

Vinicius Morais

Nashville, TN | work@vini.cloud | linkedin.com/in/vinicius-morais | github.com/vmorais17

SUMMARY

Principal engineer and AI systems architect with 20+ years in regulated financial and federal software, now building agentic AI. The through-line across two decades: pattern recognition — which framework fits, which tradeoffs matter, what the client actually needs when they walk in with an ambiguous problem. That judgment is now the product: hierarchical LLM instruction architecture, evaluation gates and guardrails, tiered-model orchestration, and the ML pipelines underneath. Built and deployed agentic AI systems (Sticky Insights, FINN), delivered ERISA and federal platforms, and translates complex AI work into executive-ready narratives. Consulting background; bilingual EN/PT

CORE SKILLS

AI / ML & Agentic Systems: LLM integration & prompt/context engineering (layered/recursive instruction design), evaluation gates & guardrails (relevance/vocabulary/degeneracy checks, gated fallback strategies, error taxonomies), tiered-model orchestration, embeddings & semantic search, agent-loop & human-in-the-loop design, prompt-session logging & auditability

Languages: C#/ .NET Core, Python (AI/ML tooling), TypeScript, JavaScript / Node.js

Architecture: Solution & system design, event-driven architecture, microservices, REST/GraphQL APIs, reliability & resilience patterns

Cloud & DevOps: AWS, Azure, Docker, CI/CD (Azure Pipelines, GitHub Actions, Jenkins), SonarQube

Domains: Regulated financial services (ERISA / retirement), federal, fintech; privacy-first / on-device ML

EXPERIENCE

NetImpact Strategies — Principal Software Engineer

Federal (USDA/NRCS) · Jul 2025 – Present · GenAI/LLM, C#/ .NET, AWS

- Built AI-assisted .NET migration assessment tooling — a risk-scoring repository scanner (dotnetupdateutility) integrated into Jenkins/Playwright/SonarQube pipelines — replacing manual, application-by-application evaluation with a repeatable, portfolio-level risk classification process across a 60-application USDA/NRCS legacy .NET portfolio targeting AWS via Terraform-based infrastructure patterns.
- Designed an AI-assisted development governance framework and presented it to senior engineering leadership (VP and Senior Manager level), establishing tiered model routing and auditable prompt-session logging to enable controlled AI adoption in a security-cleared federal engagement.

Ascensus — Principal Software Engineer

Financial services · Oct 2021 – May 2025 · C#/ .NET Core

- Hired, mentored and led a seven-engineer team (four nearshore LATAM, two US-based), growing engineers into end-to-end owners of production features across EIA and PAS and setting shared design-review and delivery standards across both locations.
- Architected Employer-Initiated Amendments (EIA), replacing a fax/mail qualified-plan amendment process with digital self-service. This eliminated manual re-keying for online submissions and improved data accuracy and traceability against a published two-week processing window.
- Defined the modernization strategy for UPDF (Unified Plan Document Fulfillment): assessed .NET Framework, Windows Services, SOAP and Sybase constraints, then designed a containerized .NET 8/CoreWCF architecture on Kubernetes that preserved existing service contracts and enabled new external integrations.
- Redesigned part of the Plan Adoption System (PAS) document-generation model with reusable components and Excel-based configuration shared across business units, so regulatory changes (SECURE 2.0, CARES Act) could ship without changes to core code, reducing regression risk and rollout time.

Fanvest — Co-Founder & Chief Engineer

Fintech · Apr 2019 – Dec 2022 (concurrent venture)

- Co-founded Fanvest and served as sole Chief Engineer for a real-time, WebSocket-based wagering platform, owning engineering end-to-end from architecture through two beta NFL product launches.
- Built and operated the platform through beta, growing to ~6,500 user signups against a ~7K registration ceiling while the company raised ~\$375K in funding; the company has since wound down.

Core10 / Insight — Lead & Senior Software Engineer

Fintech / enterprise (consulting) · Oct 2017 – Oct 2021

- Client-facing delivery across fintech and enterprise customers; rapid prototyping and ramp-up across a broad set of stacks. For TechData, optimized build/deploy on a large monolith (MSBuild, Azure Pipelines).

SELECTED AI / ML PROJECTS

Sticky Insights — Browser-native applied ML / GenAI

TypeScript, Transformers.js, WebAssembly · 2026 · [stickyinsights.vercel.app](#)

- Built a fully client-side, four-stage NLP pipeline — MiniLM sentence embeddings → dual k-means/agglomerative clustering selected by silhouette score → multi-signal keyphrase extraction → quantized small-LM (LaMini-Flan-T5) theme labeling — with zero server calls, no API keys, and no backend.
- Designed a three-gate validation layer (degeneracy, cosine relevance, vocabulary) that screens every generated label before display, falling back to the top-scoring keyphrase on failure; roughly 50–80% of generations pass on the first attempt.
- Delivered a privacy-by-architecture guarantee — no text ever leaves the browser tab, model weights cache locally for fully offline reuse after first load — aimed at UX research and regulated-data contexts (healthcare, finance, legal) where research data can't cross a network boundary.

FINN — Portable Agent Orchestration Framework

Markdown-native agents, MCP, multi-runtime · 2026

- Architected a hierarchical orchestration framework in which a triage agent routes each request to one of eight domain specialists. 13 of its 14 skills declare their dependencies as MCP resources and tools. Thin adapter files keep all runtime-specific details out of the markdown core, so the same core runs unchanged in Claude Code, Claude Cowork and Cursor. Bindings for ChatGPT and a prompt assembler for local models (Ollama, LM Studio) are also documented.
- Designed context infrastructure where plain markdown/TSV files are the source of truth and SQLite (WAL mode, continuous Litestream backup) is a rebuildable index. Retrieval combines FTS5 full-text search, vector embeddings and a knowledge graph. Agent context stays human-readable, reviewable in git, and can be regenerated from the markdown at any time. Design decisions are recorded in 41 ADRs.
- Enforced a hard boundary between a public open-source framework repo and a private instance repo. Personal memory, databases, secrets and runtime state never enter version control. Model API keys are resolved from 1Password or Keychain at runtime, never stored in files. Outbound model calls can be held for manual approval and are recorded in a tamper-evident egress log. Pre-commit hooks block commits that fail the error-handling test harness or a byte-level healthy-path check.

EDUCATION & CERTIFICATIONS

M.S., Artificial Intelligence — University of Colorado Boulder

B.S., Computer Science — Universidade da Cidade do Rio de Janeiro

AWS Certified Cloud Practitioner · AWS Certified AI Practitioner · Certified AI-Empowered SAgE® Scrum Master