Episode 078 Artifact

Thinking Fast: Inside Jev

TypeSafe AI's “System One” Decision Model, October 2026



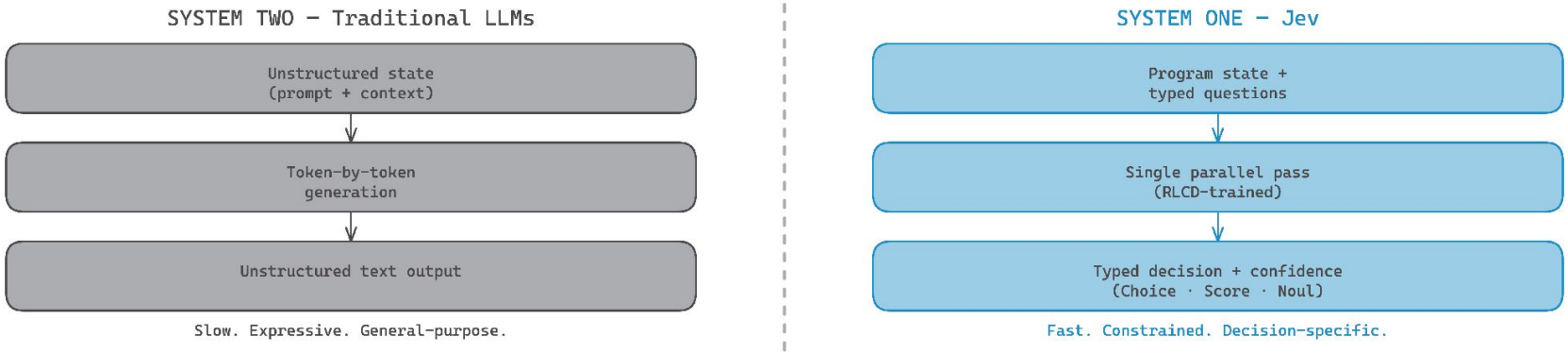
Thinking Fast: Inside Jev

TypeSafe AI's "System One" Decision Model — October 2026

LLMs generate text. Jev generates decisions.
That's not a smaller model — it's a different kind of model.

What Is Jev?

A new model type, not a smaller one



What Are the Limits?

- **Narrow by design** — it only answers pre-defined typed questions (Choice, Score, Noul). It doesn't generate open-ended text, so it can't replace an LLM for conversation or creative work.
- **Only as good as the question** — if the typed schema doesn't capture the real decision space, the output is confidently wrong in a defined shape instead of obviously wrong.
- **Text-only input** — accepts only string, JSON object, or array-of-text values. No native support for images, audio, or other modalities.
- **Depends on good upstream state** — a fast decision is still only as good as the program state fed into it.

cirrius solutions

Why Is It a Big Deal?

The cost of a decision just collapsed

$0.042 / M tokens

Input tokens – output tokens are free

40–400x cheaper than frontier LLMs*

... and fast: 70–500ms per decision (40–200x faster*)

* Self-reported by TypeSafe AI – not independently verified

✓ No Hallucination Risk

Outputs are constrained to pre-defined schemas – nothing to parse, no "wrong shape" of answer.

When a decision costs a fraction of a cent, you stop sampling and start checking everything.

How Could It Be Used?

Five places a cheap, fast decision helps

USE CASE	WHY IT FITS	BENEFIT
Content Moderation & Safety Review	Checks every post or comment instead of sampling a few	5–18x faster* (reported)
LLM Output Monitoring	Catches jailbreaks and errors in real time, in production	Cheap enough to run on every response
Model Routing	Picks which specialized model should handle a given request	Near-zero added latency or cost
Business Process Automation	Email triage and similar structured decision workflows	10–20x cheaper*
Latency-Sensitive Systems	Games, robotics, simulations — where milliseconds matter	70–500ms per decision

* Reported figures (Vercel engineering benchmark; TypeSafe AI)

Decisions, not conversations — that's the layer Jev is building underneath everything else.

Thanks for watching.

Explore all three Cirrius Talk series:

Agentic AI
Designing, building, and operating AI agents that do real work, minus the hype.

Salesforce at Scale
Real-world lessons for building and running Salesforce that scales with the business.

The Military Minute
Veterans and military families on service, transition, and the people who help.

Questions, need help, or a topic idea? Reach out: CirriusTalk@CirriusSolutions.com