PURETENSOR
INTELLIGENCE DISTILLED

Click Accuracy Picks the Wrong Browser Agent

Choosing a self-hosted computer-use model on 39 live tasks

Paper: PT-R-2026-028 v1.0

Author: PureTensor Ltd

Date: 26 September 2026

Status: Published

Abstract

We needed to choose a self-hosted, open-weights model to drive our browser agents on Linux or macOS. The ranking rule was deliberately narrow: does it run on our hardware, and how well does it perform. Model origin and licence were recorded as notes and did not enter the ranking. After a literature and model-card survey removed candidates that could not be served locally, we compared five setups on one evening, 26 September 2026. We first measured single-click targeting on a corpus of 72 live targets. We then ran a bake-off of 39 live tasks on public websites, scored only by verified end state, with no language-model judge.

The click corpus separated the three leading vision models by 3 targets out of 72. UI-Venus-2-9B hit 95.8%, Holo3.1-35B-A3B 93.1%, and Qwen3.8-27B 91.7%. A fourth model, Fara1.5-9B, hit 41.7% and was dropped. On the live tasks the order changed. Holo3.1-35B-A3B finished last at 33/39, 4 tasks behind the next setup and 5 behind the best vision setup. Qwen3.8-27B in a lean screenshot loop reached 38/39 at a median of 12.7 s per task. The DOM-only incumbent, DeepSeek-V4.1-Flash inside browser-use, scored 39/39, but it needs four GPUs across two nodes. Most of Holo's failures were errors of judgement and recovery rather than missed pixels.

The agent loop moved results as far as the choice of model did. The same Qwen3.8-27B weights scored 37/39 at a median of 38.2 s inside browser-use with vision on, 3.0x slower at the median than in the lean loop. A single viewport constant moved Holo's standard tier from 17/26 to 23/26. The cause is that a headless Chrome given a 1280x800 window and attached over its DevTools protocol reports a 1280x713 page, so a harness that assumes 800 places every click about 12% too low.

Every screenshot-only agent failed the one multi-field form with autocomplete, and both setups that could read the DOM passed it. The top four setups (39, 38, 37 and 37) cannot be separated on this sample: one run each, 39 tasks, and 2.6 percentage points per task. The findings that hold up are these: click accuracy mis-ranked Holo by a margin larger than the spread among the top four, the same-model harness speed gap is large, and the viewport trap has a measured cost. No ordering among the top four is supported.

1. The Question and the Rule

The question was practical. Our browser agents need a model to look at pages and decide what to do next, and we wanted that model to run on our own GPUs, served locally from open weights. An early framing weighed the model's origin and licence alongside performance. We dropped that, and the rule we used is simpler: a candidate either runs on our hardware or it does not, and among those that do, we rank by task performance. Origin and licence are recorded where relevant and ignored for ordering.

That rule shaped everything that followed. It meant a model with excellent published numbers but no downloadable weights was out, however good it looked on paper. It also meant we could not rely on published leaderboards to settle the order. They measure different tasks under different harnesses, and, as this paper shows, the harness moves the result.

2. Narrowing the Field

A literature and model-card survey came first, split across ten research sub-agents. It produced two exclusion lists. The first held candidates that were not runnable locally, because weights were unavailable, access was API-only, or the model could not be served on our stack: UI-Venus-2-27B, Qwen-UI-Agent-27B, UI-TARS-2, Holo3-122B, and Holo4 (API only). The second held models judged too large or too slow for an interactive agent loop: Kimi K2.6 and K3.

The survey also produced four false claims, and we caught each by checking primary sources. It asserted that UI-TARS-2 weights are open, that a ScaleCUA-Qwen3.5-9B checkpoint is on Hugging Face, that Qwen3.8 lacks vision, and that UI-Venus-2 weights are Apache-licensed. None survived verification. We record this because a delegated survey is a fast way to build a candidate list and an unreliable way to establish facts about it. Every claim that affected eligibility was checked against the model card or repository before we acted on it.

3. Setups, Hardware and Harnesses

We compared five setups. Here a harness (or agent loop) is the code that turns a model's outputs into browser actions and feeds observations back.

Label	Model	Agent loop	Observation
DOM incumbent	DeepSeek-V4.1-Flash	browser-use, as our in-house wrapper deploys it	DOM / accessibility text only
Qwen lean	Qwen3.8-27B NVFP4	H Company's published cookbook browser agent	screenshots only
UI-Venus	UI-Venus-2-9B	its own vendor browser agent	screenshots only
Qwen in browser-use	Qwen3.8-27B NVFP4	browser-use with vision mode on	DOM plus screenshots
Holo	Holo3.1-35B-A3B NVFP4	H Company cookbook agent (its native harness)	screenshots only

The DOM is the page's document tree. The accessibility tree is the structured, labelled view of it that assistive technology reads. The incumbent sees only this text and never a screenshot.

The principle was that each model runs in the harness it ships with, or in the one we already run. We kept H Company's lean screenshot agent verbatim in everything that touches the model. It uses the same system prompt and five tools: click at x,y with an element description, type with optional Enter, scroll up or down, go to URL, and answer. Coordinates are in a 0 to 1000 normalised space. Only the last 3 screenshots stay in context, with older ones evicted to a text stub. Temperature is 0.8, thinking is on, and a tool call is required every turn. Our changes were confined to plumbing. The endpoint is read from the environment. The agent attaches to

the runner's Chrome instead of launching its own. A failed action returns an error string rather than crashing. The agent prints a final JSON line for the scorer. We also applied the viewport fix described in Section 8. The UI-Venus agent behaves differently. It keeps every turn's text but images only for the latest turns, runs at temperature 0, has a note-taking action, and measures the real viewport from the page.

Every vision model was served with vLLM (nightly build) on one GPU of a two-GPU compute node of the NVIDIA RTX PRO 6000 Blackwell class. That GPU was shared with production work, leaving about 34 GB free. For the models we launched, serving limits were a maximum model length of 32,768 tokens, at most 4 concurrent sequences, and at most 5 images per prompt. UI-Venus-2-9B took 18 GB in BF16 (16-bit brain floating point) weights. Holo3.1-35B-A3B in NVFP4 (a 4-bit floating-point weight format) took 24 GB. Qwen3.8-27B NVFP4 was already running as a production serving seat and needed no new deployment. The incumbent ran on DeepSeek-V4.1-Flash served across four RTX PRO 6000 GPUs on two nodes, using tensor and expert parallelism of width 4 (splitting each layer's matrices, and the mixture-of-experts blocks, across the four cards). We switch this mode on for heavy work.

4. The Task Set, Scoring and Timeline

The bake-off used 39 live tasks on public websites. Each task got a fresh headless Chrome with a fresh profile, a limit of 20 agent steps, and a 600 s wall-clock limit. Each setup ran 4 tasks in parallel, with a UK English locale and a fixed desktop user-agent string. The window was 1280x800 for four setups and 1024x768 for the UI-Venus vendor agent.

Scoring used verified end state only. A task passes if any open tab's final URL matches a regular expression, or if the agent's final answer matches one (case-insensitive). 31 tasks are URL-checked and 8 are answer-checked.

The standard tier has 26 tasks of three kinds. The 7 search-and-open tasks include finding the Hallgrimskirkja article on Wikipedia, searching IKEA UK for "BILLY bookcase", and finding the GOV.UK adult passport renewal page. The 13 navigate-to-a-section tasks include Hacker News newest, Ask HN, the Issues tab of the vLLM repository, PEP 8 starting from python.org, arXiv new cs.CL submissions, and the Firefox download page from mozilla.org. The 6 find-and-answer tasks have fixed answers: the height of Hallgrimskirkja (74.5 m), the founding year of the University of Iceland (1911), the vLLM licence (Apache), the module providing Counter (collections), the UK standard VAT rate (20%), and an arXiv title by identifier.

The hard tier has 13 tasks involving widgets, maps, forms, dropdowns, media viewers and multi-hop lookups. Nine test controls and widgets: zooming OpenStreetMap to level 15 or closer with central Reykjavik in view, switching the Python docs version selector to 3.12, sorting Hugging Face models by likes, and opening a Wikipedia infobox photo, among others. Two are multi-field forms. The first is OpenStreetMap walking directions from Hallgrimskirkja to Harpa concert hall, which needs two autocomplete place fields and a travel-mode selector; it is verified by a routing engine name containing "foot" and both endpoints lying in central Reykjavik. The second is arXiv Advanced Search for "computer use agent" restricted to titles. The last two are live-answer lookups: the latest vLLM release tag (v0.30.0 at test time, checked against the GitHub API) and the latest playwright version on PyPI (1.63.0, checked against PyPI JSON).

One check changed during the evening. mozilla.org's Firefox download pages now 301-redirect to firefox.com, so at 21:04 we widened that check to accept firefox.com. All results were rescored against the final task files, and runs started after 21:04 used the widened check live.

Time (UK, 26 September 2026)	Event
20:30 to 20:31	click corpus built from live pages (72 targets, 13 pages)

Time (UK, 26 September 2026)	Event
20:32 to 20:46	click probes, including the Fara rescore at 20:46
20:55 to 20:57	two-task smoke runs of each harness
20:57 to 20:58	standard tier: incumbent and UI-Venus
21:04	Firefox check widened; 21:06 all runs rescored
21:05	Holo standard run with the viewport bug (17/26, invalid)
21:10	viewport fixed; Holo standard rerun (23/26)
21:11 to 21:30	remaining hard and standard tiers run sequentially on one GPU
21:40	pipeline complete

5. Single-Click Targeting

Before the live runs we measured grounding: whether a model, given a screenshot and a description of an element, returns a point on that element. The corpus consisted of live 1280x800 screenshots of 13 public home or index pages. Three further sites were dropped at build time for having too few usable targets or failing to load. We sampled up to 6 targets per page with a fixed seed from visible, unobscured, uniquely labelled links and buttons, giving 72 targets (57 links, 15 buttons). Ground truth is the element's real DOM bounding box, and a hit is a predicted point inside the box with 3 px tolerance. Each model got its vendor's own documented grounding prompt and output format at temperature 0, with thinking off for the three leaders and one unscored warm-up call each. Fara1.5 used its own system prompt and tool-call format with a 1000x1000 virtual screen.

Model	Hits / 72	Accuracy	Parse failures	Latency p50 / p90	Median prompt tokens
UI-Venus-2-9B	69	95.8%	0	0.268 s / 0.340 s	1,086.5
Holo3.1-35B-A3B	67	93.1%	0	0.129 s / 0.536 s	1,181.5
Qwen3.8-27B	66	91.7%	0	0.280 s / 0.700 s	1,086.5
Fara1.5-9B	30	41.7%	1	0.954 s / 1.756 s	2,738.5

The three leaders are close. The misses overlap: a Hugging Face size-filter button labelled "< 1B" defeated all three, OpenStreetMap controls accounted for two misses each for UI-Venus and Qwen, and Hacker News usernames and comment counts cost Holo and Qwen. Fara1.5-9B scored 0/6 on each of the arXiv, Wikipedia and Hacker News pages. Rescoring it with its coordinates read as raw pixels rather than 0 to 1000 gave 22/72, which is worse. It often emitted mouse-move actions narrated as "drag start point" instead of clicks. We dropped it from the live bake-off.

Taken alone, this table would suggest UI-Venus first, Holo second and Qwen third, all fit for the job.

6. Live Task Results

All scores below were recomputed by us from the per-task result files using the final checks. They match the harness's own consolidation script exactly. p90 is nearest-rank over the 39 per-task wall times, meaning it is the actual observed time at the 90th-percentile position rather than an interpolation. Wall time includes Chrome start-up, page loads over the public

internet, and model time, with 4 tasks sharing the model server.

Setup	Standard (26)	Hard (13)	Total (39)	Median s/task	p90 s/task	Median s, std / hard	Footprint
DOM incumbent	26	13	39/39	20.7	93.7	20.4 / 22.3	four GPUs, two nodes
Qwen lean	26	12	38/39	12.7	24.6	11.8 / 15.1	already-running seat, one GPU
UI-Venus	25	12	37/39	16.0	30.4	15.2 / 16.5	18 GB
Qwen in browser-use	25	12	37/39	38.2	129.3	34.5 / 48.7	already-running seat
Holo	23	10	33/39	12.1	36.1	12.1 / 12.6	24 GB

Holo, second on single clicks, is last here, 4 tasks behind the next setup. Qwen, third on single clicks, leads the screenshot agents. The top four lie within 2 tasks of a perfect score, and we do not treat their order as a ranking (Section 10). They differ more clearly in cost and speed. The only 39/39 needs the four-GPU mode. The best screenshot setup reached 38/39 on a model we were already serving, at the lowest median time of any setup that scored above 33.

Summed wall times need care. The lean Qwen setup totalled 1,205.6 s, of which one timed-out task contributed 600.1 s, so its mean of 30.9 s per task describes that single hang more than the other 38 runs. The incumbent's mean was 31.3 s, UI-Venus 19.2 s, Qwen in browser-use 55.8 s and Holo 16.0 s. Holo is fast partly because it gave up or declared success early.

The same Qwen3.8-27B weights scored 38/39 at a median of 12.7 s in the lean loop and 37/39 at 38.2 s inside browser-use vision mode. The median per-step time was 4.9 / 4.0 s (standard / hard) in the lean loop against 11.2 / 10.7 s in browser-use. For comparison, the incumbent ran at 5.9 / 4.9, UI-Venus at 3.5 / 3.6 and Holo at 3.1 / 2.9. On the 8 answer-checked tasks, the incumbent, Qwen lean and UI-Venus each scored 8/8, while Qwen in browser-use and Holo each scored 7/8.

Setup	Median steps	p90	Max	Tasks at 20-step cap	Median steps std / hard
DOM incumbent	4	7	19	0	3.5 / 4
Qwen lean	2.5	5	11	0 (one 600 s timeout)	2.0 / 3.0
UI-Venus	4	7	20	1	4.0 / 5
Qwen in browser-use	3	7	11	0	3.0 / 4
Holo	4	15	20	2	4.0 / 4

Step counts are not comparable across harnesses. A browser-use step can batch several actions, while one lean-agent step is exactly one action. Within a setup, the tail is informative. Holo's p90 of 15 steps and two capped tasks show loops that did not converge. The incumbent's hardest task, the walking directions, took 19 steps and 130.6 s, one step inside the cap.

7. Where Each Setup Failed

Setup	Failed task (tier)	What happened
DOM incumbent	none	39/39

Setup	Failed task (tier)	What happened
Qwen lean	OSM walking directions (hard)	two actions (open directions, pick walking mode), then hung to the 600 s timeout
UI-Venus	PEP 8 from python.org (std)	opened an old "Python style guide" essay page, not PEP 8, and reported success
UI-Venus	OSM walking directions (hard)	used all 20 steps; autocomplete chose a "Harpa" far outside Reykjavik (about 60.16 N, 1.31 W)
Qwen in browser-use	University of Iceland founding year (std)	reached the right article in 4 steps; final answer malformed ("founded in 19"); cause not established
Qwen in browser-use	arXiv Advanced Search (hard)	ran the search correctly but quoted the phrase, so the strict URL check failed
Holo	Hacker News newest (std)	landed on the front page and reported it as newest
Holo	Ask HN (std)	clicked "ask", page did not change, gave up after 3 steps with a plan instead of an action
Holo	arXiv new cs.CL (std)	opened the recent listing rather than new
Holo	OSM walking directions (hard)	20 steps typing into the wrong field, leaving text such as "Harpa concert hHarpaHarpa"
Holo	Python docs version selector (hard)	20 steps alternating between opening the dropdown and clicking "3.12" without effect
Holo	latest vLLM release (hard)	answered "v0.30.1rc0", a pre-release, in 2 steps

Two patterns stand out. First, the walking-directions form failed for every screenshot-only agent (Holo, Qwen lean and UI-Venus) and passed for both setups that could read the DOM. No other task failed for more than one setup. On this set, a form with autocomplete fields is where pixels alone ran out.

Second, Holo's six failures are mostly failures of judgement and recovery, not of pointing. It declared the wrong page done twice, gave up once, repeated a non-working dropdown interaction to the cap, and chose a pre-release tag. Only the directions form involves anything like a targeting error, and even there the problem was persistence in the wrong field, not a single missed click. This is why the click corpus could not see it: a first-glance grounding test asks whether the model can hit a named element, and never asks whether it knows which element to want, or notices that nothing happened.

Neither of Qwen-in-browser-use's failures is a navigation failure. One is a malformed answer after reaching the correct page. The other is a quoting choice our check did not accept, which we regard as arguably a checker artefact.

8. The Headless Viewport Trap

A Chrome started headless with a 1280x800 window and attached over the DevTools protocol (the remote-control interface Chrome exposes to automation tools) reports a page viewport of 1280x713, not 1280x800. The screenshot the model sees is therefore 713 px tall. A harness that converts the model's normalised y coordinate (0 to 1000) to pixels using the window

constant of 800 places every click at 800/713 = 1.12 times its intended height. That is about 12% too low, and the error grows with distance down the page.

We measured its cost directly. The first Holo standard-tier run, at 21:05, had this bug and scored 17/26, with 5 tasks burning all 20 steps. After we set the viewport explicitly to 1280x800, the rerun at 21:10 scored 23/26 with no task reaching the cap. It was the same model on the same tasks, and one constant made a difference of 6 tasks. The buggy run is retained, marked invalid, and excluded from every table above.

There are two fixes: set the page viewport size explicitly, or measure the page's inner width and height at each step. The UI-Venus vendor agent already measures. The H Company cookbook agent, as published, assumes it launched the browser and set the viewport itself. That assumption does not hold when it attaches to an existing headless browser, which is how a shared runner operates.

9. Why the Bake-off Drives Chrome Directly

Nine days earlier, on 17 September 2026, we evaluated Cua Driver 0.28.2, an open-source (MIT) desktop-automation driver that speaks the Model Context Protocol (a standard interface through which models call external tools). It installed cleanly on all 7 Linux nodes we probed, each meeting glibc 2.31 or newer, Python 3.10 or newer, and a present virtual input device, and it exposed 62 tools on Linux.

Under Xvfb (a headless virtual X display server) with a window manager, inspection worked. That covered screenshots, a 135-element accessibility tree of a test application, a clipboard round-trip, trajectory recording, and text setting through the accessibility API, which was the only exact-target text route we observed. Input did not work. The driver's headline feature, background input that does not steal focus, was refused headless for key presses, hotkeys, scroll, right-click, double-click and drag, with the error "the requested target has no focus-free input backend". Multi-pointer drag timed out because Xvfb does not hot-plug the virtual pointer. A click did not move keyboard focus, so typed text landed in a different widget from the one clicked, in both background and foreground modes. Without a window manager, every input tool failed for lack of a window process ID.

Upstream's own action ledger shows the same pattern. Linux/X11 refuses 41 of 116 actions (35%), against 23 of 122 on Windows (19%) and 6 of 145 on macOS (4%), and it lists a real Xorg session as "not yet validated". Telemetry was on by default and had registered the install before we disabled it. Operating-system-level computer use on headless Linux is not yet a dependable substrate for us. That is why the bake-off drives Chrome over its DevTools protocol.

10. What These Numbers Will Not Carry

Every setup ran each task once. The cookbook agent ran at temperature 0.8, so a repeat could differ. There are no repeats and no confidence intervals. With 39 tasks, one task is 2.6 percentage points, and the top four setups span 2 tasks. Their order (39, 38, 37, 37) is inside what a single rerun could plausibly reshuffle, and we do not read it as a ranking.

The sites were live and public. The Firefox redirect changed a check mid-evening. Live answers such as the latest release tag and the latest package version move over time. Pages also differ by time and region. These results are a snapshot of 26 September 2026 from a UK locale.

The checks are regular expressions on URLs and answers, and they are strict. At least one failure, the quoted arXiv phrase, is arguably a checker artefact. A pass verifies the end state, not that the path to it was sensible. Answer checks accept any matching substring, so they are lenient in the other direction.

The setups did not share identical conditions. The UI-Venus agent ran at a 1024x768 window and the others at 1280x800, while the click corpus used 1280x800 screenshots for every model. Timings are confounded by four tasks in parallel per setup, vision models sharing one GPU with production traffic, the incumbent running on a different four-GPU deployment, and page-load variance on the public internet. Median seconds are indicative only. The roughly threefold same-model harness gap is large enough to survive these confounds, but sub-second differences are not. Step counts mean different things in different harnesses and should not be compared across them.

The click corpus is 72 targets on 13 pages. It contains links and buttons only, with no text fields, sliders or menus, and it measures first-glance grounding, not recovery. The task set was written by us and is not a published benchmark. The hard tier has only 13 tasks, and only one of them is a multi-field autocomplete form. The observation that screenshot agents fail autocomplete forms therefore rests on one task and three failures.

Three claims survive these caveats. Click accuracy mis-ranked Holo against multi-step results, by a margin (5 tasks behind the best vision setup) larger than the whole spread among the top four. The same-model harness speed gap is real. The viewport trap exists and has the measured cost reported in Section 8. No ordering among the top four survives, and neither does any general statement about screenshot agents and forms beyond "on this one task, all three failed".

11. What We Are Changing

The decision drawn from this, not yet implemented at publication, has three parts. The default will be Qwen3.8-27B behind a lean screenshot loop, on the seat we already serve. The DOM agent stays for form-heavy work when the four-GPU mode is up. UI-Venus-2-9B, at 18 GB, is the small-footprint and Mac-capable option. We deleted the Holo3.1 and Fara1.5 weights after they lost. Every harness we run will set or measure the viewport explicitly. We will also stop treating single-click accuracy as a proxy for agent quality when choosing between models that all click well. The next experiment needs repeated runs and more form tasks, which is what separating the top four would take.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-028. The click-targeting corpus and the 39-task live bake-off were measured on the evening of 26 September 2026, roughly 20:30 to 21:40 UK time, with context from a desktop-driver evaluation on 17 September 2026. Raw artefacts, including per-task result files and the invalid viewport-bug run, are retained.