR	ΔSR	#P	#E
only execution exemplars	0.340	0.0	0	171
only planning exemplars	0.373	+0.033	35	0
VeSX	0.412	+0.072	35	171

Table 4: Ablation study on exemplar bank. SR represents the success rate. #P refers to the number of planning exemplars and #E refers to the number of planning exemplars.

Methods	SR	ΔSR
w/o abstraction	0.275	0.0
w/o action context	0.333	+0.058
abstraction & action context	0.412	+0.137

Table 5: Ablation study on building metadata. SR represents the success rate.

As shown in Table 1, VeSX consistently outperforms the publicly known best-reported methods without human guidance in all five different scenarios. The improvements are particularly significant in the Shop, CMS, and Red scenarios, with relative enhancements of 47%, 51%, and 169%, respectively, benefiting from our robust error detection and correction capabilities. Additionally, compared to the method in Sodhi et al. (2024), which uses expert-written examples for each scenario, our framework—entirely reliant on autonomous planning and execution by the LLM—demonstrates competitive performance, achieving superior results in two out of five scenarios, which are Shop and CMS. The hard tasks in Shop and CMS often require multi-step reasoning, creating a great opportunity for VeSX to showcase its error correction capabilities.

3.3 ABLATION STUDIES

To evaluate the impact of each part of VeSX, we conduct a comprehensive ablation study in this section. Our experiments are mainly conducted in the Shop scenario, whose tasks often involve multi-step reasoning and operations, well-suited for evaluating the effectiveness of our various methods. SR means the success rate of all tasks in this scenario, which corresponds to the column of Shop in Table 1. These findings validate the effectiveness of the subgoal-guided verification, hierarchical self-correction, and exemplar bank.

378 3.3.1 ABLATING SUBGOAL-GUIDED VERIFICATION
379

380 We conducted a comparative experiment to assess the implementation and methodologies of ver-
 381 ification. verification will not be performed for the implementation of no verification because it
 382 does not have the ability to detect the error. As shown in Table 2, with the implementation of both
 383 verification schemes, the success rate has significantly improved. It is because the two verification
 384 methods can both enable the framework the error-detection ability. Moreover, using external verifi-
 385 cation provides additional support compared to self-verification, resulting in even more pronounced
 386 enhancements.

387 3.3.2 ABLATING HIERARCHICAL SELF-CORRECTION
388

389 We conducted ablation experiments to assess the impact of various components of self-correction on
 390 overall performance. We tested three implementations: using only reflection, using only replanning,
 391 and employing reflection without generating expectations during execution.

392 As illustrated in Table 3, it is evident that removing any component significantly affects the final
 393 performance. In comparison to replanning, the removal of reflection has a more substantial impact,
 394 indicating that errors generated during the execution phase are more frequent or potentially easier
 395 to rectify. Nonetheless, the absence of replanning also leads to a rapid decline in the success rate,
 396 underscoring its necessity. Moreover, the success rate decreases when expectations are not utilized,
 397 demonstrating that expectations indeed enhance VeSu’s ability to detect discrepancies between en-
 398 vironmental changes and action predictions.

399 400 3.3.3 ABLATING EXEMPLAR BANK
401

402 We performed ablation experiments on the components of the exemplar bank and evaluated vari-
 403 ous strategies for generating metadata to determine each element’s contribution to overall system
 404 performance. The construction of the exemplar bank follows the same format as the main results.
 405 We sampled 60 tasks from the shopping experiment and employed a heuristic approach, allowing
 406 the LLM to evaluate its own execution results. Ultimately, we extracted 35 trajectories that were
 407 deemed correctly completed, thereby forming the exemplar bank.

408 As demonstrated in Table 5, excluding either the planning or execution exemplars led to a decrease
 409 in the success rate. Notably, utilizing only planning exemplars, without the inclusion of execution
 410 exemplars, achieved a higher success rate despite relying on fewer data samples. Based on the obser-
 411 vations found in Section 3.3.2, this phenomenon suggests that for each subtask, generating accurate
 412 execution sequences requires dynamic adjustments. Even with a limited number of samples, the
 413 LLM can accomplish tasks through self-correction, indicating a higher dependency on an effective
 414 error-correction mechanism rather than on the volume of data. Conversely, errors in the planning
 415 phase seem to pose a more complex challenge, requiring data as a reference for effective reasoning.
 416 This also implies that, compared to execution exemplars, planning exemplars are more efficient in
 417 improving performance.

418 We conducted an ablation study on two techniques: abstracting subtasks and using action context.
 419 Our results indicate that both techniques are essential for enhancing the effectiveness of in-context
 420 learning. Moreover, omitting abstraction results in a more significant loss in accuracy compared to
 421 omitting action context. This may be attributed to the system’s challenges in retrieving the most
 422 relevant examples from the exemplar bank and learning from them without abstraction.

423 424 4 RELATED WORK
425

426 **Web automation tasks.** Web automation is a sequential decision process, where agents are given
 427 a task that can be carried out on a series of websites, such as buying products, booking a ticket,
 428 and summarizing opinions on forums. The earliest research in this area often formulated the task
 429 as a Markov Decision Process and employed reinforcement learning methods Liu et al. (2018); Gur
 430 et al. (2018); Jia et al. (2018); Humphreys et al. (2022). However, with the increase in the scale of
 431 websites and tasks, these methods suffer for they are hard to comprehend and process the long web
 documents and in lack of generalization to different tasks (Gur et al., 2023).

Large language model for web automation. Language models have been used as text decoders since the era reinforcement learning is the prevalent solution to web automation tasks (Humphreys et al., 2022). As the ability to reasoning and acting of language models soared, using language models to make decisions in a web environment shows better performance, particularly on the real-world scale in recent years. Most of these methods use in-context learning paradigm (Kim et al., 2024; Zheng et al., 2023) or deal with other difficulties like long document (Zheng et al., 2023; Deng et al., 2024). However, applying this paradigm to multi-step reasoning tasks often leads to suboptimal performance. Other approaches decompose these tasks into predefined subtasks (Sodhi et al., 2024) or sequences of actions (Kim et al., 2024), which require manual design and may introduce inaccuracies. There also be works using finetuning to improve language models’ ability of reasoning or understanding of web pages (Gur et al., 2023; Lai et al., 2024), which is orthogonal to our method.

Large language model agent workflow. LLM agents are artificial entities that can sense their environment, make decisions, and take actions. They leverage the capabilities of large language models (LLMs) to perform tasks autonomously (Xi et al., 2023). The general conceptual framework for LLM-based agents includes three main components: brain (primarily composed of an LLM), perception (multimodal inputs), and action (expanding the action space of LLM-based agents). There are many development in the research of how to construct LLM agent to interact with wild outer world, such as (Li et al., 2023; Wang et al., 2023a;b) or construct LLM agent to follow human’s instructions to complete a given task with external tools, such as (Lightman et al., 2023; Hong et al., 2023; Lai et al., 2024; Significant Gravitas; Agashe et al., 2024). Web automation is a task that necessitates LLM the ability to interact with the web environment with the target given by human.

Planning with large language model. How to augment language models with reasoning skills and ability to act is of great significance when using language models in real world. Recent researches find that providing a series of intermediate reasoning steps can help LLM perform complex reasoning (Wei et al., 2022). Previous works augment the reasoning steps with the observations of the external tools (Yao et al., 2023; 2024). These prompt strategies try to decomposing the task into a sequence of actions. However, when the tasks become complex and need more steps to solve, it will be hard for LLM to generate correct action each step and the reasoning steps are hard to learn through in-context learning. Some other methods try to decompose the tasks into predefined subtasks, which can be guided to finish through in-context learning (Khot et al., 2022; Zhou et al., 2022; Sodhi et al., 2024). These methods need to manually define and write prompts for subtasks, and when the tasks become more and more various, it will be time-consuming and costly. Researches in other areas demonstrated that decomposing the task into the subtasks and using the verification module to ensure the effectiveness of the subtasks perform well in complex reasoning tasks (Wang et al., 2023b; Gou et al., 2023; Zhou et al., 2023a). Due to the strong interactivity and complexity of web automation tasks, simply transplanting these solutions is not applicable.

5 LIMITATIONS AND DISCUSSION

VeSX is an innovative LLM workflow designed for handling web automation tasks, characterized by subgoal-guided verification, hierarchical self-correction, and an exemplar bank. During our research process, we identified the following limitations of VeSX. First, its performance depends on the accuracy of verification and the success rate of reflection and replanning corrections. Although VeSX has achieved significant improvements by relying on these modules, there is still considerable room for enhancement to fully solve web automation tasks. Second, data in the web automation field is relatively scarce. While VeSX has implemented sophisticated designs for the exemplar bank to increase data utilization, generating high-quality data automatically without human annotation remains a significant challenge.

Additionally, the workflow concepts in VeSX may lead to opportunities in other fields. The novel approaches to verification and self-correction could be beneficial for tasks that require long reasoning steps and facilitate the verification of intermediate results. For instance, this could apply to complex mathematical logic reasoning or coding tasks, as well as long-term tasks in simulation environments like Minecraft. VeSX also represents a technological solution for building LLM agents with error-correction capabilities. Our future work will first focus on making new improvements

486 and contributions in data generation for web automation tasks, and then explore how the VeSX
 487 framework can be implemented and utilized in other tasks.
 488

489 6 CONCLUSION

490 In this paper, we introduce VeSX, a framework of LLM agent for web automation tasks, featured
 491 with subgoal-guided verification, hierarchical self-correction, and exemplar bank to deal with the
 492 problem about infeasible subtasks and lack of data. Our experiments on the five different scenarios
 493 show VeSX have achieved state-of-the-art success rate and significantly outperform existing meth-
 494 ods without human guidance. Our work shows the great potential of the workflow paradigm that
 495 armed with the planning-execution workflow with verification and self-correction. Moreover, VeSX
 496 proposes a new method to generate exemplars' descriptive metadata and constructs a memory sys-
 497 tem named exemplar bank for in-context learning in planning and execution phase. The ablation
 498 studies on the three contributions validate their effectiveness.

501 REFERENCES

- 502 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
 503 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
 504 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 505 Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent s: An
 506 open agentic framework that uses computers like a human. *arXiv preprint arXiv:2410.08164*,
 507 2024.
- 508 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
 509 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
 510 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 511 Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su.
 512 Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing
 Systems*, 36, 2024.
- 513 Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom
 514 Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are
 515 web agents at solving common knowledge work tasks? In *Forty-first International Conference
 516 on Machine Learning*, 2024.
- 517 Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen,
 518 et al. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint
 519 arXiv:2309.17452*, 2023.
- 520 Izzeddin Gur, Ulrich Rueckert, Aleksandra Faust, and Dilek Hakkani-Tur. Learning to navigate the
 521 web. In *International Conference on Learning Representations*, 2018.
- 522 Izzeddin Gur, Hiroki Furuta, Austin V Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and
 523 Aleksandra Faust. A real-world webagent with planning, long context understanding, and pro-
 524 gram synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.
- 525 Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang,
 526 Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-
 527 agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- 528 Peter C Humphreys, David Raposo, Tobias Pohlen, Gregory Thornton, Rachita Chhaparia, Alistair
 529 Muldal, Josh Abramson, Petko Georgiev, Adam Santoro, and Timothy Lillicrap. A data-driven
 530 approach for learning to control computers. In *International Conference on Machine Learning*,
 531 pp. 9466–9482. PMLR, 2022.
- 532 Sheng Jia, Jamie Ryan Kiros, and Jimmy Ba. Dom-q-net: Grounded rl on structured language. In
 533 *International Conference on Learning Representations*, 2018.

- 540 Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish
 541 Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The*
 542 *Eleventh International Conference on Learning Representations*, 2022.
- 543
- 544 Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks.
 545 *Advances in Neural Information Processing Systems*, 36, 2024.
- 546 Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language
 547 model agents. *arXiv preprint arXiv:2407.01476*, 2024.
- 548
- 549 Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen
 550 Zhang, Xiaohan Zhang, Yuxiao Dong, et al. Autowebglm: Bootstrap and reinforce a large lan-
 551 guage model-based web navigating agent. *arXiv preprint arXiv:2404.03648*, 2024.
- 552 Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem.
 553 Camel: Communicative agents for "mind" exploration of large language model society, 2023.
 554 URL <https://arxiv.org/abs/2303.17760>.
- 555
- 556 Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan
 557 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint*
 558 *arXiv:2305.20050*, 2023.
- 559
- 560 Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement
 561 learning on web interfaces using workflow-guided exploration. In *International Conference on*
 562 *Learning Representations*, 2018.
- 563
- 564 Significant Gravitas. AutoGPT. URL <https://github.com/Significant-Gravitas/AutoGPT>.
- 565
- 566 Paloma Sodhi, S. R. K. Branavan, Yoav Artzi, and Ryan McDonald. Step: Stacked llm policies for
 567 web actions, 2024.
- 568 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-
 569 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda-
 570 tion and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- 571
- 572 Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan,
 573 and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models.
 574 *arXiv preprint arXiv:2305.16291*, 2023a.
- 575
- 576 Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai
 577 Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents.
Frontiers of Computer Science, 18(6):186345, 2024.
- 578
- 579 Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe,
 580 explain, plan and select: Interactive planning with large language models enables open-world
 581 multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023b.
- 582
- 583 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
 584 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
585 neural information processing systems, 35:24824–24837, 2022.
- 586
- 587 Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Jun-
 588 zhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao
 589 Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou,
 590 Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuan-
 591 jing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey,
 592 2023. URL <https://arxiv.org/abs/2309.07864>.
- 593
- 594 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.
 595 React: Synergizing reasoning and acting in language models. In *International Conference on*
596 Learning Representations (ICLR), 2023.

594 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik
595 Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Ad-*
596 *vances in Neural Information Processing Systems*, 36, 2024.

597 Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar
598 prompting with memory for computer control. In *The Twelfth International Conference on Learn-*
599 *ing Representations*, 2023.

600 Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia,
601 Linqi Song, Mingjie Zhan, et al. Solving challenging math word problems using gpt-4 code
602 interpreter with code-based self-verification. In *The Twelfth International Conference on Learning*
603 *Representations*, 2023a.

604 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuur-
605 mans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. Least-to-most prompting enables complex
606 reasoning in large language models. In *The Eleventh International Conference on Learning Rep-*
607 *resentations*, 2022.

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

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648 A VeSX

649

650 A.1 WORKFLOW

651

652 A.1.1 FLOWCHART

653

654 To help readers better understand VeSX, we have provided a simplified algorithm flowchart of the
 655 VeSX workflow. We have emphasized only the most critical modules of the algorithm. Some vari-
 656 ables and dependencies are not depicted in the flowchart, as they would make the overall presentation
 657 overly lengthy, such as the transmission of error messages after verification, and the control for the
 658 environment by reflection and replanning, among others. The construction of the Exemplar Bank is
 659 also not represented in this inference workflow.

660

661 **Algorithm 1:** Simplified VeSX Workflow

662 **Input:** Web Automation Task Query Q , initial state s_0 , LLM agent for planning
 663 $LLM_{PLANNING}$, LLM agent for action prediction LLM_{ACTION} , LLM agent for reflection
 664 $LLM_{REFLECTION}$, LLM agent for replanning $LLM_{REPLANNING}$, verification module
 665 $f_{VERIFICATION}$, exemplar-bank retrieval module R

666 **Output:** Answer for the query A

667 1 $\mathcal{S} := [(t_1, g_1), (t_2, g_2), \dots, (t_n, g_n)] \leftarrow f_{PLANNING}(Q, s_0)$ \triangleright Initial Planning

668 2 $s \leftarrow s_0, U \leftarrow []$

669 3 **foreach** (t^i, g^i) in \mathcal{S} **do**

670 4 $C \leftarrow [], m \leftarrow False, v \leftarrow False$

671 5 **while** not v **do**

672 6 $r \leftarrow R(t, c), a \leftarrow LLM_{ACTION}(t, s, r)$ \triangleright Retrieval and Action Prediction

673 7 $c \leftarrow A + a$ \triangleright Action Chain (“+” indicates concatenate)

674 8 **if** $a = FINISH_SUBTASK$ **then**

675 9 $v \leftarrow f_{VERIFICATION}(g, s, c)$ \triangleright Verification

676 10 **if** not v **then**

677 11 **if** not m **then**

678 12 $c, s \leftarrow LLM_{REFLECTION}(t, c, s)$ \triangleright Reflection

679 13 $m \leftarrow True$

680 14 **end**

681 15 **else**

682 16 $[(t^{i+1}, g^{i+1}), \dots] \leftarrow LLM_{REPLANNING}(\mathcal{S})$ \triangleright Replanning

683 17 $\mathcal{S}.update([(t^{i+1}, g^{i+1}), \dots])$

684 18 **break**

685 19 **end**

686 20 **end**

687 21 **end**

688 22 **else**

689 23 $| s \leftarrow ENV(s, a)$

690 24 **end**

691 25 **end**

692 26 $u \leftarrow LLM_{SUMMARY}(t, c, s)$ \triangleright Subtask Summary

693 27 $U \leftarrow U + u$

694 28 **end** \triangleright Get Final Answer

695 29 $A \leftarrow LLM_{ANSWER}(U, s)$

696 30 **return** A

697

698 A.1.2 ONE EXAMPLE

699

700

701

702 Here is a complete execution example of the workflow to clearly illustrate the VeSX process. The
 703 section for input prompts is presented in a simplified format to save space and enhance readability.
 704 For detailed information on the inputs and outputs of each module, please refer to Appendix A.2.

702

One Complete Example

703

Query Intent:

704

What is the price range of Canon photo printer in the One Stop Market?

705

Ground Truth

706

Final Answer: 'must_include': ['2.56', '649.99']

707

708

Planning

709

1. Navigate to the "Search" function on the One Stop Market website.

710

- Objective: check_in_webpage("You have located the 'Search' function")

711

2. Enter "Canon photo printer" into the search bar and initiate the search.

712

- Objective: check_in_url("search") | OR|check_in_webpage("Search results for 'Canon photo printer' displayed")

713

3. Filter the search results to display only products related to "Canon photo printer".

714

- Objective: check_in_webpage("Search results filtered to display only 'Canon photo printer' products")

715

4. Identify the prices of all products listed under the "Canon photo printer" category.

716

- Objective: check_in_webpage("Prices of all 'Canon photo printer' products identified")

717

5. Determine and display the price range (minimum and maximum prices) for the "Canon photo printer" products.

718

- Objective: check_in_history("Price range for 'Canon photo printer' products determined and displayed")

719

720

721

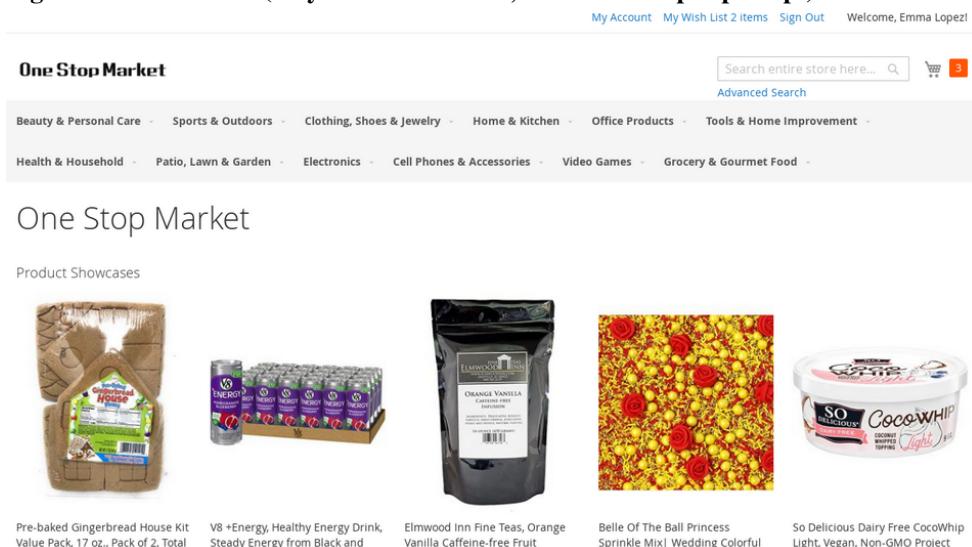
722

723

Subtask 1

1. Navigate to the "Search" function on the One Stop Market website.

Figure of Observation (only for visualization, not used in input prompt):



Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows the One Stop Market homepage with a search combobox and a disabled search button. The first step is to type "white desk" into the search combobox to enable the search button.

Next Action: According to the thoughts, the next action I will perform is "type [1697] [white desk]" Target: After typing "white desk" into the search combobox, the search button should be enabled. Summary: The current observation shows the One Stop Market homepage with a search combobox and a disabled search button. The next step is to type "white desk" into the search combobox to enable the search button. The expected outcome is to enable the search button.

755

756

Action: type [1697] [white desk] where [1697] is combobox 'ue615 Search' autocomplete: both hasPopup: listbox required: False expanded: False

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

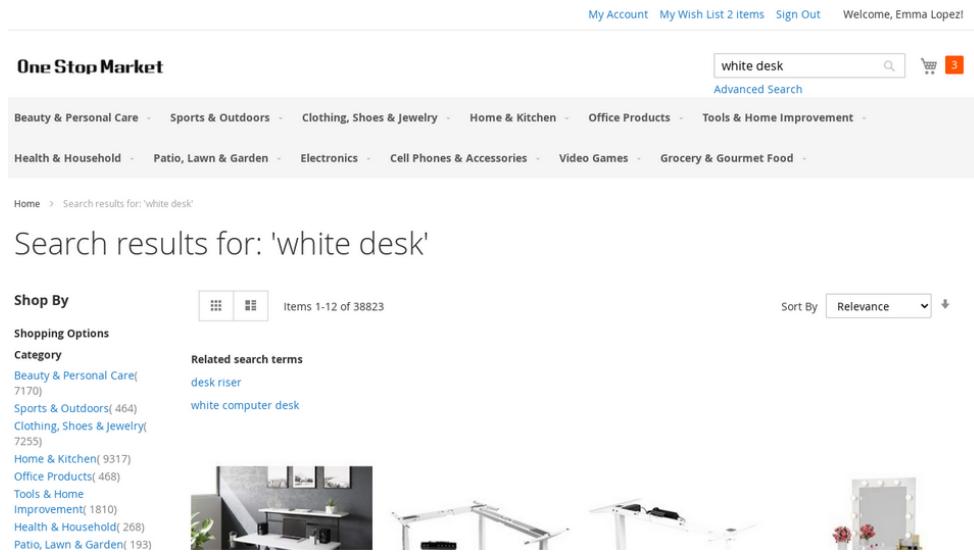
776

777

778

779

Figure of Observation (only for visualization, not used in input prompt):



Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows the search results page for "white desk" on the One Stop Market website. This indicates that the search function has already been navigated to and the search has been performed. Therefore, the subtask to navigate to the "Search" function on the One Stop Market website has been completed.

Next Action: According to the thoughts, the next action I will perform is "finish_subtask".

Target: Confirm that the search function has been navigated to and the search results for "white desk" are displayed.

Summary: The current observation shows the search results page for "white desk" on the One Stop Market website. This indicates that the search function has already been navigated to and the search has been performed. Therefore, the subtask to navigate to the "Search" function on the One Stop Market website has been completed.

Action: finish_subtask []

Verification:

check_in_webpage("You have located the 'Search' function")

1.0 To verify if the 'Search' function has been located, I need to check if there is an element related to searching on the webpage. This could include a search bar, search button, or any other search-related elements.

Verification Pass

Subtask 2

Enter "Canon photo printer" into the search bar and initiate the search.

Figure of Observation (only for visualization, not used in input prompt):

800

801

802

803

804

805

806

807

808

809

810
 811
 812
 813
 814
 815
 816
 817
 818
 819
 820
 821
 822
 823
 824
 825
 826
 827
 828
 829
 830
 831
 832
 833
 834
 835
 836
 837
 838
 839
 840
 841
 842
 843
 844
 845
 846
 847
 848
 849
 850
 851
 852
 853
 854
 855
 856
 857
 858
 859
 860
 861
 862
 863

My Account My Wish List 2 items Sign Out Welcome, Emma Lopez!

One Stop Market

Canon photo printer 3
[Advanced Search](#)

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -
 Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

Home > Search results for: 'Canon photo printer'

Search results for: 'Canon photo printer'

Shop By Items 1-12 of 6444 Sort By Relevance

Shopping Options

Category

- [Beauty & Personal Care\(457\)](#)
- [Sports & Outdoors\(90\)](#)
- [Clothing, Shoes & Jewelry\(543\)](#)
- [Home & Kitchen\(1796\)](#)
- [Office Products\(485\)](#)
- [Tools & Home Improvement\(111\)](#)
- [Health & Household\(10\)](#)
- [Patio, Lawn & Garden\(12\)](#)






Analyze and Predict Action:
 Thought: Let's think step-by-step. The current observation shows the search results page for "white desk" on the One Stop Market website. The search combobox is present, and the search button is currently disabled. The first step is to type "Canon photo printer" into the search combobox to enable the search button.
 Next Action: According to the thoughts, the next action I will perform is "type [4312] [Canon photo printer] [clear_previous_content=1] [press_enter_after=1]"
 Target: After typing "Canon photo printer" into the search combobox and pressing enter, the search results for "Canon photo printer" should be displayed.
 Summary: The current observation shows the search results page for "white desk" on the One Stop Market website. The next step is to type "Canon photo printer" into the search combobox and press enter to initiate the search. The expected outcome is to display the search results for "Canon photo printer".
Action: type [4312] [Canon photo printer] where [4312] is combobox 'ue615 Search' autocomplete: both hasPopup: listbox required: False expanded: False

Figure of Observation (only for visualization, not used in input prompt):

My Account My Wish List 2 items Sign Out Welcome, Emma Lopez!

One Stop Market

Canon photo printer 3
[Advanced Search](#)

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -
 Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

Home > Search results for: 'Canon photo printer'

Search results for: 'Canon photo printer'

Shop By Items 1-12 of 6444 Sort By Relevance

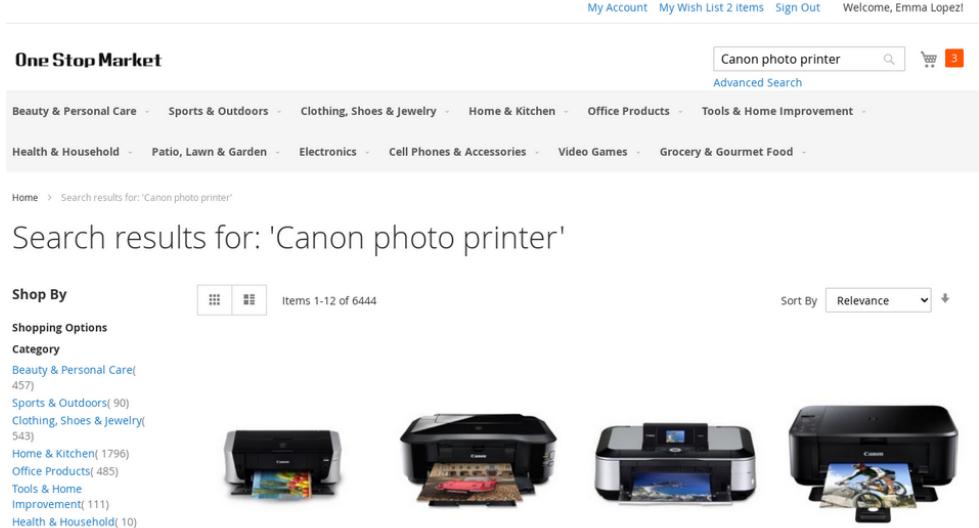
Shopping Options

Category

- [Beauty & Personal Care\(457\)](#)
- [Sports & Outdoors\(90\)](#)
- [Clothing, Shoes & Jewelry\(543\)](#)
- [Home & Kitchen\(1796\)](#)
- [Office Products\(485\)](#)
- [Tools & Home Improvement\(111\)](#)
- [Health & Household\(10\)](#)
- [Patio, Lawn & Garden\(12\)](#)






864
 865 **Analyze and Predict Action:**
 866 Thought: Let's think step-by-step. The current observation shows the search results for
 867 "Canon photo printer" on the One Stop Market website. This indicates that the search func-
 868 tion has already been navigated to and the search has been performed. Therefore, the subtask
 869 to enter "Canon photo printer" into the search bar and initiate the search has been completed.
 870 Next Action: According to the thoughts, the next action I will perform is "finish_subtask".
 871 Target: Confirm that the search results for "Canon photo printer" are displayed.
 872 Summary: The current observation shows the search results for "Canon photo printer" on the
 873 One Stop Market website. This indicates that the search function has already been navigated
 874 to and the search has been performed. Therefore, the subtask to enter "Canon photo printer"
 875 into the search bar and initiate the search has been completed.
 876 **Action: finish_subtask []**
 877
 878 **Verification:**
 879 Objective: check_in_url("search") |OR|check_in_webpage("Search results for 'Canon photo
 880 printer' displayed")
 881 **check_in_url("search")**
 882 1.0 URL: http://127.0.0.1:7770/catalogsearch/result/?q=Canon+photo+printer, "search" is
 883 in URL
 884 **check_in_webpage("Search results for 'Canon photo printer' displayed")**
 885 1.0 The content of the webpage includes multiple instances of the phrase "Search results for:
 886 'Canon photo printer'" and related elements such as headings and static text that confirm the
 887 search results are displayed
 888 **Verification Pass**
 889
 890 **Subtask 3**
 891 Filter the search results to display only products related to "Canon photo printer".
 892
 893 **Figure of Observation (only for visualization, not used in input prompt):**
 894
 895
 896
 897
 898
 899
 900 
 901
 902 **Analyze and Predict Action:**
 903 Thought: Let's think step-by-step. The current observation shows that I am on the search
 904 results page for the query "Canon photo printer." To filter the search results to display only
 905 products related to "Canon photo printer," I need to ensure that the search results are already
 906 filtered correctly. The page title and the search results indicate that the products displayed
 907 are related to "Canon photo printer."
 908 Next Action: According to the thoughts, the next action I will perform is "finish_subtask".
 909
 910
 911
 912
 913
 914
 915
 916
 917

918
919
920
921
922
923
924
925

Target: Confirm that the search results are already filtered to display only products related to "Canon photo printer." Summary: The current observation shows that the search results page is already displaying products related to "Canon photo printer." Therefore, the subtask has been completed.

Action: finish_subtask []

Verification:

check_in_webpage("Search results filtered to display only 'Canon photo printer' products")

0.0 I need to verify if the search results are specifically filtered to display only 'Canon photo printer' products. This can be determined by checking the search results and ensuring that they are relevant to 'Canon photo printer'.

Verification Fail

Reflection

Reflection: The execution process correctly identified that the search results page is displaying products related to "Canon photo printer." However, the verification result indicates that there needs to be a check to ensure that the search results are specifically filtered to display only "Canon photo printer" products.

The mistake lies in not explicitly verifying the relevance of each product in the search results to "Canon photo printer." To solve the task, I should include a step to verify that all the products listed in the search results are indeed related to "Canon photo printer." This can be done by checking the titles and descriptions of the products to ensure they match the search criteria.

Since the mistake occurred after identifying the search results, I should reload from the point just before finishing the subtask to include the verification step.

Mistakes happened during verification. There is no need to retry.

Verification (changed to) Pass

Subtask 4

Identify the prices of all products listed under the "Canon photo printer" category.

Figure of Observation (only for visualization, not used in input prompt):

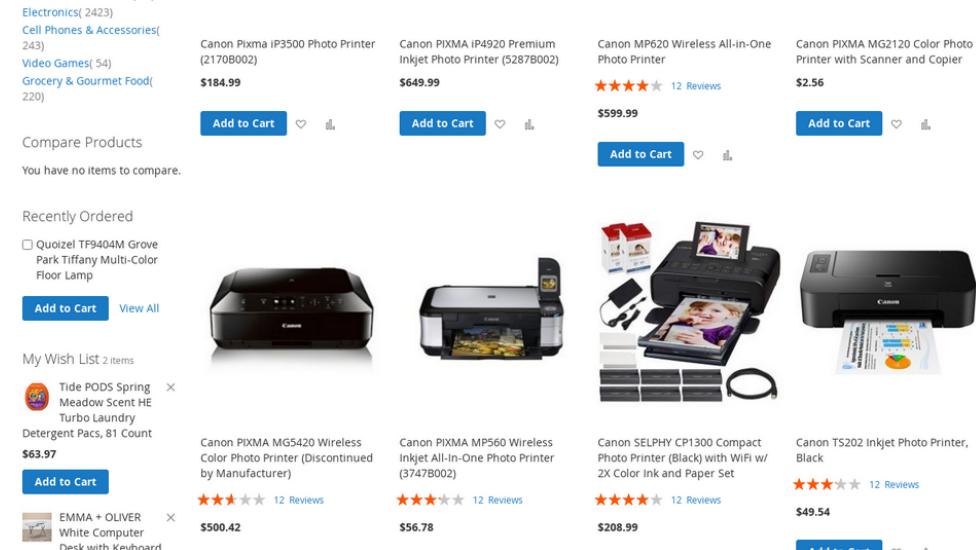
Analyze and Predict Action:

Thought: Let's think step-by-step. The current subtask is to identify the prices of all products listed under the "Canon photo printer" category. The current observation shows the search

971

972
 973 results for "Canon photo printer" on the One Stop Market website. I need to locate the prices
 974 of the products listed under this category.
 975 Next Action: According to the thoughts, the next action I will perform is "scroll [direc-
 976 tion=down]"
 977 Target: Scroll down the page to view more products and their prices.
 978 Summary: I will scroll down the page to view more products listed under the "Canon photo
 979 printer" category and their prices.
 980 **Action: scroll [down]**
 981

Figure of Observation (only for visualization, not used in input prompt):



Analyze and Predict Action:

Thought: Let's think step-by-step. I have identified several products under the "Canon photo printer" category along with their prices. Here are the products and their prices: 1. Canon Pixma iP3500 Photo Printer - \$184.99 2. Canon PIXMA iP4920 Premium Inkjet Photo Printer - \$649.99 3. Canon MP620 Wireless All-in-One Photo Printer - \$599.99 4. Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier - \$2.56 5. Canon PIXMA MG5420 Wireless Color Photo Printer - \$500.42 6. Canon PIXMA MP560 Wireless Inkjet All-In-One Photo Printer - \$56.78 7. Canon SELPHY CP1300 Compact Photo Printer - \$208.99 8. Canon TS202 Inkjet Photo Printer - \$49.54

I have identified the prices of all products listed under the "Canon photo printer" category. Next Action: According to the thoughts, the next action I will perform is "finish_subtask [I have identified the prices of all products listed under the "Canon photo printer" category: Canon Pixma iP3500 Photo Printer - \$184.99, Canon PIXMA iP4920 Premium Inkjet Photo Printer - \$649.99, Canon MP620 Wireless All-in-One Photo Printer - \$599.99, Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier - \$2.56, Canon PIXMA MG5420 Wireless Color Photo Printer - \$500.42, Canon PIXMA MP560 Wireless Inkjet All-In-One Photo Printer - \$56.78, Canon SELPHY CP1300 Compact Photo Printer - \$208.99, Canon TS202 Inkjet Photo Printer - \$49.54]"

Target: I expect to complete the subtask of identifying the prices of all products

Action: finish_subtask [I have identified the prices of all products listed under the "Canon photo printer" category: Canon Pixma iP3500 Photo Printer - \$184.99, Canon PIXMA iP4920 Premium Inkjet Photo Printer - \$649.99, Canon MP620 Wireless All-in-One Photo Printer - \$599.99, Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier - \$2.56, Canon PIXMA MG5420 Wireless Color Photo Printer - \$500.42, Canon PIXMA MP560 Wireless Inkjet All-In-One Photo Printer - \$56.78, Canon SELPHY CP1300 Compact Photo Printer - \$208.99, Canon TS202 Inkjet Photo Printer - \$49.54]

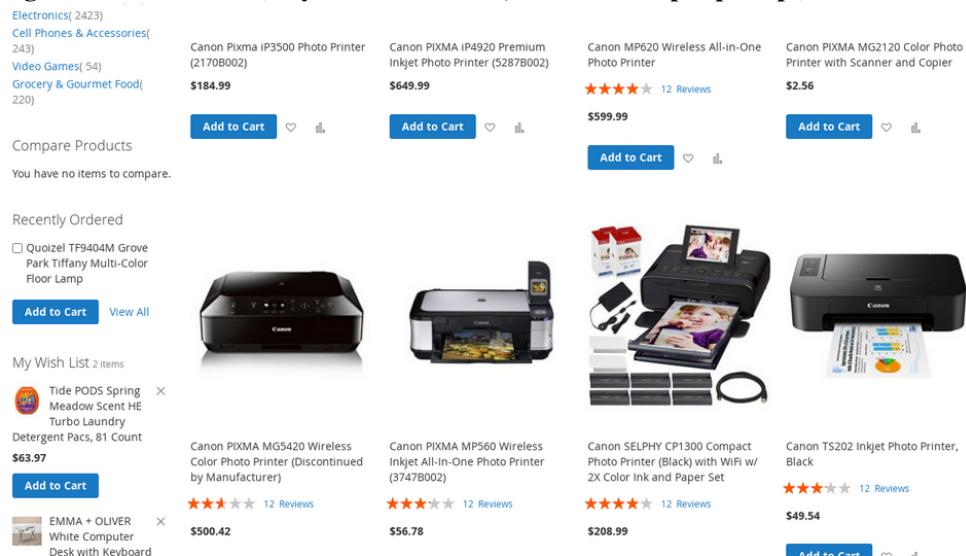
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

Verification:**check_in_webpage("Prices of all 'Canon photo printer' products identified")**

1.0 Prices of all 'Canon photo printer' products have been listed and identified.

Verification Pass**Subtask 5**

Determine and display the price range (minimum and maximum prices) for the "Canon photo printer" products.

Figure of Observation (only for visualization, not used in input prompt):**Analyze and Predict Action:**

Thought: Let's think step-by-step. The current subtask is to determine and display the price range (minimum and maximum prices) for the "Canon photo printer" products. The current observation shows multiple Canon photo printers with their respective prices. I need to identify the minimum and maximum prices from the listed products.

Here are the prices of the listed Canon photo printers: - \$184.99 - \$649.99 - \$599.99 - \$2.56 - \$500.42 - \$56.78 - \$208.99 - \$49.54

The minimum price is \$2.56, and the maximum price is \$649.99.

Next Action: According to the thoughts, the next action I will perform is "finish_subtask [2.56, 649.99]"

Target: Confirm the price range (minimum and maximum prices) for the "Canon photo printer" products.

Summary: The prices of the listed Canon photo printers have been identified, with the minimum price being \$2.56 and the maximum price being \$649.99. Therefore, the subtask has been completed.

Action: **finish_subtask [2.56, 649.99]**

Verification:**check_in_history("Price range for 'Canon photo printer' products determined and displayed")**

1.0 The subgoal target is to determine and display the price range for 'Canon photo printer' products. The summary indicates that the prices have been identified, with a minimum price of \$2.56 and a maximum price of \$649.99. This information aligns with the subgoal target.

Verification Pass

1080
 1081 **Answer:**
 1082 The price range of Canon photo printers in the One Stop Market is from \$2.56 to \$649.99.
 1083
 1084

1085 **A.2 DETAILS OF EACH MODULE**

1086
 1087 We provided a smooth example of the VeSX workflow for handling a web automation task to help
 1088 readers understand it better. The query for this task is ****. From the workflow below, we can see
 1089 how VeSX breaks down the task into individual subtasks, addresses each one separately, and verifies
 1090 the results.

1091
 1092 **A.2.1 EXAMPLE OF PLANNING PHASE**

1093
 1094 **Example of Planning Phase**

1095 **Query Intent**

1096 List out reviewers, if exist, who mention about complain of the customer service.

1097
 1098 **System Prompt:**

1099 You are an autonomous intelligent agent tasked with navigating a web browser. Your responsi-
 1100 bility entails performing specific web-based tasks, beginning from a defined starting point
 1101 on a particular webpage. To ensure the effectiveness and tractability of your operations, you
 1102 are to break down each given task into manageable and coherent sub-tasks.

1103 Each subtask you delineate should be precise, methodical, and feasible, though not necessarily
 1104 to the extent of a single action or click. You should not use one action like click something
 1105 as your subtask, but use navigate to some website or locate some button or function. Aim to
 1106 limit the total number of sub-tasks to a maximum of five.

1107 Additionally, for each sub-task stated, be sure to set clear and verifiable objectives. These
 1108 objectives will act as benchmarks for assessing the successful completion of each step and
 1109 written after each sub-task. The objective should be a form of function call, where function
 1110 should be one of the following three:

1111 1. `check_in_url(part of url)`: You think you will check in the url. This often be used when
 1112 you want to navigate to some website. Its parameter is the fragment you expect to be the
 1113 part of the url.

1114 2. `check_in_webpage(objective)`: You think you will check the objective in the last webpage.
 1115 This often be used when the last webpage has the information for your sub-task. Its param-
 1116 eter is the objective you want to achieve in the sub-task and can be check in the webpage.

1117 3. `check_in_history(objective)`: You think you should check in your action history. This
 1118 often be used when your sub-task is to get some information. Its parameter is the objective
 1119 you want to achieve in the sub-task and can be check in the history.

1120 You can use ‘|OR|’ to connect two objective if you want to use multiple methods to assess the
 1121 result of subtasks. You should add objectives after each sub-task by starting with ‘Objective’.
 1122 In situations where authentication is necessary, account credentials will be provided to you
 1123 subsequently.

1124 Please proceed to methodically dissect the following task into its constituent sub-tasks, and
 1125 articulate verifiable objectives for each to ensure their effective execution.

1126 **Example User:**

1127 Start webpage: Current Observation: The webpage does not provide any details about re-
 1128 viewers mentioning the price being unfair. There is no information related to reviewers or
 1129 their opinions on prices in the provided representation.

1130 URL: <http://onestopmarket.com/haflinger-men-s-wool-felt-open-back-slippers-beige-550-peat-us-7.html>

1131 Task: List out reviewers, if exist, who mention about price being unfair

1132 Now, try to break this task down into several sub-tasks and generate related verifiable
 1133 objectives:

1134
 1135
 1136 **Example Assistant:**
 1137 1. **Navigate to the Reviews Section** - Locate and click on the "2 Reviews" link to access
 1138 the reviews section.
 1139 - Objective: 'check_in_url("reviews") |OR|check_in_webpage("Reviews")'
 1140 2. **Extract All Reviews** - Identify and extract the text of all reviews present on the page.
 1141 - Objective: 'check_in_webpage("Review Text")'
 1142 3. **Search for Mentions of Price Being Unfair** - Analyze the extracted reviews to find
 1143 any mentions of the price being unfair.
 1144 - Objective: 'check_in_history("price being unfair")'
 1145 4. **List Out Reviewers** - Identify and list the reviewers who mentioned the price being
 1146 unfair.
 1147 - Objective: 'check_in_history("Reviewers mentioning price being unfair")'
 1148 5. **Compile and Present the Information** - Compile the information into a coherent list
 1149 and present it.
 1150 - Objective: 'check_in_history("Compiled list of reviewers")'

1151 **Example User:** Start webpage: Current Observation: The webpage mentions "Current Ob-
 1152 servation: One Stop" but does not provide a customer service number in the given text span.
 1153 URL: <http://onestopmarket.com/>
 1154 Task: Which number to call for the customer service? Now, try to break this task down into
 1155 several sub-tasks and generate related verifiable objectives:

1156
 1157
 1158 **Example Assistant:**
 1159 1. Navigate to the "Contact Us" or "Customer Service" section of the One Stop Market
 1160 website.
 1161 Objective: check_in_url("contact") |OR|check_in_webpage("Customer Service")
 1162 2. Locate the customer service phone number on the "Contact Us" or "Customer Service"
 1163 page.
 1164 Objective: check_in_webpage("Customer service phone number located")
 1165 3. Verify the customer service phone number by cross-referencing it with any additional
 1166 information provided on the page (e.g., business hours, email support).
 1167 Objective: check_in_webpage("Customer service phone number verified with additional in-
 1168 formation")

1169 **Example User:** Start webpage: Current Observation: The actree representation includes
 1170 customer reviews for the Amazon Echo Dot 3rd generation. Key points from the reviews
 1171 highlight positive aspects such as improved sound quality, ease of setup, and useful integra-
 1172 tion with other smart devices. Customers appreciate the compact design and functionality,
 1173 particularly the enhancement over previous versions. Some reviews also mention the de-
 1174 vice's good value for the price. Criticisms in the reviews include occasional issues with
 1175 voice recognition and connectivity problems. Overall, the reviews provide a generally fa-
 1176 vorable outlook on the Amazon Echo Dot 3rd generation.
 1177 URL: <http://onestopmarket.com/>
 1178 Task: Summarize customer reviews for Amazon Echo Dot 3rd generation. Now, try to
 1179 break this task down into several sub-tasks and generate related verifiable objectives:
 1180
 1181 **Example Assistant:**
 1182 1. Navigate to the product page for the Amazon Echo Dot 3rd generation on the
 1183 One Stop Market website. Objective: check_in_url("amazon-echo-dot-3rd-generation")
 1184 |OR|check_in_webpage("Amazon Echo Dot 3rd generation product page")
 1185 2. Locate the section containing customer reviews for the Amazon Echo Dot 3rd generation.
 1186 Objective: check_in_webpage("Customer reviews for Amazon Echo Dot 3rd generation lo-
 1187 cated")

1188
 1189 3. Read through the customer reviews to identify key positive aspects mentioned. Objective:
 1190 check_in_webpage("Positive aspects of Amazon Echo Dot reviews summarized")
 1191 4. Read through the customer reviews to identify key criticisms mentioned. Objective:
 1192 check_in_webpage("Criticisms of Amazon Echo Dot reviews summarized")
 1193 5. Compile the summarized information into a coherent overview of customer reviews for
 1194 the Amazon Echo Dot 3rd generation. Objective: check_in_webpage("Customer reviews
 1195 summary for Amazon Echo Dot 3rd generation compiled")
 1196
User:
 1197 Start webpage: Current Observation: Tab 0 (current): Epson WorkForce WF-3620 WiFi Di-
 1198 rect All-in-One Color Inkjet Printer, Copier, Scanner, Amazon Dash Replenishment Ready
 1199 [1] RootWebArea 'Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet Printer,
 1200 Copier, Scanner, Amazon Dash Replenishment Ready' focused: True
 1201 [1428] link 'My Account'
 1202 [1424] link 'My Wish List 1 item'
 1203 [1430] link 'Sign Out'
 1204 [1433] StaticText 'Welcome, Emma Lopez!'
 1205 [1330] link 'Skip to Content'
 1206 [1339] link 'store logo'
 1207 [1436] img 'one_stop_market_logo'
 1208 [1437] link 'ue611 My Cart 4 4items'
 1209 [2496] StaticText 'Search'
 1210 [1756] combobox 'ue615 Search' autocomplete: both hasPopup: listbox required: False
 expanded: False
 1211 [2499] link 'Advanced Search'
 1212 [1759] button 'Search' disabled: True
 1213 [1135] tablist " multiselectable: False orientation: horizontal
 1214 [1137]tabpanel "
 1215 [79] menu " orientation: vertical
 1216 [830] menuitem 'ue622 Beauty & Personal Care' hasPopup: menu
 1217 [895] menuitem 'ue622 Sports & Outdoors' hasPopup: menu
 1218 [905] menuitem 'ue622 Clothing, Shoes & Jewelry' hasPopup: menu
 1219 [919] menuitem 'ue622 Home & Kitchen' hasPopup: menu
 1220 [956] menuitem 'ue622 Office Products' hasPopup: menu
 1221 [964] menuitem 'ue622 Tools & Home Improvement' hasPopup: menu
 1222 [969] menuitem 'ue622 Health & Household' hasPopup: menu
 1223 [975] menuitem 'ue622 Patio, Lawn & Garden' hasPopup: menu
 1224 [980] menuitem 'ue622 Electronics' hasPopup: menu
 1225 [1041] menuitem 'ue622 Cell Phones & Accessories' hasPopup: menu
 1226 [1056] menuitem 'ue622 Video Games' hasPopup: menu
 1227 [1069] menuitem 'ue622 Grocery Gourmet Food' hasPopup: menu
 1228 [1365] link 'Home'
 1229 [1368] StaticText 'Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet
 Printer, Copier, Scanner, Amazon Dash Replenishment Ready'
 1230 [5] main "
 1231 [1369] heading 'Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet
 Printer, Copier, Scanner, Amazon Dash Replenishment Ready'
 1232 [11] generic 'Availability'
 1233 [13] StaticText 'IN STOCK'
 1234 [1445] StaticText 'SKU'
 1235 [1715] StaticText 'B00JXLGF06'
 1236 [1376] LayoutTable "
 1237 [1717] StaticText 'Rating:'
 1238 [1448] generic '35%'
 1239 [2469] StaticText '% of'
 1240 [2472] StaticText '100'
 1241 [1449] link '12xa0 Reviews '

```

1242 [1450] link 'Add Your Review'
1243 [1452] StaticText '$607.39'
1244 [1391] LayoutTable "
1245     [1731] StaticText 'Qty'
1246     [1732] spinbutton 'Qty' required: False valuemin: 0 valuemax: 0 valuetext:
1247     [1733] button 'Add to Cart'
1248 [1393] link 'Add to Wish List'
1249 [1394] link 'Add to Compare'
1250 [1399] link 'Skip to the end of the images gallery'
1251 [1156] button 'Previous'
1252 [1158] generic 'Image'
1253     [2502] img 'Image'
1254     [1157] button 'Next'
1255 URL: http://onestopmarket.com/epson-workforce-wf-3620-wifi-direct-all-in-one-color-
1256 inkjet-printer-copier-scanner-amazon-dash-replenishment-ready.html
1257 Task: List out reviewers, if exist, who mention about complain of the customer service
1258 Now, try to break this task down into several sub-tasks and generate related verifiable objec-
1259 tives: "

```

Output:

1. **Navigate to the Reviews Section** - Locate and click on the "12 Reviews" link to access the reviews section.
- Objective: 'check_in_url("reviews")' |OR| check_in_webpage("Reviews")'
2. **Extract All Reviews** - Identify and extract the text of all reviews present on the page.
- Objective: 'check_in_webpage("Review Text")'
3. **Search for Mentions of Customer Service Complaints** - Analyze the extracted reviews to find any mentions of complaints about customer service.
- Objective: 'check_in_history("customer service complaints")'
4. **List Out Reviewers** - Identify and list the reviewers who mentioned complaints about customer service.
- Objective: 'check_in_history("Reviewers mentioning customer service complaints")'
5. **Compile and Present the Information** - Compile the information into a coherent list and present it.
- Objective: 'check_in_history("Compiled list of reviewers")'

A.2.2 EXAMPLE OF EXECUTION PHASE**Example of Execution Phase****System Prompt:**

You are an autonomous intelligent agent tasked with navigating a web browser. You will be given a web-based task. This task will be decomposed into some sub-tasks and you will achieve the final objective of this final task by solving these sub-tasks one by one. These sub-tasks will be accomplished through the use of specific actions you can issue.

Here's the information you'll have:

The user's objective: This is the task you're trying to complete.

The current web page's accessibility tree: This is a simplified representation of the webpage, providing key information.

The current web page's URL: This is the page you're currently navigating.

The open tabs: These are the tabs you have open.

The webpage-browsing and the action history: This part contains the history of webpage you browsed and action you performed on them. The contents of these webpages are refined to decrease the context length. It may be helpful to track your progress and check whether you encountered any obstacles.

The actions you can perform fall into several categories:

Page Operation Actions:

1296
 1297 ‘click [id]’: This action clicks on an element with a specific id on the webpage.
 1298 ‘type [id] [content] [clear_previous_content=0—1] [press_enter_after=0—1]’: Use this to
 1299 type the content into the field with id. By default, the web browser will clear the previous
 1300 content. If you want to input something to replace the original content, for instance when
 1301 you need to change a value, set clear_previous_content to 1; when you want to add some-
 1302 thing, set it to 0. By default, the “Enter” key is pressed after typing unless press_enter_after
 1303 is set to 0.
 1304 ‘hover [id]’: Hover over an element with id.
 1305 ‘press [key_comb]’: Simulates the pressing of a key combination on the keyboard (e.g.,
 1306 Ctrl+v).
 1307 ‘scroll [direction=down—up]’: Scroll the page up or down.
 1308 Tab Management Actions:
 1309 ‘new_tab’: Open a new, empty browser tab.
 1310 ‘tab_focus [tab_index]’: Switch the browser’s focus to a specific tab using its index.
 1311 ‘close_tab’: Close the currently active tab.
 1312 URL Navigation Actions:
 1313 ‘go_back’: Navigate to the previously viewed page.
 1314 ‘go_forward’: Navigate to the next page (if a previous go_back action was performed).
 1315 ‘reload’: Reload this web page (if you encountered an error, e.g. 500, 502, in this web page).
 1316 Completion Action:
 1317 ‘finish_subtask [answer]’: Issue this action when you believe this subtask is finished and
 1318 then you will check the correction and completion of this subtask. If the objective is to find
 1319 a text-based answer, provide the answer in the bracket.
 1320 ‘stop [answer]’: Issue this action when you believe the task is complete. If the objective is
 1321 to find a text-based answer, provide the answer in the bracket.
 1322 Homepage:
 1323 If you want to visit other websites, check out the homepage at <http://homepage.com> by using
 1324 ‘goto [http://homepage.com]’. It has a list of websites you can visit.
 1325 <http://homepage.com/password.html> lists all the account name and password for the web-
 1326 sites. You can use them to log in to the websites.
 1327 To be successful, it is very important to follow the following rules:
 1328 1. You should only issue an action that is valid given the current observation and you should
 1329 only issue one action at a time. Your action must be one of the actions listed before. When
 1330 you want to propose an action that is the same as previous steps, consider whether it is still
 1331 necessary, as it has already been attempted.
 1332 2. You should follow the examples to reason step by step and then issue the next action. PAY
 1333 ATTENTION!!! : You should first consider whether the subtask has been completed by pre-
 1334 vious actions or it is unnecessary to tackle with. If it is completed, please promptly issue
 1335 finish_subtask. When you have completed a series of actions and largely achieved the sub-
 1336 task, but feel uncertain or have doubts about the results, you can also trigger finish_subtask to
 1337 perform verification first. Especially when you want to re-execute some actions you have
 1338 done in the past, please be particularly cautious. Do not simply redo the actions because of
 1339 doubts; you need to confirm that the previous actions had no effect before proceeding.
 1340 3. Issue ‘finish_subtask’ action when you think you have accomplished the sub-task. Issue
 1341 stop action when you think you have achieved the final objective of the whole task. When
 1342 you think you have accomplished the subtask and the whole task at the same time, issue
 1343 step action to generate your final answer. Sometimes the current subtask has already been
 1344 completed in a distant previous action, not necessarily in recent actions. Feel free to issue
 1345 “finish subtask” when you deem it appropriate.
 1346 4. The history contains two parts: global history is the summaries of the previous sub-tasks
 1347 aimed to accomplish the final whole task, and local history is the summaries of the actions
 1348 you issued to accomplish this sub-task. You should pay more attention to the last step of
 1349 local history and you should check if there is a target in it.
 1350 5. You should solve the sub-task with these parts: Thought, Next Action, Target and Sum-
 1351 mary. ‘Thought’ can reason the current situation step by step and help you generate next
 1352 action. And if there is a target in the last previous local history, you should check whether
 1353 it is achieved according to the observation in this part ‘Thought’. ‘Next Action’ is the action

1350
 1351 you choose to take, which are listed before. PAY ATTENTION!!! : You should generate
 1352 the action in the correct format faithfully. Start with a "According to the thoughts, the next
 1353 action I will perform is" phrase, followed by action inside ““”. For example, "According to
 1354 the thoughts, the next action I will perform is ““click [1234]”“". The action in wrong format
 1355 will be rejected and cause the process failed, so PLEASE GENERATE IN THE CORRECT
 1356 FORMAT. 'Target' is what you expect to do or what content you expect to see after the 'Next
 1357 Action' be performed. 'Summary' is the summary of this step, and it should contains the
 1358 observation of the web page, the thoughts, the predicted action and the target you generated.
 1359 6. When you need to publish certain content, especially when posting on Reddit or creating
 1360 a forum, you must fill out all the required fields first. These fields often appear in the format
 1361 of textbox: "xxx", required: True. You need to complete all these fields before making the
 1362 submission. Relevant examples include: when creating a website, the required fields are
 1363 Name, Title, Description, and Sidebar; when posting, the required fields are Title, Body, and
 1364 Forum.
 1365 7. If you need to sign in, use these accounts:
 [omit]

1366
 1367 **Example User:**
 1368 Current Subtask:
 1369 1. **Navigate to the "My Account" section** - Click on the "My Account" link.
 1370 Recap of Recent Actions:
 1371 There is no local history. This is the first step to solve the current subtask.
 1372 Current Observation:
 1373 The webpage titled "One Stop Market" showcases an e-commerce platform with various
 1374 sections and features accessible via links and menu items. Key elements include:
 1375 - **Header Links**:
 1376 - "My Account" (Link to user account)
 1377 - "My Wish List 2 items"
 1378 - "Sign Out"
 1379 - "Welcome, Emma Lopez!" (Greeting text)
 1380 - **Navigational Links**:
 1381 - "Skip to Content"
 1382 - "store logo" (with an embedded store logo image)
 1383 - "My Cart 3 items"
 1384 - **Search Functionality**:
 1385 - "Search" field (with options for autocomplete and advanced search)
 1386 - "Search" button (currently disabled)
 1387 - **Tab List and Menu**: A horizontal tab panel with various categories such as Beauty &
 1388 Personal Care, Sports & Outdoors, and more, each with a submenu.
 1389 - **Main Content Area**:
 1390 - "One Stop Market" heading
 1391 - "Product Showcases" section featuring several product links and images.
 1392 To **navigate to the "My Account" section**, the user should click on the "My Account"
 1393 link, which is represented as:
 1394 "plaintext
 1395 [1296] link My Account
 1396 "
 1397 URL: <http://onestopmarket.com/>"
 1398

1399 **Example Assistant**
 1400 Thought: Let's think step-by-step. The current observation indicates that I am on the One
 1401 Stop Market homepage. The page contains a My Account link with ID [1296]. To navigate
 1402 to the "My Account" page, I need to click on this link.
 1403 Next Action: According to the thoughts, the next action I will perform is ““click [1296]”“
 Target: Navigate to the "My Account" page.

1404

1405 Summary: The current observation shows that I am on the One Stop Market homepage, and
 1406 there is a My Account link. I will click on this link to navigate to the "My Account" page.
 1407

1408 **Example User**

1409 Current Subtask:

- 1410 1. **Navigate to the "My Account" section**
 1411 - Click on the "My Account" link.

1412 Recap of Recent Actions:

1413 Step: 1

1414 The summary of this step is:

1415 The current observation shows that I am on the One Stop Market homepage, and there is a
 1416 My Account link. I will click on this link to navigate to the "My Account" page.
 1417

1418 You have taken action 'click [button_id] where [button_id] is link My Account' in this step.
 1419 Current Observation:

1420 The webpage appears to be structured as a user account management page titled "My Ac-
 1421 count," which is already focused. Key elements include various links and static text related
 1422 to user account information and navigation.

- 1423 1. **Navigate to the "My Account" section**

- 1424 - To navigate to the "My Account" section, you would click on the link labeled "My Ac-
 1425 count" ([4153] link My Account).

1426 Currently, the page already shows the "My Account" section with relevant account infor-
 1427 mation for the user Emma Lopez, including contact details, billing and shipping addresses,
 1428 and links to managing various aspects of the account, such as editing contact information,
 1429 changing the password, and managing addresses ([4064] group for Default Billing Address
 1430 and [4081] group for Default Shipping Address). The page navigation elements offer access
 1431 to the wish list, newsletter subscriptions, stored payment methods, and product reviews.

1432 URL: <http://onestopmarket.com/customer/account/>

1433

1434 **Example Assistant:**

1435 Thought: Let's think step-by-step. The current observation indicates that I have successfully
 1436 navigated to the "My Account" page. The page contains various elements such as account
 1437 information, contact information, address book, and links to other account-related sections.
 1438 Since the objective was to navigate to the "My Account" page, I have achieved this subtask.

1439 Next Action: According to the thoughts, the next action I will perform is "finish_subtask".

1440 Target: Confirm that the subtask of navigating to the "My Account" page is complete.

1441 Summary: The current observation shows that I am on the "My Account" page, which
 1442 contains various account-related information and links. Therefore, I will finish the subtask
 1443 as it has been successfully completed.

1444

1445 **Example User:**

1446 Current Subtask:

- 1447 1. **Navigate to the Reviews Section**
 1448 - Locate and click on the "2 Reviews" link to access the reviews section.

1449 Recap of Recent Actions:

1450 Step: 1

1451 The summary of this step is:

1452 The current observation shows that I am on the product page for "HAFLINGER Mens Wool
 1453 Felt Open Back Slippers," and there is a link labeled "2 Reviews" with ID [1357]. Therefore,
 1454 I will click on this link to navigate to the reviews section.

1455 You have taken action 'click [button_id] where [button_id] is link 2xa0 Reviews' in this step.

1456 Current Observation:

1457 Summary of Webpage for Task Requirements

1458 **Page Title:**

1459 - HAFLINGER Mens Wool Felt Open Back Slippers, Beige 550 Peat, US:7

1460 **Content Structure:**

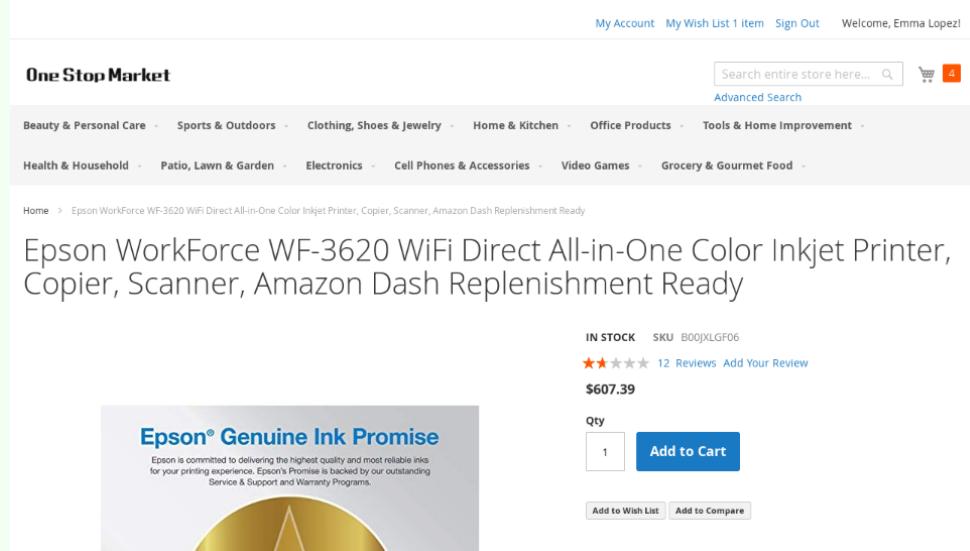
1461

1458
 1459 1. **Main Web Area**
 1460 - **Tab List Details:**
 1461 - There are two main tabs:
 1462 - **Details**
 1463 - Controlled by a link labeled Details
 1464 - Not expanded or selected
 1465 - **Reviews (2)**
 1466 - Controlled by a link labeled Reviews (2)
 1467 - Currently expanded and selected
 1468 2. **Reviews Section**
 1469 - **Customer Reviews Panel:**
 1470 - Contains StaticText elements:
 1471 - "Customer Reviews"
 1472 - "Wonderful!"
 1473 - Specific review details like "Rating: 100%", "Nothing....everyday use."
 1474 - Reviewers: Jayánd Josef Bürger
 1475 - Review dates: "4/18/23"
 1476 - Statements such as "Die Haflinger sind wie beschrieben und erwartet sehr gut"
 1477 - **Review Form:**
 1478 - Group titled: "Youfe reviewing: HAFLINGER Men's Wool Felt Open Back Slippers, Beige
 1479 550 Peat, US:7"
 1480 - Rating options (1 to 5 stars) with radio buttons
 1481 - Required textbox for Nickname labeled "Emma"
 1482 **Task Compliance:**
 1483 To navigate to the Reviews Section:
 1484 - **Locate and click the "Reviews (2)" link (Index [1374])**, which controls the Reviews
 1485 tabpanel. This will display the reviews section.
 1486 **Key Actions for Users:**
 1487 1. Identify the "Reviews (2)" tab under the tablist.
 1488 2. Ensure the "Reviews (2)" is expanded and selected to access customer reviews.
 1489 This structure ensures that the reviews, including the two customer feedback entries and
 1490 required rating fields for submitting a new review, are reached effectively.
 1491 URL: <http://onestopmarket.com/haflinger-men-s-wool-felt-open-back-slippers-beige-550-peat-us-7.html>
 1492
Example Assistant:
 1493 Thought: Let's think step-by-step. The current observation shows that I am on the product
 1494 page for "HAFLINGER Men's Wool Felt Open Back Slippers," and the "Reviews (2)" tab is
 1495 expanded and selected. The reviews section is visible, showing customer reviews and a form
 1496 for submitting a new review. This means that the "2 Reviews" link has been successfully
 1497 clicked and I have navigated to the reviews section. Therefore, I have completed the subtask.
 1498 Next Action: According to the thoughts, the next action I will perform is "finish_subtask"
 1499 Target: Confirm that the subtask of navigating to the reviews section by clicking on the "2
 1500 Reviews" link is completed.
 1501 Summary: The current observation shows that I am on the product page for "HAFLINGER
 1502 Men's Wool Felt Open Back Slippers," and the "Reviews (2)" tab is expanded and selected.
 1503 The reviews section is visible, showing customer reviews and a form for submitting a new
 1504 review. This confirms that the "2 Reviews" link has been successfully clicked and I have
 1505 navigated to the reviews section. Therefore, I will finish the subtask.
 1506
User:
 1507 Current Subtask:
 1508 1. **Navigate to the Reviews Section**
 1509 - Locate and click on the "12 Reviews" link to access the reviews section.
 1510 Recap of Recent Actions:
 1511

1512
 1513 There is no local history. This is the first step to solve the current subtask.
 1514 Current Observation:
 1515 Tab 0 (current): Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet Printer,
 1516 Copier, Scanner, Amazon Dash Replenishment Ready
 1517 [1] RootWebArea Épson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet Printer,
 1518 Copier, Scanner, Amazon Dash Replenishment Ready focused: True
 1519 [1428] link My Account
 1520 [1424] link My Wish List 1 item
 1521 [1430] link Sign Out
 1522 [1433] StaticText Welcome, Emma Lopez!
 1523 [1330] link Skip to Content
 1524 [1339] link store logo
 1525 [1436] img óne_stop_market_logo
 1526 [1437] link úe611 My Cart 4 items
 1527 [2496] StaticText Search
 1528 [1756] combobox úe615 Searchautocomplete: both hasPopup: listbox required: False
 expanded: False
 1529 [2499] link Advanced Search
 1530 [1759] button Searchdisabled: True
 1531 [1135] tablist `` multiselectable: False orientation: horizontal
 1532 [1137]tabpanel ``
 1533 [79] menu `` orientation: vertical
 1534 [830] menuitem úe622 Beauty & Personal Care hasPopup: menu
 1535 [895] menuitem úe622 Sports & Outdoors hasPopup: menu
 1536 [905] menuitem úe622 Clothing, Shoes & Jewelry hasPopup: menu
 1537 [919] menuitem úe622 Home & Kitchen hasPopup: menu
 1538 [956] menuitem úe622 Office Products hasPopup: menu
 1539 [964] menuitem úe622 Tools & Home Improvement hasPopup: menu
 1540 [969] menuitem úe622 Health & Household hasPopup: menu
 1541 [975] menuitem úe622 Patio, Lawn & Garden hasPopup: menu
 1542 [980] menuitem úe622 Electronics hasPopup: menu
 1543 [1041] menuitem úe622 Cell Phones & Accessories hasPopup: menu
 1544 [1056] menuitem úe622 Video Games hasPopup: menu
 1545 [1069] menuitem úe622 Grocery & Gourmet Food hasPopup: menu
 1546 [1365] link Home
 1547 [1368] StaticText Épson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet
 1548 Printer, Copier, Scanner, Amazon Dash Replenishment Ready
 1549 [5] main ``
 1550 [1369] heading Épson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet
 1551 Printer, Copier, Scanner, Amazon Dash Replenishment Ready
 1552 [11] generic Availability
 1553 [13] StaticText ÍN STOCK
 1554 [1445] StaticText SKU
 1555 [1715] StaticText B00JXLGF06
 1556 [1376] LayoutTable ``
 1557 [1717] StaticText Rating:
 1558 [1448] generic 35%
 1559 [2469] StaticText % of
 1560 [2472] StaticText 100
 1561 [1449] link 12
 1562 xao Reviews
 1563 [1450] link Add Your Review
 1564 [1452] StaticText \$607.39
 1565 [1391] LayoutTable ``

1566
 1567 [1731] StaticText Qty
 1568 [1732] spinbutton Qtyrequired: False valuemin: 0 valuemax: 0 valuetext:
 1569 [1733] button Add to Cart
 1570 [1393] link Add to Wish List
 1571 [1394] link Add to Compare
 1572 [1399] link Skip to the end of the images gallery' [1156] button Previous
 1573 [1158] generic Image
 1574 [2502] img Image
 1575 [1157] button Next
 1576 URL: http://onestopmarket.com/epson-workforce-wf-3620-wifi-direct-all-in-one-color-
 1577 inkjet-printer-copier-scanner-amazon-dash-replenishment-ready.html
 1578

Figure of Observation (only for visualization, not used in input prompt):



Output:

Thought: Let's think step-by-step. The current observation indicates that I am on the product page for "Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet Printer, Copier, Scanner, Amazon Dash Replenishment Ready." The page contains a link labeled "12 Reviews" with ID [1449]. To navigate to the reviews section, I need to click on this link.

Next Action: According to the thoughts, the next action I will perform is "click [1449]"

Target: Navigate to the reviews section by clicking on the "12 Reviews" link.

Summary: The current observation shows that I am on the product page for "Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet Printer, Copier, Scanner, Amazon Dash Replenishment Ready," and there is a link labeled "12 Reviews." I will click on this link to navigate to the reviews section.

A.2.3 EXAMPLE OF REFLECTION PHASE

Example of Reflection Phase

System Prompt:

You are an autonomous intelligent agent tasked with navigating a web browser. You encountered an error while performing a task between web pages. You need to reflect on the mistakes made in your previous execution process and select a checkpoint to reload to retry. Remember that in the reflection, if there could be multiple errors or points that are incon-

1620

sistent with your expectations, you should choose the earliest one to reload from. You will be provided with the following information: Task objective, execution process, verification target, and verification result. You should answer in the following format:

Reflection: (This is the reflection on why the previous execution process meets the mistake, what is your alternative way to solve the task and why you think your new method can solve the problem, which checkpoint you choose to restart from and why.)

Reload checkpoint: (The checkpoint you choose to reload. It should be an integer indicates the step. If you feel that you have actually achieved the task objectives, but the verification method is unreasonable, you should set the value to 0.)

1630

Example User:

Task Objective:

4. Find the most recent out of delivery order from the filtered list.

Execution Process:

Step: 1

The summary of this step is:

The current observation shows a table listing various orders with their statuses. There are no orders with the status "Out of Delivery" in the current list. I will now finish this subtask by indicating that there are no such orders.

You have taken action 'finish_subtask [No out of delivery orders found]' in this step.

Verification:

Target: check_in_webpage("You have found the most recent out of delivery order")

Result: [To verify if the content satisfies the subgoal of finding the most recent out-of-delivery order, I need to check the status of the orders listed and identify the most recent one that is not marked as "Complete" or "Canceled."]

1645

1646

Example Assistant:

Reflection: The execution process involved checking the current list of orders and concluding that there are no orders with the status "Out of Delivery." However, the verification result indicates that the content needs to be checked to ensure that the most recent order with a status other than "Complete" or "Canceled" is identified.

The mistake lies in not thoroughly verifying the status of each order to ensure that none of them are "Out of Delivery." Additionally, the verification should include checking for the most recent order that is not marked as "Complete" or "Canceled."

To solve the task, I should include a step to verify the status of each order on the webpage, ensuring that none of them are "Out of Delivery." If there are no such orders, I should then identify the most recent order that is not marked as "Complete" or "Canceled."

Since the mistake occurred after checking the current list of orders, I should reload from the point just before finishing the subtask to include the verification step.

Reload checkpoint: 0

1659

Example User:

Task Objective:

4. Switch to the tab with the product that has the lowest per unit price.

Execution Process:

Step: 1

The summary of this step is:

I need to check the price and quantity of the Ciclon Energy Drink in the current tab (Tab 0) to determine the per unit price.

You have taken action 'finish_subtask [Ciclon Energy Drink]' in this step.

Verification:

Target: check_in_url("URL of the tab with the lowest per unit price") —OR— check_in_webpage("You have switched to the tab with the lowest per unit price")

Result: [Subgoal url url of the tab with the lowest per unit price is not in current url http://127.0.0.1:7770/ciclon-energy-drink-regular-24-cans-8-3oz.html; To determine if the

1674
 1675 content satisfies the subgoal of switching to the tab with the lowest per unit price, I need to
 1676 compare the prices of the items in both tabs. The first tab contains "Cyclon Energy Drink -
 1677 Regular 24 cans (8.3oz)" priced at \$36.99. The second tab contains "V8 +Energy, Healthy
 1678 Energy Drink, Steady Energy from Black and Green Tea, Pomegranate Blueberry, 8 Ounce
 1679 Can, Pack of 24". I need to find the price of the V8 +Energy drink to make the comparison.]
 1680
 1681

Example Assistant:

Reflection: The execution process involved checking the price and quantity of the Cyclon Energy Drink in the current tab and concluding that it has the lowest per unit price. However, the verification result indicates that the content needs to be checked to ensure that the tab with the lowest per unit price has been correctly identified. The mistake lies in not comparing the per unit prices of the items in both tabs before making a decision.

To solve the task, I should include a step to check the price and quantity of the V8 +Energy drink in the second tab to compare the per unit prices of both products. This comparison will ensure that I switch to the tab with the actual lowest per unit price.

Since the mistake occurred after checking the price of the Cyclon Energy Drink but before comparing it with the V8 +Energy drink, I should reload from the point just before finishing the subtask to include the comparison step.

Reload checkpoint: 0'

Example User:

Task Objective:

3. **Initiate the search**

- Click the search button or press Enter to start the search.

Execution Process:

Step: 1

The summary of this step is:

The current observation shows the search results page for "EYZUTAK" on the One Stop Market website. The search has already been initiated, and the results are displayed. Therefore, the subtask to initiate the search for "EYZUTAK" has been completed.

You have taken action 'finish_subtask []' in this step.

Verification:

Target: : check_in_url(search) —OR— check_in_webpage(Search results for "usb wifi" displayed)

Result: [”subgoal url searchis not in current url http://127.0.0.1:7770/catalogsearch/result/?q=ezytak”, The content provided shows search results for the term "EYZUTAK" and not for "usb wifi". The subgoal target is to verify if the search results for "usb wifi" are displayed.]

Example Assistant:

Reflection: The execution process correctly identified that the search results for "EYZUTAK" were displayed, but the task objective was to initiate the search for "usb wifi." The verification result indicates that the search results for "usb wifi" were not displayed, and the URL did not contain the expected search term.

The mistake lies in assuming that the search for "EYZUTAK" was relevant to the task objective. To solve the task, I should ensure that the search term "usb wifi" is entered and the search is initiated correctly.

Since the mistake occurred at the very beginning of the task, I should reload from the initial step to correct the search term and initiate the search for "usb wifi."

Reload checkpoint: 1

User

Task Objective:

4. **Identify the best storage option for 11 game cards**

1728
 1729 - Review the search results to find a storage option that can fit at least 11 Nintendo Switch
 1730 game cards.
 1731 - **
 1732 Execution Process:
 1733 Step: 1
 1734 The summary of this step is:
 1735 The current observation shows the search results page for "Nintendo Switch game card stor-
 1736 age" on the One Stop Market website. I will scroll down to view more details about the
 1737 storage options.
 1738 You have taken action 'scroll [down]' in this step.
 1739 Step: 2
 1740 The summary of this step is:
 1741 The current observation shows the search results for "Nintendo Switch game card storage."
 1742 The best storage option for 11 game cards is the "HEIYING
 1743 You have taken action 'finish_subtask [The best storage option for 11 game cards is the
 1744 HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD
 1745 Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and
 1746 24 Micro SD Card Slots. (Black) priced at \$14.69]' in this step.
 1747 Verification:
 1748 Target: **: 'check_in_webpage("Storage option for 11 Nintendo Switch game cards")'
 1749 Result: [The search results provide various storage options for Nintendo Switch game cards.
 1750 I need to verify if any of these options specifically mention the capacity to store 11 game
 1751 cards.]

1752 **Output:**
 1753 Reflection: The execution process involved scrolling down to view more details about the
 1754 storage options and then selecting the "HEIYING Game Card Case" as the best option for 11
 1755 game cards. However, the verification result indicates that the content needs to be checked
 1756 to ensure that the selected storage option specifically mentions the capacity to store 11 game
 1757 cards.
 1758 The mistake lies in not explicitly verifying that the selected storage option can fit at least 11
 1759 game cards. The selected option has 24 game card slots, which meets the requirement, but
 1760 this detail was not clearly verified in the process.
 1761 To solve the task, I should include a step to explicitly verify that the selected storage option
 1762 can fit at least 11 game cards by checking the product details.
 1763 Since the mistake occurred after selecting the storage option but before verifying its capacity,
 1764 I should reload from the point just before finishing the subtask to include the verification
 1765 step.
 1766 Reload checkpoint: 1

A.2.4 EXAMPLE OF REPLANNING PHASE

1767 Example of Replanning Phase

1768

1769 **System Prompt:**

1770 You are an autonomous intelligent agent tasked with navigating a web browser. To achieve
 1771 the final objective, you have broken it down into a series of subtasks to be tackled one by
 1772 one. However, when you encounter errors while executing a particular subtask, and retrying
 1773 within this subtask does not resolve the issue, you decide to re-decompose and create new
 1774 subtasks. You will be provided with the following information:
 1775 Final Objective: Intent: The ultimate goal;
 1776 Old Scheme: The previously generated set of subtasks;
 1777 Global history: The execution history of the subtasks that have been successfully completed
 1778 before the error occurred; these subtasks are done and do not need to be changed or regen-
 1779 erated;
 1780 Current subtask: The subtask that encountered the error;

1782
 1783 Local history: The execution history of the subtask that encountered the error;
 1784 Verification target: The indicator for verification of the error;
 1785 Verification result: The error message.
 1786 Please note:
 1787 1. The subtasks executed before the error do not need to be modified, as they are already
 1788 completed. You only need to generate replacements for the new subtasks following the one
 1789 that encountered the error;
 1790 2. When creating new subtasks, you also need to establish methods of verification and
 1791 verification targets. Please follow the following format for specifics.
 1792 Your verification method should be one of the three:
 1793 1. check_in_url (part of url): You think you will check in the url. This often be used when
 1794 you want to navigate to some website. Its parameter is the fragment you expect to be the
 1795 part of the url.
 1796 2. check_in_webpage (objective): You think you will check the objective in the last web-
 1797 page. This often be used when the last webpage has the information for your sub-task.
 1798 Its parameter is the objective you want to achieve in the sub-task and can be check in the
 1799 webpage.
 1800 3. check_in_history (objective): You think you should check in your action history. This
 1801 often be used when your sub-task is to get some information. Its parameter is the objective
 1802 you want to achieve in the sub-task and can be check in the history.
 1803 You can use ‘|OR|’ to connect two objective if you want to use multiple methods to assess
 1804 the result of subtasks. You should add objectives by starting with Objectiveáfter each sub-
 1805 task. For example, in Find the latest post. Objective: check_in_webpage (“This is the latest
 1806 post”); Find the latest post.ís your verification method. After it is your verification method
 1807 check_in_webpageánd your objective of your method This is the latest post.
 1808 3. Your response should contain two parts. One is your thoughts of why the old subtasks
 1809 encountered error and why your new subtasks can solve the problem. The other is your
 1810 generated new subtasks, and this part should start with New Subtasks:
 1811 4. You need to generate objectives for every subtasks. It is forbidden to generate a subtask
 1812 without an objective.

Eexample User:

1813 Final Objective: Intent: Buy the highest rated product from the Beauty & Personal Care
 1814 category within a budget under 20.

1815 To achieve the final objective, I decompose it into these subtasks:
 1816 Old Scheme:

1. Navigate to the "Beauty & Personal Care" category.
2. Sort the products by rating in descending order.
3. Filter the products to show only those priced under \$20.
4. Identify the highest-rated product within the filtered results.
5. Add the highest-rated product to the shopping cart.

1817 Previous Subtasks and the Summary of Their Execution History:
 1818 Global History:

1. Navigate to the "Beauty & Personal Care" category.

1819 Summary: The navigation to the "Beauty & Personal Care" category was successful, as
 1820 confirmed by the URL indicating the correct category page.
 1821 Current Subtask:

2. Sort the products by rating in descending order.

1822 Local History:
 1823 Previous Steps:

1824 Step: 1
 1825 The summary of this step is:

1826 The current observation shows a "Sort By" combobox that is currently focused and ex-
 1827 panded. I will click on this combobox to view the sorting options.
 1828 You have taken action ‘click [13132] where [13132] is combobox Šort Byfocused: True
 1829 hasPopup: menu expanded: True‘ in this step.

1836
 1837 Step: 2
 1838 The summary of this step is:
 1839 The current observation shows that the "Sort By" combobox is focused but not expanded.
 1840 The next step is to expand the combobox to view the sorting options.
 1841 You have taken action 'click [13132] where [13132] is combobox Sort Byfocused: True
 1842 hasPopup: menu expanded: False' in this step.
 1843 Step: 3
 1844 The summary of this step is:
 1845 The current observation shows that the "Sort By" combobox is focused and expanded, dis-
 1846 playing the sorting options. The next step is to select the option to sort the products by rating
 1847 in descending order.
 1848 You have taken action 'click [13133] where [13133] is link \ue613 Set Descending Direction'
 1849 in this step.
 1850 Step: 4
 1851 The summary of this step is:
 1852 The current observation shows that the "Sort By" combobox is focused but not expanded,
 1853 and the sorting direction is set to descending. I need to ensure that the products are sorted
 1854 by rating in descending order. Therefore, I will expand the "Sort By" combobox to view the
 1855 sorting options.
 1856 You have taken action 'click [16146] where [16146] is combobox Sort ByhasPopup: menu
 1857 expanded: False' in this step.
 1858 Step: 5
 1859 The summary of this step is:
 1860 The current observation shows that the "Sort By" combobox is focused and expanded, dis-
 1861 playing the sorting options. The next step is to select the option to sort the products by rating
 1862 in descending order.
 1863 You have taken action 'click [16147] where [16147] is link \ue614 Set Ascending Direction'
 1864 in this step.
 1865 Step: 6
 1866 The summary of this step is:
 1867 The current observation shows that the "Sort By" combobox is focused but not expanded,
 1868 and the sorting direction is set to descending. I need to ensure that the products are sorted
 1869 by rating in descending order. Therefore, I will expand the "Sort By" combobox to view the
 1870 sorting options.
 1871 You have taken action 'click [19160] where [19160] is combobox Sort ByhasPopup: menu
 1872 expanded: False' in this step.
 1873 Verification:
 1874 Target: : check_in_webpage("Products sorted by rating")
 1875 Result: [To determine if the products are sorted by rating, I need to look for any indication
 1876 that the sorting criteria is based on ratings. This could be a label or a dropdown selection
 1877 indicating "Sort by Rating" or similar wording.]
 1878 You should not modify the already completed subtasks, namely
 1879 1. Navigate to the "Beauty & Personal Care" category.
 1880 You should reflect on the original mistakes (i.e., the local history mentioned above) and
 1881 generate subsequent new subtasks and verification plans based on the completed results
 1882 (i.e., the global history mentioned above). Note that your new sub-tasks index should be
 1883 started from the 'Current Subtask', that is 1, and your new subtasks should not exceed 4.
 1884
1885 Example Assistant:
 1886 The error in the previous execution was due to the repeated attempts to expand the "Sort
 1887 By" combobox and select the sorting option without successfully verifying that the products
 1888 were sorted by rating in descending order. The verification method requires checking for
 1889 any indication that the sorting criteria are based on ratings.
 To resolve this, the new subtasks must ensure that the products are sorted by rating in de-
 scending order and that this sorting is verified before proceeding to the next steps.

1890
 1891 New Subtasks:
 1892 2. Sort the products by rating in descending order.
 1893 Objective: check_in_webpage("Products sorted by rating in descending order")
 1894 3. Filter the products to show only those priced under \$20.
 1895 Objective: check_in_webpage("Products filtered by price under \$20")
 1896 4. Identify the highest-rated product within the filtered results.
 1897 Objective: check_in_webpage("Highest-rated product identified within filtered results")
 1898 5. Add the highest-rated product to the shopping cart.
 1899 Objective: check_in_webpage("Highest-rated product added to the shopping cart")
 1900 These new subtasks should ensure that the products are sorted by rating in descending order,
 1901 filtered by price, and that the highest-rated product is identified and added to the shopping
 1902 cart. This should resolve the previous issues and achieve the final objective.

1903 **Example User:**
 1904 Final Objective: Intent: Buy the highest rated product from the Beauty & Personal Care
 1905 category within a budget under 20.
 1906 To achieve the final objective, I decompose it into these subtasks:
 1907 Old Scheme:
 1908 1. Navigate to the "Beauty & Personal Care" category.
 1909 2. Sort the products by rating in descending order.
 1910 3. Filter the products to show only those priced under \$20.
 1911 4. Identify the highest-rated product within the filtered results.
 1912 5. Add the highest-rated product to the shopping cart.
 1913 Previous Subtasks and the Summary of Their Execution History:
 1914 Global History:
 1915 1. Navigate to the "Beauty & Personal Care" category.
 1916 Summary: The navigation to the "Beauty & Personal Care" category was successful, as
 1917 confirmed by the URL indicating the correct category page.
 1918 Current Subtask:
 1919 2. Sort the products by rating in descending order.
 1920 Local History:
 1921 Previous Steps:
 1922 Step: 1
 1923 The summary of this step is:
 1924 The current observation shows a page listing beauty and personal care products. To sort the
 1925 products by rating in descending order, I need to interact with the "Sort By" combobox. I
 1926 will start by clicking on the combobox to open it and view the sorting options.
 1927 You have taken action 'click [19160] where [19160] is combobox Sort Byfocused: True
 1928 hasPopup: menu expanded: True' in this step.
 1929 Step: 2
 1930 The summary of this step is:
 1931 The current observation shows that the "Sort By" combobox is focused but not expanded.
 1932 The next step is to click on the "Sort By" combobox to expand it and view the sorting
 1933 options.
 1934 You have taken action 'click [19160] where [19160] is combobox Sort Byfocused: True
 1935 hasPopup: menu expanded: False' in this step.
 1936 Step: 3
 1937 The summary of this step is:
 1938 The current observation shows that the "Sort By" combobox is now expanded, and the sorting
 1939 options are visible. The next step is to select the option to sort by rating in descending
 1940 order by clicking on the "Set Descending Direction" link.
 1941 You have taken action 'click [19161] where [19161] is link \ue613 Set Descending Direction'
 1942 in this step.
 1943 Step: 4
 1944 The summary of this step is:

1944
 1945 The current observation shows that the "Sort By" combobox is not expanded, and the sorting
 1946 options are not visible. The next step is to click on the "Sort By" combobox to expand it and
 1947 view the sorting options.
 1948 You have taken action 'click [22174] where [22174] is combobox Šort ByhasPopup: menu
 1949 expanded: False' in this step.
 1950 Step: 5
 1951 The summary of this step is:
 1952 The current observation shows that the "Sort By" combobox is expanded, and the sorting
 1953 options are visible. The next step is to select the option to sort by rating in descending order
 1954 by clicking on the appropriate option.
 1955 You have taken action 'click [22175] where [22175] is link ūe614 Set Ascending Direction'
 1956 in this step.
 1957 Step: 6
 1958 The summary of this step is:
 1959 The current observation shows that the "Sort By" combobox is not expanded, and the sorting
 1960 options are not visible. The next step is to click on the "Sort By" combobox to expand it and
 1961 view the sorting options.
 1962 You have taken action 'click [25188] where [25188] is combobox Šort ByhasPopup: menu
 1963 expanded: False' in this step.
 1964 Verification:
 1965 Target: : check_in_webpage("Products sorted by rating in descending order")
 1966 Result: [To verify if the products are sorted by rating in descending order, I need to check
 1967 for any indication that the sorting criteria is based on rating and that it is in descending order.
 1968 The content should show a sorting option or label indicating "rating" and "descending".]
 1969 You should not modify the already completed subtasks, namely
 1970 1. Navigate to the "Beauty & Personal Care" category.
 1971 You should reflect on the original mistakes (i.e., the local history mentioned above) and
 1972 generate subsequent new subtasks and verification plans based on the completed results
 1973 (i.e., the global history mentioned above). Note that your new sub-tasks index should be
 1974 started from the 'Current Subtask', that is 1, and your new subtasks should not exceed 4.
 1975
Example Assistant:
 1976 The error in the previous execution was due to the repeated attempts to interact with the
 1977 "Sort By" combobox without successfully selecting the option to sort by rating in descending
 1978 order. The verification method requires checking for any indication that the sorting criteria
 1979 is based on rating and that it is in descending order.
 1980 To resolve this, the new subtasks must ensure that the products are sorted by rating in de-
 1981 scending order and that this sorting is verified.
 1982 New Subtasks:
 1983 2. Sort the products by rating in descending order.
 1984 Objective: check_in_webpage("Products sorted by rating in descending order")
 1985 3. Filter the products to show only those priced under \$20.
 1986 Objective: check_in_webpage("Products filtered to show only those priced under \$20")
 1987 4. Identify the highest-rated product within the filtered results.
 1988 Objective: check_in_webpage("Highest-rated product identified within the filtered results")
 1989 5. Add the highest-rated product to the shopping cart.
 1990 Objective: check_in_webpage("Highest-rated product added to the shopping cart")
 1991 These new subtasks should ensure that the products are sorted correctly, filtered by price,
 1992 and that the highest-rated product is identified and added to the shopping cart. This should
 1993 resolve the previous issues and achieve the final objective.
 1994
Example User:
 1995 Final Objective: Intent: Tell me the status of my latest order and when will it arrive
 1996 To achieve the final objective, I decompose it into these subtasks:
 1997 Old Scheme:

1998
 1999 1. Navigate to the "My Account" page.
 2000 2. Locate the "Orders" section within the "My Account" page.
 2001 3. Identify the list of orders and find the latest order.
 2002 4. Retrieve the status of the latest order.
 2003 5. Verify the presence or absence of the estimated arrival date for the latest order.
 2004 Previous Subtasks and the Summary of Their Execution History:
 2005 Global History:
 2006 1. Navigate to the "My Account" page.
 2007 Summary: The "My Account" link was clicked, successfully navigating to the "My Account" page, as confirmed by the URL and the presence of account-related information on the page.
 2008
 2009 2. Locate the "Orders" section within the "My Account" page.
 2010 Summary: The "My Orders" link was clicked, successfully navigating to the "Orders" section within the "My Account" page, as confirmed by the detailed order table present on the page.
 2011
 2012 3. Identify the list of orders and find the latest order.
 2013 Summary: The latest order, 000000170 dated 5/17/23, was identified, but the verification step to confirm it as the most recent order on the webpage was missed. The process needs to be reloaded to include this verification step.
 2014
 2015 4. Retrieve the status of the latest order.
 2016 Summary: The status of the latest order (000000170) was identified as "Canceled," but the verification step to check the webpage content for this status was missed. The process needs to include a step to verify the presence of the latest order's status on the webpage.
 2017 Current Subtask:
 2018 5. Verify the presence or absence of the estimated arrival date for the latest order.
 2019 Local History:
 2020 Previous Steps:
 2021 Step: 1
 2022 The summary of this step is:
 2023 The current observation shows the details of order 000000170. To verify the presence or absence of the estimated arrival date, I need to scroll down the page to check for any mention of it. The expected outcome of this action is to see if there is any information regarding the estimated arrival date for this order.
 2024 You have taken action 'scroll [down]' in this step.
 2025 Step: 2
 2026 The summary of this step is:
 2027 The current observation shows the details of order 000000170, including items ordered, order information, shipping address, shipping method, billing address, and payment method.
 2028 There is no mention of the estimated arrival date in the visible details. Therefore, I will finish the subtask to confirm that the estimated arrival date is absent for the latest order.
 2029 You have taken action 'finish_subtask []' in this step.
 2030 You have encountered error in previous trial, this is your reflection:
 2031 Reflection: The execution process involved checking the details of order 000000170 to verify the presence or absence of the estimated arrival date. The steps included scrolling down the page and observing the order details, ultimately concluding that there was no mention of the estimated arrival date. The verification result indicates that the content does not include any information about the estimated arrival date and suggests checking for any mention of an estimated arrival date or an indication of order cancellation.
 2032 The mistake lies in not explicitly verifying the webpage content for any mention of an estimated arrival date or an indication of order cancellation. To solve the task, I should include a step to verify the webpage content for any mention of an estimated arrival date or an indication that the order has been canceled.
 2033 Since the mistake occurred after observing the order details and not verifying the absence of the estimated arrival date or order cancellation, I should reload from the point just before finishing the subtask to include the verification step.
 2034 And you choose to restart from this state.
 2035 Step: 3

2052
 2053 The summary of this step is:
 2054 The current observation shows the details of order 000000170, including items ordered,
 2055 order information, shipping address, shipping method, billing address, and payment method.
 2056 There is no mention of the estimated arrival date in the visible details. The next step involves
 2057 scrolling down the page to check for any mention of the estimated arrival date. The expected
 2058 outcome of this action is to see if there is any information regarding the estimated arrival
 2059 date for this order.
 2060 You have taken action ‘scroll [down]’ in this step.
 2061 Step: 4
 2062 The summary of this step is:
 2063 The current observation shows the details of order 000000170, including items ordered, or-
 2064 der information, shipping address, shipping method, billing address, and payment method.
 2065 There is no mention of the estimated arrival date in the visible details. The next step in-
 2066 volves confirming the absence of the estimated arrival date for the latest order. The expected
 2067 outcome of this action is to confirm that the estimated arrival date is absent for the latest
 2068 order.
 2069 You have taken action ‘finish_subtask [The estimated arrival date is absent for the latest
 2070 order.]’ in this step.
 2071 Verification:
 2072 Target: : check_in_webpage(“Estimated arrival date for order 000000170”)
 2073 |OR|check_in_webpage(“No estimated arrival date due to order cancellation”)
 2074 Result: [The content provided does not include any information about the estimated arrival
 2075 date for order 000000170. It contains details about the order information, shipping address,
 2076 shipping method, billing address, payment method, and some additional elements like re-
 2077 cently ordered items and newsletter sign-up, but no mention of the estimated arrival date.;
 2078 The content provided does not mention anything about an estimated arrival date or order
 2079 cancellation. It primarily includes order information, shipping and billing addresses, ship-
 2080 ping method, payment method, and some additional elements like a wish list and newsletter
 2081 sign-up.]
 2082 You should not modify the already completed subtasks, namely
 2083 1. Navigate to the “My Account” page.
 2084 2. Locate the “Orders” section within the “My Account” page.
 2085 3. Identify the list of orders and find the latest order.
 2086 4. Retrieve the status of the latest order.
 2087 You should reflect on the original mistakes (i.e., the local history mentioned above) and
 2088 generate subsequent new subtasks and verification plans based on the completed results
 2089 (i.e., the global history mentioned above). Note that your new sub-tasks index should be
 2090 started from the ‘Current Subtask’, that is 4, and your new subtasks should not exceed 1.
 2091
Example Assistant:
 2092 The error in the previous execution was due to the lack of explicit verification for the pres-
 2093 ence or absence of the estimated arrival date or an indication of order cancellation. The
 2094 verification method requires checking the webpage content for any mention of the estimated
 2095 arrival date or an indication that the order has been canceled. To resolve this, the new subtask
 2096 must ensure that the webpage content is thoroughly checked and verified for this informa-
 2097 tion.
 2098 New Subtasks:
 2099 5. Verify the presence or absence of the estimated arrival date for the latest order and check
 2100 for any indication of order cancellation.
 2101 Objective: check_in_webpage(“Estimated arrival date for order 000000170”)
 2102 |OR|check_in_webpage(“No estimated arrival date due to order cancellation”)
 2103 This new subtask should ensure that the webpage content is thoroughly checked and verified
 2104 for the presence or absence of the estimated arrival date or any indication of order cancella-
 2105 tion. This will ensure that the task objective is fully met and verified.

2106
 2107 **User:**
 2108 Final Objective: Intent: I have a lot of Nintendo Switch game cards now, help me find the
 2109 best storage option to fit all 23 cards
 2110 To achieve the final objective, I decompose it into these subtasks:
 2111 Old Scheme:
 2112 1. **Navigate to the "Electronics" section**
 2113 - Locate and click on the "Electronics" menu item to explore related products.
 2114 2. **Locate the "Video Games" subcategory**
 2115 - Within the "Electronics" section, find and click on the "Video Games" subcategory to
 2116 narrow down the search.
 2117 3. **Search for "Nintendo Switch game card storage"**
 2118 - Use the search bar within the "Video Games" section to search for "Nintendo Switch game
 2119 card storage".
 2120 4. **Identify the best storage option for 23 game cards**
 2121 - Review the search results and identify a storage option that can fit at least 23 Nintendo
 2122 Switch game cards.
 2123 5. **Verify the product details**
 2124 - Click on the identified product to verify its details, ensuring it meets the requirement of
 2125 fitting 23 game cards.
 2126 Previous Subtasks and the Summary of Their Execution History:
 2127 Global History:
 2128 1. **Navigate to the "Electronics" section**
 2129 - Locate and click on the "Electronics" menu item to explore related products.
 2130 Summary: The "Electronics" menu item was successfully clicked, leading to the "Electron-
 2131 ics" section, as confirmed by the URL containing "electronics".
 2132 2. **Locate the "Video Games" subcategory**
 2133 - Within the "Electronics" section, find and click on the "Video Games" subcategory to
 2134 narrow down the search.
 2135 Summary: The "Video Games" subcategory was successfully located and clicked within the
 2136 "Electronics" section, leading to the correct "Video Games" page as confirmed by the URL.
 2137 3. **Search for "Nintendo Switch game card storage"**
 2138 - Use the search bar within the "Video Games" section to search for "Nintendo Switch game
 2139 card storage".
 2140 Summary: The search query "Nintendo Switch game card storage" was successfully entered
 2141 and the search button was clicked, resulting in the display of relevant search results. The
 2142 search bar was verified to contain the correct query text.
 2143 4. **Identify the best storage option for 23 game cards**
 2144 - Review the search results and identify a storage option that can fit at least 23 Nintendo
 2145 Switch game cards.
 2146 Summary: Detailed Summary of Steps and Verification Result
 2147 **Step 1:**
 2148 - **Summary:** The current observation shows the search results page for "Nintendo
 2149 Switch game card storage" on the One Stop Market website. The next step is to scroll
 2150 down to view more details about the search results, with the expected outcome of seeing
 2151 more details about the storage options, including their capacities.
 2152 - **Action Taken:** Scrolled down to view more details.
 2153 **Reflection:**
 2154 - **Summary:** The execution process involved identifying the "HEIYING Game Card
 2155 Case" as the best storage option for 23 game cards based on the search results. However, the
 2156 verification result indicates that the content needs to be checked to ensure that the identified
 2157 product meets the specific requirement of fitting at least 23 Nintendo Switch game cards. The
 2158 mistake was not thoroughly verifying the details of the identified product and not checking
 2159 other potential options.
 2160 - **Action Taken:** Decided to restart from the point just before finishing the subtask to
 2161 include the verification step.
 2162 **Step 3:**

2160
 2161 - **Summary:** The current observation shows the search results for "Nintendo Switch
 2162 game card storage" on the One Stop Market website. The next step is to click on the "HEIY-
 2163 ING Game Card Case" to navigate to its product page and verify its details.
 2164 - **Action Taken:** Clicked on the link for "HEIYING Game Card Case."
 2165 **Step 4:**
 2166 - **Summary:** The current observation confirms that the "HEIYING Game Card Case"
 2167 has 24 game card slots, making it a suitable option for storing at least 23 Nintendo Switch
 2168 game cards. Therefore, the subtask is finished.
 2169 - **Action Taken:** Finished the subtask.
 2170 **Verification Result:**
 2171 - **Target:** The webpage should describe a storage option for 23 Nintendo Switch game
 2172 cards.
 2173 - **Result:** The content describes a game card
 Current Subtask:
 2174 5. **Verify the product details**
 2175 - Click on the identified product to verify its details, ensuring it meets the requirement of
 2176 fitting 23 game cards.
 2177 - **Local History:
 Previous Steps:
 2178 Step: 1
 2179 The summary of this step is:
 2180 The current observation shows the product page for the "HEIYING Game Card Case for
 2181 Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch
 2182 Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots." I will
 2183 click on the product to verify its details and ensure it meets the requirement of fitting 23
 2184 game cards.
 2185 You have taken action 'click [16505] where [16505] is StaticText HEIYING Game Card
 2186 Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards,Portable
 2187 Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots.
 2188 (Black)' in this step.
 2189 You have encountered error in previous trial, this is your reflection:
 2190 Reflection: The execution process involved clicking on the product and confirming that it
 2191 meets the requirement of fitting 23 game cards. However, the verification target requires ex-
 2192 plicitly checking the product details to confirm the storage capacity for 23 Nintendo Switch
 2193 game cards. The mistake lies in not explicitly verifying the product description for this
 2194 specific detail.
 2195 To resolve this, I should reload from the point before finishing the subtask and explicitly
 2196 check the product description on the webpage to confirm that it mentions storage for 23
 2197 Nintendo Switch game cards.
 2198 And you choose to restart from this state.
 2199 Step: 3
 2200 The summary of this step is:
 2201 The current observation confirms that the "HEIYING Game Card Case for Nintendo Switch
 2202 & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory
 2203 Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots" has 24 game card slots,
 2204 which meets the requirement of fitting 23 game cards. Therefore, I will finish the subtask.
 2205 You have taken action 'finish_subtask [The product has 24 game card slots, which meets the
 2206 requirement of fitting 23 game cards]' in this step.
 2207 Verification:
 2208 Target: **: 'check_in_webpage("Product details confirm storage for 23 Nintendo Switch
 2209 game cards")'
 2210 Result: [The subgoal target is to confirm whether the product details mention storage for 23
 2211 Nintendo Switch game cards. I need to verify if the product description includes this specific
 2212 storage capacity.]
 2213 You should not modify the already completed subtasks, namely
 1. **Navigate to the "Electronics" section**
 - Locate and click on the "Electronics" menu item to explore related products.

2214
 2215 2. **Locate the "Video Games" subcategory**
 2216 - Within the "Electronics" section, find and click on the "Video Games" subcategory to
 2217 narrow down the search.
 2218 3. **Search for "Nintendo Switch game card storage"**
 2219 - Use the search bar within the "Video Games" section to search for "Nintendo Switch game
 2220 card storage".
 2221 4. **Identify the best storage option for 23 game cards**
 2222 - Review the search results and identify a storage option that can fit at least 23 Nintendo
 2223 Switch game cards.
 2224 You should reflect on the original mistakes (i.e., the local history mentioned above) and
 2225 generate subsequent new subtasks and verification plans based on the completed results (i.e.,
 2226 the global history mentioned above). Note that your new sub-tasks index should be started
 2227 from the 'Current Subtask', that is 4, and your new subtasks should not exceed 1.

2228 **Output:**

2229 The error in the previous execution was due to not explicitly verifying the product descrip-
 2230 tion for the specific detail of fitting 23 Nintendo Switch game cards. The verification method
 2231 requires checking the product details on the webpage to confirm this specific storage capac-
 2232 ity.

2233 To resolve this, the new subtask must ensure that the product description is thoroughly
 2234 checked and verified for the storage capacity of 23 Nintendo Switch game cards.

2235 New Subtasks: 5. Verify the product details to ensure it meets the requirement of fitting 23
 2236 game cards. Objective: check.in_webpage("Product details confirm storage for 23 Nintendo
 2237 Switch game cards")

2238 This new subtask should ensure that the product description is thoroughly checked and ver-
 2239 ified for the specific storage capacity of 23 Nintendo Switch game cards. This will ensure
 2240 that the task objective is fully met and verified.

2242 B MORE EXPERIMENTS

2244 B.1 COMPUTATIONAL COST

2248 The computational cost is an important criterion for evaluating the efficiency of an LLM workflow.
 2249 In Table 6, we report the computational cost incurred by each module in VeSX and visualize the
 2250 proportion of each module's contribution in Figure 5, Figure 6, Figure 7. Although we implemented
 2251 some design changes in prompts for planning and execution, the count of tokens used is small.
 2252 The additional cost resulting from process improvements can largely be attributed to verification,
 2253 reflection, and replanning.

2254 Considering that the computational cost is determined by the number of tokens, the verification,
 2255 reflection and replanning modules account for approximately 10.7% of the input tokens (Figure 5).
 2256 For output tokens, the three modules account for about 16.1% of the tokens (Figure 6). Moreover, the
 2257 count of LLM calls in the three modules account for about 22.3%. This indicates that the three main
 2258 modules in VeSX workflow, verification, reflection, and replanning, do not introduce significant
 2259 more computational cost.

2260 We also tested the token usage for examples retrieved from the exemplar bank. The examples for
 2261 execution accounted for 33.9% of the total input tokens, while the examples for planning also ac-
 2262 counted for 2.4% of the input tokens. This is a considerable proportion, but given the widespread
 2263 adoption of ICL, most methods Sodhi et al. (2024) Koh et al. (2024) Drouin et al. (2024) will incur
 2264 this overhead. And the computational cost for these examples can be further saved due to the fact
 2265 found in Appendix B.2 that fewer examples can still maintain a high level of accuracy.

2266 Due to the significant time and cost involved in reproducing other approaches, we did not fully
 2267 reproduce all of them. We compared our method with the Sodhi et al. (2024) approach, which
 2268 uses human-labeled subtasks to guide the LLM in breaking down and solving tasks. Their average

Module	planning	execution	verification	reflection	replanning	summary&answer
Input Token Count	7743	69045	5229	1034	4076	10311
Output Token Count	188	2095	247	98	184	475
Module Call Count	1	13.1	4.1	0.60	0.68	4.6

Table 6: Computational cost statistics of each module in VeSX workflow.

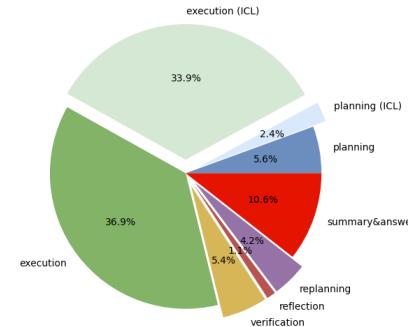


Figure 5: Proportions of input tokens for each module.

Module	Instruction	Exemplar
Planning	4147	3596
Execution	18583	50462

Table 7: Statistics of token counts for instructions and exemplars in planning and execution module inputs.

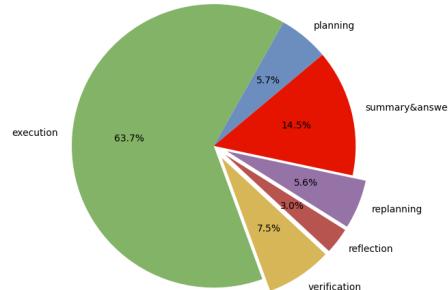


Figure 6: Proportions of output tokens for each module.

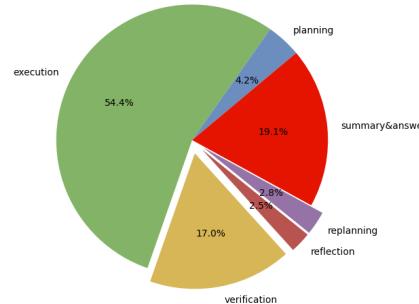


Figure 7: Proportions of module calls for each module.

action step is 9.1, while our average action step is 13.1. From this perspective, VeSX requires approximately 43.9% more actions compared to human-guided workflows to achieve competitive results.

B.2 SCALING OF EXEMPLARS

We first analyzed the size of the exemplar bank in various scenarios, as shown in the Table ??.

We also conducted experiments to test the scalability of the exemplars, primarily using shopping as the experimental scenario. In Table 9, we tested the impact of using different numbers of exemplars during ICL on the final end-to-end results. The SR improves when the count of ICL exemplars increases, underlying the potential of the scaling of exemplars during the inference process. In Table 10, we examined how the SR changes under different settings of the count of planning and execution exemplars. We set four settings: (1) sampling 70% of the whole exemplar bank, (2) using the whole execution exemplar bank but not the planning exemplar, (3) using the whole planning exemplar bank but not the execution exemplar, and (4) using the whole exemplar bank. The results shows that using only 70% of the whole bank causes a obvious decrease of SR comparing to even removing the whole planning exemplars or execution exemplars. It indicates the more diversity of the execution or the planning exemplars will bring more increase and the potential of the scaling of the exemplar bank.

Scenario	Shop.	CMS	Red.	Git.	Map
Planning	35	24	30	29	20
Execution	171	123	211	193	87

Table 8: Statistics of exemplar banks for the five scenarios.

#P	#E	SR
26	120	0.334
0	171	0.340
35	0	0.373
35	171	0.412

Table 10: Success rates under different settings of the count of planning and execution exemplars. SR represents the success rate. #P refers to the number of planning exemplars and #E refers to the number of planning exemplars.

Condition	SR	Proportion
Verification pass	0.43	38%
Verification fail	0.40	62%

Table 12: Success rates and proportion of different modules works in scenario Shop. SR represents the success rate. The “verification pass” means that all verifications in this process pass, while the “verification fail” means at least one verification does not pass.

B.3 PERFORMANCE OF VERIFICATION, REFLECTION AND REPLANNING

We analyzed the performance of the success rate after verification, reflection, and replanning occurred, and we calculated the proportions of each situation to test the capabilities and performance of each module. Table 11 presents the performance across all five scenarios. Here, the “verification pass” is defined as the instance that all verifications in this process pass, while the “verification fail” is defined as the instance where at least one verification does not pass. It can be seen that the success rate of “verification pass” reaches 0.384, while that of “verification fail” achieves 0.318 after undergoing reflection and replanning. When reflection is triggered without the need for replanning, the success rate can reach 0.661, and this portion of examples accounts for 12.5%. This is because these examples are relatively simple, making errors easier to correct (C.1 provides such an example). The cases that trigger replanning account for 53.6%, with a corresponding success rate of only 0.229, indicating that these examples are difficult and highlighting the challenge of having the LLM replan correctly globally.

Additionally, we analyzed the success rate and its proportion in the shopping scenario when verification passes and fails, as shown in Table 12. We previously conducted an ablation study on the cases without verification. Combining these two parts of the experiment, we can roughly estimate the confusion matrix of verification, as shown in Table 13. The false positive rate can be estimated as 0.57. This result demonstrates that the verification may not be reliable when classified as a positive sample. The false negative rate needs to be estimated using the success rate when not verifying and the success rate when verification fails, and can be estimated as 0.20. This result demonstrates that

Num of ICL exemplars	SR
3	0.412
5	0.458

Table 9: Success rates under different settings of the count of in-context-learning exemplars. SR represents the success rate.

Condition	SR	Proportion
Reflection	0.386	43.3%
Replanning	0.229	53.6%
Reflection (w/o replanning)	0.661	12.5%
Verification pass	0.384	33.8%
Verification fail	0.318	66.2%

Table 11: Success rates and proportion of different modules works. SR represents the success rate. The “verification pass” means that all verifications in this process pass, while the “verification fail” means at least one verification does not pass.

	Label=Pos	Label=Neg
Pred=Pos	0.16	0.22
Pred=Neg	0.13	0.49

Table 13: The confusion matrix of verification in scene Shop. Pos represents positive samples and Neg represents negative samples.

2376 verification rarely classifies positive samples as negative, which does not incur much unnecessary
 2377 self-correction overhead. The true negative rate can be estimated as 0.79. This result demonstrates
 2378 that the verification has high rate to capture the error in the process.
 2379

2380 C CASE STUDIES

2392 C.1 REFLECTION

2397 One Complete Example

2398 **Query Intent:**

2399 I have a lot of Nintendo Switch game cards now, help me find the best storage option to fit
 2400 all 11 cards.

2401 **Ground Truth**

2402 URL Match: 'reference_url': 'http://127.0.0.1:7770/heiyng-game-card-case-for-nintendo-
 2403 switch-switch-oled-game-card-or-micro-sd-memory-cards-portable-switch-game-memory-
 2404 card-storage-with-24-game-card-slots-and-24-micro-sd-card-slots-black.html'

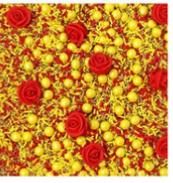
2406 **Planning**

- 2407 1. **Navigate to the "Electronics" section** - Locate and click on the "Electronics" menu
 2408 item to access the electronics category.
 - **Objective**: 'check_in_url("electronics")'
- 2409 2. **Locate the "Video Games" subcategory** - Within the "Electronics" section, find and
 2410 click on the "Video Games" subcategory to narrow down the search to video game-related
 2411 products.
 - **Objective**: 'check_in_url("video-games")'
- 2412 3. **Search for "Nintendo Switch game card storage"** - Use the search bar within the
 2413 "Video Games" section to search for "Nintendo Switch game card storage".
 - **Objective**: 'check_in_webpage("Search bar is populated with 'Nintendo Switch game
 2414 card storage'")'
- 2415 4. **Identify the best storage option for 11 game cards** - Review the search results to find
 2416 a storage option that can fit at least 11 Nintendo Switch game cards.
 - **Objective**: 'check_in_webpage("Storage option for 11 Nintendo Switch game cards")'
- 2417 5. **Navigate to the product page of the identified storage option** - Click on the identified
 2418 storage option to navigate to its product page for more details.
 - **Objective**: 'check_in_url("nintendo-switch-game-card-storage")'
 2419 |OR|check_in_webpage("You are on the product page of the identified storage option")'

2425 **Subtask 1**

- 2426 1. **Navigate to the "Electronics" section** - Locate and click on the "Electronics" menu
 2427 item to access the electronics category.

2428 **Figure of Observation (only for visualization, not used in input prompt):**

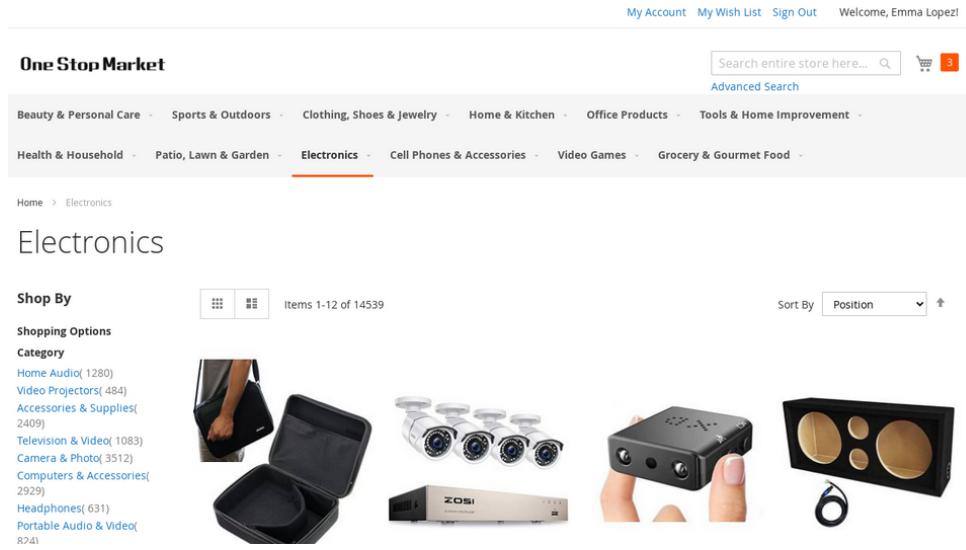
2430
 2431
 2432
 2433 **One Stop Market**
 2434 [My Account](#) [My Wish List](#) [Sign Out](#) Welcome, Emma Lopez!
 2435 Search entire store here... 3
 2436 Advanced Search
 2437 Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -
 2438 Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -
 2439
 2440 One Stop Market
 2441 Product Showcases
 2442   
 2443 Pre-baked Gingerbread House Kit Value Pack, 17 oz., Pack of 2, Total 2
 2444 V8 +Energy, Healthy Energy Drink, Steady Energy from Black and
 2445 Elmwood Inn Fine Teas, Orange Vanilla Caffeine-free Fruit
 2446  
 2447 Belle Of The Ball Princess Sprinkle Mix | Wedding Colorful Light, Vegan, Non-GMO Project
 2448 So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project
 2449 **Analyze and Predict Action:**
 2450 Thought: Let's think step-by-step. The current observation indicates that I am on the One Stop Market homepage. The page contains a vertical menu with various categories, including "Electronics" with ID [911]. To navigate to the "Electronics" section, I need to click on this menu item.
 2451 Next Action: According to the thoughts, the next action I will perform is "click [911]"
 2452 Target: Navigate to the "Electronics" section.
 2453 Summary: The current observation shows that I am on the One Stop Market homepage, and there is an "Electronics" menu item. I will click on this menu item to navigate to the "Electronics" section.
 2454 **Action: click [911] where [911] is menuItem 'ue622 Electronics' hasPopup: menu**
 2455
 2456
 2457
 2458
 2459
 2460 **Figure of Observation (only for visualization, not used in input prompt):**
 2461 [My Account](#) [My Wish List](#) [Sign Out](#) Welcome, Emma Lopez!
 2462 Search entire store here... 3
 2463 Advanced Search
 2464 Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -
 2465 Health & Household - Patio, Lawn & Garden - **Electronics** - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -
 2466
 2467 Home > Electronics
 2468 Electronics
 2469
 2470 Shop By Items 1-12 of 14539 Sort By Position ↑
 2471 Shopping Options
 2472 Category
 2473 Home Audio(1280)
 2474 Video Projectors(484)
 2475 Accessories & Supplies(2409)
 2476 Television & Video(1083)
 2477 Camera & Photo(3512)
 2478 Computers & Accessories(2929)
 2479 **Headphones(631)**
 2480 Portable Audio & Video(824)
 2481
 2482 **Analyze and Predict Action:**
 2483 Thought: Let's think step-by-step. The current observation shows that I am on the "Electronics" page, which lists various items in this category. This means that the "Electronics" menu item has been successfully clicked and I have navigated to the correct page. Therefore, I have completed the subtask.

2484
 2485 Next Action: According to the thoughts, the next action I will perform is ““finish_subtask““
 2486 Target: Confirm that the subtask of locating and clicking on the “Electronics” menu item is
 2487 completed.
 2488 Summary: The current observation shows that I am on the “Electronics” page, which lists
 2489 various items in this category. This confirms that the “Electronics” menu item has been
 2490 successfully clicked and I have navigated to the correct page. Therefore, I will finish the
 2491 subtask.
 2492 **Action:** `finish_subtask []`
 2493

2494 **Verification:**
 2495 `check_in_url("electronics")`
 2496 1.0 “electronics” is in the URL: <http://127.0.0.1:7770/electronics.html>
 2497 **Verification Pass**

2498
 2499 **Subtask 2**
 2500 2. **Locate the “Video Games” subcategory** - Within the “Electronics” section, find and
 2501 click on the “Video Games” subcategory to narrow down the search to video game-related
 2502 products.

2503 **Figure of Observation (only for visualization, not used in input prompt):**



2512 **Analyze and Predict Action:**
 2513 Thought: Let’s think step-by-step. The current observation shows that I am on the “Elec-
 2514 tronics” page, and there is a menu item labeled “Video Games” with ID [3789]. Therefore,
 2515 I will click on this menu item to navigate to the “Video Games” subcategory page.
 2516 Next Action: According to the thoughts, the next action I will perform is ““click [3789]““
 2517 Target: Navigate to the “Video Games” subcategory page.

2518 Summary: The current observation shows that I am on the “Electronics” page, and there is
 2519 a menu item labeled “Video Games” with ID [3789]. Therefore, I will click on this menu
 2520 item to navigate to the “Video Games” subcategory page.
 2521 **Action:** `click [3789] where [3789] is [3789] menuitem 'ue622 Video Games' hasPopup: menu`

2522 **Figure of Observation (only for visualization, not used in input prompt):**

2538

2539

2540

2541

2542

2543

2544

2545

2546

2547

2548

2549

2550

2551

2552

2553

2554

2555

2556

2557

2558

2559

2560

2561

2562

2563

2564

2565

2566

2567

2568

2569

2570

2571

2572

2573

2574

2575

2576

2577

2578

2579

2580

2581

2582

2583

2584

2585

2586

2587

2588

2589

2590

2591

My Account My Wish List Sign Out Welcome, Emma Lopez!

One Stop Market

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

Home > Video Games

Video Games

Shop By

Category

- Xbox One(158)
- PC(181)
- Legacy Systems(293)
- PlayStation 4(233)
- Nintendo Switch(51)

Price

- \$0.00 - \$999.99(930)
- \$1,000.00 and above(2)

Items 1-12 of 932

Sort By Position ↑

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that I am on the "Video Games" page, which lists various items in this subcategory. This means that the "Video Games" menu item has been successfully clicked and I have navigated to the correct page. Therefore, I have completed the subtask.

Next Action: According to the thoughts, the next action I will perform is "finish_subtask"

Target: Confirm that the subtask of locating and clicking on the "Video Games" subcategory is completed.

Summary: The current observation shows that I am on the "Video Games" page, which lists various items in this subcategory. This confirms that the "Video Games" menu item has been successfully clicked and I have navigated to the correct page. Therefore, I will finish the subtask.

Action: finish_subtask []

Verification:

```
check_in_url("video-games")
check_in_url("video-games")
```

1.0 "video-games" is in the URL: http://127.0.0.1:7770/video-games.html

Verification Pass

Subtask 3

3. **Search for "Nintendo Switch game card storage"** - Use the search bar within the "Video Games" section to search for "Nintendo Switch game card storage".

Figure of Observation (only for visualization, not used in input prompt):

2592
 2593
 2594
 2595
 2596
 2597
 2598
 2599
 2600
 2601
 2602
 2603
 2604
 2605
 2606
 2607
 2608
 2609
 2610
 2611
 2612
 2613
 2614
 2615
 2616
 2617
 2618
 2619
 2620
 2621
 2622
 2623
 2624
 2625
 2626
 2627
 2628
 2629
 2630
 2631
 2632
 2633
 2634
 2635
 2636
 2637
 2638
 2639
 2640
 2641
 2642
 2643
 2644
 2645

My Account My Wish List Sign Out Welcome, Emma Lopez!

One Stop Market

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - **Video Games** - Grocery & Gourmet Food -

Home > Video Games

Video Games

Shop By Items 1-12 of 932 Sort By Position ↑

Category

- Xbox One(158)
- PC(181)
- Legacy Systems(293)
- PlayStation 4(233)
- Nintendo Switch(51)

Price

- \$0.00 - \$999.99(930)
- \$1,000.00 and above(2)



Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that I am in the "Video Games" section on the One Stop Market website. The search bar is present but the search button is currently disabled. To search for "Nintendo Switch game card storage," I need to type the query into the search bar and press Enter.

Next Action: According to the thoughts, the next action I will perform is "type [7445] [Nintendo Switch game card storage] [clear_previous_content=1] [press_enter_after=1]"

Target: After typing "Nintendo Switch game card storage" into the search bar and pressing Enter, the search results for "Nintendo Switch game card storage" should be displayed.

Summary: The current observation shows that I am in the "Video Games" section on the One Stop Market website. The next step is to use the search bar to search for "Nintendo Switch game card storage." The expected outcome is to display the search results for "Nintendo Switch game card storage."

Action: type [7445] [Nintendo Switch game card storage] where [7445] is combobox 'ue615 Search' autocomplete: both hasPopup: listbox required: False expanded: False

Figure of Observation (only for visualization, not used in input prompt):

Nintendo Switch game carc Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

Home > Search results for: 'Nintendo Switch game card storage'

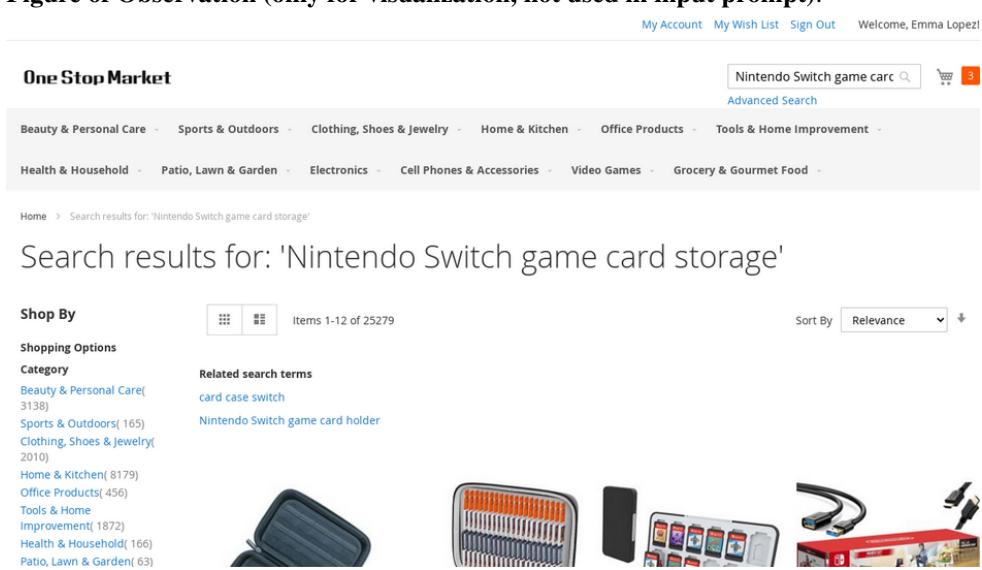
Search results for: 'Nintendo Switch game card storage'

Shop By Items 1-12 of 25279 Sort By Relevance ↑

Related search terms

- card case switch
- Nintendo Switch game card holder



2646
 2647 **Analyze and Predict Action:**
 2648 Thought: Let's think step-by-step. The current observation shows the search results page for
 2649 "Nintendo Switch game card storage" on the One Stop Market website. This indicates that
 2650 the search function has already been navigated to and the search has been performed. There-
 2651 fore, the subtask to search for "Nintendo Switch game card storage" has been completed.
 2652 Next Action: According to the thoughts, the next action I will perform is "finish_subtask".
 2653 Target: Confirm that the search function has been navigated to and the search results for
 2654 "Nintendo Switch game card storage" are displayed.
 2655 Summary: The current observation shows the search results page for "Nintendo Switch
 2656 game card storage" on the One Stop Market website. This indicates that the search function
 2657 has already been navigated to and the search has been performed. Therefore, the subtask to
 2658 search for "Nintendo Switch game card storage" has been completed.
 2659 **Action:** finish_subtask []
 2660
 2661
 2662 **Verification:**
 2663 **check_in_webpage("Search bar is populated with 'Nintendo Switch game card stor-
 age'")**
 2664 1.0 I need to verify if the search bar contains the text 'Nintendo Switch game card storage'.
 2665 The relevant information should be found in the description of the search bar or its associated
 2666 elements.
 2667 **Verification Pass**
 2668
 2669 **Subtask 4**
 2670 4. **Identify the best storage option for 11 game cards** - Review the search results to find
 2671 a storage option that can fit at least 11 Nintendo Switch game cards.
 2672
 2673 **Figure of Observation (only for visualization, not used in input prompt):**
 2674
 2675 
 2676
 2677 **Analyze and Predict Action:**
 2678 Thought: Let's think step-by-step. The current observation shows the search results page for
 2679 "Nintendo Switch game card storage" on the One Stop Market website. To identify the
 2680 best storage option for 11 game cards, I need to review the search results and find a storage
 2681 option that can fit at least 11 Nintendo Switch game cards. The next step is to scroll down
 2682 to view more details about the search results.
 2683 Next Action: According to the thoughts, the next action I will perform is "scroll [down]".
 2684 Target: View more details about the storage options, including their capacities.
 2685
 2686
 2687
 2688
 2689
 2690
 2691
 2692
 2693
 2694
 2695
 2696
 2697
 2698
 2699

2700

2701 Summary: The current observation shows the search results page for "Nintendo Switch game
 2702 card storage" on the One Stop Market website. I will scroll down to view more details about
 2703 the storage options.

2704 **Action:** scroll [down]

2705

2706 **Figure of Observation (only for visualization, not used in input prompt):**

2707

2708

2709

2710

2711

2712

2713

2714

2715

2716

2717

2718

2719

2720

2721

2722

2723

2724

2725

2726

2727

2728

2729

2730

2731

2732

2733

2734

2735

2736

2737

2738

2739

2740

2741

2742

2743

2744

2745

2746

2747

2748

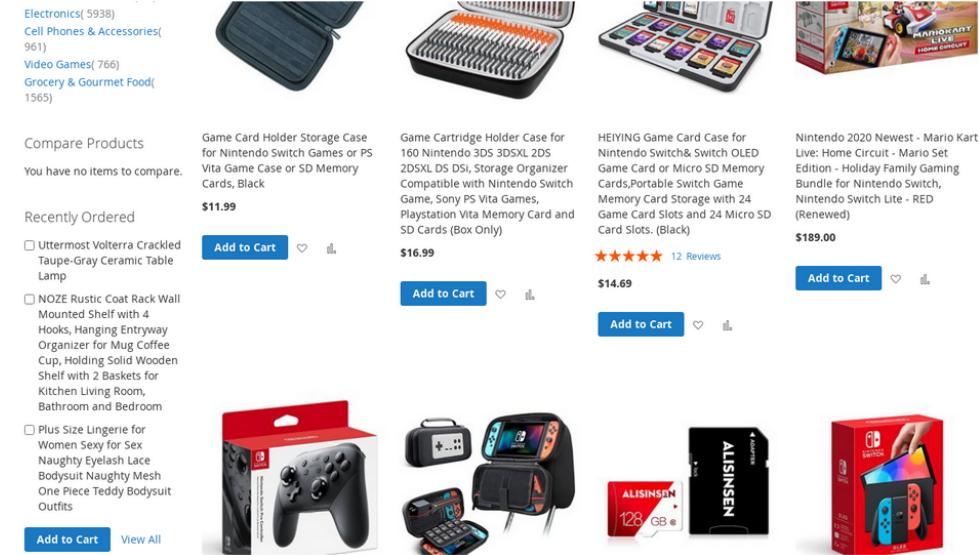
2749

2750

2751

2752

2753



Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows the search results for "Nintendo Switch game card storage" on the One Stop Market website. I need to identify a storage option that can fit at least 11 Nintendo Switch game cards. From the search results, I can see the following options:

1. **Game Card Holder Storage Case for Nintendo Switch Games or PS Vita Game Case or SD Memory Cards, Black** - \$11.99
2. **Game Cartridge Holder Case for 160 Nintendo 3DS 3DSXL 2DS 2DSXL DS DSI, Storage Organizer Compatible with Nintendo Switch Game, Sony PS Vita Games, Playstation Vita Memory Card and SD Cards (Box Only)** - \$16.99
3. **HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black)** - \$14.69

The third option, "HEIYING Game Card Case," has 24 game card slots, which is sufficient for storing 11 game cards. This makes it the best storage option among the listed products.

Next Action: According to the thoughts, the next action I will perform is "finish_subtask [The best storage option for 11 game cards is the HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black) priced at \$14.69]"

Target: Confirm that the subtask of identifying the best storage option for 11 game cards has been completed.

Summary: The current observation shows the search results for "Nintendo Switch game card storage." The best storage option for 11 game cards is the "HEIYING

Action: finish_subtask [The best storage option for 11 game cards is the HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black) priced at \$14.69]

Verification:

2754
 2755 **Objective: check_in_webpage("Storage option for 11 Nintendo Switch game cards")**
 2756 0.0 The search results provide various storage options for Nintendo Switch game cards. I
 2757 need to verify if any of these options specifically mention the capacity to store 11 game
 2758 cards.

2759 **Verification Fail**

2760 **Reflection**

2761 Reflection: The execution process involved scrolling down to view more details about the
 2762 storage options and then selecting the "HEIYING Game Card Case" as the best option for 11
 2763 game cards. However, the verification result indicates that the content needs to be checked
 2764 to ensure that the selected storage option specifically mentions the capacity to store 11 game
 2765 cards.

2766 The mistake lies in not explicitly verifying that the selected storage option can fit at least 11
 2767 game cards. The selected option has 24 game card slots, which meets the requirement, but
 2768 this detail was not clearly verified in the process.

2769 To solve the task, I should include a step to explicitly verify that the selected storage option
 2770 can fit at least 11 game cards by checking the product details.

2771 Since the mistake occurred after selecting the storage option but before verifying its capacity,
 2772 I should reload from the point just before finishing the subtask to include the verification
 2773 step.

2774 **Reload checkpoint: 1**

2775 **Figure of Observation (only for visualization, not used in input prompt):**

My Account My Wish List Sign Out Welcome, Emma Lopez!

One Stop Market

Nintendo Switch game card Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

Home > Search results for: 'Nintendo Switch game card storage'

Search results for: 'Nintendo Switch game card storage'

Shop By

Items 1-12 of 25279

Sort By Relevance

Related search terms: card case switch, Nintendo Switch game card holder

Category: Beauty & Personal Care(3138), Sports & Outdoors(165), Clothing, Shoes & Jewelry(2010), Home & Kitchen(8179), Office Products(456), Tools & Home Improvement(1872), Health & Household(166), Patio, Lawn & Garden(63)

2795 **Analyze and Predict Action:**

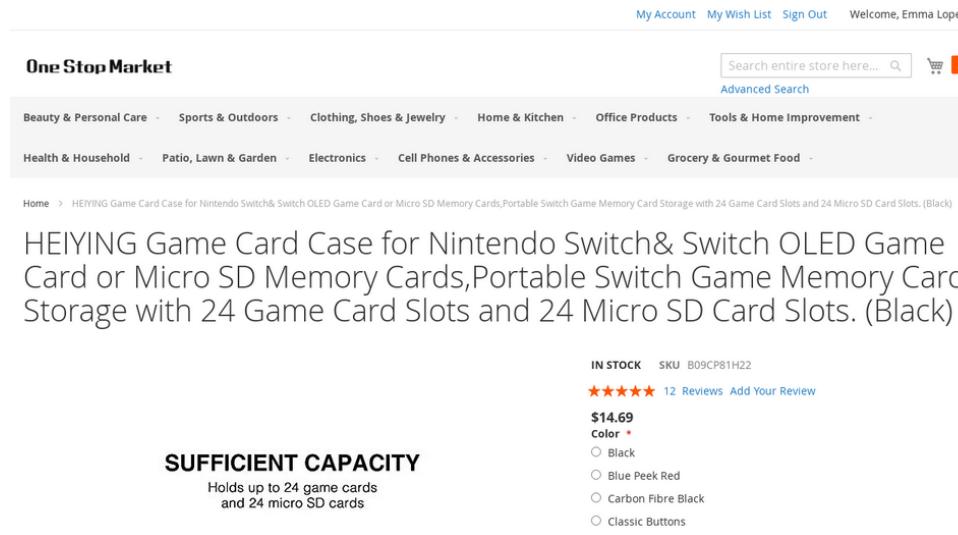
2796 Thought: Let's think step-by-step. The current observation shows the search results for
 2797 "Nintendo Switch game card storage" on the One Stop Market website. I need to identify a
 2798 storage option that can fit at least 11 Nintendo Switch game cards. From the search results,
 2799 I can see the following options:

- 2800 1. **Game Card Holder Storage Case for Nintendo Switch Games or PS Vita Game Case or SD Memory Cards, Black** - \$11.99
- 2801 2. **Game Cartridge Holder Case for 160 Nintendo 3DS 3DSXL 2DS 2DSXL DS DSi, Storage Organizer Compatible with Nintendo Switch Game, Sony PS Vita Games, Playstation Vita Memory Card and SD Cards (Box Only)** - \$16.99
- 2802 3. **HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black)** - \$14.69

2803
 2804
 2805
 2806
 2807

2808
 2809 The third option, "HEIYING Game Card Case," has 24 game card slots, which is sufficient
 2810 for storing 11 game cards. This makes it the best storage option among the listed products.
 2811 Next Action: According to the thoughts, the next action I will perform is "click [14241]"
 2812 Target: Navigate to the product page of the "HEIYING Game Card Case for Nintendo
 2813 Switch & Switch OLED Game Card or Micro SD Memory Cards,Portable Switch Game
 2814 Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black)"
 2815 Summary: The current observation shows the search results for "Nintendo Switch game
 2816 card storage" on the One Stop Market website. I will click on the link for the "HEIYING
 2817 Game Card Case" to navigate to its product page and verify its details.
 2818 **Action: click [14241] where [14241] is link 'HEIYING Game Card Case for Nintendo**
 2819 **Switch& Switch OLED Game Card or Micro SD Memory Cards,Portable Switch**
 2820 **Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots.**
 2821 **(Black)'**

Figure of Observation (only for visualization, not used in input prompt):



Analyze and Predict Action:

Thought: Let's think step-by-step. The current page provides detailed information about the "HEIYING Game Card Case for Nintendo Switch& Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots." This confirms that the product meets the requirement of fitting at least 11 game cards, as it has 24 game card slots. I will now finish the subtask.

Next Action: According to the thoughts, the next action I will perform is "finish_subtask [The best storage option for 11 game cards is the HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black) priced at \$14.69]"

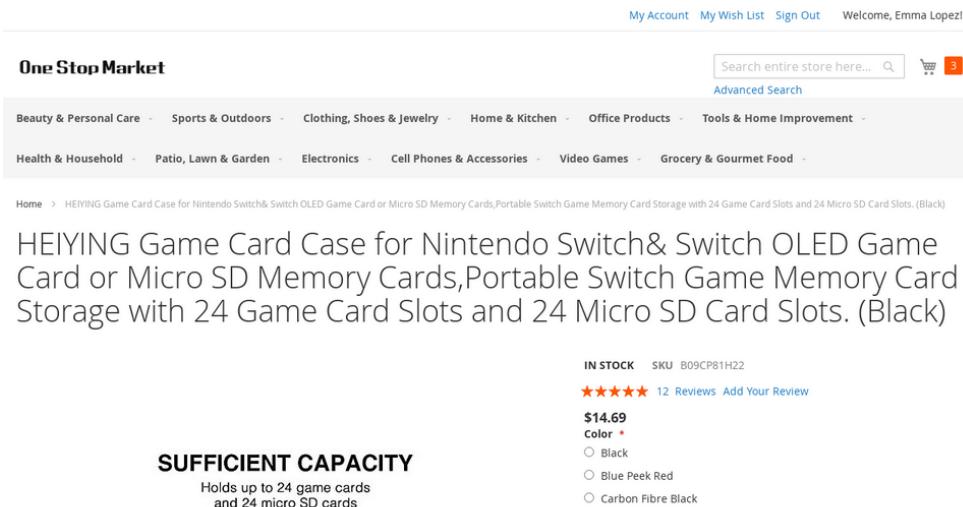
Target: Confirm that the subtask of identifying the best storage option for 11 game cards has been completed.

Summary: The current observation confirms that the "HEIYING Game Card Case" has 24 game card slots, making it a suitable option for storing at least 11 Nintendo Switch game cards. Therefore, I will finish the subtask.

Action: finish_subtask [The best storage option for 11 game cards is the HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards, Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black) priced at \$14.69]

2862
 2863 **Verification:**
 2864 **Objective: check_in_webpage("Storage option for 11 Nintendo Switch game cards")**
 2865 1.0 The content describes a game card case for the Nintendo Switch that has 24 game card
 2866 slots and 24 micro SD card slots. The subgoal target is to find a storage option for 11
 2867 Nintendo Switch game cards.
 2868 **Verification Pass**

2869 **Subtask 5**
 2870 5. **Navigate to the product page of the identified storage option** - Click on the identified
 2871 storage option to navigate to its product page for more details.
 2872 **Figure of Observation (only for visualization, not used in input prompt):**

2873 

2874
 2875 **Analyze and Predict Action:**
 2876 Thought: Let's think step-by-step. The current observation indicates that I am already on
 2877 the product page for the "HEIYING Game Card Case for Nintendo Switch& Switch OLED Game
 2878 Card or Micro SD Memory Cards,Portable Switch Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots. (Black)". Therefore, the subtask to
 2879 navigate to the product page has already been completed.
 2880 Next Action: According to the thoughts, the next action I will perform is "finish_subtask".
 2881 Target: Confirm that I have navigated to the product page of the identified storage option.
 2882 Summary: I am already on the product page for the "HEIYING Game Card Case for Ninten-
 2883 do Switch& Switch OLED Game Card or Micro SD Memory Cards,Portable Switch
 2884 Game Memory Card Storage with 24 Game Card Slots and 24 Micro SD Card Slots.
 2885 (Black)". I will now confirm that this subtask is complete.
 2886 **Action: finish_subtask []**

2887
 2888 **Verification:**
 2889 check_in_url("nintendo-switch-game-card-storage") |OR|check_in_webpage("You are on
 2890 the product page of the identified storage option")
 2891 **check_in_url("nintendo-switch-game-card-storage")**
 2892 1.0 "nintendo-switch-game-card-storage" is in the URL: http://127.0.0.1:7770/heiyng-
 2893 game-card-case-for-nintendo-switch-switch-oled-game-card-or-micro-sd-memory-cards-
 2894 portable-switch-game-memory-card-storage-with-24-game-card-slots-and-24-micro-sd-
 2895 card-slots-black.html
 2896 **check_in_webpage("You are on the product page of the identified storage option")**
 2897 1.0 The content provided includes detailed information about a product, specifically the
 2898 "HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro
 2899 SD Memory Cards." This indicates that I am indeed on a product page for a storage option.

2916

2917

2918

2919

Verification Pass

2920

Answer:

2921

The best storage option for your 11 Nintendo Switch game cards is the **HEIYING Game Card Case for Nintendo Switch & Switch OLED Game Card or Micro SD Memory Cards**. This case offers 24 game card slots, which is more than sufficient for your needs. It is priced at \$14.69 and is available in multiple colors. You can find it on the One Stop Market website. **Last URL:** <http://127.0.0.1:7770/heiying-game-card-case-for-nintendo-switch-switch-oled-game-card-or-micro-sd-memory-cards-portable-switch-game-memory-card-storage-with-24-game-card-slots-and-24-micro-sd-card-slots-black.html>

2922

2923

2924

2925

2926

2927

2928

2929

2930

2931

2932

2933

2934

2935

2936

2937

C.2 REPLANNING

2938

2939

2940

2941

2942

2943

One Complete Example

2944

Query Intent:

2945

Show me the billing address for order number 00178.

2946

Ground Truth

2947

Final Answer: 'must_include': ['101 S San Mateo Dr', 'San Mateo', 'California', '94010', 'United States']

2948

2949

Planning

2950

1. **Navigate to the "My Account" section** - Click on the "My Account" link.

2951

- Objective: 'check_in_url("account")' |OR| check_in_webpage("You have arrived at the 'My Account' page")'

2952

2. **Locate and access the "Order History" or "Orders" section** - Find and click on the link or tab that leads to order history or orders.

2953

- Objective: 'check_in_url("order-history")' |OR| check_in_webpage("Order History")'

2954

3. **Search for order number 00178 in the list of orders** - Look for order number 00178 in the order history list.

2955

- Objective: 'check_in_webpage("Order number 00178")'

2956

4. **Access the details of order number 00178** - Click on the order number 00178 to view its details.

2957

- Objective: 'check_in_webpage("Order details for 00178")'

2958

5. **Identify and display the billing address for order number 00178** - Locate the billing address section within the order details and display it.

2959

- Objective: 'check_in_webpage("Billing address for order number 00178")'

2960

Subtask 1

2961

1. **Navigate to the "My Account" section** - Click on the "My Account" link.

2962

2963

2964

Figure of Observation (only for visualization, not used in input prompt):

2970
 2971
 2972
 2973
 2974
 2975
 2976
 2977
 2978
 2979
 2980
 2981
 2982
 2983
 2984
 2985
 2986
 2987
 2988

One Stop Market

My Account My Wish List 1 item Sign Out Welcome, Emma Lopez!

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

One Stop Market

Product Showcases

Pre-baked Gingerbread House Kit Value Pack, 17 oz., Pack of 2, Total V8 +Energy, Healthy Energy Drink, Steady Energy from Black and Elmwood Inn Fine Teas, Orange Vanilla Caffeine-free Fruit Belle Of The Ball Princess Sprinkle Mix, Wedding Colorful So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project

Analyze and Predict Action:
 Thought: Let's think step-by-step. The current observation indicates that I am on the One Stop Market homepage. The page contains a 'My Account' link with ID [1317]. To navigate to the "My Account" page, I need to click on this link.
 Next Action: According to the thoughts, the next action I will perform is "click [1317]"
 Target: Navigate to the "My Account" page.
 Summary: The current observation shows that I am on the One Stop Market homepage, and there is a 'My Account' link. I will click on this link to navigate to the "My Account" page.
Action: click [1317] where [1317] is [1317] link 'My Account'

2999
 3000
 3001
 3002
 3003
 3004
 3005
 3006
 3007
 3008
 3009
 3010
 3011
 3012
 3013
 3014
 3015
 3016
 3017
 3018
 3019
 3020
 3021
 3022
 3023

Figure of Observation (only for visualization, not used in input prompt):

My Account My Wish List 1 item Sign Out Welcome, Emma Lopez!

One Stop Market

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

My Account

My Orders
My Downloadable Products
My Wish List

Address Book
Account Information
Stored Payment Methods

My Product Reviews
Newsletter Subscriptions

Compare Products

My Account

Account Information

Contact Information
Emma Lopez
emma.lopez@gmail.com
[Edit Change Password](#)

Newsletters
You aren't subscribed to our newsletter.
[Edit](#)

Address Book [Manage Addresses](#)

Default Billing Address
Emma Lopez
222 Redwood Rise, Suite 300
San Mateo, California, 94010
United States

Default Shipping Address
Emma Lopez
222 Redwood Rise, Suite 300
San Mateo, California, 94010
United States

Analyze and Predict Action:
 Thought: The current observation indicates that I have successfully navigated to the "My Account" page. The page contains various elements such as account information, contact information, address book, and links to other account-related sections. Since the objective was to navigate to the "My Account" page, I have achieved this subtask.
 Next Action: According to the thoughts, the next action I will perform is "finish_subtask"

3024
 3025 Target: Confirm that the subtask of navigating to the "My Account" page is complete.
 3026 Summary: The current observation shows that I am on the "My Account" page, which
 3027 contains various account-related information and links. Therefore, I will finish the subtask
 3028 as it has been successfully completed.

3029 **Action:** `finish_subtask []`

3030

3031 **Verification:**

3032 **'check_in_url("account")'**

3033 1.0 "account" is in the URL: <http://127.0.0.1:7770/customer/account/>

3034 **check_in_webpage("You have arrived at the 'My Account' page")**

3035 1.0 The content of the webpage includes multiple elements that are typically found on a 'My
 3036 Account' page, such as account information, contact information, address book, and links to
 3037 manage various account-related settings. Additionally, the heading 'My Account' is present,
 3038 which confirms the page's identity.

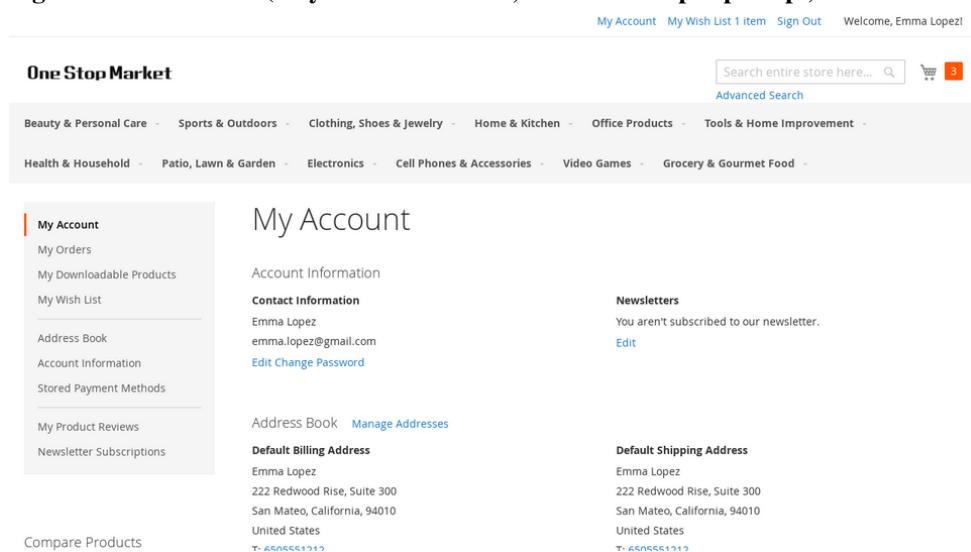
3039 **Verification Pass**

3040

3041 **Subtask 2**

3042 **Locate and access the "Order History" or "Orders" section** - Find and click on the link
 3043 or tab that leads to order history or orders.

3044 **Figure of Observation (only for visualization, not used in input prompt):**



3052 **Analyze and Predict Action:**

3053 Thought: The current observation shows that I am on the "My Account" page. There is a
 3054 link labeled "My Orders" with ID [4527]. Clicking on this link should navigate to the "Order
 3055 History" or "Orders" section.

3056 Next Action: According to the thoughts, the next action I will perform is "click [4527]"
 3057 Target: Navigate to the "Order History" or "Orders" section.

3058 Summary: The current observation shows that I am on the "My Account" page, and there
 3059 is a link labeled "My Orders" with ID [4527]. Clicking on this link should navigate to the
 3060 "Order History" or "Orders" section.

3061 **Action:** `click [4527] where [4527] is link 'My Orders'`

3062

3063 **Figure of Observation (only for visualization, not used in input prompt):**

3078
 3079
 3080
 3081
 3082
 3083
 3084
 3085
 3086
 3087
 3088
 3089
 3090
 3091
 3092
 3093
 3094
 3095
 3096

[My Account](#) [My Wish List](#) [1 item](#) [Sign Out](#) Welcome, Emma Lopez! 3

One Stop Market

Advanced Search

[Beauty & Personal Care](#) - [Sports & Outdoors](#) - [Clothing, Shoes & Jewelry](#) - [Home & Kitchen](#) - [Office Products](#) - [Tools & Home Improvement](#) -

[Health & Household](#) - [Patio, Lawn & Garden](#) - [Electronics](#) - [Cell Phones & Accessories](#) - [Video Games](#) - [Grocery & Gourmet Food](#) -

My Orders

Order #	Date	Order Total	Status	Action
000000170	5/17/23	\$365.42	Canceled	View Order Reorder
000000189	5/2/23	\$754.99	Pending	View Order Reorder
000000188	5/2/23	\$2,004.99	Pending	View Order Reorder
000000187	5/2/23	\$1,004.99	Pending	View Order Reorder
000000180	3/11/23	\$65.32	Complete	View Order Reorder
000000166	3/10/23	\$17.99	Complete	View Order Reorder
000000161	2/27/23	\$762.18	Complete	View Order Reorder
000000156	2/24/23	\$231.54	Canceled	View Order Reorder
000000158	2/11/23	\$174.99	Canceled	View Order Reorder

Compare Products

Analyze and Predict Action:
 Thought: Let's think step-by-step. The current observation shows that I am now in the "My Orders" section within the "My Account" page. The heading "My Orders" and the table listing various orders confirm this. Therefore, the subtask of locating and accessing the "Order History" or "Orders" section has been completed.
 Next Action: According to the thoughts, the next action I will perform is "finish_subtask"
 Target: Confirm that the "Order History" or "Orders" section has been located and accessed within the "My Account" page.
 Summary: The current observation confirms that we are in the "My Orders" section within the "My Account" page, as indicated by the heading "My Orders" and the table listing various orders. Therefore, the subtask of locating and accessing the "Order History" or "Orders" section has been completed.

Action: `finish_subtask []`

Verification:
 Objective: `check_in_url("search")` |OR|`check_in_webpage("Search results for 'Canon photo printer' displayed")`
`check_in_url("order-history")`
 1.0 "order-history" is in the URL: <http://127.0.0.1:7770/sales/order/history/>
`check_in_webpage("Order History")`
 1.0 The content provided is a detailed view of the "My Orders" page, which includes a table listing various orders with their respective order numbers, dates, totals, statuses, and actions. This matches the expected content for an "Order History" page
Verification Pass

Subtask 3
 3. **Search for order number 00178 in the list of orders** - Look for order number 00178 in the order history list.

Figure of Observation (only for visualization, not used in input prompt):

3132
 3133
 3134
 3135
 3136
 3137
 3138
 3139
 3140
 3141
 3142
 3143
 3144
 3145
 3146
 3147
 3148
 3149
 3150

One Stop Market

My Account My Wish List 1 item Sign Out Welcome, Emma Lopez!

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

My Account
My Orders
 My Downloadable Products
 My Wish List
 Address Book
 Account Information
 Stored Payment Methods
 My Product Reviews
 Newsletter Subscriptions

My Orders

Order #	Date	Order Total	Status	Action
000000170	5/17/23	\$365.42	Canceled	View Order Reorder
000000189	5/2/23	\$754.99	Pending	View Order Reorder
000000188	5/2/23	\$2,004.99	Pending	View Order Reorder
000000187	5/2/23	\$1,004.99	Pending	View Order Reorder
000000180	3/11/23	\$65.32	Complete	View Order Reorder
000000166	3/10/23	\$17.99	Complete	View Order Reorder
000000161	2/27/23	\$762.18	Complete	View Order Reorder
000000156	2/24/23	\$231.54	Canceled	View Order Reorder
000000158	2/11/23	\$174.99	Canceled	View Order Reorder

Compare Products

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that I am on the "My Orders" page, and the list of orders is displayed. However, order number "00178" is not visible in the current view. There are multiple pages of orders, and I need to scroll down to check if it is listed further down or navigate to subsequent pages if necessary.

Next Action: According to the thoughts, the next action I will perform is "scroll [down]"

Target: Scroll down the "My Orders" page to check if order number "00178" is listed further down.

Summary: The current observation shows that order number "00178" is not visible in the current view of the "My Orders" page. I will scroll down to check if it is listed further down. The expected outcome of this action is to see more orders listed on the page.

Action: scroll [down]

3161
 3162
 3163
 3164
 3165
 3166
 3167
 3168
 3169
 3170
 3171
 3172
 3173
 3174
 3175
 3176
 3177
 3178
 3179
 3180
 3181
 3182
 3183
 3184
 3185

Figure of Observation (only for visualization, not used in input prompt):

You have no items to compare.

Recently Ordered

- Plus Size Lingerie for Women Sexy for Sex Naughty Eyelash Lace Bodysuit Naughty Mesh One Piece Teddy Bodysuit Outfits
- Uttermost Volterra Crackled Taupe-Gray Ceramic Table Lamp
- NOZE Rustic Coat Rack Wall Mounted Shelf with 4 Hooks, Hanging Entryway Organizer for Mug Coffee Cup, Holding Solid Wooden Shelf with 2 Baskets for Kitchen Living Room, Bathroom and Bedroom

Add to Cart View All

My Wish List 1 item

Baking Soda Whitening Toothpaste,Intensive Stain Removal Whitening Toothpaste Strengthening Toothpastes Strong Cleaning Power Natural Stain Remover Fluoride-Free Toothpaste(Blueberry+Passion Fruit)
\$19.86

Add to Cart

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that order number "00178" is not visible in the current view of the "My Orders" page. There are multiple

3186
 3187 pages of orders, and I need to navigate to the next page to check if order number "00178" is
 3188 listed there. I will click on the link for "Page 2" to navigate to the next page of orders.
 3189 Next Action: According to the thoughts, the next action I will perform is "click [7801]" "
 3190 Target: Navigate to the next page of orders to locate order number "00178".
 3191 Summary: The current observation shows that order number "00178" is not visible on the
 3192 current page of "My Orders". There are multiple pages of orders, and I will click on the link
 3193 for "Page 2" to navigate to the next page and check if order number "00178" is listed there.
 3194 The expected outcome of this action is to navigate to the next page of orders.
 3195 **Action: click [7801] where [7801] is link 'Page 2'**
 3196

3197 **Figure of Observation (only for visualization, not used in input prompt):**

Order #	Date	Order Total	Status	Action
000000148	1/29/23	\$440.64	Complete	View Order Reorder
000000163	1/16/23	\$132.24	Complete	View Order Reorder
000000154	12/18/22	\$97.15	Complete	View Order Reorder
000000184	12/14/22	\$20.49	Complete	View Order Reorder
000000162	12/12/22	\$53.29	Complete	View Order Reorder
000000174	12/4/22	\$32.47	Complete	View Order Reorder
000000164	11/26/22	\$218.17	Complete	View Order Reorder
000000171	11/20/22	\$133.07	Complete	View Order Reorder
000000183	11/11/22	\$51.94	Complete	View Order Reorder

3215 **Analyze and Predict Action:**

3216 Thought: Let's think step-by-step. The current observation shows that I am on the "My Orders" page, specifically on Page 2. The table lists various orders, but order number "00178" 3217 is not visible on this page. There are multiple pages of orders, and I need to navigate to the 3218 next page to check if order number "00178" is listed there. I will click on the link for "Page 3" to navigate to the next page of orders.
 3219

3220 Next Action: According to the thoughts, the next action I will perform is "click [7802]" "
 3221 Target: Navigate to the next page of orders to locate order number "00178".
 3222

3223 Summary: The current observation shows that order number "00178" is not visible on the
 3224 current page of "My Orders". There are multiple pages of orders, and I will click on the link
 3225 for "Page 3" to navigate to the next page and check if order number "00178" is listed there.
 3226 The expected outcome of this action is to navigate to the next page of orders.
 3227 **Action: click [7802] where [7802] is No match found**
 3228

3229 **Figure of Observation (only for visualization, not used in input prompt):**

3240
 3241
 3242
 3243
 3244
 3245
 3246
 3247
 3248
 3249
 3250
 3251
 3252
 3253
 3254
 3255
 3256
 3257
 3258
 3259
 3260
 3261
 3262
 3263
 3264
 3265
 3266
 3267
 3268
 3269
 3270
 3271
 3272
 3273
 3274
 3275
 3276
 3277
 3278
 3279
 3280
 3281
 3282
 3283
 3284
 3285
 3286
 3287
 3288
 3289
 3290
 3291
 3292
 3293

My Account My Wish List 1 item Sign Out Welcome, Emma Lopez!

One Stop Market

Search entire store here... 3

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

My Account

- My Orders**
- My Downloadable Products
- My Wish List
- Address Book
- Account Information
- Stored Payment Methods
- My Product Reviews
- Newsletter Subscriptions

My Orders

Order #	Date	Order Total	Status	Action
000000148	1/29/23	\$440.64	Complete	View Order Reorder
000000163	1/16/23	\$132.24	Complete	View Order Reorder
000000154	12/18/22	\$97.15	Complete	View Order Reorder
000000184	12/14/22	\$20.49	Complete	View Order Reorder
000000162	12/12/22	\$53.29	Complete	View Order Reorder
000000174	12/4/22	\$32.47	Complete	View Order Reorder
000000164	11/26/22	\$218.17	Complete	View Order Reorder
000000171	11/20/22	\$133.07	Complete	View Order Reorder
000000183	11/11/22	\$51.94	Complete	View Order Reorder

Compare Products

Analyze and Predict Action:

Thought: Let's think step-by-step. The current page is "My Orders" and it shows a list of orders. The order number "00178" is not visible in the current view. The previous attempts to navigate to the next pages failed because the elements with IDs [7802] were not found. Since the current page is "My Orders" and it shows a list of orders, I will scroll down to check if the order number "00178" is listed further down.

Next Action: According to the thoughts, the next action I will perform is "scroll [down]"

Target: Scroll down the page to locate the order number "00178". Summary: The current observation shows that order number "00178" is not visible in the current view of the "My Orders" page. I will scroll down to check if it is listed further down. The expected outcome of this action is to see more orders listed on the page.

Action: scroll [down]

Figure of Observation (only for visualization, not used in input prompt):

You have no items to compare.

Recently Ordered

- Plus Size Lingerie for Women Sexy for Sex Naughty Eyelash Lace Bodysuit Naughty Mesh One Piece Teddy Bodysuit Outfit
- Uttermost Volterra Crackled Taupe-Gray Ceramic Table Lamp
- NOZE Rustic Coat Rack Wall Mounted Shelf with 4 Hooks, Hanging Entryway Organizer for Mug Coffee Cup, Holding Solid Wooden Shelf with 2 Baskets for Kitchen Living Room, Bathroom and Bedroom

[Add to Cart](#) [View All](#)

My Wish List 1 item

Baking Soda Whitening Toothpaste,Intensive Stain Removal Whitening

Toothpastes Strengthening Toothpastes Strong Cleaning Power Natural Stain Remover Fluoride-Free Toothpaste(Blueberry+Passion Fruit)

\$19.86

[Add to Cart](#)

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that order number "00178" is not visible on the current page of "My Orders". There are multiple pages of

3294
 3295 orders, and I will click on the link for "Page 3" to navigate to the next page and check if
 3296 order number "00178" is listed there. The expected outcome of this action is to navigate to
 3297 the next page of orders.
 3298 Next Action: According to the thoughts, the next action I will perform is "click [10365]"
 3299 Target: Navigate to Page 3 of the order history to check for order number "00178".
 3300 Summary: The current observation shows that order number "00178" is not visible on the
 3301 current page of "My Orders". There are multiple pages of orders, and I will click on the link
 3302 for "Page 3" to navigate to the next page and check if order number "00178" is listed there.
 3303 The expected outcome of this action is to navigate to the next page of orders.
 3304 **Action: click [10365] where [10365] is [10365] link 'Page 3'**
 3305
 3306 Reached step limit within one subtask. Enter verification.
 3307 **Verification:**
 3308 **check_in_webpage("Order number 00178")**
 3309 0.0 I need to verify if the order number "00178" is present in the content provided. I will
 3310 look for the specific order number in the table of orders.
 3311 **Verification Fail**
 3312
 3313 **Replanning**
 3314 The error in the previous execution was due to the discrepancy between the order number
 3315 formats "00178" and "000000178." The verification result indicates that the order number
 3316 "000000178" should be confirmed as the target order number "00178" despite the difference
 3317 in format. To resolve this, the new subtasks must include steps to verify the order number
 3318 and then proceed to identify and display the billing address.
 3319 New Subtasks: 3. Verify that order number "000000178" corresponds to the target order
 3320 number "00178" despite the difference in format. Objective: check_in_webpage("Order
 3321 number 00178")
 3322 4. Access the details of order number "000000178." Objective: check_in_webpage("Order
 3323 000000178")
 3324 5. Identify and display the billing address for order number "000000178." Objective:
 3325 check_in_webpage("Billing Address")
 3326 These new subtasks should ensure that the order number is verified correctly and that the
 3327 billing address for the order is identified and displayed. This should resolve the previous
 3328 issues and achieve the final objective.
 3329
 3330 **New Subtask 3**
 3331 3. Verify that order number "000000178" corresponds to the target order number "00178"
 3332 despite the difference in format.
 3333 **Figure of Observation (only for visualization, not used in input prompt):**
 3334
 3335
 3336
 3337
 3338
 3339
 3340
 3341
 3342
 3343
 3344
 3345
 3346
 3347

3348
 3349
 3350
 3351
 3352
 3353
 3354
 3355
 3356
 3357
 3358
 3359
 3360
 3361
 3362
 3363
 3364
 3365
 3366
 3367
 3368
 3369
 3370
 3371
 3372
 3373
 3374
 3375
 3376
 3377
 3378
 3379
 3380
 3381
 3382
 3383
 3384
 3385
 3386
 3387
 3388
 3389
 3390
 3391
 3392
 3393
 3394
 3395
 3396
 3397
 3398
 3399
 3400
 3401

My Account My Wish List 1 item Sign Out Welcome, Emma Lopez!

One Stop Market

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

[My Account](#)

My Orders

- My Downloadable Products
- My Wish List
- Address Book
- Account Information
- Stored Payment Methods
- My Product Reviews
- Newsletter Subscriptions

[Compare Products](#)

My Orders

Order #	Date	Order Total	Status	Action
000000178	10/21/22	\$345.84	Complete	View Order Reorder
000000177	10/18/22	\$2,126.32	Canceled	View Order Reorder
000000185	10/3/22	\$18.99	Complete	View Order Reorder
000000179	9/29/22	\$2,890.53	Complete	View Order Reorder
000000175	9/1/22	\$133.85	Complete	View Order Reorder
000000172	8/12/22	\$77.66	Canceled	View Order Reorder
000000155	8/10/22	\$38.99	Complete	View Order Reorder
000000165	8/4/22	\$36.99	Complete	View Order Reorder
000000149	7/25/22	\$354.66	Canceled	View Order Reorder

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that we are on the "My Orders" page, and order number "000000178" is visible in the list. To access the details of this order, I need to click on the "View Order" link associated with order number "000000178".

Next Action: According to the thoughts, the next action I will perform is "click [12075]"

Target: Access the details of order number "000000178". Summary: The current observation shows that order number "000000178" is visible in the list on the "My Orders" page. I will click on the "View Order" link associated with this order to access its details. The expected outcome of this action is to navigate to the order details page for order number "000000178".

Action: click [12075] where [12075] is link 'View Order'

Figure of Observation (only for visualization, not used in input prompt):

My Account My Wish List 1 item Sign Out Welcome, Emma Lopez!

One Stop Market

Search entire store here... Advanced Search

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

[My Account](#)

My Orders

- My Downloadable Products
- My Wish List
- Address Book
- Account Information
- Stored Payment Methods
- My Product Reviews
- Newsletter Subscriptions

[Compare Products](#)

Order # 000000178

Complete
Order Date: October 21, 2022

[Reorder](#) [Print Order](#)

Items Ordered

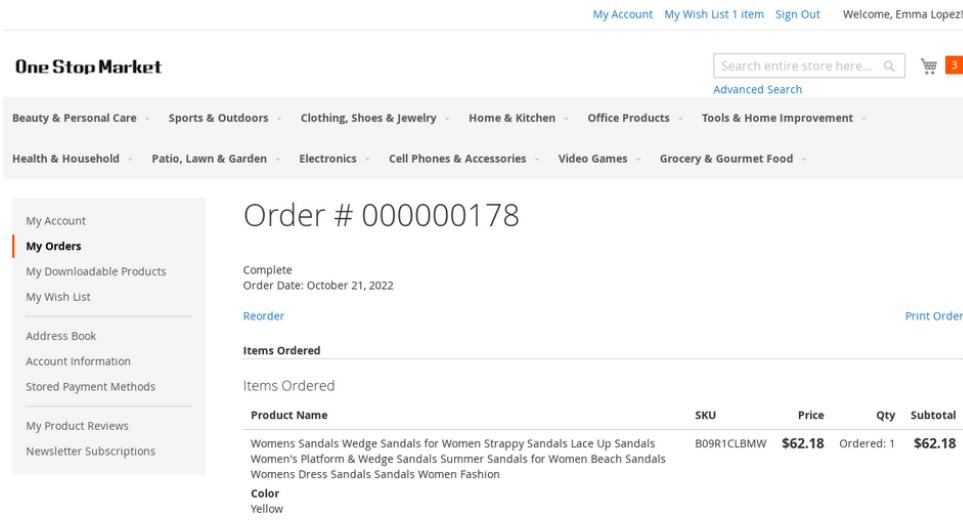
Items Ordered

Product Name	SKU	Price	Qty	Subtotal
Womens Sandals Wedge Sandals for Women Strappy Sandals Lace Up Sandals Women's Platform & Wedge Sandals Summer Sandals for Women Beach Sandals Womens Dress Sandals Sandals Women Fashion	B09R1CLBMW	\$62.18	Ordered: 1	\$62.18
Color Yellow				

Size
7 R Narrow

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows the details of order number "000000178". To verify that this order corresponds to the target order number "00178" despite the difference in format, I need to compare the significant digits of both order num-

3402
 3403 numbers. The significant digits of "000000178" are "178", which matches the target order number "00178".
 3404
 3405 Next Action: According to the thoughts, the next action I will perform is "“finish_subtask [The order number "000000178" corresponds to the target order number "00178" despite the difference in format.]“
 3406
 3407 Target: Confirm that the order number "000000178" corresponds to the target order number "00178".
 3408
 3409 Summary: The current observation shows the details of order number "000000178".
 3410 The significant digits of "000000178" are "178", which matches the target order number "00178". The next step involves confirming that the order number "000000178" corresponds to the target order number "00178" despite the difference in format. The expected outcome of this action is to confirm that the order number "000000178" corresponds to the target order number "00178".
 3411
 3412 **Action: finish_subtask [The order number "000000178" corresponds to the target order number "00178" despite the difference in format.]**
 3413
 3414
 3415
 3416
 3417
 3418 **Verification:**
 3419 **Objective: check_in_webpage("Order number 00178")**
 3420 1.0 The subgoal target is to verify if the order number is 00178. The content provided shows the order number as 000000178.
 3421
 3422 **Verification Pass**
 3423
 3424 **New Subtask 4**
 3425 4. Access the details of order number "000000178."
 3426
 3427 **Figure of Observation (only for visualization, not used in input prompt):**

 3428
 3429
 3430
 3431
 3432
 3433
 3434
 3435
 3436
 3437
 3438
 3439
 3440
 3441
 3442
 3443
 3444
 3445
 3446 **Analyze and Predict Action:**
 3447 Thought: Let's think step-by-step. The current observation shows that I am already on the "Order 000000178" page, which displays the details of order number 178. The details include the order date, status, items ordered, and the total cost. I have achieved the objective of accessing the details of order number 178. I will issue the 'finish_subtask' action to confirm this.
 3448
 3449
 3450
 3451 Next Action: According to the thoughts, the next action I will perform is "“finish_subtask“
 3452 Target: Confirm that the details of order number 178 have been accessed.
 3453 Summary: The current observation section specifies elements present on the 'Order 000000178' page, including the order date, status, items ordered, and the total cost. The next step involves confirming that the details of order number 178 have been accessed. The
 3454
 3455

3456

expected outcome of this action is to confirm that the details of order number 178 have been accessed.

Action: finish_subtask []

3460

3461

Verification:

Objective: check_in_webpage("Order number 00178")

1.0 The content provided is a detailed webpage structure for Order 000000178. It includes various elements such as links, static texts, headings, and a table with order details. The presence of the heading "Order 000000178" and the detailed order information confirms that the content is indeed related to Order 000000178.

Verification Pass

3468

New Subtask 5

Identify and display the billing address for order number "000000178."

Figure of Observation (only for visualization, not used in input prompt):

Product Name	SKU	Price	Qty	Subtotal
Womens Sandals Wedge Sandals for Women Strappy Sandals Lace Up Sandals Women's Platform & Wedge Sandals Summer Sandals for Women Beach Sandals Womens Dress Sandals Sandals Women Fashion	B09R1CLBMW	\$62.18	Ordered: 1	\$62.18

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows the order details page for order 000000178. To identify and display the billing address for this order, I need to locate the section that contains the billing address information. This information is typically found in the order details page, often near the shipping address or payment method sections.

Next Action: According to the thoughts, the next action I will perform is "scroll [down]"

Target: Locate the section that contains the billing address information for order 000000178.

Summary: The current observation shows the order details page for order 000000178. To identify and display the billing address for this order, I will scroll down the page to locate the section that contains the billing address information. The expected outcome of this action is to find the billing address information for order 000000178

Action: scroll [down]

3502

3503

3504

3505

3506

3507

3508

3509

Figure of Observation (only for visualization, not used in input prompt):

3510
 3511
 3512
 3513
 3514
 3515
 3516
 3517
 3518
 3519
 3520
 3521
 3522
 3523
 3524
 3525
 3526
 3527
 3528
 3529
 3530
 3531
 3532
 3533
 3534
 3535
 3536
 3537
 3538
 3539
 3540
 3541
 3542
 3543
 3544
 3545
 3546
 3547
 3548
 3549
 3550
 3551
 3552
 3553
 3554
 3555
 3556
 3557
 3558
 3559
 3560
 3561
 3562
 3563

You have no items to compare.

Recently Ordered

<input type="checkbox"/> Plus Size Lingerie for Women Sexy for Sex Naughty Eyelash Lace Bodysuit Naughty Mesh One Piece Teddy Bodysuit Outfits	Tea Drops Organic Citrus Ginger Tea - Bulk Pack of 20 Lightly Sweetened, Loose Leaf, Pressed Organic Tea - Herbal Citrus Ginger Tea Blend To Go - Iced or Hot Bagless Tea Gift	B07ZG4X2X1	\$22.99	Ordered: 1	\$22.99
<input type="checkbox"/> Uttermost Volterra Cracked Taupe-Gray Ceramic Table Lamp	Flavor Name Strawberry Shortcake				
<input type="checkbox"/> NOZE Rustic Coat Rack Wall Mounted Shelf with 4 Hooks, Hanging Entryway Organizer for Mug Coffee Cup, Holding Solid Wooden Shelf with 2 Baskets for Kitchen Living Room, Bathroom and Bedroom	Size 2-Pack (30 Servings)				
<input type="checkbox"/> Boost Pudding, Chocolate, 5-Ounce Tins (Pack of 48)		B001W6RGV2	\$76.72	Ordered: 1	\$76.72
<input type="checkbox"/> Ad Tec 6in Women Certified Safety Work Boots - Waterproof Grain Oiled Leather with Composite Toe	Color Brown	B08JZ95LM8	\$89.95	Ordered: 1	\$89.95

Add to Cart **View All**

My Wish List 1 item

 Baking Soda Whitening Toothpaste,Intensive Stain Removal Whitening	Subtotal	\$320.84
Toothpaste Strengthening	Shipping & Handling	\$25.00
Toothpastes Strong Cleaning Power	Grand Total	\$345.84
Natural Stain Remover Fluoride-Free		
Toothpaste(Blueberry+Passion Fruit)		

\$19.86

Add to Cart

Order Information

Shipping Address

Shipping Method

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that I have scrolled down the order details page for order 000000178. However, the billing address information is not yet visible. I need to scroll down further to locate the section that contains the billing address information.

Next Action: According to the thoughts, the next action I will perform is “scroll [down]”

Target: Locate the billing address information for order 000000178.

Summary: The current observation shows that I have scrolled down the order details page for order 000000178, but the billing address information is not yet visible. I need to scroll down further to locate the section that contains the billing address information. The expected outcome of this action is to find the billing address information for order 000000178.

Action: scroll [down]

Figure of Observation (only for visualization, not used in input prompt):

\$19.86

Add to Cart

Go to Wish List

Order Information

Shipping Address

Emma Lopez
101 S San Mateo Dr
San Mateo, California, 94010
United States
[T: 6505551212](#)

Shipping Method

Flat Rate - Fixed

Billing Address

Emma Lopez
101 S San Mateo Dr
San Mateo, California, 94010
United States
[T: 6505551212](#)

Payment Method

Check / Money order

Privacy and Cookie Policy

Search Terms

Advanced Search

Contact Us

Copyright © 2013-present Magento, Inc. All rights reserved.
Help Us Keep Magento Healthy [Report All Bugs](#)

Enter your email address **Subscribe**

Analyze and Predict Action:

Thought: Let's think step-by-step. The current observation shows that I have successfully located the billing address information for order number "000000178". The billing address

3564

is listed under the "Billing Address" section and includes the following details: - Name: Emma Lopez - Address: 101 S San Mateo Dr - City: San Mateo, California, 94010 - Country: United States - Phone: 6505551212

Next Action: According to the thoughts, the next action I will perform is ““finish_subtask [Emma Lopez, 101 S San Mateo Dr, San Mateo, California, 94010, United States, 6505551212]““

Target: Confirm that the billing address for order number ”000000178” has been identified and displayed.

Summary: The current observation confirms that the billing address for order number ”000000178” has been located and includes the following details: Emma Lopez, 101 S San Mateo Dr, San Mateo, California, 94010, United States, 6505551212. Therefore, the subtask of identifying and displaying the billing address for order number ”000000178” has been completed.

Action: finish_subtask [Emma Lopez, 101 S San Mateo Dr, San Mateo, California, 94010, United States, 6505551212]

Verification:

check_in_webpage("Billing Address")

1.0 The subgoal target is to determine and display the price range for 'Canon photo printer' products. The summary indicates that the prices have been identified, with a minimum price of \$2.56 and a maximum price of \$649.99. This information aligns with the subgoal target.

Verification Pass

Answer:

The billing address for order number 00178 is: **Emma Lopez** 101 S San Mateo Dr San Mateo, California, 94010 United States T: [6505551212](tel:6505551212)

3590

3591

3592

3593

3594

3595

3596

3597

3598

3599

3600

3601

3602

3603

3604

3605

3606

3607

3608

3609

3610

3611

3612

3613

3614

3615

3616

3617