Nikita Baklazhenko

Senior AI Engineer · Agentic Systems, Evaluation & Production ML

Barcelona · +34 663 686 669 · baklazhenkonikita@gmail.com · linkedin.com/in/nikita-baklazhenko · baklazhenko.dev

Senior AI Engineer with a track record of taking agentic and ML systems from prototype to production. Built the agent architecture, shared platform and evaluation harness that the Novartis finance data-science department builds its agents on, after delivering ML forecasts covering €20B+ in yearly costs across 30+ countries. Outside work, built and launched production multi-agent apps and open-source agent evaluation tooling.

EXPERIENCE

Novartis Barcelona, Spain — Sep 2023 – Present

Senior AI Expert · leading the agentic platform since Apr 2026 — Oct 2026 – Present

- Build the backbone architecture and shared platform for agentic apps: agent-to-agent design patterns, reusable Python packages, internal integrations and CI/CD templates so teams ship agents consistently and faster.
- Design LangGraph multi-agent finance systems on AWS: orchestrator–subagent patterns, semantic models over SQL data, reusable subagents (knowledge retrieval, SQL research, dynamic charts) and human-in-the-loop.
- Run a production finance agent over **100M+ data points across 100+ tables**, holding hallucinations **under 1%** across live user interactions and tester sessions.
- Develop the evaluation harness used by every agentic app: LLM-as-a-judge and deterministic metrics on fixture, ground-truth and synthetic test sets. Architecture, model or context changes ship only on measurable gains.
- Ship full-stack agentic apps (React/TypeScript, FastAPI on AWS Lambda, ECR, S3, DynamoDB) with hallucination guardrails, tracing, staged fallbacks and continuous feedback loops.

Expert Data Scientist · joined as Data Scientist, promoted Jun 2025 — Sep 2023 – Oct 2026

- Replaced a quarterly, Excel-based finance forecast with daily automated ML forecasts covering **€20B+ in yearly costs across 30+ countries** — now the baseline for analysts and senior finance managers.
- Cut per-country quarterly MAPE to **under 1% — 50%+ better than analyst forecasts**, with significantly larger gains at finer granularity — saving **150+ analyst hours a month**.
- Product owner of the gross-to-net and tech-ops forecasting products: took them from POC to production and ran the roadmap, stakeholder alignment and quarterly business reviews.
- Shipped agentic products into finance workflows: multi-agent systems on AWS Bedrock and OpenAI Agents SDK, function-calling agents over Snowflake/Databricks, RAG over policy corpora and MCP servers for forecasting.
- Owned LLM rollout: model selection (Bedrock, OpenAI, Anthropic, Ollama), Langfuse tracing, compliance review.

Citibeats · Data Scientist Barcelona, Spain — Mar 2023 – Jul 2023

- Classified **millions of social-media posts** for WHO and UNDP with fine-tuned BERT/Llama in 10+ languages.

Earlier: Transfer Pricing Consultant, PwC (2021) · Teaching Assistant, Statistics & Microeconomics, HSE (2019–22)

PROJECTS & OPEN SOURCE

Periplo — solo-built AI trip planner, **~500 monthly users** — *Web · iOS · Android · periploapp.com*

- **5 orchestrators and 10+ agents** on a FastAPI/GCP backend with SSE streaming, token budgets and staged fallbacks; OAuth-secured MCP servers for Claude and ChatGPT; React Native apps on the App Store and Google Play.

PolicyArena — open-source agent evaluation engine — *pip install policy-arena · policyarena.dev*

- Benchmarks rule-based, RL and LLM agents across **15+ game-theoretic environments** under repeatable conditions; provider-agnostic across OpenAI, Anthropic, Gemini, DeepSeek and Ollama.

Also: **PasteReady** — agentic document assistant that reads email and documents and triggers workflows such as hotel and flight check-ins or ID renewals · **Hoodwise** (neighbourhood scoring) · **ROIInvest** (property analytics).

SKILLS

Agentic AI & LLMs: LangGraph, LangChain, OpenAI Agents SDK, MCP servers (FastMCP, OAuth 2.1), agent evaluation (LLM-as-a-judge), agent memory, RAG, AWS Bedrock, OpenAI / Anthropic / Gemini SDKs, LoRA fine-tuning, vLLM

ML: Time-series forecasting, gradient boosting (XGBoost, LightGBM, CatBoost), NLP, PyTorch, scikit-learn

Engineering & MLOps: Python, TypeScript, React, FastAPI, Docker, Jenkins, GitHub Actions, unit/integration/E2E testing, AWS (Lambda, ECR, S3, DynamoDB, SageMaker, Step Functions), GCP Cloud Run, Terraform, MLflow, Langfuse

Data: SQL, Snowflake, Databricks, PySpark, Polars

Languages: English (fluent), Russian (native), Spanish (intermediate)

EDUCATION

MSc Data Science · Barcelona School of Economics · top 20% — 2022 – 2023

BSc Economics · HSE University, Saint Petersburg · top 10% · Erasmus year at UCL — 2018 – 2022