NICHOLAS PREVITALI

Bergamo, Italy • nicholasprevitali96@gmail.com • LinkedIn • GitHub • Portfolio

AI Engineering Lead | Production RAG, Multi-Agent Systems & LLM Evaluation

AI Engineering Lead with 5 years of experience designing and running production AI systems, from an HR assistant used by 12,000 employees worldwide to multi-agent and call-analysis platforms. I lead the AI unit at S.I. 2001 (team of 5+) and work hands-on on retrieval (hybrid search, reranking, RAG), multi-agent systems (LangGraph, MCP), evaluation and secure deployment on Azure, for clients in HR, retail, financial services, industry, legal and e-commerce.

Work Experience

S.I. 2001 SpA, Bergamo

Head of AI Business Unit | Jan 2023 – Present

Leading the AI division end-to-end: strategy, architecture, delivery, and client management across multiple concurrent enterprise projects.

- **HR assistant used by 12,000 employees** for a European data & AI consultancy (architecture and most of the development, 2023– present): topic classification with hybrid search on pgvector (per-language full-text + vectors, RRF, LLM reranker, permission filters), document RAG with a faithfulness judge, LangGraph memory, multi-client deployment on AKS. Used by 12,000 employees worldwide in 5 languages; 88.5% top-1 topic accuracy on 655 synthetic queries over a 131-topic catalogue; up to 97% of required facts in RAG answers in internal evaluation.
- **Multi-agent planogram assistant** for a global system integrator’s category-management product: LangGraph supervisor with 5 specialist agents and 27 tools, “no data, no opinion” design on top of an OR-Tools CP-SAT optimiser (contributor). Architecture and development of the LLM orchestrator.
- **New in 2026: AI process platform for software teams** (in development): one workspace for PMs, engineers and developers that unifies docs, boards, tracking and Git, with Claude as the engine for task checks, Gantt planning and PR/MR pre-reviews, and Turbocharge as the standard, verifiable workflow with tests and reviews.
- **Sales-call analysis** for an international credit insurer: nightly transcription, speaker identification, objection extraction, scoring per campaign and coaching, plus a LangGraph tutor chatbot. Coaching aimed at the wrong speaker cut from 25% to 0% on the same 12 calls.
- **Privacy-first employee assistant** for an Italian supermarket chain under a cyber-security clearance: no memory or RAG, in-pod anonymiser before every model call, deployed inside the client’s Kubernetes cluster.
- **Technical RAG** over industrial manuals and standards with page-level citations: corrective RAG, Document Intelligence parsing with table re-joining, cost-based routing across three model sizes.
- **GEO/SEO structured data**: multi-tenant Shopify app (JSON-LD, FAQs), WooCommerce JSON-LD pipeline with conflict detection and validation, JSON-LD identity and implementation brief for a credit insurer’s website.
- Also: recommender + RAG for a platform linking companies and non-profits (co-developer), competitive intelligence on ad campaigns with a vision LLM (sole developer), text and image search for a fashion catalogue, contributions to a legal RAG and document-generation platform used by several Italian law firms (300K+ documents, growing).
- Reliability layers across LLM systems: anonymisers, guardrails, fallback policies, tool-call constraints, structured outputs, and tracing of latency, cost and quality (Langfuse).

Senior AI Engineer & Team Lead | Apr 2022 – Dec 2022

Managed 5+ engineers while maintaining hands-on architecture and delivery.

- **Built CNN web page classifier** for e-commerce SEO: full ML pipeline (data collection, labeling, training, validation, monitoring) achieving 98% accuracy and F1 0.95. Eliminated 100% of manual classification across thousands of web pages, enabling downstream SEO automation.
- Delivered corporate AI training to ~50 employees. Scaled team from scratch, establishing Agile/Scrum workflows and delivery standards for enterprise clients.

AI Engineer | Apr 2021 – Mar 2022

First AI hire. Built and fine-tuned CNN-based computer vision models (image classification, object detection, similarity search) using transfer learning. Integrated AI workflows into production using FastAPI microservices, Docker, and Azure AKS with load balancing, caching, and distributed inference. Built monitoring pipelines tracking latency, throughput, model accuracy, and failure modes. Translated business processes into automated ML workflows by collaborating with operations, HR, finance, and technical teams.

Technical Skills

AI & LLMs: LangChain, LangGraph, MCP, Azure OpenAI (GPT-5 / GPT-6 family), Claude, Gemini • Multi-Agent Orchestration, Guardrails, HITL • Prompt Engineering, Function Calling, Structured Outputs

RAG & Search: Hybrid Search (Dense + BM25, RRF), Reranking, Knowledge Graphs • pgvector, Supabase, Pinecone, Qdrant, Weaviate • Docling, Document Ingestion & Chunking

Evaluation: RAGAS (faithfulness, answer relevance, context precision/recall) • Recall@k, MRR, nDCG • LLM-as-Judge, Golden Sets with Domain Experts, Regression Testing

ML & CV: CNNs, Transfer Learning • LLM Fine-tuning (OpenAI models, Flan-T5) • PyTorch, TensorFlow • Recommendation Systems (Two-Tower, Collaborative Filtering)

AI-Assisted Engineering: Claude Code (daily use in delivery) • Plugin development: skills, subagents, hooks, slash commands • MCP integration

Infra: Python, FastAPI • Docker, Kubernetes (Azure AKS) • Azure Cloud • CI/CD, Git • LLMOps, Observability & Tracing (Langfuse)

Education

MSc Computer Engineering, Summa Cum Laude (110/110) | Università di Bergamo | 2019–2022
Thesis: *Extraction of Semantic Topological Maps from 2D Occupational Maps.* Co-authored with C. Celestino, B. Caglioni, D. Brugali.

BSc Computer Engineering | Università di Bergamo | 2016–2019

Research & Open Source

LLM-powered recommendation and agentic retrieval: 50+ papers studied (neural CF, retrieval-augmented recommendations).

Turbocharge (github.com/nicodiansk/turbocharge): open-source Claude Code plugin I built and published (10 skills, 6 agents, 2 hooks): brainstorm → plan → build → review pipeline for test-driven AI-assisted development.

Italian (Native) • **English** Professional working proficiency (Cambridge First Certificate)