Google

Published/updated OCT 07, 2025

Gemini 2.5 Computer Use Additional Information

Google DeepMind – Gemini 2.5 Computer Use

Because computer use evals are sensitive to the agent's environment and system instructions, this document provides details on how we evaluated the Gemini 2.5 Computer Use model, including environments, prompts, and methodologies.

Published: October 7, 2025

Environment

We evaluated our model in three distinct settings:

1. Online-Mind2Web, following an industry standard for evaluating web agents
2. WebVoyager**,** following an industry standard for evaluating web agents
3. AndroidWorld, testing generalization to another surface and action space

Some evaluations were conducted by Browserbase on its harness, which allows for normalizing for variations across APIs and eval prompt sets.

Prompting & Methodologies

Online-Mind2Web
For Online-Mind2Web, one of the requirements is to use the specified starting website, and *not* to use a search engine or a navigate tool to switch directly to an alternative site, which are two of the Supported UI Actions. To abide by this, we set the excluded functions parameter in the open source reference implementation to:

excluded_predefined_functions = ["search", "navigate"]

We also deleted all sentences referring to either of these actions from the system instruction. Besides these deletions, we made no other benchmark specific modifications or additions to the system instruction.

The evaluations were run using Anchor for the environment and actuation. All sampling parameters were left at the default settings in the API reference implementation (temperature=1,

include_thoughts=True, etc.). After collecting the trajectories via the API (autoregressive sampling, pass@1), we gathered per-trajectory human judgments for the question:

> Did the agent complete the task successfully? Take a close look at the original goal, the steps the AI agent took, and the final answer. In your assessment, did the AI agent follow all the instructions, complete the task successfully and give a correct answer?

We gathered three independent human judgments per task, computed success as the majority vote, and had the benchmark organizers independently validate the results. The model's final success rate was 69.0%, comparing favorably to all prior leaderboard submissions.

WebVoyager

The original WebVoyager dataset consisted of 643 task queries. Many of these queries were time-sensitive to particular dates in the past (e.g. *Show me the list of one-way flights today (February 17, 2024) from Chicago to Paris.)* or information no longer available (e.g. *Check the price for an Apple iPhone 14 Plus with 256GB storage in Purple color.)*. Following industry practice, for the self-reported results we used a set for which any explicitly mentioned dates were updated to be in the future, and any tasks that were still infeasible were removed, leaving 559 remaining, date-edited tasks. As different models of this benchmark are run at different points of time and exclude different tasks, self-reported results are difficult to compare. In contrast, the Browserbase results are done over the same query sets with the same websites accessed on the same day. Actuation, sampling, settings, and human evaluation followed the same process as for Online-Mind2Web above.

AndroidWorld

Adaptations for mobile with the Gemini 2.5 Computer Use API

AndroidWorld requires operating in a mobile rather than web environment. Assessing the Gemini 2.5 Computer Use API and model on this benchmark required modifying the action space and the SI to fit the mobile environment. The below predefined functions were excluded:

excluded_predefined_functions = ["open_web_browser", "search", "navigate", "hover_at", "scroll_document", "go_forward", "key_combination", "drag_and_drop"]

The below functions were added using the custom_functions functionality of the API:

["open_app", "long_press_at", "go_home"]

with declarations for each function including name, description, and expected input parameters. Input parameters were defined as closely as possible to how the native action space was defined. For example, long_press_at is defined in terms of x, y parameters similarly to the native click_at action in the Gemini 2.5 Computer Use API.

The system prompt was configured to be:

```
SYSTEM_PROMPT = """You are operating an Android phone. Today's date is October 15, 2023, so
```

```
ignore any other date provided.
* To provide an answer to the user, *do not use any tools* and output your answer on a separate line.
IMPORTANT: Do not add any formatting or additional punctuation/text, just output the answer by
itself after two empty lines.
* Make sure you scroll down to see everything before deciding something isn't available.
* You can open an app from anywhere. The icon doesn't have to currently be on screen.
* Unless explicitly told otherwise, make sure to save any changes you make.
* If text is cut off or incomplete, scroll or click into the element to get the full text before providing an
answer.
* IMPORTANT: Complete the given task EXACTLY as stated. DO NOT make any assumptions that
completing a similar task is correct.  If you can't find what you're looking for, SCROLL to find it.
* If you want to edit some text, ONLY USE THE `type` tool. Do not use the onscreen keyboard.
* Quick settings shouldn't be used to change settings. Use the Settings app instead.
* The given task may already be completed. If so, there is no need to do anything.
"""
```

Claude's computer use tool with Sonnet 4 and Sonnet 4.5

We also tried the [Claude computer use tool] against the AndroidWorld benchmark. To discourage the use of tools unavailable in mobile, the Android system prompt above used for the Gemini 2.5 Computer Use tool was augmented to also include:

```
* Because you are operating an Android phone, the following tools are not implemented and
shouldn't be used: key, mouse_move, left_click_drag, right_click, middle_click, double_click,
cursor_position, left_mouse_down, left_mouse_up, hold_key, and triple_click.
```

And the line about formatting a response to the user was adjusted to:

```
* If you want to provide an answer to the user, *do not use any tools* and instead provide your
answer as a new paragraph as plain text. Do not add any formatting or additional punctuation/text,
just output the answer.
```

The following functions were added as custom functions:

[open_app, long_press, go_home, go_back]

As for the Gemini 2.5 Computer Use API, input parameters were defined as closely as possible to native tools, so long_press is defined in terms of a coordinate array rather than x, y integers, to match the Claude computer use tool's click function.

For both models, the evaluations on AndroidWorld were run on a pool of Pixel 6 emulators running Android 13 (API level 33). Parameters such as the maximum number of steps and the random seed

were kept at the default settings. The experiments are done in the screenshot-only setting; no accessibility tree was provided to the models.

Browserbase evaluations and comparisons
Browserbase provided additional evaluations across each of the Google, Anthropic, and OpenAI computer use APIs, measuring both accuracy and latency. By using the same third-party platform and prompt set for each of the experiments, areas of potential variance such as access date, actuation type, frequency of blocking, etc. were as normalized as possible. Browserbase measured accuracies on WebVoyager and Online-Mind2Web, two prominent academic benchmarks using identical harnesses for each API. All accuracy numbers reported are based on human evaluations.

Acknowledgements

Various teams contributed to Computer Use API across Google. These include Google Deepmind, Research, Labs, Core, Global Affairs, Trust and Safety, and many more. Thank you to all of our human AI trainers and testers, as well as the following individuals for their contributions to the Model, API and System Card:

Contributors: Salah Ahmed, Isabela Albuquerque, Ankesh Anand, Shereen Ashraf, Anton Bakalov, Galen Ballew, Elisa Bandy, Parker Barnes, Shrestha Basu Mallick, Harleen Batra, Tarun Bharti, Fabien Blanc-paques, Tim Blyth, Peter de Boursac, Demetra Brady, Jenny Brennan, Geoff Brown, Sasha Brown, Michael Buettner, Folawiyo Campbell-Ajala, Ming-Wei Chang, Saikat Chatterjee, JD Chen, Wei Chen, Xi Chen, Jeremy Cole, Josh Cowls, Max Curran, Benoît Dancoisne, Eric Deng, Vishal Dharmadhikari, Eric Dong, Madeleine Elish, Aldi Fahrezi, David Gaddy, James Gan, Frankie Garcia, Kanav Garg, Jason Gelman, Sahra Ghalebikesabi, Badih Ghazi, Steren Giannini, Megha Goel, Mark Graham, Sarah Murphy Gray, Anmol Gulati, Ryan Guo, Yoni Halpern, Julie Jin, Mandar Joshi, Anna Kelly, Vaishakh Keshava, Shadi Khalek, Mina Khan, Urvashi Khandelwal, Tushar Khot, Sneha Kudugunta, Chinmay Kulkarni, Avery Lamp, Kenton Lee, Kevin Lee, Ving Ian Lei, Alice Li, Simon Li, Fei Liu, Frederick Liu, Aroma Mahendru, Justin Mahood, Sam McCauley, Antoine Miech, Shikhar Murty, Nikita Namjoshi, Dev Nath, Karolina Netolicka, Linda Nyberg, Wojciech Opydo, Karthika Pai, Lisa Patel, Miteyan Patel, Simon Pelchat, Florence Perot, Borja De Balle Pigem, Emily Pitler, Lev Proleev, Mateo Quiros, Chris Rawles, Aniket Ray, William Ren, Oriana Riva, Matthew Robertson, Sohan SM, Omar Sanseviero, Omkar Savant, Sambuddha Sen, Abhanshu Sharma, Pete Shaw, Tianze Shi, Lei Shu, Anu Sinha, Lars Lowe Sjoesund, Rachel Soh, Ben Solis-Cohen, Eric Stavarache, Ruoxi Sun, Srinivas Sunkara, Andrea Tacchetti, Satish Tallapaka, Melissa Tan, Santhosh Thangaraj, Metin Toksoz-Exley, Kristina Toutanova, Marcella Valentine, Maria Wang, Wudi Wang, Zhengdong Wang, Tianyi Wang, Sarah Weldon, Mateo Wirth, Elung Wu, Xiao Wu, Xinyi Wu, Yao Xiao, Da Yu, Junwei Yuan, Wenjie Yuan, Ming Zhang, Wangxing Zhao, Jack Zhou

---