Ha-Quang LE

📍 Paris, France | 📧 winlp4ever@gmail.com | 🔗 linkedin.com/in/ha-quang-le | 💻 github.com/winlp4ever

PROFESSIONAL SUMMARY

AI systems engineer and architect with 5+ years of experience designing, deploying, and leading large-scale production AI systems. Currently Chief AI Officer at Stellia.ai, a French startup building next-generation educational AI assistants. Proven ability to bridge applied research and scalable infrastructure — retrieval-augmented generation (RAG), agentic reasoning, cloud-agnostic AI architecture, and high-performance model serving at scale.

PROFESSIONAL EXPERIENCE

Stellia.ai — Paris, France

Chief AI Officer (formerly CTO)
Jun 2019 – Present

In 2019, I helped found ProfessorBob.ai (later rebranded as Stellia.ai) and have since led its AI team, defining the vision, architecture, and roadmap for intelligent educational assistants. We successfully raised €4M from Innovacom and INCO Ventures — see announcement.

At Stellia, I worked on:

- Implemented the backbone of an end-to-end **Graph-RAG** framework — the core of our AI educational assistant — covering document ingestion and hierarchical parsing to agentic question answering and personalized exercise recommendations; scaled to ~2K concurrent LLM requests with 1–3 s TTFT in production.
- Led the development and benchmarking of Graph-RAG models achieving +7.2% to +10.1% **DeepEval** accuracy over LangChain/OpenAI RAG baselines, with stronger long-context reasoning.
- Defined product vision and roadmap for initiatives like **Knowledge Graph** construction and **Exercise Generation**, boosting the assistant's domain understanding and task performance.
- Led the development of a **math-solver AI assistant** for Arizona State University (ASU), deployed on their e-learning platform (calculus, statistics); used each semester by 600–1,000 students.
- Deployed production assistants serving thousands of users via client platforms including Galileo, ASU, and Enedis.
- Built and operated the **search cluster** (Elasticsearch, Qdrant) with per-client customization; handled 4–5K RPS with ~31 ms average latency, using Redis for caching, NGINX for load balancing, and Prefect for task orchestration.
- Built a dedicated **model-serving cluster** for LLMs and text-embedding models using vLLM / Text-Embedding-Inference frameworks, serving models such as Qwen, Llama 3, and embedding models like BGE, Jina-Embedding-V2, and E5; supported **2–3 K concurrent requests** with autoscaling through Kubernetes HPA and AWS Batch/Lambda.
- Won the "A.B Code" project (Bpifrance): fine-tuned GPT-2 on coding problems and shipped an AI tutor that helps beginners learn programming.
- Collaborated with **CNRS (Paris-Saclay)** on dataset annotation and fine-tuning T5, BART (text generation) and ColBERT, Sparse (embeddings) for educational tasks (exercise/FAQ generation), achieving +16.2% accuracy and +9.4% "interestingness" on human evaluations; co-authored resulting NLP papers.
- Managed and mentored a 7-person AI team composed of top profiles from leading French programs (École Polytechnique, ENS, Paris-Saclay), overseeing hiring, delivery, and research-to-production integration.
- Co-developed a new **agentic product**: a multi-step reasoning AI agent with a deep-research mode and tool calling (web search, code execution, browser navigation, MCP connectors to Notion, GitHub, Slack, etc.); backend built with OpenAI Agents SDK and LiteLLM, frontend in React (Zustand, React Router, React Flow), including **mind map/schema generation and visualization** for exploratory learning.

Technicolor R&D — Rennes, France

Deep Learning Research Intern
Apr 2018 – Sep 2018
Researched neural architectures for image and audio style-transfer (Gatys, Fader Networks, Adversarial Autoencoders).
Technologies: PyTorch, TensorFlow, WaveNet, NSynth.

BioSerenity — Paris, France

Java Backend Developer Intern
Jun 2017 – Sep 2017
Developed real-time signal-processing pipelines for medical IoT systems, including data filtering and spectrogram generation.
Technologies: Java, AWS, Maven, Spring Boot.

EDUCATION

Master of Science, Data Science — *Université Paris-Saclay, France* (2018–2019)
Engineering Diploma, Cycle Ingénieur Polytechnicien (Machine Learning & Computer Vision) — *École Polytechnique, France* (2015–2018)
Bachelor in Mathematics — *University of Natural Sciences, Hanoi, Vietnam* (2012–2014)

TECHNICAL SKILLS

AI & ML Systems: Graph-RAG, Retrieval-Augmented Generation, Agentic AI, LLM fine-tuning, Text Embeddings, NLP, Information Retrieval
Frameworks: PyTorch, TensorFlow, Hugging Face Transformers/Text-Embeddings-Inference, vLLM, OpenAI Agents SDK, SmolAgents
Infrastructure & Orchestration: Docker, Kubernetes, NGINX, Prefect, Airflow
Databases & Search Engines: PostgreSQL, Elasticsearch, Qdrant, Milvus, Redis, MongoDB
Cloud & Platforms: Production deployments across AWS, GCP, Scaleway, and Azure
• AWS: EC2, RDS, S3, Batch, Lambda, Bedrock, SageMaker
• GCP: Vertex AI, Compute Engine, Cloud Storage
• IaC/automation: Terraform
MLOps & Observability: model registry, eval pipelines, feature store, drift detection, canary deploys, Prometheus, Grafana
Languages: Python, Go, Java, TypeScript, Node.js
Frontend: React, Zustand, TanStack React Query, React Router, React Flow
Others: Linux, Git, CI/CD, REST & GraphQL APIs

HONORS & AWARDS

- 1st Prize — National Mathematics Olympiad for University Students, Vietnam (2014)
- 2nd Prize — National Mathematics Olympiad for High School Students, Vietnam (2012)

LANGUAGES

- Vietnamese: Native
- French: Fluent
- English: Fluent