

Anurag Ravi Nimonkar

Machine Learning / AI Engineer · Agentic Systems · Evaluation & Guardrails · RAG

Dublin, Ireland | +353 87 485 7439 | nimonkaranurag@gmail.com | linkedin.com/in/anurag-nimonkar | github.com/nimonkaranurag

SUMMARY

I'm an AI engineer at IBM, two patents (one pending) into my first eighteen months. I lead the squad moving Banco do Brasil onto watsonx Orchestrate, where I build the multi-agent systems that run the bank's PIX, cards/credit and customer account management workflows — and I work/build/architect the stack beneath them end to end, from the agent runtime of Orchestrate to its proprietary evaluation framework to Orchestrate's agent knowledge infrastructure. On my own time, I ship Agent Policy Layer (APL), an open-source guardrails protocol that drops into any agent unmodified.

TECHNICAL SKILLS

Languages	Python, Java, C++14, SQL, Bash, Scala
AI / ML	LLM Evaluation, Agentic AI, Multi-Agent Systems, Retrieval-Augmented Generation (RAG), Prompt Engineering, Guardrails / AI Safety (PII redaction, prompt-injection defense), PyTorch, TensorFlow, Hugging Face
Agent Stack	LangChain, LangGraph, LangSmith, LlamaIndex, Model Context Protocol (MCP), FastMCP, Langfuse, Unitxt
Infra / Data	Docker, Kubernetes, OpenShift, AWS (EC2, EMR), IBM Cloud, Tekton & GitHub Actions CI/CD, REST / gRPC, Microservices, Prometheus, Grafana; PostgreSQL, MongoDB, OpenSearch / Pinecone / Milvus (vector DBs), Spark
Engineering	Distributed Systems, System Design, BDD / TDD (Cucumber, pytest), CUDA / GPGPU, OpenMP, High-Performance Computing, OAuth

EXPERIENCE

IBM	Dublin, Ireland
AI Engineer — watsonx Orchestrate & Watson Assistant	Jan 2025 – Present

- **Squad lead** for migrating Banco do Brasil's Watson Assistant deployment into the watsonx Orchestrate agentic ecosystem. Built the production **multi-agent systems** behind the bank's core financial workflows — PIX (Brazil's instant-payments gateway), Cards/Credit, and Accounts — including the **MCP tool-kits, orchestration, security guardrails, evaluation harness, etc.**
- Resolved a **Sev-1 launch blocker for Banco do Brasil in <12 hours**: a flaw in how IBM's customer-care agent pattern integrated with the Orchestrate runtime was driving **supervisor-agent hallucinations** and had frozen every deployment. Re-architected the agent execution graph (**LangGraph**) around a more robust conversation-context model, eliminating the hallucinations entirely and unblocking the rollout.
- Resolved a **second Banco do Brasil Sev-1 in <4 hrs**: made toolkit registration idempotent and single-writer after collaborator agents re-registering records the supervisor already owned triggered unique-key violations that froze all production deployments.
- Established the migration squad's **engineering standards** for delivery — **Tekton CI/CD** with **regression-gated agent performance thresholds**, toolkit-image versioning, secret-scanning with pre-commit linting, and CODEOWNERS review — setting the contribution baseline the migration team builds on.
- Pioneered the (**patent-pending**) agent-evaluation methodology behind IBM's agent-ops framework. Rebuilt its **LLM user-simulator** — the component that drives automated multi-turn testing — fixing the hallucination, role-confusion, and goal-drift that made results unreliable and lifting conversation-success from **30–40% to 77–86%**.
- Built a **record/replay proxy** that eliminated **\$2.5-3K+/year** in eval-framework CI inference costs and cut test pipelines from **10–15 min to ~30 sec**: 60+ live model calls per commit were served from a cache that refreshes only when a prompt changes, making CI deterministic and resilient to upstream model outages.
- Own the production **RAG stack** behind agent knowledge in Orchestrate: architected the ingestion and retrieval system on **OpenSearch** and led the migration off **Milvus**. Built a [unitxt](#)-based evaluation harness that benchmarked the ingestion path (chunking, embedding) on retrieval and agent-response quality, proving the migration introduced **no regressions**, with a measured 0.75–2% gain across all metrics.
- Identified accuracy gap in evaluation framework's text-matching step that caused success metric regressions; presented findings to stakeholders, implemented BLEU-based semantic analyzer replacing LLM calls with local computation — **eliminated 100+ API calls per test run, reducing operational costs by \$35,000- \$70,000 per year.**
- Resolve production escalations across Watson Assistant and watsonx Orchestrate, owning customer-blocking issues end-to-end for marquee accounts like **DISH, Etihad Airways, and UPS.**

- Ported the computational backend of an astrophysical simulation library (PION) onto a GPU/CPU hybrid mainframe for a **2.5× speedup** on nebula-formation simulations, parallelizing its **C++14** ODE solver.

- Built and benchmarked deep-learning models (**SegNet**, **dilated CNN** + **BiLSTM**) for a smart stethoscope’s real-time heart-sound segmentation, with a full **DSP** feature pipeline (MFCC, spectrograms, wavelets).

PROJECTS

- Designed and shipped a runtime-agnostic guardrails protocol for AI agents — the safety counterpart to MCP: it intercepts an agent at every lifecycle boundary and enforces policy (PII redaction, prompt-injection defense, tool-call approval) via five verdicts (allow / deny / modify / escalate / observe), fail-closed by default, so one policy guards an OpenAI, Anthropic, or LangGraph agent unmodified. On PyPI, Apache-2.0.
- Designed the verdict-composition algebra (deny-overrides, unanimous, weighted) and one-line auto-instrumentation that patches major LLM SDKs in place; every architectural decision is documented as an ADR.

- A Claude Skill that evaluates any deployed agent by running test-agent conversations in parallel and checking behavioral assertions (tool-call expectations, routing, step budgets) against its execution traces; runtime-agnostic via a declarative manifest, with a browser-automation fallback for chat-only deployments.

EDUCATION

Stanford CS224N (self-study): built BERT and a minimal GPT-2 from scratch.
Coursework: Distributed Systems, Cloud, Big Data, NLP, ML, Deep Learning, DSP.

PATENTS & RECOGNITION

- **Patent (filed, IBM 2025):** “Traditional Assistant as a White-Box Tool for Agents” — in prosecution (“Search 2”).
- IBM Bronze Learner badge — 120+ hours of non-mandated professional development.
- AI Mentor, 2025 IBM WatsonX Challenge; led Agentic AI workshops (25+ engineers).