neysa | Jocata

Jocata powers credit, fraud, and compliance decisions with AI

Jocata runs a fleet of AI agents across underwriting, document processing, credit assessment, and fraud checks for India's banks and lenders on Neysa Velocis.

NEYSA SETUP

Service

Velocis Managed Inference

Model

Llama 3.2 3B, Llama 3.1 8B, Qwen 32B, Mistral Small 3.1 24B, GPT-OSS, embedding and reranking models, Llama Guard, Prompt Guard.

OUTCOME

1 to 5 sec

Response on simple conversational queries

Up to ~50 sec

Response on complex multi-step workflows

ABOUT JOCATA

Jocata is a BFSI digital transformation partner that powers digital lending for 50+ financial institutions across India, ASEAN, and the Middle East, processing millions of transactions a day across credit automation, risk, and compliance.

OVERVIEW

Jocata built GRID to put AI at the center of lending, fraud, and compliance for India's banks and lenders - a platform of specialized agents that cover document processing, underwriting, credit assessment, fraud checks, and regulatory workflows.

As GRID moved from early use cases into production, its workload changed shape: more agents running at once, answering in real time, against decisions an auditor may revisit years later. That set a high bar for the infrastructure underneath - it had to run many agents at production-speed and scale, keep Jocata in custody of the exact models in production, and hold every request inside a data boundary its clients' compliance teams could approve.

Jocata weighed its options against that bar and chose Neysa Velocis. Velocis meets those requirements as a single managed platform, letting Jocata scale GRID in production without trading speed for control, or control for compliance.

> " The frontier models are good, that was never the question. Our need is custody. For every credit decision, we should be able to point to the exact model weights that produced it, reproduce it for an auditor years later, and show the data stayed within dedicated, in-country infrastructure. Owning open weights on compute reserved for us was the cleanest way to do that. "
>
> **Sundari Vedula - CTO, Jocata**

CHALLENGE

As the GRID platform scaled in usage, it needed to run many more agents at once at production speed - without loosening Jocata's grip on the models or the data. Meeting both at once came down to three requirements.

01. Performance and capacity at production scale

Running a fleet of agents in production means a serving stack that can sustain the load with headroom to add more, and offers an audit-ready, secure setup.

02. Custody of the models in production

In regulated credit decisioning, Jocata needs to know the exact model behind every output, pin it to a version, fine-tune it for domain tasks, and reproduce it for an auditor on demand.

03. A managed serving stack with guardrails in the path

Running a fleet of agents in production means a serving stack that is managed and offers safety-checked on every request, without Jocata assembling and operating all of it in-house.

> " Running inference at this scale for regulated financial workflows requires the whole stack to work together - compute, models, safety layers, data isolation. Neysa gave us that as a single, managed layer. "
>
> **Saidulu Yerpula - Associate Senior Engineering Manager, Jocata**

TRANSFORMATION

A fleet of AI agents now runs in production, served through vLLM on dedicated Neysa Velocis compute across underwriting, document processing, support, business rule execution, customer assistance, case management, and faster access to insights, which ultimately supports quicker and better decision-making.

01.Dedicated compute for a multi-agent production stack

Their agents run concurrently on dedicated GPU nodes. vLLM's continuous batching handles the concurrency without stacking latency. General financial queries return in under 5 seconds.

02. A model stack built for domain precision

Open-weight models run version-pinned under Jocata's control, so every output traces to exact weights it can fine-tune for domain tasks and reproduce for an auditor.

03. Safety and sovereignty baked into the serving layer

A fully managed serving stack applies Llama Guard and Prompt Guard to every request, with safety kept in the inference path.