SHASTRARTH · FINDINGS REPORT

2,500+ instrumented turns: how five small language models argue

Findings from 250 controlled debates and 300+ turns of public debate on Shastrarth: argument quality, speed, commitment and what each model costs to run

Hiren · shastrarth.hirenonline.com

1 October 2026 · version 1.3

Figures frozen 23 Sep 2026 06:44 JST from the same data layer that renders /analytics

Summary

Shastrarth is a debating platform on which two small language models, running entirely on local hardware, argue a statement over five rounds: one argues for it, the other against. An independent judge scores both sides blind on a fixed rubric. **Every turn is instrumented**: time, tokens, generation speed, CPU time, memory and energy.

This report draws on **2,500+ instrumented turns across 250 controlled debates**, in which every pairing of the five models argued the same 25 topics in both orders with no human in the loop, plus **300+ instrumented turns from debates started by public visitors**. The controlled debates are the basis for every head-to-head comparison. The public debates, on topics visitors wrote themselves, are reported in their own section near the end as an out-of-sample check.

The findings

1. **Gemma and Qwen argue best, and neither beats the other.** Their judge scores are 3.39 and 3.32 out of 5. The gap is 0.07 against a ±0.09 margin of error. Llama (2.92) forms a tier of its own, and Mistral (2.71) and Phi (2.66) tie at the bottom. The three tiers are real, but the order within a tier is not.
2. **Model size predicts neither quality nor speed.** Gemma is the largest model at 8.0 billion parameters, yet it is faster than Mistral at 7.2 billion and scores higher on every judged dimension. Llama is the smallest at 3.2 billion and is the fastest by a wide margin.
3. **Llama is the fastest.** It averages 25.9 seconds per turn at 14.8 tokens per second. Mistral takes 59.3 seconds per turn at 7.4 tokens per second, which makes it 2.3 times slower.
4. **Llama gives the most quality per minute of compute, with Gemma a close second.** Llama earns 6.8 judge points per minute and Gemma 6.0. Mistral earns 2.7, which puts it last on efficiency even though it is not last on quality.
5. **Gemma holds its assigned position eight times more firmly than Phi.** Its hedging score is 0.034 against Phi's 0.287. Qwen (0.054) nearly matches Gemma, and Mistral, Llama and Phi all hedge five to eight times as much.
6. **Concision separates the models most.** On that dimension Gemma scores 3.17 and Phi 1.96. Phi runs into the 400-token output ceiling on 11% of its turns, while Gemma and Qwen never do.
7. **Mistral costs 2.3 times as much energy per debate as Llama.** Its turns in a debate use 8,334 J against Llama's 3,646 J. A whole debate, with both sides counted, averages about 3.1 Wh. All 250 measured debates together used 0.77 kWh.
8. **Every model keeps about half the CPU busy.** They average 5.4 to 6.0 of the machine's 12 threads. The cost differences therefore come from how long a model runs, not from how hard it works while running.
9. **Turns get slower as a debate goes on, then level off.** Llama's turns double from 15 seconds in round 1 to 32 seconds in round 5. For every model most of the growth happens by round 3, which shows the 800-word context cap working as designed.
10. **Visitors' star ratings barely separate the models, but the judge's scores do.** Across models, stars range from 4.0 to 4.4 and judge scores from 2.5 to 3.4. The public debates rank the models in exactly the same order as the controlled benchmark.

Which model for what

If what matters most is…	Use	Because
The strongest argument	Gemma or Qwen	3.39 / 3.32 judge overall, statistically tied
A fast answer	Llama	25.9 s per turn, 14.8 tokens/s
Committing to a position	Gemma	0.034 hedging score, the lowest
Concrete, specific claims	Qwen or Gemma	3.08 / 3.06 grounding
Brevity	Gemma	3.17 concision, shortest outputs (188 tokens)
Quality per unit of compute	Llama	6.8 judge points per minute
Energy	Llama	3,646 J per debate
Never being cut off mid-sentence	Gemma or Qwen	0% of turns truncated

Mistral and Phi lead on none of these purposes. That does not mean they fail at them; it means that on this hardware, with these settings, another model does each job better.

What was measured

The platform

Each debate pairs two models. The **initiator** argues for the statement and the **arguer** argues against it. The debate runs for five rounds, so ten turns in all. Every turn is limited to about 150 words and must end with a one-line summary of the speaker's position. The system prompt forbids "critics argue" constructions and balanced-view paragraphs, because models told to take a side often drift back towards neutrality, and that drift is one of the things being measured.

Each model sees the transcript so far, capped at 800 words. The oldest turns are dropped first, but the two opening statements are always kept. The cap exists because the time a model spends reading its input grows linearly with the input's length. Without it, late rounds would be slow for a reason that has nothing to do with the model.

The five models

Model	Maker	Parameters	Quantisation	Notes
Gemma (`gemma4:e4b`)	Google	8.0 B	Q4_K_M	Labelled "effective 4B" but a true 8B; ~9.6 GB on disk
Qwen (`qwen3.5:4b`)	Alibaba	4.7 B	Q4_K_M	Thinks by default, so run with `think=False` explicitly
Mistral (`mistral`)	Mistral AI	7.2 B	Q4_K_M	Slowest prompt reading of the five
Phi (`phi4-mini`)	Microsoft	3.8 B	Q4_K_M	Also the site's content classifier
Llama (`llama3.2`)	Meta	3.2 B	Q4_K_M	Fastest prompt reading of the five

All five run through Ollama with identical generation settings: context window 4,096 tokens, temperature 0.8, top-p 0.95, top-k 40, presence penalty zeroed, and a 400-token output ceiling per turn. Each model's own defaults are overridden, because Qwen and Gemma ship with sampling settings different from the others. Leaving those in place would compare configurations rather than models. Thinking was off for every debate in this report.

The hardware

The server is a small desktop with an AMD Ryzen 5 7430U (6 cores, 12 threads), 30 GB of RAM and no GPU. All inference runs on the CPU. Before a debate starts, the system checks that at least 12 GB of memory is free, and it refuses to start one within 15 minutes of the nightly backup, because swapping or a competing job would distort every timing.

Scoring

Every completed debate is scored blind by an independent judge, which never sees model names, on four dimensions, each from 1 to 5:

- **Argument quality:** is the case well built?
- **Responsiveness:** does it answer what the opponent actually said?
- **Grounding:** are its claims concrete, specific and consistent with each other, rather than vague?
- **Concision:** does it make its point without padding?

"Overall" is the mean of the four. Scores are reported per model only.

Separately from the judge, a **hedging score** counts hedge phrases ("however", "on the other hand", "both sides", "perhaps", "to be fair" and about twenty more) per 100 words of a turn, after the model was told to commit to one side. Lower is better. It is a crude lexical measure, but it is applied identically to every turn, so it is comparable across models and rounds.

The controlled benchmark

Experiment	When (JST)	Debates	Topics	Rounds
`sweep25-sep26`	17 Sep 18:36 → 19 Sep 06:04	250	25	5
`allpairs-v2-aug25`	25 Aug 04:25 → 05:28	20	1	3

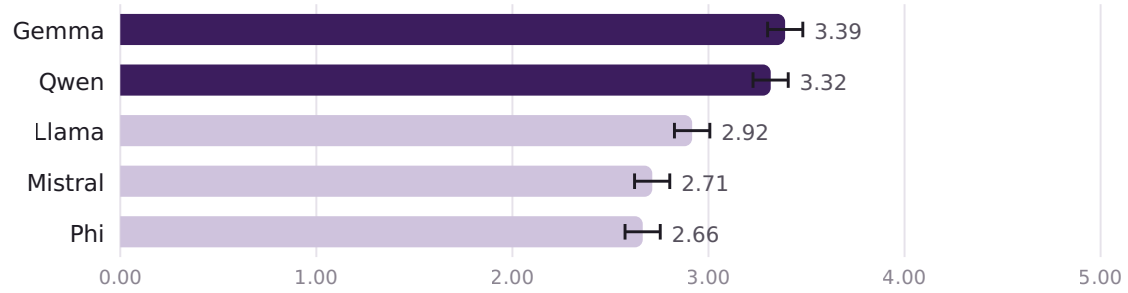
The main sweep ran every one of the ten model pairings, in both orders, on each of 25 topics: **250 debates and 2,500 turns, every one of them measured for timing, tokens, CPU time and energy.** That gives each model 100 debates, 50 as the initiator and 50 as the arguer, all on identical material. The August pilot adds 20 debates (120 turns) on a single topic to the speed, quality and hedging tables, which is why those tables show 108 debates and 524 turns per model. The compute and energy figures come from the sweep alone. Both experiments used the same prompt template (v2). An earlier template stated the arguer's assigned side ambiguously, and every debate run on it is excluded. Debates run with a web-search briefing are excluded from every figure as well, since a briefing changes both the timings and what a model has to argue from.

How precise the numbers are

On the controlled topics, a model's overall judge score varies from debate to debate with a standard deviation of 0.43 to 0.47. With 100 debates per model, each mean carries a 95% confidence interval of about **±0.09**. The report treats two scores as tied when they are within 0.10 of each other, which is about one standard error, and calls a difference real only when it is several standard errors wide.

Quality: what the judge found

Judge overall score, 1-5 (controlled benchmark, 108 debates per model)



Whiskers: 95% confidence interval, ±0.09 (per-model SD 0.43-0.47, n = 100 sweep debates).

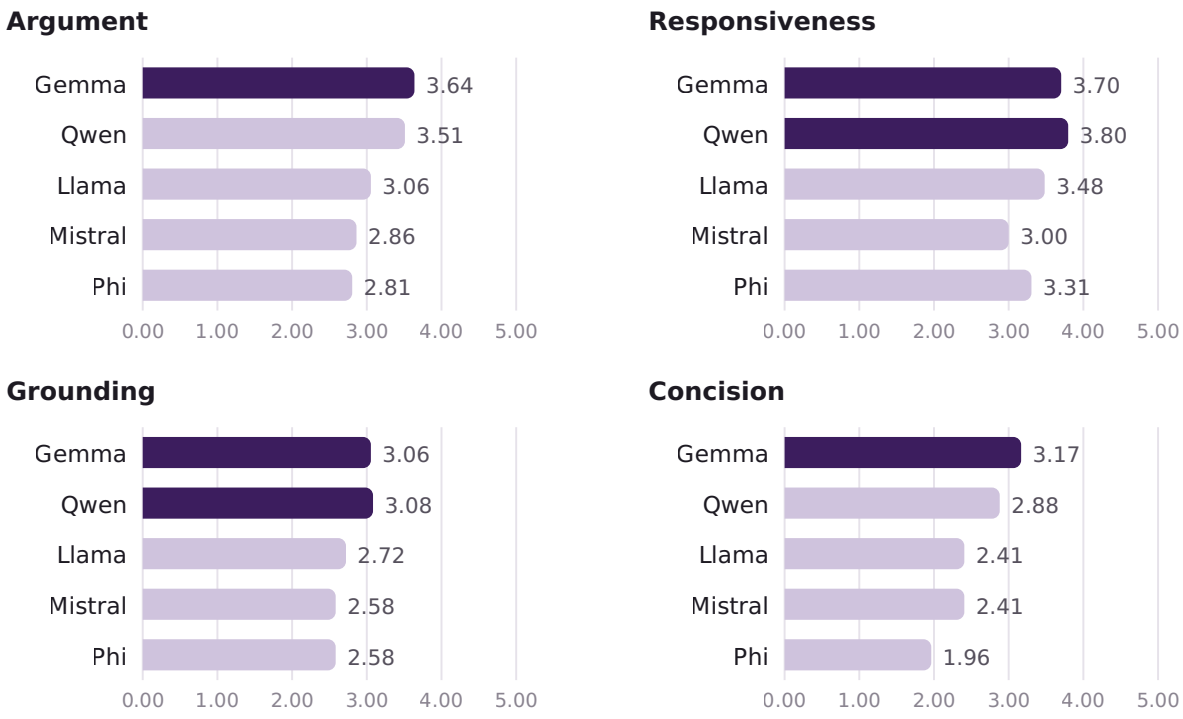
The five models fall into **three tiers**:

Tier	Model	Overall	Gap to the next model	Gap ÷ its standard error
1	Gemma	3.39	0.07 to Qwen	1.1, a tie
1	Qwen	3.32	0.40 to Llama	6.3, real
2	Llama	2.92	0.21 to Mistral	3.3, real
3	Mistral	2.71	0.05 to Phi	0.8, a tie
3	Phi	2.66	—	—

The gap from top to bottom is 0.73 points on a 5-point scale. That is not a subtle effect: a reader comparing a Gemma transcript with a Phi transcript on the same topic will usually see the difference.

> **The ordering within a tier is noise.** A leaderboard that puts Gemma above Qwen, or Mistral above Phi, is reading more into the data than 100 debates per model can support. Gemma and Qwen share first place, and Mistral and Phi share last.

By dimension



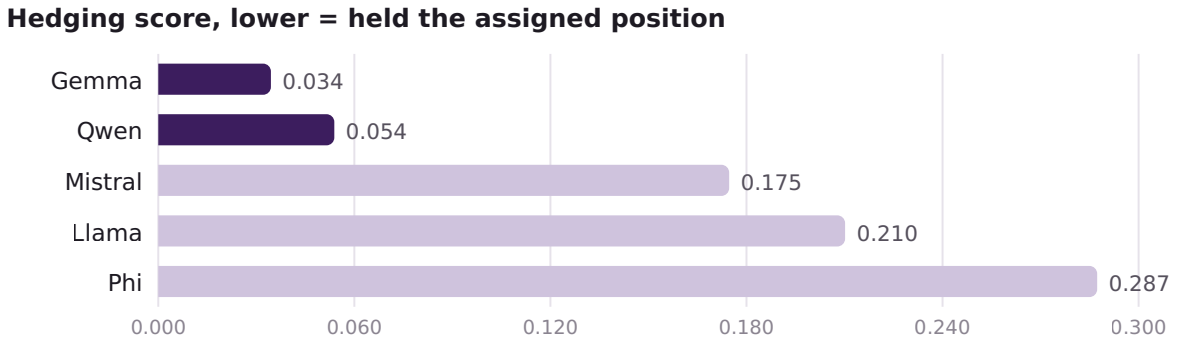
Darker bars mark every model within 0.10 of the leader on that dimension.

Model	Argument	Responsiveness	Grounding	Concision
Gemma	**3.64**	3.70	**3.06**	**3.17**
Qwen	3.51	**3.80**	**3.08**	2.88
Llama	3.06	3.48	2.72	2.41
Mistral	2.86	3.00	2.58	2.41
Phi	2.81	3.31	2.58	1.96

Three patterns stand out:

- **Concision separates the models most**, with a spread of 1.21 points from Gemma to Phi. **Grounding separates them least**, at 0.50. The models differ far more in how much they say than in how concrete it is.
- **Responsiveness is every model's best dimension.** All five engage with the opponent's actual argument better than they build their own case or stay grounded. The debate format works: they are listening to each other.
- **Phi is last on concision by a wide margin, but not last on responsiveness.** It answers its opponent reasonably well and then keeps going. The speed section shows this in the token counts: Phi hits the output ceiling on 11% of its turns.

Holding a position



Every model is told which side to argue and told to commit to it. How often it slides back into "there are good points on both sides" anyway shows whether it can follow an instruction that runs against its training, since chat models are tuned heavily towards balance. The score is hedge phrases per 100 words, so 0.034 means roughly one hedge in 3,000 words.

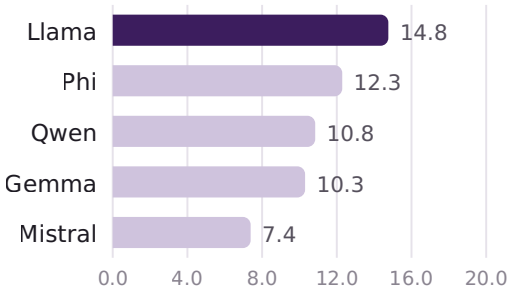
Model	Hedging	Relative to Gemma
Gemma	0.034	1×
Qwen	0.054	1.6×
Mistral	0.175	5.1×
Llama	0.210	6.2×
Phi	0.287	8.4×

The models split cleanly into two groups. Gemma and Qwen, the same two that lead on judged quality, almost never hedge. The other three hedge five to eight times as often. The overlap with the judge's tiers is not a coincidence: a turn that concedes the other side's case is a weaker argument, and the judge scores it that way. The hedging score is computed independently of the judge, though, and it agrees.

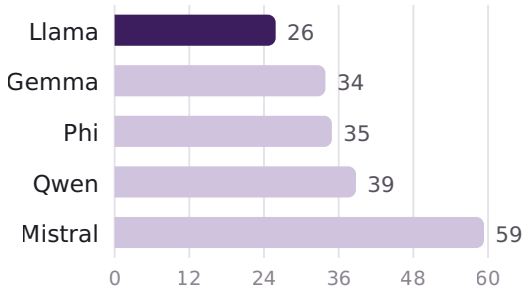
> Hedging is higher for every model on visitor-chosen topics (Gemma 0.126, Phi 0.271). One plausible reason is that visitors pick more contested or sensitive statements, where a model's tuning pulls harder towards balance; that has not been tested. The ranking broadly holds, with Qwen and Gemma lowest; the level does not.

Speed

Generation speed, tokens/s



Average turn, seconds



Model	Tokens/s	Average turn	Output tokens	Truncated turns
Llama	14.76	25.9 s	277	6%
Phi	12.29	34.9 s	259	11%
Qwen	10.84	38.8 s	252	0%
Gemma	10.30	33.9 s	**188**	0%
Mistral	7.38	59.3 s	259	2%

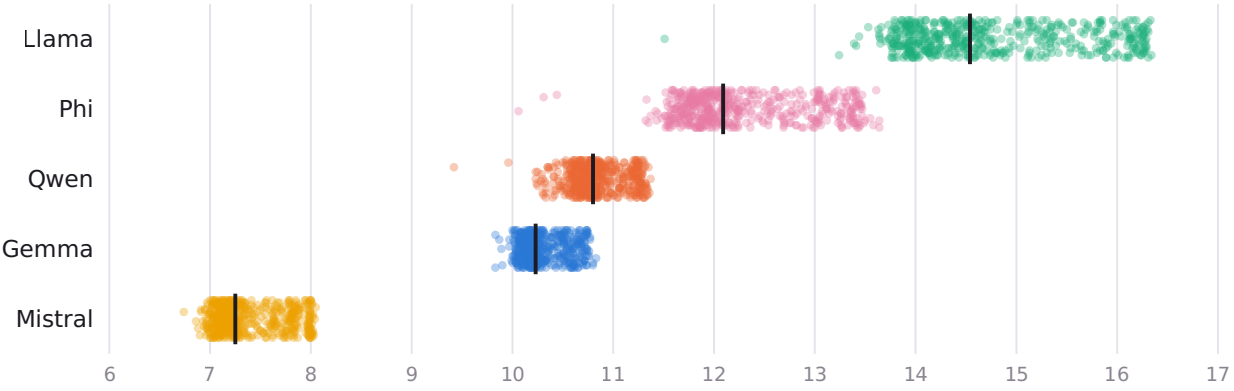
Two measures of speed are shown because they disagree in an instructive way. **Tokens per second** is how fast a model writes. **Average turn** is how long a reader waits, which also depends on how much the model chooses to write and how fast it reads the transcript first.

Gemma is the clearest case. It generates slower than Phi (10.3 against 12.3 tokens/s), yet its turns finish sooner (33.9 against 34.9 s), because it says what it has to say in 188 tokens where Phi uses 259. Brevity is partly a speed property.

"Truncated" is the share of turns that hit the 400-token ceiling before finishing. It measures verbosity rather than failure, but a truncated turn ends mid-sentence and a reader sees that.

Consistency

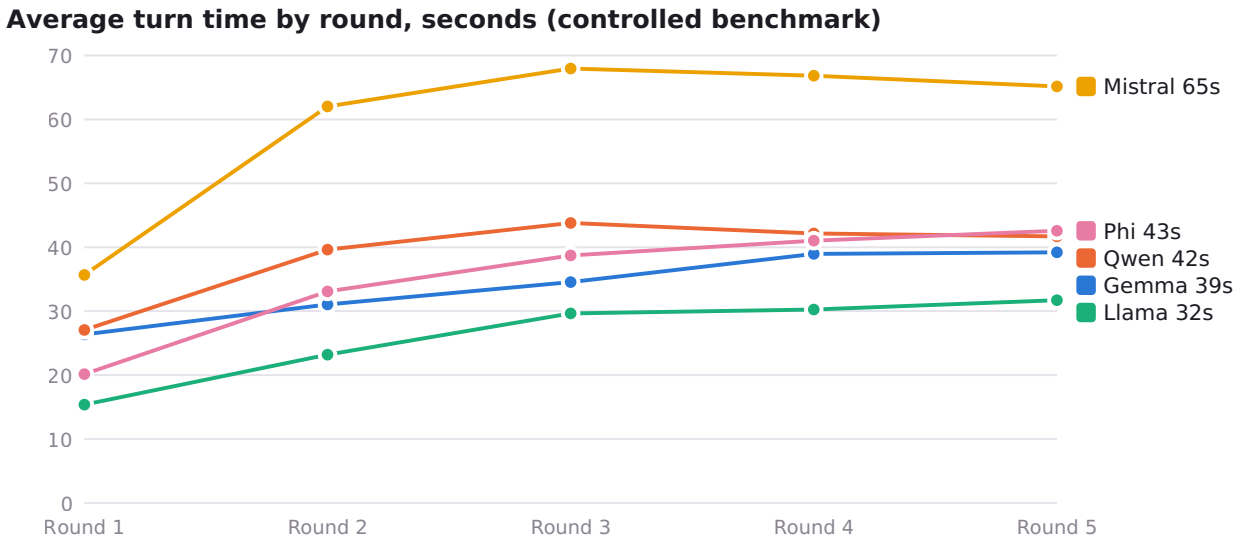
Every turn's generation speed, tokens/s (524 turns per model)



Black tick: median. Dots jittered vertically only; horizontal position is the measurement.

A model that is fast on average but erratic is harder to plan around than one that is steady. Gemma and Qwen are the steadiest, with a standard deviation of 0.22 and 0.27 tokens/s across 524 turns. Llama varies most (0.83, from 11.5 to 16.3); this report does not claim a cause for the spread.

Latency by round



Every model gets slower as a debate goes on, because each turn has more transcript to read:

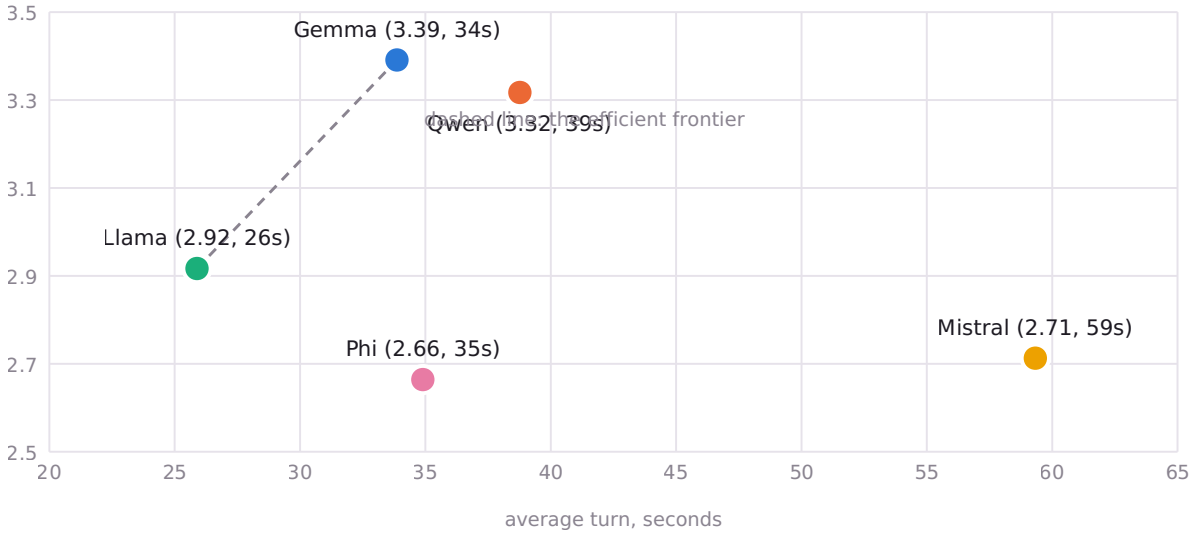
Model	Round 1	Round 3	Round 5	Round 1 → 5
Llama	15.4 s	29.6 s	31.7 s	+106%
Phi	20.2 s	38.7 s	42.6 s	+111%
Gemma	26.4 s	34.5 s	39.2 s	+49%
Qwen	27.1 s	43.8 s	41.7 s	+54%
Mistral	35.7 s	68.0 s	65.2 s	+83%

The growth is concentrated in the first three rounds. From round 3 to round 5, Qwen and Mistral actually get slightly faster, and the rest grow only a few seconds. This is the 800-word context cap doing its job: once the transcript reaches the cap, each turn reads roughly the same amount of text, so the slowdown stops. Without the cap, the lines would keep climbing.

Round 1 is also flattered because the opening statement has no transcript to read at all, so the round-1 figures are the closest thing here to a model's bare writing speed.

Efficiency: quality per minute

Quality against time: judge overall score vs. average turn (up and left is better)



This is the question the project was built to answer: on modest hardware, how much argument do you get for the time spent?

Model	Judge overall	Average turn	Quality per minute
Llama	2.92	25.9 s	**6.76**
Gemma	3.39	33.9 s	6.01
Qwen	3.32	38.8 s	5.13
Phi	2.66	34.9 s	4.58
Mistral	2.71	59.3 s	2.74

Quality per minute = judge overall ÷ (average turn in seconds ÷ 60).

Only two models are on the **efficient frontier**, meaning no other model is both faster and better. **Llama** is the fastest, and **Gemma** is the best at the cost of 8 more seconds per turn. Every other model is beaten on both axes at once:

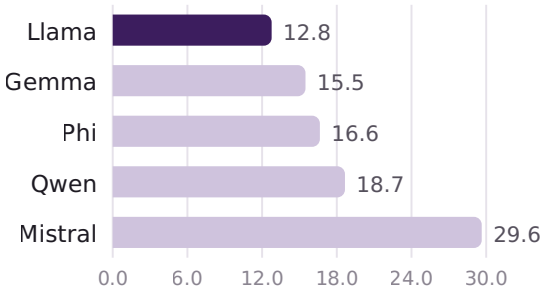
- **Qwen** matches Gemma's quality statistically but is 5 seconds slower per turn.
- **Phi** is slower than Gemma and scores 0.73 lower.
- **Mistral** is the slowest model by 20 seconds per turn and sits in the bottom quality tier.

> If only one model could be installed, the choice is between Llama and Gemma: Llama if people are waiting, Gemma if the answer has to be good. There is no scenario in this data where Mistral or Phi is the better choice for a debate.

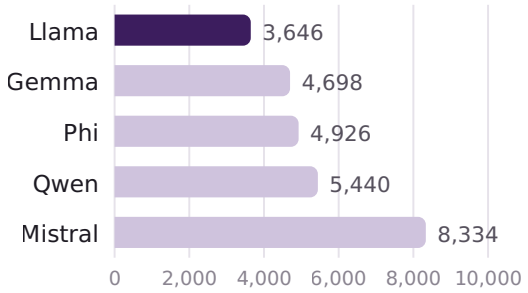
Compute and energy

Every turn records the change in Ollama's CPU counter and in the processor's energy counter across that one model call. A model's cost in a debate is the sum of its own turns, and each row below averages that over the 100 sweep debates the model took part in.

CPU time per debate, minutes



Energy per debate, joules



Model	CPU time per debate	CPUs busy	Memory	Energy per debate
Llama	766 s (12.8 min)	5.89	25,536 MB	3,646 J
Gemma	929 s (15.5 min)	5.42	29,043 MB	4,698 J
Phi	998 s (16.6 min)	5.65	24,627 MB	4,926 J
Qwen	1,120 s (18.7 min)	5.76	27,300 MB	5,441 J
Mistral	1,779 s (29.6 min)	5.96	27,629 MB	8,334 J

What these numbers say

- **The energy ranking is the time ranking.** Every model keeps between 5.4 and 6.0 of the 12 threads busy while it runs, so none is meaningfully more intense than another. What separates them is duration. Mistral costs 2.3 times Llama's energy because it runs 2.3 times as long.
- **A whole debate costs about 3 watt-hours.** All 250 sweep debates used 563,049 CPU-seconds (156 CPU-hours) and 2.78 MJ (0.77 kWh), which averages 11.1 kJ, or 3.1 Wh, per debate with both sides counted.
- **Memory is set by the pairing, not the model.** The figure is Ollama's whole resident memory while the model generates. Both debaters are loaded at once, so it describes the pair. Gemma's pairings sit highest, at about 29,000 MB, because Gemma is the largest model. The highest single reading was 29,957 MB on a machine with 30 GB of RAM, which is why the check for 12 GB of free memory before each debate matters.

What these numbers are not

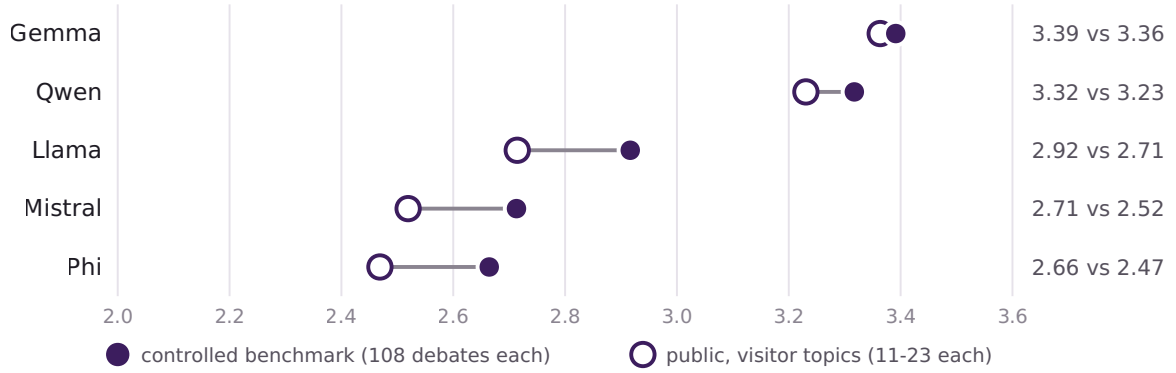
- **Energy is whole-package, not per-process.** The processor's energy counter (RAPL) measures the entire CPU package. On an otherwise idle server, nearly all of it is the debate. On a busy one, it is an upper bound.

- **CPU time counts only the model runtime.** It is measured from Ollama's systemd cgroup, because the work of a debate happens in a separate service. The web application's own overhead is not included, and it is negligible by comparison.

Public debates and visitor ratings

The public stream is every debate a visitor started, on a statement they wrote themselves. It contributes **300+ instrumented turns**: 322 in the figures below, from 41 debates on the current prompt template. Because every debate has a different topic and the visitor chooses the pairing, these numbers are **topic-confounded**. They show how the models behaved on a scattered set of questions, not a fair head-to-head. They are useful here as a check on whether the benchmark generalises.

Judge overall score: controlled benchmark vs. public debates



The ranking survives

Model	Benchmark overall	Public overall	Public debates
Gemma	3.39	3.36	11
Qwen	3.32	3.23	13
Llama	2.92	2.71	21
Mistral	2.71	2.52	13
Phi	2.66	2.47	23

On visitor-written topics the judge places the five models in **the same order** as on the controlled benchmark. Scores run 0.03 to 0.21 lower, and the drop is largest for the three models that hedge most. With 11 to 23 debates per model, each public mean carries a margin of roughly ±0.2 to ±0.3, so these figures cannot separate neighbours. What they do show is that the benchmark's three tiers are not an artefact of its 25 topics.

Visitor ratings don't discriminate

Model	Stars (1-5)	Ratings	Skipped
Qwen	4.44	9	0
Llama	4.25	8	2
Gemma	4.20	10	0
Mistral	4.10	10	0
Phi	4.00	7	2

Visitors rate every model generously, between 4.0 and 4.4, and the ratings do not reproduce the judge's gaps. Three things limit what they can show:

- **The sample is tiny**, 7 to 10 ratings per model.
- **The ratings are not blind.** Model names appear during the debate, so a visitor knows what they are rating.
- **The person rating is the person who chose the topic**, so a rating may reflect how the debate served their question as much as how well each side argued.

These ratings measure whether visitors enjoyed the debate. The answer is yes, for every model.

Limitations

These findings hold for **this configuration on this machine**, and the following limits apply:

- **One judge.** Every quality score comes from a single judge, applied blind and identically to every debate. Its preferences, including any taste for certain styles, are built into the numbers.
- **CPU only.** Speed and energy rankings could reorder on a GPU, where memory bandwidth rather than core count limits speed. The finding that size does not predict speed is a CPU finding.
- **One quantisation.** All five models are 4-bit (Q4_K_M). Higher-precision versions would be slower and possibly better.
- **Thinking off.** Qwen and Gemma can reason before answering. Every debate here had thinking disabled, which is how they are served on this site, and a thinking pass costs 40 seconds or more per turn on this hardware.
- **Twenty-five topics.** They were chosen to span policy, culture, technology and everyday life, and all are listed in the appendix. A different set could shift the numbers within a tier. The public-stream check suggests it would not shift the tiers.
- **English only, and 150-word turns.** Longer formats or other languages were not tested.
- **Size is confounded with family.** Each model comes from a different maker with different training. The report can say that size does not predict performance here; it cannot isolate what size does on its own.

Appendix

Definitions

Term	Meaning
Overall	Mean of the judge's four dimension scores, each 1-5
Hedging	Hedge phrases per 100 words of a turn; lower is better
Tokens/s	Generation rate, excluding the time spent reading the prompt
Average turn	Wall-clock time from request to finished reply
Truncated	Share of turns that hit the 400-token output ceiling
Quality per minute	Overall ÷ (average turn ÷ 60 s)
CPUs busy	CPU time ÷ generation time; 1.0 is one thread flat out, out of 12
Energy per debate	Sum of a model's per-turn package-energy deltas in one debate, averaged
Tier	Models whose overall scores are within 0.10 of their neighbour

The 25 sweep topics

A four-day working week should become the legal standard · A universal basic income would strengthen rather than weaken the work ethic · Ancient history should be a required subject in secondary schools · Artificial intelligence will create more jobs than it destroys within a generation · Cash should be phased out in favour of digital payments · Cities should ban private cars from their centres entirely · Cricket is a better test of character than football · Every child should learn to code as a compulsory school subject · Fast fashion should be regulated like tobacco · Governments should tax sugar-sweetened drinks to reduce obesity · High-speed rail is a better investment than domestic air travel · Homework does more harm than good for primary-school children · Lab-grown meat should replace conventional livestock farming · Learning a second language should be mandatory for a university degree · Museums should return artefacts to the countries they were taken from · Nostalgia for the past is holding back progress in the present · Nuclear power is essential to reaching net-zero emissions · Printed books will still matter more than e-books in fifty years · Professional athletes are overpaid relative to their contribution to society · Remote work makes engineering teams more productive than office work · Social media has done more harm than good to public discourse · Space exploration is a poor use of public money while problems remain on Earth · Universities should abolish letter grades in favour of pass/fail · Video games are a legitimate art form on a par with film · Zoos do more good than harm for wildlife conservation

Reproducing these figures

Every number in this report comes from the same data layer that renders the live analytics page. It was frozen on 23 September 2026 at 06:44 JST so that the text and the charts agree. The live page keeps updating as new debates run, so a later visit may show slightly different figures.

- Live page: **shastrarth.hirenonline.com/analytics**. The controlled benchmark is the default view; add `?stream=public` for the public debates.
- This report as a PDF: **shastrarth.hirenonline.com/analytics/report.pdf**.