NICHOLAS PREVITALI

Bergamo, Italy • +39 380 126 2131 • nicholasprevitali96@gmail.com

LinkedIn: linkedin.com/in/nicholas-previtali-660b83190

Generative AI / ML Engineer

AI-focused Software Engineer and AI Unit Lead with 4+ years of experience in NLP, RAG architectures, Generative AI, and Computer Vision. Led a team of 5+ engineers delivering 10+ enterprise AI solutions. Expert in production-grade multi-agent systems, advanced RAG pipelines, and GEO/SEO AI optimization for e-commerce. Proven track record reducing manual workflows by up to 100% and boosting developer productivity by ~60-70% through AI-powered automation and multi-agent orchestration.

Work Experience

S.I. 2001 SpA

Head of Artificial Intelligence Business Unit | Bergamo | Apr 2021 – Present

Team & Delivery Leadership

- Led and scaled an AI team of 5+ engineers, managing high-impact AI/ML projects for enterprise clients using Agile/Scrum methodologies.
- Owned end-to-end delivery of 10+ AI solutions, driving automation efficiencies and improved customer experience across legal, e-commerce, industrial, and HR domains.

RAG Systems & Knowledge Architecture

- Architected multimodal RAG systems across legal, industrial, and e-commerce domains, implementing hybrid search (semantic + lexical), knowledge graphs, domain-specific embeddings, and vector databases (Pinecone, Qdrant, Weaviate).
- Designed context window management strategies and retrieval optimization for large-scale document processing (100K+ documents).

AI Agents & Automation

- Designed and deployed LLM-powered AI agents automating business workflows (customer support, document processing, HR operations) using LangGraph, LangChain, OpenAI Function Calling, MCP, and custom multi-agent orchestration.
- Architected MCP-based integrations connecting Confluence, Slack, and Jira to establish Single Source of Truth (SSOT) documentation, enabling AI coding assistants (Claude Code, Cursor) to access real-time project context.
- Designed and deployed custom AI agents with single-scope responsibilities (code review, documentation ingestion, context building) using advanced prompt engineering techniques, orchestrating multi-model cooperation (Claude + Gemini) to improve team development velocity by ~60-70%.
- Translated complex business processes into agentic flows by collaborating with operations, HR, finance, and technical teams to map edge cases, decision points, and success criteria.
- Implemented reliability layers including guardrails, fallback policies, tool-call constraints, structured outputs, and human-in-the-loop escalation for high-risk actions.

GEO & SEO AI Optimization

- Developed AI-powered product feed optimization pipelines using latest OpenAI models for GEO (Generative Engine Optimization) and SEO discoverability, reducing content optimization time by ~80%.
- Built hybrid recommendation systems combining semantic ranking, embeddings, vector search, and knowledge-graph signals for personalized product discovery.

Computer Vision & MLOps

- Developed and fine-tuned CNN-based computer vision models (image classification, object detection, similarity search) using transfer learning and custom architectures.
- Integrated AI workflows into production using FastAPI microservices, Docker, and Azure AKS, ensuring scalable deployments with load balancing, caching, and distributed inference.
- Developed monitoring pipelines tracking latency, cost, hallucination rate, success rate, and failure modes; continuously optimized prompts, routing,

and agent architecture.

Key Projects

AI-Powered Developer Productivity Platform

Architected MCP-based system integrating Confluence, Slack, and Jira as SSOT for AI coding assistants. Designed custom agents with single-scope responsibilities (code review, docs ingestion, context building) using advanced prompt engineering, orchestrating Claude + Gemini cooperation for context-aware development. Achieved ~60-70% improvement in team development velocity.

E-commerce GEO/SEO Optimization Platform

Built end-to-end AI pipeline for automatic product enhancement: descriptions, FAQs, structured data (JSON-LD), and multilingual tone-of-voice adaptation. Reduced manual optimization time by ~80% while maintaining brand consistency across languages.

Legal Document RAG System

Architected RAG platform for ~100K Italian legal documents with cross-citation handling and hierarchical retrieval. Achieved 91% retrieval recall@10 and 89% answer relevance, evaluated via LLM-as-judge and legal expert validation.

E-commerce Page Classification (CNN)

Developed full ML pipeline—data collection, labeling, training, validation, and monitoring—for classifying e-commerce page types. Eliminated ~100% of manual classification work across thousands of web pages, enabling downstream SEO automation.

Employee Self-Service Assistant

Built conversational AI chatbot enabling employees to access HR and company information in natural language: PTO balance queries, room booking requests, company newsletter summaries, and policy lookups. Integrated with internal systems via API, with guardrails and human-in-the-loop escalation for sensitive requests.

Hybrid Recommendation System

Built low-latency recommendation engine combining collaborative filtering with vector similarity search. Optimized for minimal compute overhead while maintaining sub-100ms response times.

Technical Skills

LLMs & Generative AI

OpenAI (GPT-4, GPT-4o, o1, o3), Claude, Gemini • Claude Code, Cursor • LangChain, LangGraph, MCP • RAG, Hybrid Search, Knowledge Graphs • Prompt Engineering, Function Calling, Structured Outputs • Multi-Agent Orchestration, Guardrails, Human-in-the-Loop • GEO & SEO AI Optimization

Machine Learning & Computer Vision

CNNs, Transfer Learning, Fine-tuning • Image Classification, Object Detection, Similarity Search • NLP, Embeddings, Semantic Search • PyTorch, TensorFlow • Recommendation Systems • ML Evaluation & Monitoring

Infrastructure & MLOps

Python, FastAPI • Docker, Kubernetes (Azure AKS) • Vector Databases (Pinecone, Qdrant, Weaviate, Supabase) • CI/CD Pipelines, Git • Azure Cloud Services • LLMOps, Observability • Load Balancing, Caching, Distributed Inference

Leadership & Collaboration

Team Leadership (5+ engineers) • Agile/Scrum • Confluence, Slack, Jira • End-to-End Project Delivery • Client Collaboration • Technical Architecture

Education

Master’s Degree in Computer Engineering

Università degli Studi di Bergamo | Bergamo | Sep 2019 – Mar 2022

Summa cum laude (110/110 with honors)

Thesis: Extraction of Semantic Topological Maps from 2D Occupational Maps

Developed algorithms for semantic topological mapping to improve robot navigation through graph-based spatial representations.

Bachelor’s Degree in Computer Engineering

Università degli Studi di Bergamo | Bergamo | Sep 2016 – Dec 2019

Thesis: Iridology Analysis Using Computer Vision for Eye Feature Detection

Applied computer vision for automated iris analysis, implementing feature extraction for pattern and anomaly detection.

Languages

- Italian (Native)
- English (Professional)