

Hamza Benyamina

Senior Machine Learning Engineer
Riyadh, Saudi Arabia

+966 54 743 5472
✉ hamza.benyamina.official@gmail.com
in LinkedIn Profile

SUMMARY

Senior Machine Learning Engineer with **5+ years** building production AI; currently leading the AI direction at a Saudi legal-tech company whose platform delivers the most accurate legal AI in the Kingdom. Specializes in LLMs, AI agents, and end-to-end deployment.

TECHNICAL SKILLS

Core Competencies: Deep Learning • LLMs • AI Agents • Generative AI • Natural Language Processing (NLP) • MLOps

Agentic AI Patterns: Multi-agent orchestration • Skills • Tool use • Streaming (SSE) • Canvas mode

Programming & Frameworks: Python (FastAPI, Celery, asyncio) • PyTorch • PyTorch Lightning •

HuggingFace (Transformers, PEFT, TRL) • LangChain / LangGraph • JavaScript (Next.js)

Databases: Neo4j • Elasticsearch • Vector Databases (Milvus, ChromaDB) • PostgreSQL/Supabase • Redis

Deployment & Infrastructure: vLLM • HuggingFace TEI • Triton Inference Server • ONNX • CTranslate2 • llama.cpp • Microservices • Model serving • GCP • Ray • SLURM

WORK EXPERIENCE

• Senior Machine Learning Engineer

November 2024 – Present

Qanoniah

Riyadh / Makkah

- **Lead the AI direction at Qanoniah:** ship every AI feature (legal Q&A, document drafting), pitch performance improvements, and shape feature definition by advising on feasibility, approach, and quality.
- Built **QanoniahAI**, a production agentic legal AI platform with **modular multi-agent orchestration, skills, and tool use**; supports **deep search** (citation-verified) and **canvas** modes with background streaming and human-in-the-loop; achieves **90%** on the legal benchmark.
- Shipped the **document drafter** agent: generates Saudi-law contracts using tools and agent skills, with block-level and contract-level verifications (compliance, risk, party, and obligation analysis), placeholder handling, and **chat-driven edits to specific blocks** (no manual selection needed).
- Built a modular **LLMOps layer** with multiple providers (OpenAI, Gemini, Groq, on-prem), OpenInference tracing, cost tracking, YAML configuration, and Jinja-templated prompts.
- Built a full-stack Next.js **eval app** that scores the agent system across a curated scenario set and gates every release on the result; raised answer accuracy by **15%**.
- Modeled Saudi law as a **1.7M+ node Neo4j knowledge graph**; designed indexes and optimized Cypher across three distinct query patterns (**ANN**, **graph traversals**, and filtered/text queries); powers **hybrid retrieval** (vector + keyword + graph) with RRF and reranking.
- Addressed business blockers to legal-AI adoption: kept all inference in-Kingdom (data residency) and shipped **multilingual PII protection** (EN, AR) with transfer learning from English into Arabic.
- Fine-tuned open-source **Qwen** models for Arabic legal document embedding and reranking; deployed embedding and other models on **Triton, vLLM, and HuggingFace TEI**.
- Owned Arabic OCR end to end: continuously benchmarked open-source OCR models, built an **evaluation app** for legal experts to score outputs, and deployed the winning model **in-Kingdom on vLLM/LitServe**.

• Team Lead and Senior AI Engineer (Advisory)

August 2024 – February 2025

Kurasty

Riyadh

- Led development of an AI-driven SaaS that drafts government RFPs and auto-audits them for compliance violations.
- Shipped the first MVP end-to-end: front-end (Next.js), back-end (FastAPI), and AI pipeline, owning all three layers.
- Set up DevOps and MLOps tooling, configured cloud accounts, and provisioned infrastructure for a fast-moving startup.
- Hired and onboarded the entire founding team (engineering and operations); standardized workflow on ClickUp, Slack, and Gmail.

• Machine Learning Engineer

May 2023 – November 2024

Alwasaeet

Riyadh

- Integrated **13+ AI capabilities** (OCR, transcription, image recognition, etc.) into a Digital Asset Management platform, automating information extraction from text, audio, and video assets.
- Built a **voice-biometric system** generating persistent voice fingerprints and diarizing multi-speaker media for time-stamped attribution.
- Developed a **multimodal RAG agent** answering from text, images, audio, video (including YouTube) with **90%** accuracy globally and per-asset.
- Delivered an **enterprise AI search engine** with optimized RAG, reaching **92% faithfulness** and **97% relevance** on company-asset queries.
- Orchestrated an event-driven inference pipeline (GCP Pub/Sub, Celery, Redis) and MLOps stack (DVC, Langfuse, OpenTelemetry); accelerated inference **2–4×** with ONNX + CTranslate2.
- Created an **AI tutor** for 3rd-grade textbooks: advanced Arabic math reasoning, interactive quizzes, in-text help, and a knowledge-graph backbone for complex dependency tracking.
- Built an **LLM-powered course generator** producing videos and slides with customizable templates, animations, and natural voiceovers; **4×** faster generation.

• Machine Learning Research Engineer

February 2022 – May 2023

New York University – Abu Dhabi

Abu Dhabi

- Developed and **published** a novel embedding technique enhancing the **Tiramisu compiler**, reducing MAPE by **3.6%**.  publication
- Introduced a **novel pre-training method** that learns polyhedral geometry, then transfers this knowledge to optimize code performance for complex irregular loops.
- Engineered an alternative **reinforcement-learning method** for code optimization on the Tiramisu compiler.

• Machine Learning Research Intern

June 2020 – December 2021

New York University – Abu Dhabi

Remote

- **Pre-trained ProBERT**, a RoBERTa-based language model, on **6 million source-code files** from 14 programming languages.
- Demonstrated **cross-language generalization**: ProBERT performed well on programming languages not seen during pre-training.
- Fine-tuned ProBERT on four downstream tasks (code clone detection, code review, code search, defect detection); surpassed baseline by **+2% MRR** on code search.

• Machine Learning Engineer

January 2021 – March 2022 (concurrent)

Upwork (multiple clients)

Remote – Freelance

- Engineered a recommendation engine for a note-taking social network (for ULTANIO AG); applied **reinforcement learning** for personalized real-time suggestions and built a triplet-based knowledge graph via relation extraction for semantic relevance.
- Developed an AI-powered EEG / wrist-brain-signal interpreter (research collaboration); achieved **95%+** accuracy and **quantized + pruned** the model for deployment on a **Zynq FPGA**.
- Fine-tuned **GPT-2** for synthetic data generation; deployed on a self-hosted GPU server.
- Completed **21 projects**, maintained a **100% job-success rate**, and earned the **Top-Rated** badge.

• System Administrator

October 2019 – June 2020

AIEGS

Algiers, Algeria

- Deployed and monitored mission-critical Java services on Apache TomEE; integrated the International Postal System with Algérie Post for real-time parcel tracking.

PUBLICATIONS

- **Novel Embedding Technique for the Tiramisu Compiler** – published research reducing MAPE by **3.6%** on code-performance prediction; introduces polyhedral-geometry pre-training transferred to irregular-loop optimization.  arXiv:2403.11522
- **ProBERT** – pre-trained a RoBERTa-class language model on **6M source-code files**; generalized to programming languages not seen at pre-training; +**2%** **MRR** over baseline on code search.

EDUCATION

- **M.Sc. in Computer Science (5-year integrated engineering program)** *June 2016 – June 2021*
École Nationale Supérieure d'Informatique (ESI) Algiers
- **Coursework:** Machine Learning • Natural Language Processing • Data Analysis • Combinatorial Optimization
• Advanced Mathematics

LANGUAGES

Arabic: Native • **English:** Fluent • **French:** Fluent