Response Latency by Miner Pool

Internal study — measured chat latency of the community pool and the BYOK pool on identical workloads, for both the immediate answer and the full orchestration pipeline.

Generated: 2026-08-08 18:50 UTC

Response Latency by Miner Pool

Internal study — August 8, 2026

Summary

We measured chat response latency on Elis AI across its two personal-scope miner pools — the **community pool** (platform-operated and community-contributed miners) and the **BYOK pool** (miners backed exclusively by the customer's own provider API keys) — on identical workloads. Two latency metrics are reported for every request class, because the platform delivers answers in two stages:

- **Immediate answer** — the platform's answer-now response: a direct answer for simple questions, or a best-effort first answer for complex ones while deeper work continues.
- **Full completion** — for requests that engage the orchestration pipeline (multi-part questions requiring research), the time until all background research, retrieval, and follow-up messages have finished.

Metric (median)	Community pool	BYOK pool
Standard request — first token	3.4s	3.3s
Standard request — completed answer	5.2s	4.1s
Orchestrated request — immediate answer	4.0s	7.2s
Orchestrated request — full pipeline completion	124.8s	108.7s

Both pools deliver equivalent interactive latency. Pool scope resolution adds sub-second overhead; total latency in either pool is dominated by the serving models and, for orchestrated requests, by the depth of research the pipeline chooses to perform — not by pool machinery. Restricting an account to BYOK-only serving carries no measurable latency penalty.

Measurement definition

All timings are wall-clock, measured client-side from message submission (T+0) over the platform's live streaming interface.

- **First token** — first answer content received on the stream.

- **Completed answer / immediate answer** — the terminal completion event of the answer message delivered on the live stream.
- **Full pipeline completion** — the conversation's streaming state polled until every background orchestration run finished; the elapsed time when the conversation went quiet. Orchestrated conversations finished with 7–24 messages (progress updates, intermediate findings, final deliverables).

Every timed run's answer body was verified to be a genuine model response. Runs that returned a platform notice instead of an answer, or that were interrupted by infrastructure restarts, were excluded and re-run — never averaged in. Model and miner binding was fully dynamic in every run; no model was pinned anywhere in the harness. Web retrieval was verified healthy for all orchestrated runs reported here.

Workload

Class	Representative prompt
Trivial	"Hello, how are you?"
Easy	"What is the capital of France?"
Medium	"Compare pros and cons of solar vs wind energy."
Hard	"Analyze Apple revenue by segment for the last 3 years and summarize the trend."
Orchestrated	Compound 2–3 part requests (technology comparison + risk assessment; state-of-field summary + economic comparison + investment memo; structured multi-angle research analysis)

Standard requests (trivial–hard) are answered directly. Orchestrated requests engage the task pipeline: the platform answers immediately with its best current knowledge, then runs background research with live web retrieval and delivers findings as follow-up messages.

Community pool

Standard requests, seconds (first token / completed answer):

Speed mode	Trivial	Easy	Medium	Hard
instant	1.6 / 2.3	1.4 / 2.4	4.4 / 5.7	4.8 / 7.3
fast	1.6 / 2.6	1.3 / 3.2	5.2 / 6.1	4.9 / 4.9
medium	6.0 / 7.3	1.7 / 3.7	6.9 / 7.8	4.4 / 4.4
deep	1.3 / 1.3	4.1 / 4.1	5.4 / 5.4	21.7 / 21.8
auto	2.2 / 8.7	2.3 / 7.3	2.7 / 6.2	2.8 / 2.8

Latency scales with question difficulty rather than being flat-taxed: simple questions complete in 1–4 seconds in every mode, and the platform invests 20+ seconds only where a hard question meets deep mode.

Orchestrated requests, seconds (immediate answer / full pipeline completion):

Request	auto	deep
Comparison + risk assessment (2 parts)	9.0 / 123.1	2.5 / 73.2
Summary + economics + memo (3 parts)	3.3 / 126.4	2.8 / 84.6
Multi-angle research analysis	26.5 / 199.3	4.6 / 175.4
Median	**6.1 / 162.9**	**2.8 / 84.6**

The user is never waiting blind: an answer arrives in seconds (median 4.0s across both modes), and the full researched deliverable — built from live web retrieval across 7–11 pipeline messages — lands in one to three minutes.

BYOK pool

The same account was switched to BYOK-only serving (customer keys for two providers; 22 BYOK miners). **Standard requests**, seconds (first token / completed answer):

Speed mode	Trivial	Easy	Medium	Hard
fast	1.2 / 2.1	1.1 / 2.0	4.8 / 5.7	4.6 / 4.7
auto	1.2 / 2.0	2.1 / 3.6	12.6 / 13.4	4.5 / 4.5

Orchestrated requests (auto mode), seconds (immediate answer / full pipeline completion):

Request	auto
Comparison + risk assessment (2 parts)	7.2 / 86.1
Summary + economics + memo (3 parts)	4.2 / 108.7
Multi-angle research analysis	10.3 / 338.9
Median	**7.2 / 108.7**

The widest run (338.9s, 24 messages) reflects the pipeline choosing to pursue substantially deeper retrieval on the customer's keys — depth of research, not pool overhead. BYOK latency otherwise tracks the community pool closely: median full completion of 108.7s vs 124.8s.

BYOK serving is strictly key-scoped at all times. Provider-side throttling on a customer key affects only that key's miners: the platform benches the affected key group with escalating backoff and serves from the customer's remaining BYOK keys, without ever widening the request outside the customer's own pool.

Method notes and limits

- Single-tenant benchmark environment; one run per cell in the per-mode tables. Aggregates: 20 community and 8 BYOK standard runs; 6 community and 3 BYOK orchestrated runs (dual-metric).
- BYOK requests were paced to remain within the customer keys' provider rate limits; community-pool requests were not paced.
- The dynamic scorer may bind different models to the same question across runs, particularly in auto mode; orchestrated full-completion times vary with the depth of research the pipeline elects to perform (73–339s observed).
- Full-completion polling has ±2s resolution.

Conclusion

Pool scope is a governance choice, not a performance trade-off. On identical workloads the community pool and a well-provisioned BYOK pool deliver equivalent latency at both stages of the platform's answer model: an interactive immediate answer in single-digit seconds (medians 3.3–7.2s across

request classes), and fully researched orchestrated deliverables in one to three minutes (medians 108.7s BYOK vs 124.8s community) — while BYOK-only accounts retain strict key-scoped serving at all times.
