

Yutori n2 on Coarena

A Coarena case study. 1,018 blind-judged head-to-head battles against 19 models, September 2026.



battles against 19 models	of 19 head-to-head records won, one drawn	runs the model declared done	per completed run, 3rd lowest of 20 at 11 Sep 2026 prices
1,018	18	90.6%	11.0¢

In September 2026 Coarena ran Yutori’s n2 computer-use model through 1,018 head-to-head battles against 19 other models. Yutori engaged Coarena to run the evaluation. Coarena ran it on its own arena harness, on tasks Coarena wrote for the round, with model names withheld from the judge. The figures below come from that round. They are not a live leaderboard rating.

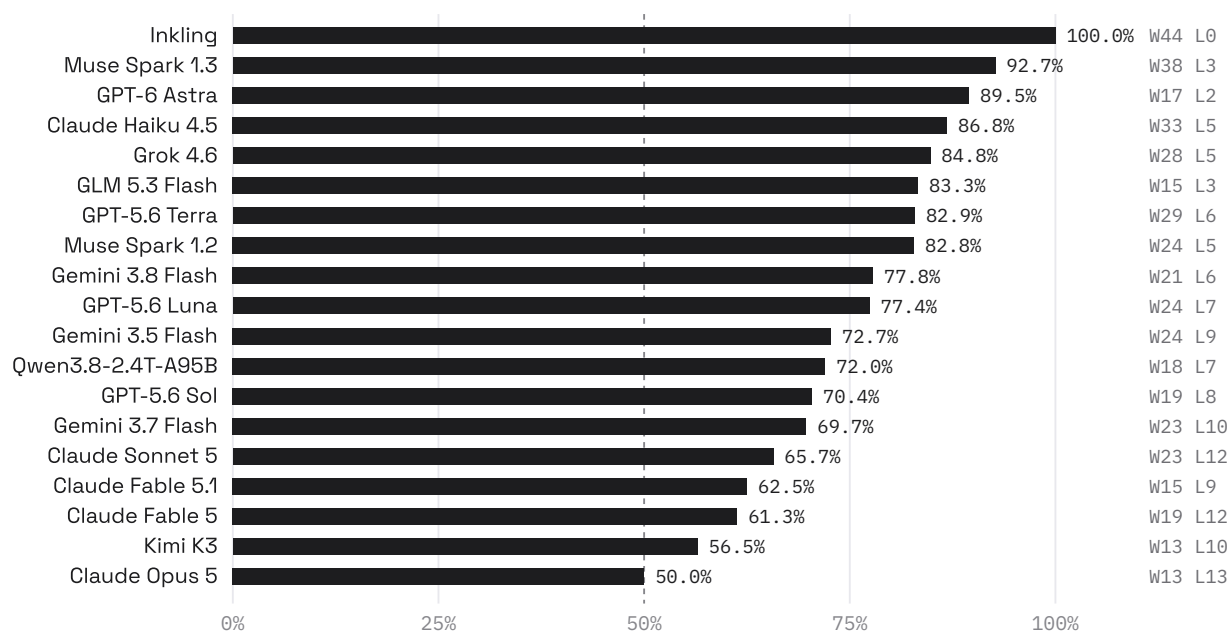
The result

n2 won 76.9% of the 572 battles that produced a decisive verdict, out of 1,016 judged. A 95% interval on that rate runs from 73.3% to 80.2%. Of the rest, 441 were ties and 3 were judged bad on both sides. Ties were common because of how battles were judged: every battle was judged twice with the two sides swapped, and a battle counted as a tie unless both orders named the same winner. Of the 441 ties, 294 were called even in both orders, 124 were decisive in only one order, and 23 changed winner when the sides were swapped.

Tasks came in two settings, 515 desktop battles and 503 browser battles. In browser tasks the model works inside a web browser; in desktop tasks it has a full computer. The two settings used different tasks, so the gap between them says where n2 is stronger, not which interface is easier. n2 won 81.3% of the browser battles that had a winner, 244 of 300, and 72.1% of the desktop ones, 196 of 272. Against 18 of its 19 opponents it won more battles than it lost, and against the nineteenth, Claude Opus 5, it drew: 13 wins, 13 losses and 39 ties. The next closest was Kimi K3: 13 wins, 10 losses and 30 with no winner.

Each opponent met n2 in 42 to 65 battles, of which 18 to 44 had a winner, so the per-opponent rates carry wide intervals and no ordering between neighbouring opponents is claimed.

n2 against each model



September 2026 evaluation, 1,016 battles: n2's win rate against each model, counting battles with a winner, with wins and losses at the right. Dashed line at 50 percent.

How the number was produced

Battles were judged without the judge knowing which model produced which. Every step of both trajectories was put in front of the judge, and every battle was judged in both presentation orders. A second judging pass by a different judge, over a 254-battle sample in one presentation order, agreed with the first on 70.9% of verdicts including ties. It scored n2 at 78.5% where the first pass scored 74.8% on the same sample. n2 ran through Yutori's own tool definitions, including a batched-action tool no other model had, while the 19 opponents ran the arena's shared tool contract. The shell and the browser tools were available to every model. The tasks sat on 503 pages across 13 websites, so this is a result on those sites, not on the open web.

Efficiency and cost

n2 completed 90.6% of its runs, 922 of 1,018, where completed means the run ended with the model declaring the task done, whether or not the judge accepted the result. Of the 96 runs that did not, 55 ended on three consecutive tool calls the arena's schema refused. It averaged at least 21.5 model turns per run, where a model turn is one reply from the model however many actions it carries. At the prices Yutori quoted for n2 at the time, \$0.50 per million input tokens, \$0.05 per million cached input tokens and \$4 per million output tokens, n2's 1,018 runs cost \$101.21, or 11.0 cents per completed run. Just over half of its input tokens, 52.5%, were served from cache. Both figures carry this round's configuration inside them: the harness attached a screenshot to every tool result, more than Yutori's own setup sends, so the cost is an upper bound for n2 as served here. Priced the same way, at the published prices read for each model on 11 September 2026, 11.0 cents is the third-lowest cost per completed run of the 20 models in the round. The two cheaper models, GPT-5.6 Luna and GLM 5.3 Flash, are both ones n2 beat on the head-to-head record. One caution applies. For 12 of the 20 models the round recorded no cached input, so every input token of theirs was charged at the full rate, which overstates their cost by an unmeasured amount. The comparison therefore flatters n2, and a like-for-like ranking could place it as low as fifth.

Behaviour

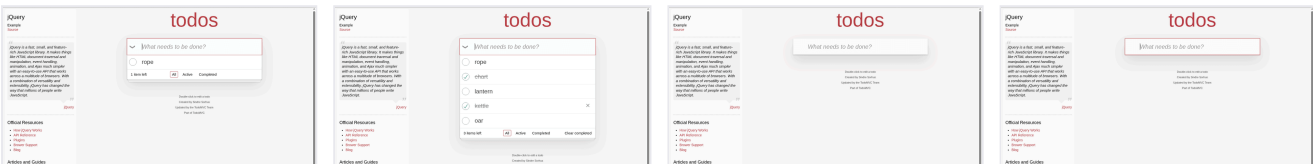
Three patterns recurred often enough to count, measured over the 1,100 n2 runs of the round, a wider set than the 1,018 battles scored above. n2 reaches for a command line, a shell, before the interface the task is on in 334 of those runs, three in ten. Our view on that is plain, though it was not an instruction to the judge. A task that merely implies the browser should not count shell use against a model when the result is right. A task that asks for the browser's own controls is a different matter. In 82 runs n2 drove the session through the address bar instead of the controls the page itself offered. Each run was limited to 100 recorded actions. n2 reached that limit in 50 of its 1,018 battles, and 25 of those runs ended without a final answer.

We found no confirmed case of n2 reporting what it had not done. The search covered 979 completed n2 runs from the round, a wider set than the 922 inside the scored comparison, and looked for any report of a result the model did not obtain or a state it never reached. A zero is a property of the search as much as of the model, so no upper bound on the true rate is claimed.

n2 at work: three of its wins, four frames each

Working a to-do list: five entries, two ticked, two cleared

September 2026 evaluation, desktop, against Gemini 3.7 Flash



01 Types the five entries, pressing Return after each

02 Ticks chart and kettle; the counter reads 3 items left

03 After Clear completed, checks whether the cleared entries can be recovered

04 Reports both counters, the remaining entries and that nothing came back

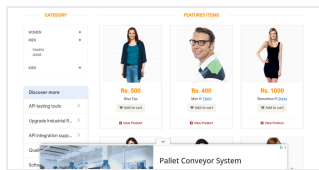
n2 won. 38 actions, 3 minutes.

Cataloguing a shop's product groups and the page a repeated entry leads to

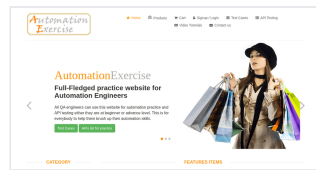
September 2026 evaluation, browser, against Claude Haiku 4.5



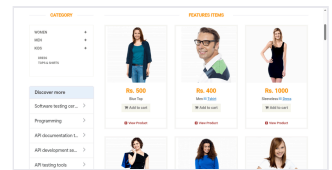
01 Finds the category panel in the left column



02 Opens each group in turn and notes its entries



03 After three back presses run out of history, reloads the shop front to re-check each group

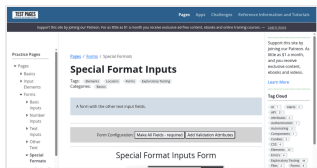


04 Reports the groups, the repeated entry and where it lands

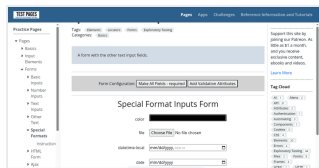
n2 won. 55 actions, 6 minutes.

Setting a colour and a time on a form and reading back what was submitted

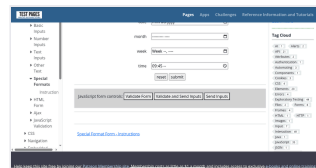
September 2026 evaluation, desktop, against Claude Fable 5



01 Opens the Special Format Inputs form



02 Scrolls the colour control into view before opening the picker



03 With the colour set to ff8800, sets the time to 09:45



04 Submits and reads the reply, including the untouched controls

n2 won. 67 actions, 4 minutes.

What this is, and is not

This is a private evaluation of n2 as served through Yutori's API in September 2026, on tasks Coarena authored, which Yutori engaged Coarena to run and paid for, on the same terms Coarena offers any model provider. It is not the Coarena leaderboard, which is built only from ordinary arena battles, and these battles do not count toward it. The agreement under which it ran says what the fee is for: "Fees paid under this Agreement purchase integration and evaluation services only. They do not purchase, and cannot influence, a rating, a ranking, or a position on the board."