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 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 architecture, and high-performance model serving at scale.

PROFESSIONAL EXPERIENCE

Stellia.ai — Paris, France

Chief AI Officer (formerly CTO)
Jun 2019 – Present

I co-founded ProfessorBob.ai (later rebranded as Stellia.ai) and lead its AI team, defining the architecture and roadmap for large-scale educational AI assistants. We raised **€4 M** from Innovacom and INCO Ventures — see announcement.

1) Educational Assistant (core product)

- Built the backbone of an end-to-end **Graph-related RAG system**, the core of our assistant, integrating document ingestion, hierarchical structure parsing, and agentic Q&A; **scaled to ~2 K concurrent LLM requests** with 1–3 s TTFT.
- Engineered the production platform:
  - **Search cluster:** Elasticsearch or Qdrant (per client); **4–5 K RPS**, ~31 ms latency, Redis caching, NGINX load balancing.
  - **Model serving:** vLLM / Text-Embedding-Inference serving Qwen, Llama 3, BGE, E5; **2–3 K concurrent requests** with ~100–200 ms latency, autoscaled via Kubernetes HPA and AWS Batch/Lambda.
  - **Orchestration:** Prefect pipelines managing complex dependencies and maximizing parallelism.
  - **MLOps:** Develop with MLflow + DeepEval to auto-evaluate/monitor and deploy ML models.
- Improved RAG accuracy by **+7.2–10.1 % (DeepEval)** over LangChain/OpenAI baselines with stronger long-context reasoning.
- Deployed assistants at scale, including a **math-solver AI assistant** for ASU (calculus, statistics; 600–1 000 students/semester) and deployments for **Galileo**, **ASU**, and **Enedis**.
- Defined roadmap for **Knowledge Graph**, **Exercise Generation**, and **Recommendation Systems**, while mentoring a 7-person AI team from École Polytechnique, ENS, and Paris-Saclay.

2) Research & initiatives

- **A.B Code (Bpifrance):** project **winner of the i-Nov concours** for innovative tech products. Fine-tuned GPT-2 (TensorFlow) on scraped public Python code to build a coding tutor that helps beginners learn programming.
- Collaborate with **CNRS (Paris-Saclay)** to fine-tune T5/BART (text-gen) and ColBERT/Sparse (embeddings) for educational tasks, improving accuracy by +16.2 % and "interestingness" by +9.4 %; co-authored NLP papers.

3) Deep Research Agent (Perplexity-style alternative)

Designed a next-gen agentic system using OpenAI Agents SDK + LiteLLM, supporting self-hosted and proprietary models; **300–700 concurrent agent requests** on a single 64 GB machine; the agent outperforms Grok-deeper-research and Sonar-reasoning-pro at first attempt (38.26 vs 38.22 vs 37.76 respectively) on DeepResearch benchmark.

- **Capabilities:** multi-step reasoning, tool use, and a Deep Research mode enabling extended exploration and synthesis across multiple sources.
- **Tools integrated:** web search, code execution, browser navigation, and MCP connectors to Notion, GitHub, Slack.
- **Memory:** persistent note/message system with Qdrant hybrid search (keyword + vector).
- **Knowledge vault:** sub-agents generate schemas and mind maps with a visualization tool to explore the evolving knowledge base.
- **Frontend:** built the user-facing interface in **React** with real-time streaming responses and an interactive **whiteboard** featuring a rich text notion-like editor for note-taking and graph visualization (Streamdown, React Router, React Flow, Zustand, Milkdown, Tiptap).

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 (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)

TECHNICAL SKILLS

AI & ML Systems: Graph-related RAG, Retrieval-Augmented Generation, Agentic AI, LLM fine-tuning, Text Embeddings, NLP, Information Retrieval
Frameworks: PyTorch, TensorFlow, Hugging Face Transformers, vLLM, Text-Embedding-Inference, OpenAI Agents SDK, SmolAgents, LangChain, LangGraph, n8n
Infrastructure & Orchestration: Docker, Kubernetes, NGINX, Prefect, Airflow
Databases & Search: PostgreSQL, Elasticsearch, Qdrant, Milvus, Redis, MongoDB
Cloud & Platforms: AWS, GCP, Scaleway, 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, MLflow, Neptune.ai
Languages: Python, Go, Java, TypeScript, Node.js
Frontend: React, Zustand, TanStack React Query, React Router, React Flow, Tiptap, Milkdown, d3.js, Tailwind
Others: Linux, Git, Github Actions, 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