Frontier-Beating Performance at the Orchestration Layer

Sai Sidhanth Manoharan Jayanthi Yash Tripathi

Senal AI

Abstract

Senal is a proprietary model-orchestration system delivered as a single OpenAI-compatible endpoint. Across ten industry-standard benchmarks it scores above the strongest generally available frontier models on eight, and above GPT 5.6 Sol (Max) on all ten, reaching 63 on the AA Intelligence Index against a prior state of the art of 60. Input costs 60% less than Fable 5 and output costs 52% less. The system retains no data at any stage and requires no code changes beyond a base URL. This report sets out the approach, the measured performance across agentic execution, scientific reasoning, long-context reasoning and factual reliability, and the conditions under which those results were produced.

Technical Report · 2026

1 Approach

The scaling hypothesis holds that capability emerges from the breadth and volume of training data. Labs across the closed-source and open-source spectrum, from Anthropic, OpenAI and Google to DeepSeek, Moonshot and others, have each demonstrated this within a single training run. Senal extends the thesis across those independently trained models. The aggregate training data represented across foundation models with distinct architectures, objectives, data sources and training regimes is orders of magnitude larger and more diverse than any single run. Exploiting that diversity at inference time is the core of what we do.

No single model is trained on all of human knowledge. Every foundation model reflects the priorities, sources and coverage decisions of the lab that built it, meaning every model carries gaps that are structural rather than incidental. The ceiling on what a single call can return is set by what that one distribution contains. Critically, these gaps are largely non-overlapping: models trained on different corpora with different objectives fail on different inputs and succeed on different reasoning paths. A question outside one model's competence frequently falls within another's.

Senal draws on that diversity at inference time, surfacing a broader hypothesis space than any single distribution supports and applying proprietary verification to identify the best-supported answer from within it. The result is intelligence not bounded by the ceiling of any individual model, and factual reliability substantially better than any single-model deployment. Drawing answers from across distinct training distributions surfaces agreement and conflict that carry reliability information no single generation produces. The 41-point margin over GPT 5.6 Sol and 7 over Fable 5 on non-hallucination is this property measured directly.

The intelligence gains follow from the approach. The cost and latency profile is a separate and harder result. Exploiting distributional diversity naively carries steep computational cost: querying multiple foundation models multiplies inference spend and compounds latency in ways that make the system commercially unviable without significant engineering. Reducing that to negligible added latency relative to a single Fable 5 call, at $4 per million input tokens and $24 per million output tokens against Fable 5 at $10 and $50, required deep optimisation work across the inference stack specific to this system. That work is proprietary. The outcome is not replicable by applying the same approach without it, and the technical depth it required is the primary reason this combination has not existed before.

Senal differs from the main alternatives by construction. Self-refinement iterates on one model's output and inherits its systematic biases on every pass. Traditional orchestration selects one model per request and is bounded by whichever it picks. Test-time compute applied to a single model raises depth within one distribution. None of these approaches can access the gains that come from operating across independently trained distributions.

Four properties hold in production. Every request receives identical computation, with no difficulty classification, no effort adaptation and no fallback tier. Every candidate is verified before dispatch. Performance scales with the breadth of available training distributions rather than any single model. No prompt, completion or intermediate state is retained at any stage.

The interface is OpenAI-compatible chat completions, with streaming, function and tool calling, structured outputs, system prompts and multi-turn state.

2 Results

Benchmark	Senal	GPT 5.6 Sol (Max)	Fable 5
GDPval-AA v2	**66.0**	62.0	63.0
τ^3-Banking	**36.9**	33.0	26.8
Terminal-Bench v2.1	**91.1**	89.5	84.6
SciCode	58.3	56.1	**60.2**
HLE	50.2	47.2	**53.3**
GPQA-Diamond	**96.5**	94.1	92.6
CritPt	**33.8**	32.3	28.6
AA Accuracy	**62.0**	59.0	61.0
AA Non-Hallucination	**52.0**	11.0	45.0
AA LCR	**78.0**	73.7	70.0
AA Intelligence Index	**63**	59	60

All figures were produced against the production API, called exactly as a customer calls it, with no benchmark detection, no per-benchmark configuration and no tuning. Each benchmark ran through its canonical public evaluation suite at a pinned commit, scored by that suite's own implementation, with baselines at maximum reasoning configuration on identical parameters and task instances. All results reflect production API performance.

2.1 What the results mean in deployment

Fewer answers that need checking. Non-hallucination measures whether a model declines instead of fabricating when it lacks grounds to answer. It is a separate property from accuracy, and it governs how much of a workflow can run without a person in it. Senal leads Fable 5 by 7 points and GPT 5.6 Sol by 41. Where every output is reviewed because any output might be invented, that margin is the difference between review as a safety net and review as a bottleneck.

Agents that finish the job. τ^3-Banking scores the end state of a database against an annotated goal under written policy. It measures work completed, not dialogue that sounds correct. Senal completes 10.1 points more of it than Fable 5, against a 6.2-point spread between the two baselines. Terminal-Bench verifies system state after multi-step shell tasks. Together these predict whether an agent deployment ships or stalls in pilot. Absolute scores remain low across the field; multi-turn tool use is unsolved, and we have not solved it.

Capability that holds at production context. Real prompts are far longer than test cases, and models that look equivalent on short inputs diverge as context grows. Senal leads Fable 5 by 8 points on AA LCR, which is the margin that matters once a system is carrying real documents instead of benchmark items.

Performance that transfers. GDPval draws from real occupational work products and is built to resist narrow optimisation. That makes it the closest available proxy for a customer's own evaluation set. Leading both baselines on it is the strongest evidence that these results will reproduce on your workload and not only on ours.

3 Cost, latency and data

	Senal	**Fable 5**	**Traditional orchestrators**
Input, per 1M tokens	**$4**	$10	Varies
Output, per 1M tokens	**$24**	$50	Varies
Added latency	Negligible	Baseline	High
Data retention	**None, at any stage**	30 days; no ZDR terms	Varies
Intelligence tier	**Frontier+**	Frontier	Frontier on some tasks

Senal is 60% below Fable 5 on input and 52% below on output, or roughly 55–60% lower blended across typical workload mixes.

Rate card understates the gap. In production, spend is dominated by retries, human review and escalation. A system that fabricates less and completes more of each task consumes fewer of all three, so the non-hallucination and τ^3-Banking results are cost results as well as quality results.

In regulated sectors, procurement requires zero data retention. Fable 5 does not offer it at any price.