

**On 5 October 2026 the researchers who maintain OSWorld, a benchmark that asks an AI system to operate a real computer, updated the official results for its new version, OSWorld 2.0. The test, built at the XLANG Lab of the University of Hong Kong with collaborators at eleven other institutions, consists of 108 long workflows of the kind an office worker does—following a tutorial, reconciling receipts against emails, filing an expense claim through a web portal—which its authors say take people a median of about 1.6 hours. The system under test receives a written task and may put questions to a simulated user, but it observes the computer only through screenshots and acts through mouse movements, clicks and keystrokes. The best official run on the full set, by Anthropic's Claude Opus 5, completed 44.3% of the workflows outright; scored instead on the share of checkpoints its final state satisfied, the same run earned 77.7%. Anthropic's Claude models are used in producing this publication.**

Those two figures, from one run, are a compact introduction to what changes when an AI system stops being a chatbot. Designs built for text have been adapted to read pictures, sound and screens with surprisingly little alteration. The checking does not come with them: whether a test of seeing actually required sight, and whether an action taken in the world achieved what it was meant to, have to be established afresh for each kind of input and output. Three ideas organise the subject—modality, perception and action, and evaluation designed for the modality being tested—and they are best taken together.

## Two readings that mislead

---

The first misreading is that AI now uses a computer better than people do. It has a number behind it. On OSWorld-Verified, the repaired 2025 edition of the benchmark first published in April 2024, the maintainers' listed results now run past 90%—the highest, 90.2%, dated 25 July 2026—against a human reference of about 72% from the original study. That reference comes from a particular group under particular conditions. The 2024 paper describes them as

computer science major college students who possess basic software usage skills but have not been exposed to the samples or software before

and the maintainers' own announcement of the repaired "Verified" edition, in July 2025, called the figure "estimated at ~72% from our original study". The first entry on the verified results to pass it, dated 11 December 2025 at 72.6%, came from a system that drew on ten attempts at each task. The newer and longer test has, as far as a search of its paper and site shows, no published human success rate at all.

The laboratories' launch pages report partial scores under conditions that differ from the board's and from each other. OpenAI's page for GPT-6 Astra, published on 3 September, headed a section "The world's best computer use model" and reported 72.6% on a row labelled "OSWorld 2.0 (v2026.08.08, offline set, partial

score)"—a partial-credit score on the 82-task subset that runs without internet access. Anthropic's page for Claude Opus 5.5, published on 22 September, reports "81.8% partial" on "OSWorld 2.1", without stating the subset in that row. Both are the companies' own runs, and on 11 October neither model had a row on the maintainers' official board.

The second misreading runs the other way: that agents which fail more than half the time are mostly hype. The workflows are long, and the failures are specific. The paper introducing OSWorld 2.0, first posted on 28 June and revised on 13 July, is explicit about where they lie:

These failures are not about basic GUI control or coding.

Agents, the authors write, "execute local actions well but cannot hold a task-level model together over a long horizon": they drop constraints they were given, miss information that arrives mid-task, guess instead of asking the user, and skip verification. The paper sorts the failures into five recurring dimensions, among them "perception–action timing", and notes that on tasks people find easy, "perceptual and interactive demands keep most workflows hard for agents". Completion, it adds, "collapses toward zero on the longest workflows even as partial scores stay high". Basic clicking and typing are no longer the main obstacle; tracking information, timing and verification over a long task are.

## An old word for a new machine

---

*Modality* is a word from the study of the senses. In 1878 Hermann Helmholtz, the German physicist and physiologist, gave an address called "The Facts of Perception" in which he distinguished two kinds of difference between sensations. Within a single sense, sensations differ in quality, as red differs from blue. Between senses—sight against taste, warmth against pitch—they differ in what, in an earlier work, he had called modality, and that difference, he said, is "so fundamental as to exclude any possible transition from one to another and any relationship of greater or less similarity". His illustration:

one cannot ask whether sweet is more like red or more like blue

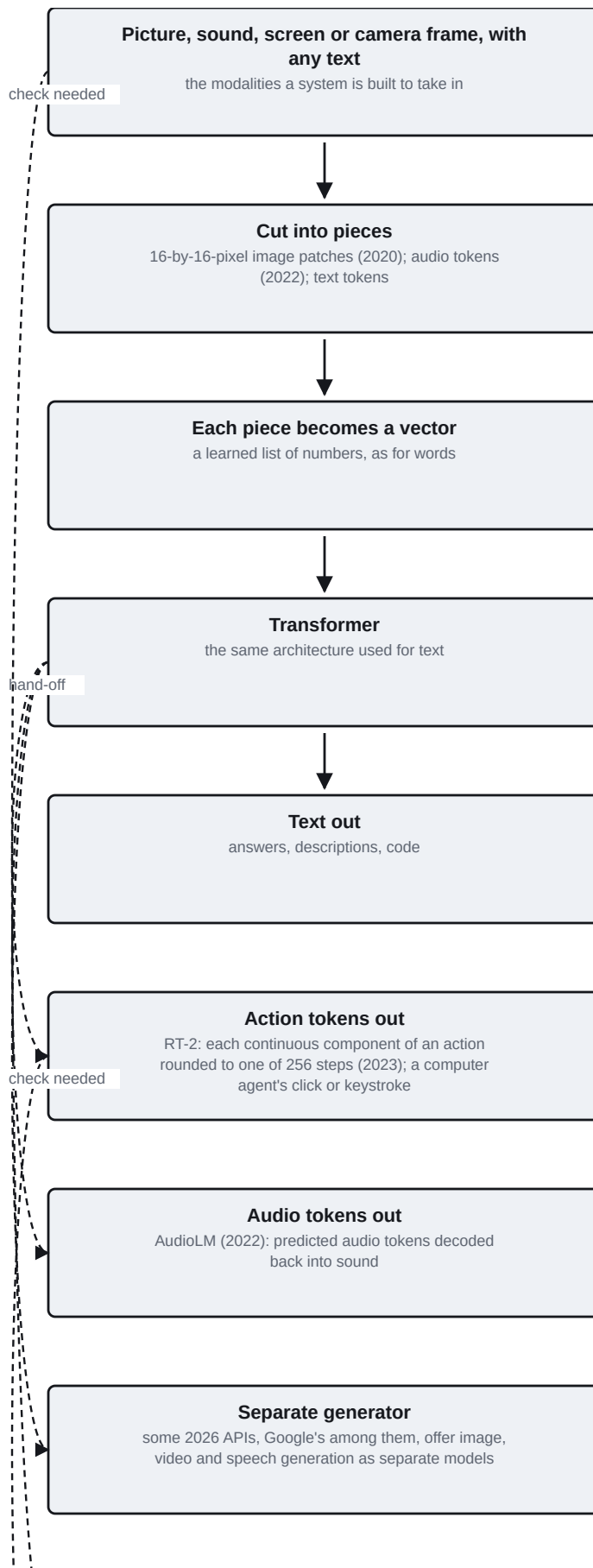
A useful way to picture what several machine-learning designs have since done, offered as an illustration rather than as anything Helmholtz claimed, is that they convert pictures, sound and even actions into the same kind of material. A language model does not take in words as such: text is cut into tokens, and each token is represented by a learned list of numbers, a vector. In October 2020 a team at Google showed that a photograph could be handled the same way. It cut images into squares 16 pixels on a side, turned each square into a vector, and gave the sequence to a standard transformer, the architecture built for text, trained to classify images. The paper's title was "An Image is Worth 16x16 Words"; its abstract reported that

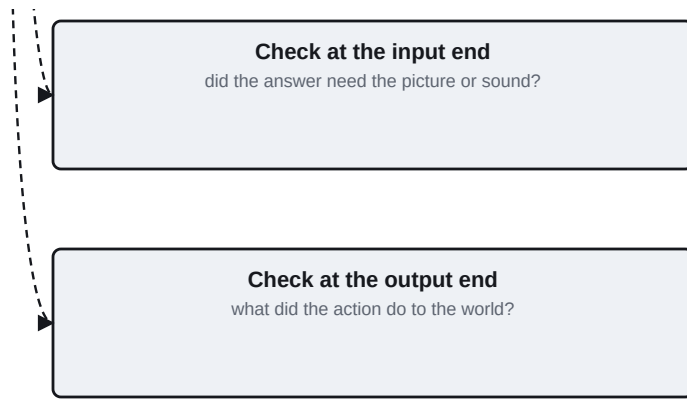
a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks

Those results, the abstract adds, came after pre-training "on large amounts of data". Sound followed. Google's AudioLM, posted in September 2022, "maps the input audio to a sequence of discrete tokens and casts audio generation as a language modeling task". So did action. In July 2023 Google DeepMind's RT-2 controlled a robot arm by rounding each continuous dimension of a movement into one of 256 steps and writing the result as tokens a language model already knew:

we express the actions as text tokens and incorporate them directly into the training set of the model in the same way as natural language tokens

**Figure 1. Pieces in, pieces out: how pictures, sound and actions pass through a language-model architecture**





*Schematic. Sources: Dosovitskiy et al., arXiv 2010.11929 (22 October 2020); Borsos et al., arXiv 2209.03143 (7 September 2022); Brohan et al., arXiv 2307.15818 (28 July 2023), section 3.2; model documentation from OpenAI, Anthropic and Google read 11 October 2026. Not every system uses every route: some robot models keep actions continuous rather than as tokens.*

#### ▼ Table view

Figure 1. Pieces in, pieces out: how pictures, sound and actions pass through a language-model architecture — stages

| #  | Stage                                                 | Note                                                                                                                   |
|----|-------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| 1  | Picture, sound, screen or camera frame, with any text | the modalities a system is built to take in                                                                            |
| 2  | Cut into pieces                                       | 16-by-16-pixel image patches (2020); audio tokens (2022); text tokens                                                  |
| 3  | Each piece becomes a vector                           | a learned list of numbers, as for words                                                                                |
| 4  | Transformer                                           | the same architecture used for text                                                                                    |
| 5  | Text out                                              | answers, descriptions, code                                                                                            |
| 6  | Action tokens out                                     | RT-2: each continuous component of an action rounded to one of 256 steps (2023); a computer agent's click or keystroke |
| 7  | Audio tokens out                                      | AudioLM (2022): predicted audio tokens decoded back into sound                                                         |
| 8  | Separate generator                                    | some 2026 APIs, Google's among them, offer image, video and speech generation as separate models                       |
| 9  | Check at the input end                                | did the answer need the picture or sound?                                                                              |
| 10 | Check at the output end                               | what did the action do to the world?                                                                                   |

Figure 1. Pieces in, pieces out: how pictures, sound and actions pass through a language-model architecture — connections

| From                                                  | To                          | Label        |
|-------------------------------------------------------|-----------------------------|--------------|
| Picture, sound, screen or camera frame, with any text | Cut into pieces             |              |
| Cut into pieces                                       | Each piece becomes a vector |              |
| Each piece becomes a vector                           | Transformer                 |              |
| Transformer                                           | Text out                    |              |
| Transformer                                           | Action tokens out           |              |
| Transformer                                           | Audio tokens out            |              |
| Transformer                                           | Separate generator          | hand-off     |
| Picture, sound, screen or camera frame, with any text | Check at the input end      | check needed |

| From              | To                      | Label        |
|-------------------|-------------------------|--------------|
| Action tokens out | Check at the output end | check needed |

Two designs coexist for joining the pieces. One bolts components together; DeepMind's Flamingo, in April 2022, was built to "bridge powerful pretrained vision-only and language-only models". OpenAI's announcement of GPT-4o in May 2024 described its earlier voice feature as "a pipeline of three separate models": one transcribed speech into text, a language model answered in text, and a third read the answer aloud, so that the central model "can't directly observe tone, multiple speakers, or background noises". The other trains a single network across modalities; for GPT-4o OpenAI said it had "trained a single new model end-to-end across text, vision, and audio", a description of its own system. Interface documentation does not show which design sits behind a product. By their makers' developer documentation, read on 11 October, OpenAI's gpt-6-astra takes images in but supports neither audio nor video; all of Anthropic's current Claude models take text and images and produce text; and Google's gemini-3.8-flash takes text, images, video, audio and PDF documents and produces text. Google separately lists models for real-time voice and for video generation.

| Selected API models (documentation read 11 October 2026) | Takes in                                       | Produces |
|----------------------------------------------------------|------------------------------------------------|----------|
| OpenAI gpt-6-astra                                       | text, images (audio and video "Not supported") | text     |
| Anthropic Claude, all current models                     | text, images                                   | text     |
| Google gemini-3.8-flash                                  | text, images, video, audio, PDF                | text     |

The ambition to make machines see is much older than the transformer, and its history carries a warning about how hard the problem looked. On 7 July 1966 Seymour Papert at the Massachusetts Institute of Technology circulated "The Summer Vision Project". Its stated aim was bounded: the project was "an attempt to use our summer workers effectively in the construction of a significant part of a visual system", chosen partly because the work could be split among individuals. Sixty years on, the authors of the newest desktop benchmark find that agents handle basic clicking and typing, and still struggle when a screen changes between a look and an action.

## The input end: did the answer use the picture?

A question about an image can often be answered from its wording, its options or what a model already knows, without the image. That is the first check a shared design does not supply: a test of seeing has to show that seeing was required. In the vocabulary of measurement this is construct validity, the question of whether a test measures what its name says, applied to a sense.

MMMU, published in November 2023 by Xiang Yue and colleagues, is a test of that kind: about 11,500 questions from college exams, quizzes and textbooks across 30 subjects, each built around an image—a chart, a diagram, a map, a musical score, a chemical structure. Four months later a team from the University of Science and Technology of China, the Chinese University of Hong Kong and the Shanghai AI Laboratory reported what happened when models were given such questions with the images removed:

Visual content is unnecessary for many samples.

Their headline example was a text-only model: Google's Gemini Pro scored 42.9% on MMMU "without any visual input". The paper's matched comparison is more telling. Given MMMU's questions with the images withheld, OpenAI's GPT-4V scored 45.1%; given the images, 53.6%. The vision version of Gemini Pro scored 39.4% and 44.4%. The benchmark's own baselines put random guessing at 22.1% and always choosing the most frequent answer at 26.8% on the validation set.

**Figure 2. The same models on MMMU with the images withheld and supplied**



Blue: the vision model given only the question and options. Orange: the same model given the images. Source: Chen et al., 'Are We on the Right Way for Evaluating Large Vision-Language Models?', arXiv 2403.20330 v1 (29 March 2024), Table 3, MMMU column (validation set). On the same table random choice scores 22.1%. The differences (8.5 and 5.0 percentage points) are this publication's subtraction. An image can raise a score without being needed for every correct answer.

#### ▼ Table view

Figure 2. The same models on MMMU with the images withheld and supplied

| Item                               | Value |
|------------------------------------|-------|
| GPT-4V, images withheld            | 45.1% |
| GPT-4V, images supplied            | 53.6% |
| Gemini Pro Vision, images withheld | 39.4% |
| Gemini Pro Vision, images supplied | 44.4% |

MMMU's authors answered with MMMU-Pro in September 2024. They asked four text-only language models every question ten times, without the images, and removed any question that at least three of the four answered correctly in most trials. Even then, they acknowledge, some questions could still be answered by text-only models, which is one reason they also widened the choices. Their paper prints two questions that a text-only model got right; in one, the model began:

I do not see the image, but the correct sequence based on the standard steps involved in bacteriophage infection is likely to be

and went on to name the right option. The rebuilt test widened the choices from about four to as many as ten, and added a setting in which the whole question arrives as a screenshot or photograph, so that the text cannot be read except through the image. The same models scored far lower.

**Figure 3. The same models on MMMU and on the filtered MMMU-Pro**

Blue: MMMU validation set. Orange: MMMU-Pro overall, the average of the ten-option setting and the setting in which the question is given as an image. Source: MMMU leaderboard data, read 11 October 2026; Yue et al., arXiv 2409.02813, Table 1. The GPT-4o and GPT-4o mini MMMU figures are marked on the leaderboard as reported by their developer. MMMU-Pro overall is the mean of the ten-option and image-only results.

#### ▼ Table view

Figure 3. The same models on MMMU and on the filtered MMMU-Pro

| Item                                    | Value |
|-----------------------------------------|-------|
| GPT-4o (May 2024), MMMU                 | 69.1% |
| GPT-4o (May 2024), MMMU-Pro             | 51.9% |
| Claude 3.5 Sonnet (June 2024), MMMU     | 68.3% |
| Claude 3.5 Sonnet (June 2024), MMMU-Pro | 51.5% |
| GPT-4o mini (July 2024), MMMU           | 59.4% |
| GPT-4o mini (July 2024), MMMU-Pro       | 37.6% |

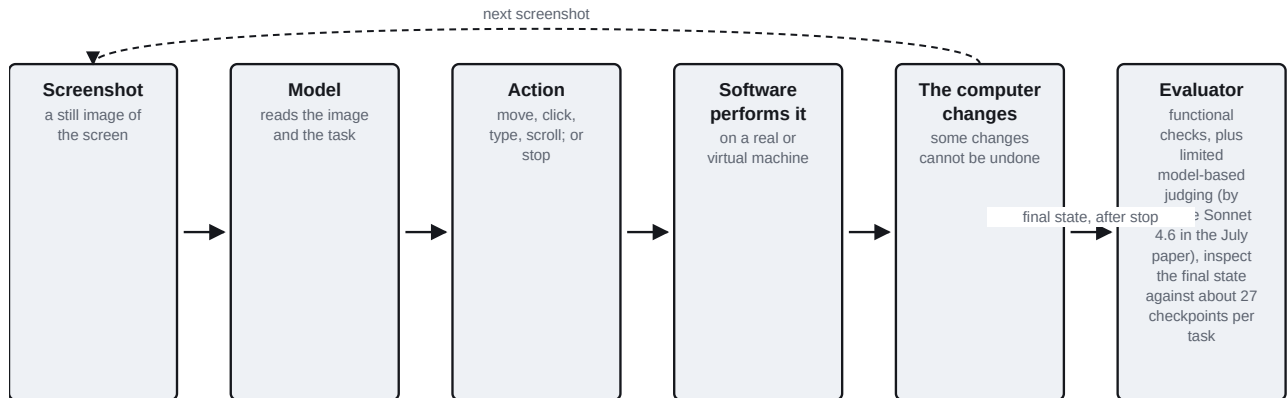
The lesson has not lost its relevance. On 11 October 2026 the highest MMMU-Pro rows on the benchmark's own leaderboard are marked as supplied by the models' developers, and the leaderboard's estimate of human expert performance on MMMU-Pro is, in its authors' words, an approximation "based on the original MMMU human evaluation data" rather than a new study. The independent evaluator Artificial Analysis runs MMMU-Pro's 1,730-question standard ten-option format; the benchmark's separate vision-only format, in which the whole question arrives as a screenshot or photograph, is not part of its published runs.

## The output end: the world answers back

The second thing that does not come with the design is responsibility for consequences. When software executes a model's output, a mistake can leave lasting changes: a message sent, a purchase made, a file altered. A computer-use agent is the familiar agent loop with eyes and hands: a screenshot comes in, the model returns an action—move to these coordinates, click, type—software around the model performs it,

and the next screenshot shows what happened. Anthropic, describing its own system in October 2024, said it worked by counting pixels to decide where to move the cursor, and by "taking screenshots and piecing them together, rather than observing a more granular video stream".

**Figure 4. The computer-use loop, and where it is checked**



*Schematic. Sources: Xie et al., arXiv 2404.07972 (OSWorld, April 2024), Figure 1; Yuan et al., arXiv 2606.29537 (OSWorld 2.0, July 2026), sections 1 and 2.1 (27.25 checkpoints per task on average; model-based evaluation contributes 11.53% of the total score); Anthropic, 'Developing a computer use model', 22 October 2024.*

#### ▼ Table view

Figure 4. The computer-use loop, and where it is checked — stages

| # | Stage                | Note                                                                                                                                                        |
|---|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | Screenshot           | a still image of the screen                                                                                                                                 |
| 2 | Model                | reads the image and the task                                                                                                                                |
| 3 | Action               | move, click, type, scroll; or stop                                                                                                                          |
| 4 | Software performs it | on a real or virtual machine                                                                                                                                |
| 5 | The computer changes | some changes cannot be undone                                                                                                                               |
| 6 | Evaluator            | functional checks, plus limited model-based judging (by Claude Sonnet 4.6 in the July paper), inspect the final state against about 27 checkpoints per task |

Figure 4. The computer-use loop, and where it is checked — connections

| From                 | To                   | Label                   |
|----------------------|----------------------|-------------------------|
| Screenshot           | Model                |                         |
| Model                | Action               |                         |
| Action               | Software performs it |                         |
| Software performs it | The computer changes |                         |
| The computer changes | Screenshot           | next screenshot         |
| The computer changes | Evaluator            | final state, after stop |

Irreversibility is the clearest example. When OpenAI tested the agent behind its Operator product before adding its safeguards, on 100 prompts resembling tasks users might give it, it counted 13 errors; eight "could be easily reversed (i.e., within a few minutes)", and

The other 5 mistakes were, to some degree, irreversible or possibly severe

among them an email sent to the wrong recipient and a medication reminder set for the wrong date, according to the Operator system card of 23 January 2025—the company's own test, on its own sample. Time is a second difference: the screen can change between the screenshot and the click, a condition OSWorld 2.0 tags as "streaming interaction" in about one task in eighteen. Data is a third. Google's RT-1 paper of December 2022 pointed to "the difficulty of collecting real-world robotic data" as the reason generalisation matters so much in robotics, and RT-2's authors were explicit about what web data could not supply: their robot "does not acquire any ability to perform new motions" from it, its physical skills remaining "limited to the distribution of skills seen in the robot data".

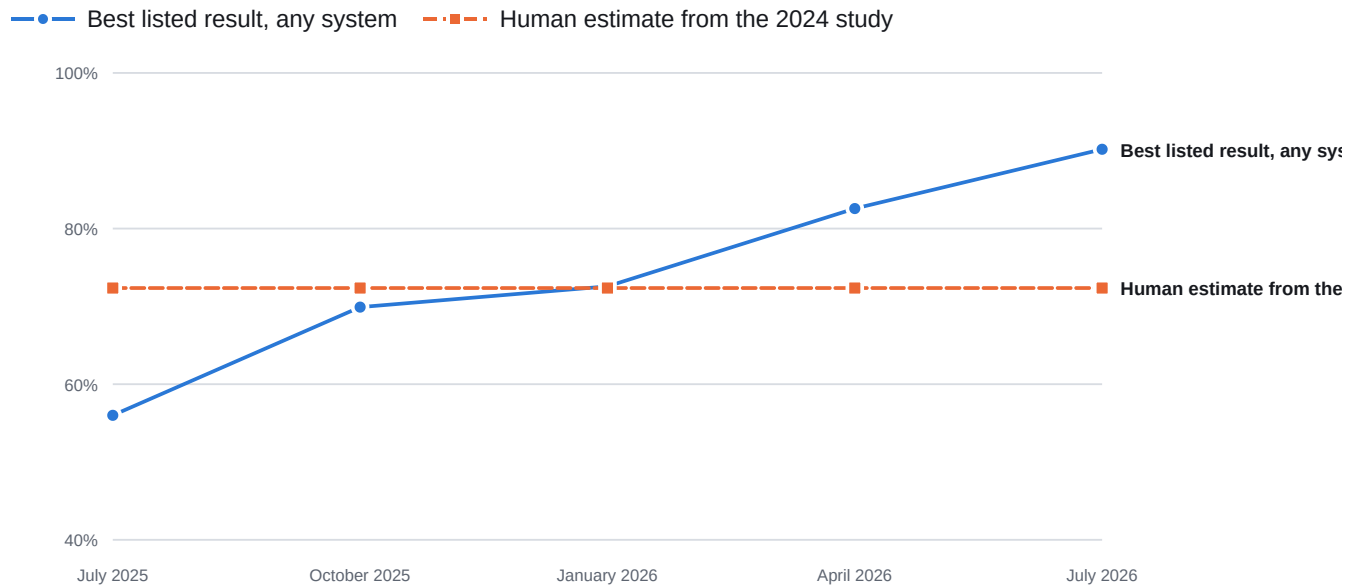
Perception and action, together, are the second idea: the system reads the world through one modality and changes it through another, and an error in either can carry through to a consequence.

## From 12% to 90%, and back to 44%

---

The history of OSWorld maps neatly onto both ends. When the benchmark appeared in April 2024, people scored about 72% and the best model about 12%. That model read no screenshots at all: it was a text-only model reading the "accessibility tree", a text description of the screen's elements. Working from screenshots alone, the strongest models scored between 5.3% and 5.8%, and the authors attributed the gap chiefly to "GUI grounding"—turning what is seen into the right place to click. In 2024, the input end was the bottleneck.

Commercial systems followed. Anthropic offered computer use in public beta on 22 October 2024, calling it "at times cumbersome and error-prone"; OpenAI released its Computer-Using Agent on 23 January 2025, describing it as trained to interact with "the buttons, menus, and text fields people see on a screen—just as humans do". On 28 July 2025 the maintainers relaunched the benchmark as OSWorld-Verified, with community-reported tasks fixed and results run under unified settings, though trusted institutions may also submit their own trajectories for checking. Scores then climbed through the human estimate.

**Figure 5. OSWorld-Verified: best listed result by selected month ends, against the 2024 human estimate**

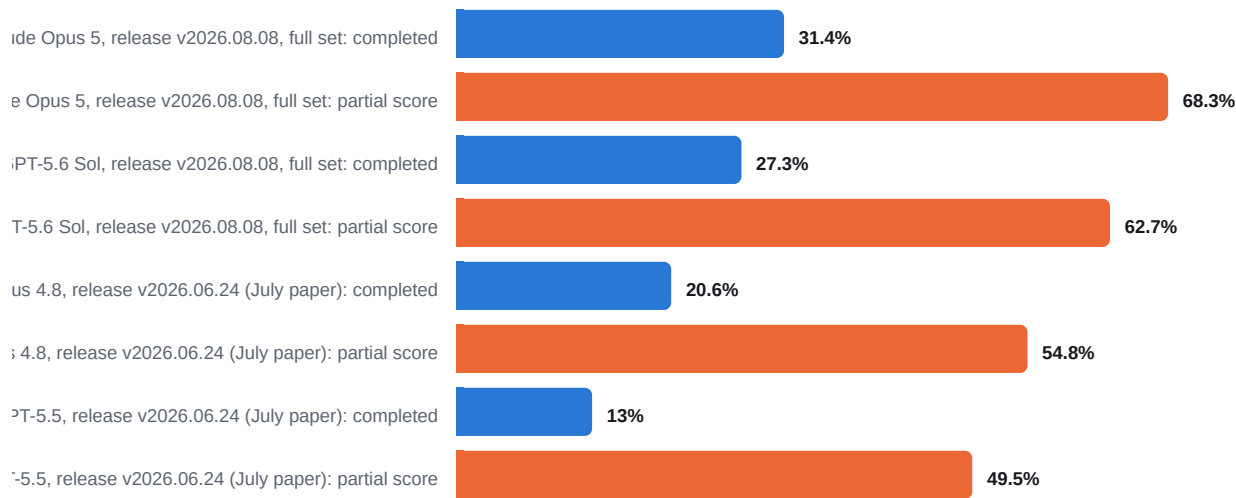
Source: the maintainers' OSWorld-Verified results file on the OSWorld project site, read 11 October 2026; human figure from Xie et al., arXiv 2404.07972, measured on computer-science students new to the software. Several leading entries are agentic frameworks that act through code as well as clicks; the December 2025 entry drew on ten attempts per task. The human figure was not re-measured on the repaired tasks.

#### ▼ Table view

Figure 5. OSWorld-Verified: best listed result by selected month ends, against the 2024 human estimate

| End of month | Best listed result, any system | Human estimate from the 2024 study |
|--------------|--------------------------------|------------------------------------|
| July 2025    | 56%                            | 72.4%                              |
| October 2025 | 69.9%                          | 72.4%                              |
| January 2026 | 72.6%                          | 72.4%                              |
| April 2026   | 82.6%                          | 72.4%                              |
| July 2026    | 90.2%                          | 72.4%                              |

By then the maintainers had built a harder test. OSWorld 2.0, released on 26 June 2026, has 108 long workflows where the original had 369 shorter tasks; one leading agent needed an average of 318 tool calls per workflow, against about 30 on the original. In the paper's July revision the best agent, Claude Opus 4.8, completed 20.6% of workflows at a partial score of 54.8%. By 5 October, on a revised release of the tasks (v2.1), the best official run completed 44.3%. Releases differ, so the two figures do not make a clean trend; what changed more clearly is the kind of failure. In 2024 models struggled to turn a screenshot into the right place to click; in 2026 the paper's authors find basic clicking and typing largely handled, and failures concentrated in tracking information, timing and verification over long tasks.

**Figure 6. OSWorld 2.0: completion and partial credit for the same runs, models paired by task release**

Blue: share of the 108 workflows completed outright. Orange: partial score, the weighted share of checkpoints satisfied by the final state. Completion and partial credit are paired within each configuration, and each pair of models shares a task release; results from different releases are not matched comparisons. Claude Opus 5's best official full-set run, on the later release v2.1 (44.33% completed, 77.67% partial), has no OpenAI counterpart on that release and is not plotted. All at maximum or extra-high reasoning settings and a 500-step budget. Sources: OSWorld 2.0 official results data, updated 5 October 2026 ([osworld-v2.xlang.ai](https://osworld.v2.xlang.ai)); Yuan et al., *arXiv 2606.29537 v2* (13 July 2026). The board averages 7 runs for Claude Opus 5 and 2 for GPT-5.6 Sol on release v2026.08.08. Lab-reported figures for GPT-6 Astra (72.6%, offline subset) and Claude Opus 5.5 (81.8%) are partial scores from the companies' own runs and do not appear on the board.

#### ▼ Table view

Figure 6. OSWorld 2.0: completion and partial credit for the same runs, models paired by task release

| Item                                                             | Value |
|------------------------------------------------------------------|-------|
| Claude Opus 5, release v2026.08.08, full set: completed          | 31.4% |
| Claude Opus 5, release v2026.08.08, full set: partial score      | 68.3% |
| GPT-5.6 Sol, release v2026.08.08, full set: completed            | 27.3% |
| GPT-5.6 Sol, release v2026.08.08, full set: partial score        | 62.7% |
| Claude Opus 4.8, release v2026.06.24 (July paper): completed     | 20.6% |
| Claude Opus 4.8, release v2026.06.24 (July paper): partial score | 54.8% |
| GPT-5.5, release v2026.06.24 (July paper): completed             | 13%   |
| GPT-5.5, release v2026.06.24 (July paper): partial score         | 49.5% |

## Two ways to touch a computer

How an agent should meet software at all is an open design question, and the evidence supports two postures. The first is to use the screen—the interface built for human eyes and hands—as OpenAI's agent was trained to, which, OpenAI said, gives it "the flexibility to perform digital tasks without using OS-or web-specific APIs". The second is to build an interface for the agent. In May 2024 the Princeton group behind the SWE-bench coding benchmark argued that

LM agents represent a new category of end users

and gave their agent purpose-built commands, among them a file viewer and an editing command, under the name "agent-computer interface"; on SWE-bench, the paper reported, the agent solved 12.5% of issues against a previous best of 3.8%, and on a 300-issue subset its interface beat a plain Linux shell by 10.7 percentage points. Neither posture has won. The original OSWorld paper found that adding the text description of the screen helped some models and misled others; several of the highest OSWorld-Verified entries are frameworks that act through code as well as clicks. The reading that fits the evidence is that screen interaction offers a route into many graphical applications, while in SWE-agent's coding experiment, commands tailored to the agent outperformed a default Linux shell. A robot, too, needs a designed control interface, and a command it issues still has to produce the intended result in the physical world.

The two robot systems discussed here report their developers' own evaluations. RT-2's paper reported about 6,000 evaluation trials, run by Google DeepMind. Google DeepMind's announcement of Gemini Robotics 2 on 30 July 2026 shows average success rates by skill category for whole-body and gripper tasks and individual results for multi-finger tasks, which it describes as still "challenging"; the post gives no trial counts, and its robot-control (VLA) and on-device models are available to early-access partners, while its reasoning model is on Google AI Studio. RoboArena, described in a June 2025 paper, offers a distributed alternative: double-blind comparisons of pairs of robot policies run by evaluators at several academic institutions.

## What to watch

---

**The OSWorld 2.0 board** ([osworld-v2.xlang.ai](https://osworld-v2.xlang.ai)), last updated on 5 October 2026. Whether official rows appear for GPT-6 Astra and Claude Opus 5.5, whose makers have published only their own runs, and for any Gemini model, which has none; and whether any run completes half of the 108 workflows outright, against 44.3% today.

**MMMU's leaderboard** (on the MMMU project site), whose newest row is dated 1 July 2026. Whether a 2026 flagship's MMMU-Pro result appears from someone other than its developer, and in particular for the setting in which the question arrives as a photograph.

**The Conference on Robot Learning, 9–11 November 2026.** Whether robot results presented there say how many trials were run, on real robots or in simulation, and who ran them.

**Gemini Robotics 2.** Whether Google DeepMind's VLA and on-device models move beyond early-access partners, or whether success rates with trial counts are published by someone other than Google.

## The idea to keep

The systems described here that are not chatbots still read pieces: patches of a picture, stretches of sound, screenshots of a desktop, quantised components of a robot's action. That design transferred from text, and it transferred remarkably well. The checking has to be rebuilt for each new kind of input and output. A test of seeing has to show that seeing was needed, as MMMU's own authors found when text-only models answered its picture questions. A test of acting has to inspect what the action did to the world, and report whether it counted completion or partial credit.

## Sources

| Source                                                                                                                                                                                            | Date                                                     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|
| Hermann Helmholtz, <i>The Facts of Perception</i> , address of 1878, English translation in <i>Selected Writings of Hermann Helmholtz</i> (Wesleyan University Press), reproduced at marxists.org | 1878; read 11 October 2026                               |
| Seymour Papert, <i>The Summer Vision Project</i> , MIT Artificial Intelligence Group, Vision Memo No. 100 (MIT DSpace)                                                                            | 7 July 1966                                              |
| Alexey Dosovitskiy and colleagues (Google), <i>An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale</i> , arXiv 2010.11929                                                  | 22 October 2020                                          |
| Zalán Borsos and colleagues (Google), <i>AudioLM: a Language Modeling Approach to Audio Generation</i> , arXiv 2209.03143                                                                         | 7 September 2022                                         |
| Anthony Brohan and colleagues (Google), <i>RT-1: Robotics Transformer for Real-World Control at Scale</i> , arXiv 2212.06817                                                                      | 13 December 2022                                         |
| Anthony Brohan and colleagues (Google DeepMind), <i>RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control</i> , arXiv 2307.15818, abstract and section 3.2                | 28 July 2023                                             |
| Xiang Yue and colleagues, <i>MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI</i> , arXiv 2311.16502                                              | 27 November 2023                                         |
| Lin Chen and colleagues, <i>Are We on the Right Way for Evaluating Large Vision-Language Models?</i> , arXiv 2403.20330, abstract                                                                 | 29 March 2024                                            |
| Tianbao Xie and colleagues, <i>OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments</i> , arXiv 2404.07972                                                  | 11 April 2024                                            |
| John Yang, Carlos E. Jimenez and colleagues (Princeton University), <i>SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering</i> , arXiv 2405.15793                          | 6 May 2024                                               |
| OpenAI, <i>Hello GPT-4o</i> (Internet Archive capture)                                                                                                                                            | 13 May 2024                                              |
| Xiang Yue and colleagues, <i>MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark</i> , arXiv 2409.02813, sections 2 and 3 and Figure 2                                    | 4 September 2024                                         |
| Anthropic, <i>Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku; Developing a computer use model</i>                                                                        | 22 October 2024                                          |
| OpenAI, <i>Computer-Using Agent; Operator System Card</i> (Internet Archive captures)                                                                                                             | 23 January 2025                                          |
| Conference on Robot Learning 2026, programme page (corl.org)                                                                                                                                      | main conference 9–11 November 2026; read 11 October 2026 |

| Source                                                                                                                                                      | Date                                         |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------|
| Pranav Atreya and colleagues, <i>RoboArena: Distributed Real-World Evaluation of Generalist Robot Policies</i> , arXiv 2506.18123                           | 22 June 2025                                 |
| XLANG Lab, <i>OSWorld-Verified</i> announcement; OSWorld-Verified results file                                                                              | 28 July 2025; read 11 October 2026           |
| Mengqi Yuan, Tianbao Xie, Tao Yu and colleagues, <i>OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks</i> , arXiv 2606.29537   | 28 June 2026; revised 13 July 2026           |
| Google DeepMind, <i>Gemini Robotics 2 brings whole-body intelligence to robots</i>                                                                          | 30 July 2026                                 |
| OpenAI, <i>GPT-6 Astra</i> launch page (Internet Archive capture)                                                                                           | 3 September 2026                             |
| Anthropic, <i>Introducing Claude Opus 5.5</i> (Internet Archive capture)                                                                                    | 22 September 2026                            |
| OSWorld 2.0 official results (osworld-v2.xlang.ai)                                                                                                          | updated 5 October 2026; read 11 October 2026 |
| MMMU leaderboard data (MMMU project site); Artificial Analysis, MMMU-Pro results and methodology                                                            | read 11 October 2026                         |
| OpenAI developer documentation, gpt-6-astra model page; Claude documentation, models overview; Google Gemini API documentation, gemini-3.8-flash model page | read 11 October 2026                         |