

Specialized Decisions, Frontier Reasoning

A 2,000-case evaluation of a purpose-built decision model

Wavera Health Engineering

Technical report • Public field note • 26 September 2026

Abstract

AI-native clinical software has at least two kinds of model work. Some tasks require broad synthesis and explanation; others require a bounded decision about where a request should go. We evaluated TypeSafe AI’s Jev decision model (`jev-1.13.0`) against the general-purpose generative model serving as our routing baseline. The study used 2,000 synthetic routing cases across ten broad task families, with 1,000 cases held out for evaluation. No patient data was used and neither model could execute an EHR action.

On the held-out set, Jev reached 96.1% exact-route accuracy versus 84.6% for the baseline, reducing incorrect routes from 154 to 39. Median latency was 168.1 ms versus 592.6 ms, and measured cost was approximately \$0.055 versus \$0.328 per 1,000 decisions. At a 0.90 confidence threshold, Jev accepted 80% of requests and was exact on all accepted cases in this sample. A simple Jev-first cascade reached 97.5% exact-route accuracy. These results support a compositional architecture: use specialized models for fast typed decisions, reserve frontier models for broad reasoning, and expose uncertainty to the surrounding software.

1 Motivation

The most capable AI system is not necessarily the one that uses the largest model at every step. In healthcare, speed, predictability, and restraint matter alongside intelligence. A general-purpose frontier model is valuable when the work requires longitudinal context, evidence synthesis, or a carefully explained recommendation. Many interactions, however, begin with a narrower question: what kind of request is this, where should it go, and is confidence high enough to proceed?

Routing is easy to hide inside an end-to-end demonstration, but it shapes the behavior of the complete product. A wrong route can invoke the wrong tool, add latency, or require recovery downstream. We therefore isolated this decision layer and measured it directly.

2 Jev and the system context

TypeSafe AI introduced Jev in September 2026 as its first public “System One” model [1]. Rather than generating open-ended prose, Jev receives unstructured state and returns a typed probabilistic decision with a confidence value. The provider’s current model documentation identifies `jev-1.13.0` as the stable release used in this study [2].

The proposition is unusually concrete: give up general generation in exchange for a bounded output that ordinary software can inspect and act upon. This makes Jev relevant to orchestration, classification, dispatch, and other small decisions that sit between a user request and deeper model work. Provider claims still require workload-specific evaluation, so we compared the pinned model against our own general-purpose routing baseline rather than extrapolating from launch materials.

3 Benchmark design

We wrote 100 semantic request patterns spanning ten broad kinds of work that an AI-native clinical system may need to distinguish. Each pattern was expressed in 20 surface forms, yielding 2,000 synthetic cases. Half were reserved for held-out evaluation. The dataset was balanced by design and measured only whether the selected destination matched the full expected route.

The 1,000 held-out cases represent 50 unseen semantic seeds expressed in 20 variants each, not 1,000 fully independent intents. We also resampled by the underlying request pattern; the direction of the measured advantage remained positive under that more conservative analysis.

Table 1: Benchmark composition and execution.

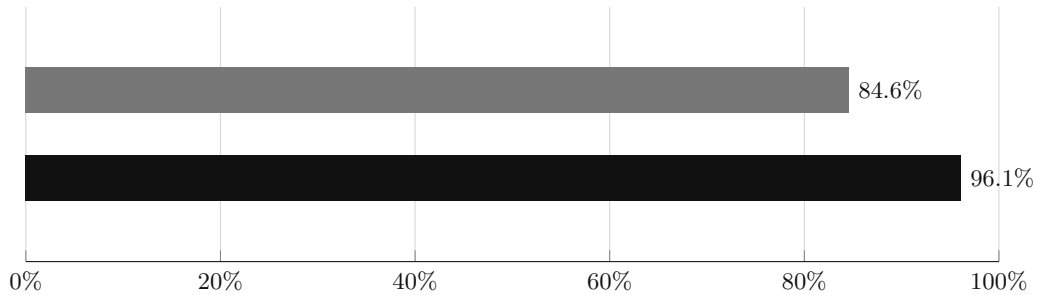
Measure	Value
Synthetic cases	2,000
Held-out cases	1,000
Task families	10
Semantic request patterns	100
Successful provider calls	4,000 / 4,000
Patient data used	None
Tested TypeSafe model	<code>jev-1.13.0</code>

4 Exact-route results

Jev reached 96.1% exact-route accuracy on the held-out set. The general-purpose baseline reached 84.6%. In raw terms, the baseline made 154 incorrect routing decisions and Jev made 39, a difference of 115 errors. This corresponds to a 74.7% relative reduction in routing errors in this balanced sample.

Table 2: Held-out benchmark results ($N = 1,000$).

Measure	General-purpose	Jev	Difference
Exact-route accuracy	84.6%	96.1%	+11.5 pp
Incorrect routes	154	39	115 fewer
Median latency	592.6 ms	168.1 ms	3.5 $\times$ faster
95th-percentile latency	780.4 ms	212.7 ms	3.7 $\times$ faster
Cost / 1,000 decisions	\$0.328	\$0.055	83.1% lower

**Figure 1:** Held-out exact-route accuracy. “Exact” means that the selected destination matched the complete expected route.

5 Operational profile

A routing decision sits at the beginning of an interaction. Every added fraction of a second contributes to the impression that an assistant is waiting before it can begin useful work. In this run, Jev’s median latency was 168.1 ms, compared with 592.6 ms for the general-purpose baseline. At the 95th percentile, the corresponding values were 212.7 ms and 780.4 ms.

Measured token cost was approximately \$0.055 per 1,000 Jev decisions and \$0.328 per 1,000 baseline decisions. These are workload-specific observations based on provider-reported token counts and public list prices at the time of the run, not universal price or performance claims.

The conclusion is not that a compact model can replace a frontier model. It cannot. The result instead suggests that a narrow decision need not consume broad reasoning capacity when a specialized component can make that decision faster, more predictably, and with an explicit representation of uncertainty.

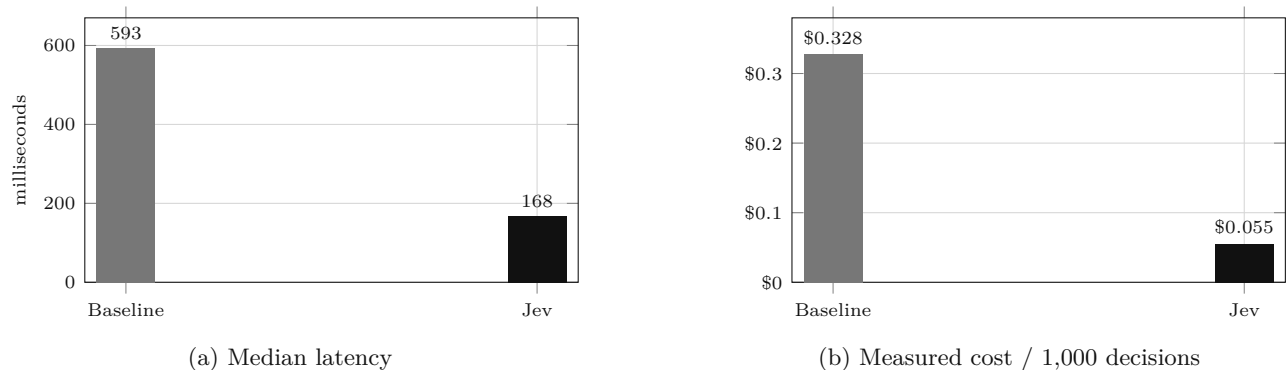


Figure 2: Observed operational results from the benchmark run.

6 Confidence gating and cascade behavior

A healthcare AI system should not turn uncertainty into confidence theater. Jev returns a confidence value with each choice. At a conservative 0.90 threshold, the model accepted 80% of held-out requests and was exact on all 800 accepted cases in this study. The remaining 20% can move to a more capable model, ask the user for clarification, or stop for review.

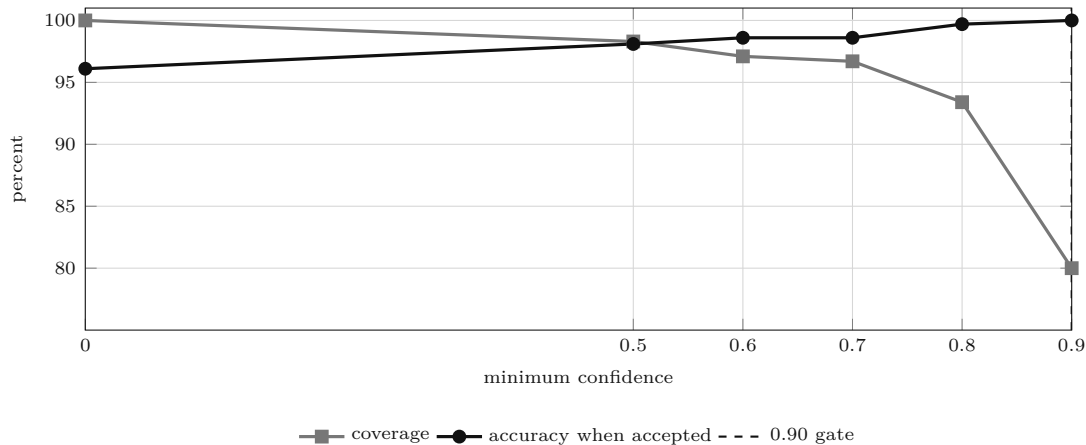


Figure 3: Coverage and accepted accuracy as the minimum confidence threshold rises. The 100% result at 0.90 describes this held-out sample and is not a guarantee of future performance.

We also measured a sequential cascade in which requests below 0.90 moved to the general-purpose model. Jev accepted 800 requests and escalated 200. The combined system reached 97.5% exact-route accuracy (975 of 1,000), with a 175 ms median and an estimated cost of \$0.120 per 1,000 decisions.

Table 3: Observed behavior of the Jev-first cascade.

	Jev accepted	Escalated	Combined
Requests	800 (80%)	200 (20%)	1,000
Exact-route result	800 / 800	—	975 / 1,000
Median latency	—	—	175 ms
95th-percentile latency	—	—	835 ms
Cost / 1,000 decisions	—	—	\$0.120

Tail latency rose because the slowest requests paid for two sequential calls. The cascade improved aggregate accuracy and average economics, but it did not improve every latency statistic. Production architecture is a set of negotiated constraints rather than a single leaderboard number.

7 Implications for an AI-native EHR

We are building Wavera around a compositional operating principle: use frontier intelligence where medicine needs context and synthesis; use specialized models where software needs a fast typed decision; and keep deterministic controls around consequential actions.

- **Decide quickly.** Classification and dispatch should be fast enough to disappear into the interaction. A specialist can turn routing from a generation problem into a bounded decision.
- **Reason deeply.** Frontier models remain essential for ambiguous, contextual work: connecting information across a chart, evaluating evidence, and communicating nuance.
- **Escalate uncertainty.** Confidence is useful only when the product acts on it. Low-confidence decisions should move to a stronger model, ask a clarifying question, or stop.
- **Measure the system.** Model scores are not enough. The complete path includes accuracy, latency, cost, traceability, and whether a consequential action is appropriately controlled.

The future EHR will not be powered by one model. It will be an orchestrated system that knows when to decide quickly, when to reason deeply, and when to stop.

8 Limitations and next steps

- **Synthetic workload.** The balanced dataset does not estimate real-world request prevalence and is not a clinical validation set.
- **Related surface forms.** The held-out set contains 50 unseen semantic seeds expressed in 20 variants each, so all 1,000 observations are not fully independent intents.
- **Pinned model.** Results apply to `jev-1.13.0`; moving aliases were not used.
- **Observed infrastructure.** Latency was measured from one workstation under concurrent load. Other locations and traffic patterns may differ.
- **Workload-specific economics.** Cost uses provider-reported token counts and public list prices at the time of the run.
- **No executed actions.** Neither model could act in an EHR. This study evaluates a routing layer, not clinical judgment or autonomous execution.

The next step is careful shadow evaluation on representative, de-identified traffic, with attention to distribution shift and to the boundary between informational and consequential requests. Confidence thresholds should be calibrated by task against the cost of each kind of error.

9 Conclusion

The benchmark argues for composition rather than replacement. A specialist made the initial routing decision with higher observed accuracy and lower operational cost; a frontier model remains available for work that requires its breadth. In a healthcare system, this is more than an optimization. It is a way to make intelligence faster, more legible, and easier to govern.

References

- [1] TypeSafe AI. “Introducing System One Models & Jev.” 15 September 2026. <https://typesafe.ai/blog/introducing-system-one-models-and-jev>
- [2] TypeSafe AI. “Models.” Accessed 26 September 2026. <https://docs.typesafe.ai/models>

About Wavera Research. We publish selected engineering and product findings from our work on AI-native clinical software. Aggregate results are shared to advance the design conversation; patient data, proprietary prompts, and sensitive implementation details are not included.