

Mohammad Bazzal

Applied AI Engineer | Agentic AI · LLMs · RAG · Machine Learning

+33 6 22 79 31 54 | mhmd.bazzal.99@gmail.com | linkedin.com/in/mohammad-bazzal-3b768b20b | github.com/bazzal99 | Brest, France

SUMMARY

AI Engineer specializing in production LLM systems, agentic AI, RAG, and reinforcement learning, with a PhD from IMT Atlantique. Experienced building AI agents and automation workflows that integrate with business platforms, APIs, and cloud services.

PROFESSIONAL EXPERIENCE

Applied AI & Machine Learning Engineer

2023 – Present

Freelance / Client Projects

Remote

- Designed and delivered custom conversational AI and knowledge assistants integrating private business data, APIs, tool calling, structured outputs, multi-turn memory, and human-in-the-loop workflows, supported by hybrid retrieval, reranking, and RAPTOR hierarchical indexing.
- Developed agentic automation workflows connecting LLMs with external APIs, email, WhatsApp, booking systems, knowledge sources, and internal tools, with state management, validation, retries, structured decision flows, and human escalation.
- Applied machine-learning models to historical business and media data, using traffic, leads, conversions, spend, and performance signals to forecast outcomes, estimate cost and cost-per-lead, and support budget-allocation decisions.
- Built AI-powered analytics and competitive-intelligence systems combining campaign KPIs, web analytics, brand metrics, sentiment, and external market signals to detect trends and anomalies, explain performance changes, and generate actionable insights.
- Developed computer-vision pipelines combining object detection, multi-object tracking, pose estimation, and spatio-temporal modelling for behavioural anomaly and weapon detection in video streams.

Doctoral Researcher — Algorithm Design & Optimisation

Nov 2022 – Dec 2025

IMT Atlantique, Lab-STICC (CNRS UMR 6285)

Brest, France

- Designed a novel decoder-free optimisation algorithm achieving a **100× runtime speedup** over state-of-the-art baselines for large-scale code-distance estimation.
- Published three peer-reviewed works and released three open-source C/MATLAB simulation toolkits, owning algorithm design, implementation, simulation, benchmarking, and reproducible evaluation.

Machine Learning R&D Engineer — Deep Reinforcement Learning

Mar 2022 – Sep 2022

Mitsubishi Electric R&D Centre Europe

Rennes, France

- Developed a CTDE multi-agent actor-critic system in PyTorch for real-time beamforming optimisation in next-generation MIMO networks.
- Reduced outage probability by **~10%** versus the evaluated brute-force baseline using a shared global critic and decentralised inference.

SELECTED AI SYSTEMS

DocuMind AI — Multi-Agent Document Intelligence Platform

github.com/bazzal99/Documind-AI

LangGraph · Qdrant · FastAPI · PostgreSQL · Redis · Docker · Prometheus/Grafana

- Architected a supervisor-led multi-agent document intelligence platform coordinating retrieval, summarisation, synthesis, and critique through stateful LangGraph workflows with bounded corrective-retry loops.
- Built hybrid BM25/vector retrieval with Reciprocal Rank Fusion and cross-encoder reranking; on a 137-question evaluation suite, improved **MRR by 6.2 points** and **nDCG@10 by 5.0 points** over dense-only retrieval.
- Evaluated end-to-end generation quality with **0.96–0.99 faithfulness**, achieving **92.9% correct abstention** and **10.0% false abstention** on unsupported questions.
- Instrumented the corrective-retry loop in production traffic, measuring a **31.6% trigger rate** and **\$0.00207 added cost/query**; identified and documented a retry-exhaustion edge case as a known system limitation.
- Hardened reliability and security with bounded retries, timeouts, Pydantic-validated outputs, citation verification, and adversarial testing; achieved **0 failures across 60 live calls**, **100% citation validity**, and **0/12 successful attacks**.
- Instrumented per-node latency and cost with Prometheus/Grafana, gated retrieval and faithfulness regressions through GitHub Actions, and deployed the platform as a **7-service containerised stack**.

MailTune — Post-Trained Email Writing Assistant

github.com/bazzal99/MailTune

Qwen2.5-1.5B · QLoRA · PEFT · Transformers · FastAPI · MCP

- Fine-tuned Qwen2.5-1.5B-Instruct with 4-bit QLoRA, training only **4.4M parameters (0.28%)** through low-rank adapters; built the full dataset-generation, training, inference, and before/after evaluation pipeline.
- Established raw-base and prompted-base baselines before evaluating fine-tuning gains; QLoRA won **100%** of pairwise comparisons against the raw model and **73%** against the prompted baseline across 120 Claude-judged comparisons on verified-unseen topics.
- Evaluated factual reliability on prompts with embedded ground-truth facts, reducing fabrication to **6%** versus **44%** for the base model and **17%** for the prompted baseline, while preserving **85%** of target facts.
- Validated generalisation across unseen email categories, reduced **p50 inference latency from 7.91s to 2.91s**, and served the shared model through FastAPI and MCP with end-to-end Claude Desktop integration.

TECHNICAL SKILLS

GenAI & Agentic AI:	LangChain, LangGraph, Hugging Face, RAG, Multi-Agent Systems, Tool Calling, MCP, Prompt Engineering, LoRA/QLoRA
Machine Learning:	PyTorch, TensorFlow, Scikit-learn, Transformers, Predictive ML, Reinforcement Learning, Computer Vision
Evaluation:	LLM-as-Judge, Golden Datasets, Retrieval Evaluation, Faithfulness/Grounding, Hallucination Analysis, Regression Testing
Backend, Data & Automation:	Python, FastAPI, REST APIs, Async Python, PostgreSQL, Redis, Qdrant, SQL, n8n
Cloud & Engineering:	Azure, Azure OpenAI, AWS, Docker, Git, GitHub Actions, Prometheus, Grafana

EDUCATION

IMT Atlantique <i>PhD in Telecommunications Engineering</i> <i>Thesis: Distance-Centric Design Methods for Parallel Concatenated Convolutional Codes</i>	Brest, France 2022 – 2025
Université Paris-Saclay <i>MSc in Machine Vision and Artificial Intelligence</i>	Paris, France 2021 – 2022
Lebanese University <i>MEng in Telecommunications Engineering</i>	Lebanon 2017 – 2022
Lebanese University <i>BSc in Computer Science</i>	Lebanon 2018 – 2021