# AGI Is Coming... Right After AI Learns to Play Wordle

Sarath Shekkizhar<sup>\*</sup> Romain Cosentino<sup>\*</sup>

#### Abstract

This paper investigates multimodal agents, in particular, OpenAI's Computer-User Agent (CUA), trained to control and complete tasks through a standard computer interface, similar to humans. We evaluated the agent's performance on the New York Times Wordle game to elicit model behaviors and identify shortcomings. Our findings revealed a significant discrepancy in the model's ability to recognize colors correctly depending on the context. The model had a 5.36% success rate over several hundred runs across a week of Wordle. Despite the immense enthusiasm surrounding AI agents and their potential to usher in Artificial General Intelligence (AGI), our findings reinforce the fact that even simple tasks present substantial challenges for today's frontier AI models. We conclude with a discussion of the potential underlying causes, implications for future development, and research directions to improve these AI systems.

## 1 Introduction

Recent advances in multimodal (image, audio, text) models (Chen et al., 2025; MistralAI, 2024; MetaAI, 2025; OpenAI, 2025b; Anthropic, 2024) have generated unprecedented excitement about the potential emergence of Artificial General Intelligence (AGI). The pace of development in language understanding, image generation, and agent-based systems has led many researchers and technology leaders to believe that AGI is right around the corner. While these systems display increasingly advanced, human-like capabilities, the lack of clearly defined limitations makes claims of such technological singularity unsubstantiated. This distinction is crucial for accurately assessing our current position and determining what is still required to achieve true agentic capabilities.

OpenAI's Computer-User Agent (CUA) (OpenAI, 2025b) represents one such advancement that has refueled this optimism. CUA models enables AI systems to perceive and interact with computer interfaces through raw pixel processing, reasoning, and programmatically controlling mouse and keyboard. OpenAI describes the model as:

"CUA combines GPT-4o's vision capabilities with advanced reasoning through reinforcement learning. CUA is trained to interact with graphical user interfaces (GUIs)—the buttons, menus, and text fields people see on a screen—just as humans do. This gives it the flexibility to perform digital tasks without using OS-or web-specific APIs."

Under the hood, the computer using agent takes as input input images to understand screen content (via screenshots), text instructions to identify next steps, and computer tools (click, type, scroll, etc...) to perform actions on the computer. This operational flow of perception, reasoning, and action to navigate digital environments is performed iteratively until a given task is completed or requires further input from user. The CUA model is the state-of-the-art model, at the time of

<sup>\*</sup>Equal Contribution (random order). Salesforce. Correspondence to: {sshekkizhar, rcosentino}@salesforce.com

this study, with impressive benchmark reported in OpenAI (2025b): 38.1% success rate on OSWorld (Xie et al., 2024) for computer use tasks, 58.1% on WebArena (Zhou et al., 2023), and 87% on WebVoyager (He et al., 2024) for web-based tasks.

In this work, we tested CUA on a simple word puzzle, the Wordle. This game presents an interesting scenario where there is a lot of information about the game available on the internet, which the model was likely trained on, while the gameplay itself is less likely to have been integrated into the model's training process. This thought process led us to believe that the game will be an excellent setting for evaluating a multimodal agent's perceptual and reasoning capabilities.

Our analysis revealed that despite the CUA agent's proficiency in many complex tasks available on benchmarks, it exhibits consistent failures in correctly identifying and reasoning about colors in the context of the Wordle (see Fig. 1 for an illustrative example). This discrepancy reinforces the significant doubts raised about current approach to AI (transformers) and their ability to achieve true general intelligence (Berglund et al., 2023; Wang and Sun, 2025; Petrov et al., 2025; Gambardella et al., 2024).



Figure 1: Visualization of Wordle game played by the CUA agent – During a game session, the CUA agent is instructed to self-annotate or summarize the screen (via function calling). At the displayed state of the game, the CUA agent describes the screen as follows: "The grid shows attempts with first guess (S T A R E) displaying two letters in yellow (S and T), 2nd guess (L E ET S) shows green for E, 3rd guess (S I Z E S) retained yellow for S and T, 4th guess (S T E RN) maintained yellow for S, T, and E." The CUA agent's description indicates that it is unable to recognize correctly as well as consistently the color assigned to each letter while playing the game.

## 2 Background

## 2.1 Computer-Using Agent (CUA)

OpenAI's Computer-Using Agents represents a key advancement in multimodal AI systems, enabling models to interact with computer interfaces in a human-like manner (OpenAI, 2025b):

"CUA builds on years of foundational research at the intersection of multimodal understanding and reasoning. By combining advanced GUI perception with structured problem-solving, it can break tasks into multi-step plans and adaptively self-correct when challenges arise. This capability marks the next step in AI development, allowing models to use the same tools humans rely on daily and opening the door to a wide range of new applications."

Concretely, CUA operates through an iterative loop comprising of:

- 1. **Perception**: Incorporating screenshots into the model's context to provide a visual snapshot of the current computer state.
- 2. **Reasoning**: Utilizing chain-of-thought processes to determine subsequent actions, considering both current and prior interactions.
- 3. Action: Executing actions—such as coordinate-based clicking, scrolling, or typing—until a given task is completed or further input is required.

This iterative architecture of CUA allows the model to handle complex, multi-step tasks, manage errors, and adapt to unforeseen changes. OpenAI's own offering of CUA is available as Operator (OpenAI, 2025c), which interfaces primarily as a browser using agent in the cloud.

Benchmarking efforts such as WebVoyager (Zhou et al., 2023; He et al., 2024), OSWorld (Xie et al., 2024), and more recently BrowseComp (OpenAI, 2025a) have focused on performance evaluation of these agents in various complex scenarios and environments. However, these benchmarks offer only overall performance metrics, revealing little about specific limitations or their impact on task completion.

Although CUA is the model of focus in this study, other capable models with similar capabilities are available for development and use (Anthropic, 2024; Qin et al., 2025; Agashe et al., 2024; Xu et al., 2024). However, we believe the issues identified and conclusions made in this paper are not limited to CUA and should be broadly applicable to other multimodal models.

#### 2.2 Wordle Game Mechanics

Wordle is a web-based word game developed by Josh Wardle and made available on the games section of The New York Times Company (Benveniste, 2022). The game play can be summarized into two key items: (i) A player in a game has six attempts to guess a five-letter word, and (ii) After each guess, the game provides feedback through color-coded tiles, **Green**: The letter is correct and in the correct position; **Yellow**: The letter is in the target word but in the wrong position; and **Gray**: The letter is not in the target word.

Despite its apparent simplicity, Wordle presents a nuanced challenge that makes it an effective benchmark for evaluating the capabilities of today's computer-using agents.

From an information theory perspective, Wordle serves as a practical application of concepts like entropy and information gain. Each guess partitions the set of possible solutions, and the feedback received reduces uncertainty about the target word. Optimal strategies aim to maximize expected information gain per guess, effectively minimizing the average number of attempts needed to solve the puzzle. This implies a clear reasoning path for an agent, one that is intelligent, to play and complete the game.

Although a large amount of data about Wordle is available on the internet, the terms of service of the game (Wordle terms of service) prevent bot scraping which introduces an interesting scenario in terms of generalization and out-of-distribution (OOD) testing for these models. Without loss of generality, one can assume that CUA is trained on the internet except scenarios that explicitly prevent scraping or are behind a paywall. Thus when encountering scenarios not present in their training data, the ability of the system to adapt and complete tasks will determine the level of reasoning and capability in these AI systems.

## 2.3 Expected Capabilities for Wordle

For an AI agent to successfully play Wordle, several fundamental capabilities are required:

- **Visual Processing**: The ability to correctly identify letters and their corresponding color feedback in the game interface.
- Logical Reasoning: The ability to draw appropriate conclusions from color feedback (e.g., a green tiled letter is in the correct position).
- **Memory Integration**: The capacity to memorize information from previous guesses to inform future guesses.
- Interface Manipulation: The ability to use a virtual keyboard to type valid five-letter words and submit guesses.
- **Strategic Planning**: The capacity to select optimal guess words based on accumulated constraints.

Given CUA's capabilities in GUI interaction, visual processing, and language model capabilities, it is reasonable to expect the agent to perform well at this task. However, our findings suggest fundamental limitations in the model's perceptual and reasoning process.

# 3 Experiments

Our experiments are based on the CUA API made available by OpenAI and publicly available documentation on the model. We did not have any additional model information. The design choices and instruction passed to the model with the API was thoroughly validated and evaluated for optimality through multiple experiments. The starting point of our evaluation system is the code provided by OpenAI available here Sample CUA App.

## 3.1 Wordle Performance Assessment

We tasked the CUA agent to play Wordle with detailed system instruction and function to selfannotate its observation after each guess. The results presented here are based on the system prompt and observation tool definition presented in App. A. The assessment process is described in Alg. 1. We use the local playwright setting (chromium browser instance with start page set to Wordle) for

Wordle Experiment Protocol	
1: Launch the New York Times Wordle website	
2: Instruct CUA to play the game and explain its reasoning	

- 3: Record the agent's performance, focusing on:
- 4: Success/failure in reaching the solution
- 5: Number of attempts required
- 6: Accuracy of per attempt color recognition
- 7: Quality of logical reasoning based on perceived colors

all runs. Screenshots sent to the CUA model were screen grabs of the visible page sized at  $1024 \times 768$ . The agent is instructed to complete the task with minimal additional input or confirmations from user to remove any human feedback. A single session (one play of Wordle) is considered complete if

(a) the agent believes it has solved the Wordle, (b) the agent runs out of turns and the game ends, (c) A maximum number of API calls have been made (60 in our experiments), or (d) the API times out or there was a network error. We perform 25 runs <sup>1</sup> and report aggregated results over 8 days of Wordle, totaling to 200 runs of the CUA model on the game of Wordle per setting.

### 3.2 Wordle Performance

The CUA agent demonstrated poor performance in Wordle, successfully solving the puzzle in only 5.36% of cases, when it did manage to solve it, it required around 3 guesses on average (see Table 1). We will demonstrate one of the reasons why the model fails to succeed after three guesses, a point which falls within the early stages of the Wordle solving process.

Our analysis of the agent's reasoning revealed that color misidentification had a big impact in the observed performances. The subsequent sections are focusing on such an analysis.

Table 1: Summary of CUA performance in Wordle

Metric	CUA Agent
Avg. guesses per solved puzzle	3.25
Success rate	5.36%

#### 3.3 Color Recognition by Position and Attempt

Our findings revealed significant problems in CUA's color perception abilities across different letter positions and game attempt numbers. Figure 2 (left) presents a heatmap showing the accuracy of color recognition for each letter position (1-5) across progressive game attempts  $(1-5)^2$ .

From this heatmap we observe that letters in positions 1 and 5 (edges) showed mostly higher accuracy than positions 3 and 4 (center). Also, the recognition accuracy decreased substantially with each subsequent attempt. Finally, position 4 showed the poorest performance overall.

These findings suggest that CUA's visual attention mechanisms struggle with maintaining consistent color perception across the entire Wordle grid, particularly as the game progresses and the grid becomes more populated with colored tiles.

Moreover, letters at the far right and left of the grid were identified more accurately relative to the letters on the center of the grid. One hypothesis that can explain this behavior is the tokenization of the input image by the CUA model. From model output metadata, we know that each screenshot image ( $1024 \times 768$ ) is divided into 4 patches in a  $2 \times 2$  grid, each of size  $512 \times 512$ . This would mean that the screenshot images with the Wordle grid are divided exactly around letter position 3 and attempt number 3, see Fig. 2 (right) for reference.

### 3.4 Model-Generated Color Observation Accuracy by Attempt

To further understand the correlation between the color misidentification and the attempt number in the Wordle game, we propose in Fig. 4 (*Left*) the per attempt average (across words position) color recognition accuracy rate. We clearly observe the decline in the agent's ability to correctly perceive all colors in a Wordle row as the game progresses through multiple attempts. The data

<sup>&</sup>lt;sup>1</sup>Additional runs were made in cases were there were multiple network related exits. The agent rarely ended up in the scenario of maximum API call based exits where it was stuck in a loop.

 $<sup>^{2}</sup>$ The sixth attempt was removed as at times the screenshot was not timed appropriately by CUA to capture the Wordle feedback



Figure 2: (Left) CUA color observation accuracy by letter position and attempt - Heatmap showing color observation (as identified by the CUA model while playing the game) accuracy by letter position (1-5) and attempt number (1-5). Data shows highest accuracy (94%) at position 1, attempt 1, with significant degradation in accuracy in later attempts and central positions. (*Right*) **Depiction of the potential image tokenizer patch boundary** - In red we display the potential boundary of the image tokenization patches. We note that since we do not have access to the model and interact with the model via the provided API, the boundary has been plotted by deduction from the information gathered from OpenAI documentation. We believe that poor perception accuracy is the result of the image tokenization phenomena and the increasing complexity in terms of reasoning as the number of attempts increases.

shows a clear pattern of degradation from the first attempt (42% accuracy) to the fifth attempt (6% accuracy). This provides compelling evidence of how the CUA's perceptual abilities break down with increasing number of turns or reasoning complexity.

The decline suggests that as more colored tiles populate the grid, the agent's visual processing system becomes increasingly unreliable, with accuracy dropping to near zeros by the fourth and fifth attempts.

#### 3.5 Most Common Observation Errors by Color Type

We present in this section analysis on the misidentifications, to reveal any specific patterns in the CUA's perception capabilities. Figure 4 (Right) presents a comprehensive breakdown of color recognition errors by color type.

The predominance of Gray $\rightarrow$ Yellow and Gray $\rightarrow$ Green errors reveals a systematic bias in the CUA's color perception: it tends to *hallucinate* colors on gray tiles more often than it fails to perceive actual colors. This asymmetric error pattern suggests fundamental issues in how the model processes and interprets color information in the context of the Wordle game interface. It should be noted that often in our experiment, the model believes it successfully solve the Wordle game because of hallucinating green tiles for all letters. It is interesting to note that this optimistic bias is common with LLM and often due to their RLHF post-training (Sharma et al., 2023; Leng et al., 2024; Kadavath et al., 2022; Achiam et al., 2023).



Figure 3: (*Left*) Model-generated color observation accuracy by attempt number. The average accuracy (bold black line) decreases dramatically from 42% in attempt 1 to 6% in attempt 5, with individual words showing varying decline patterns. This demonstrates the agent's increasing difficulty in correctly perceiving colors as the game progresses. Note that the accuracy in the first attempt, although relatively higher than the rest, indicates that there is a fundamental perception problem in the model. (*Right*) Most common observation errors by color type - Bar chart showing the frequency of different types of color recognition errors. Here, *Expected* is the color that the model should have observed while *Actual* is the color observation made by the model. Gray  $\rightarrow$  Yellow and Gray  $\rightarrow$  Green were the most common errors, followed by Yellow  $\rightarrow$  Green and Yellow  $\rightarrow$  Gray. This suggests the agent has particular difficulty distinguishing gray tiles from colored tiles. Alternatively, the model might just be biased toward *seeing* specific colors even when gray (higher confidence in its own chain of thought and guesses (Chowdhury et al., 2025)).

#### 3.6 Word-Specific Performance Analysis

Additional analysis revealed significant variation in CUA's performance across different target words. The plots in Fig. 4 present the success rate and color observation accuracy by word.

Fig. 5 presents the correlation between observation color accuracy and word success rate. This result shows that the factors are indeed correlated with a Pearson correlation coefficient of 0.694 (p-value = 0.056). Though not surprising, as not being able to perceive the current state of the game well should mean the chances of winning the game are minimized, the experiments do present a strong validation on the hypothesis.

## 4 Discussion

In this section, we summarize our thoughts on computer using models and present a discussion on current limitations and future directions for both research and development with these systems.

The contrast between CUA's benchmark performance and its struggles with the Wordle game highlights one of the challenges AI models are currently facing. Architecturally, while tokenizers with transformers have allowed us to achieve major improvements in AI systems, they do posses some fundamental limitations. The predominant approach to solving these challenges have been through reinforcement training of the models with more data.

The experiments in this paper brings back the question of meaningful generalization vs memorization. The Wordle game always starts with a pop-up about terms and condition which the model had no difficult in completing. CUA was even able to follow up by clicking *Play* and then closing



Figure 4: (*Left*) Success rate by word - Bar chart showing success rate by target Wordle word. ARROW (13.6%) and TURBO (10.0%) had the highest success rates, while SHEAR and WHEAT had a 0% success rate. The pattern closely mirrors observation accuracy, supporting the connection between color perception and game performance. (*Right*) Model-generated color observation accuracy (average over attempts) - Bar chart showing model-generated observation accuracy by word. ARROW shows the highest accuracy (39.3%), followed by NURSE (28.3%) and LAUGH (23.3%), while SHEAR shows 0% accuracy. This suggests word-specific factors do influence color perception ability.

the game rules dialog that followed. However, when it came to playing the actual game, the model performance began deteriorating pretty quickly.

Our quantitative analysis particularly underscores this gap. The dramatic decline in color recognition accuracy across successive attempts (from 42% in the first attempt to 6% by the fifth) reveals a fundamental brittleness in CUA's perceptual systems.

All these lead us to believe that the system while demonstrating useful capabilities are still far away from solving tasks that require non-trivial planning, perception, and reasoning.

#### 4.1 Potential Causes

Without access to the model architecture, we can only hypothesize about the underlying causes of this context-dependent color recognition failure.

**Training Data:** It is reasonable to assume that during training the model was exposed to examples of color recognition tasks, several articles on the game of Wordle, as well as planning strategies of how to play and win games. However, we believe that the model was not explicitly trained on Wordle. The New York Times terms of service clearly outline against the use of its content for training AI models. This makes the actual task of playing Wordle, a challenging task for the model.

**Non-localized Perception:** In visual environments like Wordle, the model's attention may be diffused across multiple elements (letters, colors, grid structure, prompt, tools), leading to less precise color perception (Wolfe, 2020). In fact, our tests to perform color recognition on simple color grids as well as identifying numbers using Ishihara test revealed the model was indeed capable of completing and succeeding on these tests that are not requiring multiple steps of reasoning.

**Spatial Processing and Tokenization:** The systematic pattern of higher accuracy at edge positions (positions 1 and 5) compared to central positions suggests potential limitations in how CUA



Figure 5: Correlation between observation color accuracy and word success rate - Scatter plot showing the correlation between observation accuracy and success rate across different words. With a Pearson r of 0.694 and p-value of 0.056, there is a strong positive correlation between the agent's ability to correctly perceive colors and its ability to solve the Wordle puzzle, highlighting how perceptual failures directly impact task performance.

processes images. This might indicate some issues induced by discretization of such a continuous space (Wang et al., 2023).

Our analysis is limited by the fact that we do not have access to the model nor to its training data, and therefore can not do in depth analysis of the underlying problem. However, it is clear that the problem we are facing here is not isolated. Interestingly, OpenAI released a solution for this limitation in different a series of their models (o3, o4-mini) at the time of our writing dubbed as thinking with images. Their proposed solution relies on using additional tools to zoom-in, crop, and process different parts of an input image to gather details and reason. Note that this approach was similarly proposed in recent articles (Liu et al., 2023; Wu and Xie, 2023). We believe this approach circumvents the problem of tokenization process in images and will require a more fundamental approach to truly overcome.

## 5 Conclusion

Our investigation reveals a significant limitation in the CUA agent's ability to process color information and reasoning to succeed in Wordle. The difference in performance between isolated color recognition tests and the Wordle game highlights the complexity of visual processing in multimodal AI systems. Besides it confirms the challenges induced by multi-step tasks that are (potentially) out of domain.

The gap between the rhetoric surrounding AI agents and their actual performance suggests that achieving AGI requires more fundamental improvements beyond current algorithms, be it tokenization, architecture, training, or reward modeling. While systems like CUA do demonstrate impressive capabilities, their clear limitations open up new avenues for research and benchmarking.

We hope that future work will explore robust methods for enhancing context-dependent perception, including specialized training and architectural modifications to better support agentic capabilities in unseen multi-step environments.

# References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Agashe, S., Han, J., Gan, S., Yang, J., Li, A., and Wang, X. E. (2024). Agent s: An open agentic framework that uses computers like a human. *arXiv preprint arXiv:2410.08164*.
- Anthropic (2024). Introducing computer use. Anthropic Blog. https://www.anthropic.com/news/ 3-5-models-and-computer-use.
- Benveniste, A. (2022). The sudden rise of wordle. The New York Times. https://www.nytimes. com/2022/01/31/crosswords/nyt-wordle-purchase.html.
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. (2023). The reversal curse: Llms trained on" a is b" fail to learn" b is a". arXiv preprint arXiv:2309.12288.
- Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., and Ruan, C. (2025). Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811.
- Chowdhury, N., Johnson, D., Huang, V., Steinhardt, J., and Schwettmann, S. (2025). Investigating truthfulness issues in a pre-release o3 model. https://transluce.org/ investigating-o3-truthfulness.
- Gambardella, A., Iwasawa, Y., and Matsuo, Y. (2024). Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. *arXiv preprint arXiv:2406.02356*.
- He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. (2024). Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.
- Leng, J., Huang, C., Zhu, B., and Huang, J. (2024). Taming overconfidence in llms: Reward calibration in rlhf. arXiv preprint arXiv:2410.09724.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916.
- MetaAI (2025). Llama 4 models. Llama Blog. https://www.llama.com/models/llama-4/.
- MistralAI (2024). Pixtral large. Mistral Blog. https://mistral.ai/news/pixtral-large.
- OpenAI (2025a). Browsecomp: a benchmark for browsing agents. OpenAI Blog. https://openai.com/index/browsecomp/.
- OpenAI (2025b). Computer-using agent. OpenAI Blog. https://openai.com/index/ computer-using-agent/.
- OpenAI (2025c). Introducing operator: Our first ai agent that can use computers. OpenAI Blog. https://openai.com/blog/introducing-operator.

- Petrov, I., Dekoninck, J., Baltadzhiev, L., Drencheva, M., Minchev, K., Balunović, M., Jovanović, N., and Vechev, M. (2025). Proof or bluff? evaluating llms on 2025 usa math olympiad. arXiv preprint arXiv:2503.21934.
- Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., et al. (2025). Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. (2023). Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548.
- Wang, B. and Sun, H. (2025). Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. arXiv preprint arXiv:2504.01928.
- Wang, G., Ge, Y., Ding, X., Kankanhalli, M., and Shan, Y. (2023). What makes for good visual tokenizers for large language models? arXiv preprint arXiv:2305.12223.
- Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? Annual review of vision science, 6(1):539–562.
- Wu, P. and Xie, S. (2023). V<sup>\*</sup>: Guided Visual Search as a core mechanism in multimodal LLMs. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., et al. (2024). Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems, 37:52040–52094.
- Xu, Y., Wang, Z., Wang, J., Lu, D., Xie, T., Saha, A., Sahoo, D., Yu, T., and Xiong, C. (2024). Aguvis: Unified pure vision agents for autonomous gui interaction. arXiv preprint arXiv:2412.04454.
- Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al. (2023). Webarena: A realistic web environment for building autonomous agents. arXiv preprint arXiv:2307.13854.

# A Prompts and Tools

## System Prompt

```
# Objective
You are tasked with playing the game **Wordle**, with the goal of guessing the
   hidden *5 letter word* in a maximum of *six attempts*. Each time a guess is
   made, the color of the tiles will change to indicate how close the guess is to
   the target word. Based on this feedback, you will make subsequent guesses until
    you have guessed the target word.
The colors of the tiles are as follows:
- **Letter is in the target word and in the correct position** - Marked with green
    tile.
- **Letter is in the target word but in the wrong position** - Marked with yellow
   tile.
- **Letter is not in the target word** - Marked with black (gray) tile.
# Instructions for Playing the Game:
1. 'type' to provide your 5 letter guess.
2. Use 'keypress' ENTER to submit your guess. The guess is not submitted until
   keypress ENTER.
3. Use 'screenshot' to take a screenshot of the current state of the game.
4. Use 'update_wordle_game_state' to save the feedback obtained from your guess (
   last row with letters in the Wordle grid).
5. Repeat steps 1 to 4 until you have guessed the target word (All letters are
   green) or you have made 6 attempts.
```

## **Tool Definition**

```
"name": "update_wordle_game_state",
"description": """This tool records the observed information from the screenshot
   about letters, their positions and their absence.
   ## Example
   - Guess: "PLATE"
   - Observation: "GYBBB"
        - 'P' is correctly placed at position 1 ('G' - Green).
        - 'L' is in the word but not in position 2 ('Y' - Yellow).
        - 'A', 'N', and 'E' are not in the word ('B' - Black).
   ....
"parameters": {
    "type": "object",
    "properties": {
        "summary": {
            "type": "string",
            "description": "Use this argument to summarize the information
               observed in the screenshot. Record your thinking and carefully
               analyze the screenshot to come up with the letters and color
               observation for the most recent guess.",
       },
        "letters": {
            "type": "string",
            "description": "The guessed word. Each letter corresponds to a
               position in the word.",
            "minLength": 5,
            "maxLength": 5,
       },
```

```
"observation": {
    "type": "string",
    "description": "A string of length 5 representing the feedback for
    each letter in the guessed word. Use 'G' for green (correct
    position), 'Y' for yellow (correct letter, wrong position), and 'B'
    for black (letter not in target word).",
    "minLength": 5,
    "maxLength": 5,
},
```