# Ha-Quang LE

**Chief AI Officer | Senior AI Systems Engineer | AI Infrastructure Architect**
📍 Paris, France | 📧 winlp4ever@gmail.com | 🔗 linkedin.com/in/ha-quang-le | 💻 github.com/winlp4ever

## PROFESSIONAL SUMMARY

AI engineer and systems architect with 5+ years of experience building large-scale, production-grade AI and data systems. Currently **Chief AI Officer at Stellia.ai** (previously CTO), a French EdTech startup specializing in agentic AI assistants. Experienced in designing and deploying multi-cloud infrastructure, leading high-performance technical teams, and developing full-stack AI systems combining retrieval, reasoning and generation (RAG + LLMs + Agents). Passionate about bridging applied research and scalable engineering.

## PROFESSIONAL EXPERIENCE

### Stellia.ai — Paris, France

**Chief AI Officer (formerly CTO)**
*Jun 2019 – Present*

**Company Overview:**
Stellia.ai builds AI-powered assistants for education, combining advanced document understanding, retrieval and reasoning capabilities. The company successfully **raised €4 M in funding** in partnership with Innovacom and INCO Ventures — see announcement.

**Key Achievements:**

- **Team Leadership:** Managed and mentored a team of 7 data scientists and engineers from leading French institutions (École Polytechnique, ENS, Paris-Saclay).
- **AI System Design:** Architected and implemented the company's core **Graph-RAG system**, handling the full AI pipeline from document ingestion to reasoning-based question answering.
- **Infrastructure:** Designed and deployed a multi-cloud AI infrastructure across AWS, GCP and Scaleway — including distributed storage (S3, RDS, Postgres), compute orchestration, and model serving with Docker & Kubernetes.
- **Scalability & Reliability:** Built high-throughput ML APIs capable of handling thousands of concurrent requests with efficient async I/O, load-balancing and caching mechanisms.
- **Core Product Development:**
    - Implemented an end-to-end graph-based hybrid retrieval system, integrating semantic (vector) and keyword search with Elasticsearch, Qdrant and Milvus.
    - Engineered the document ingestion and parsing pipeline preserving layout, tables and hierarchy (using Tesseract, MinerU, Marker).
    - Developed an agentic AI assistant system combining retrieval, reasoning and tool-usage (web search, code execution, external APIs) built with frameworks like OpenAI Agents SDK and SmolAgents.
- **Research & Model Work:** Benchmarked and fine-tuned LLMs and text-embedding models for domain-specific tasks (Q&A, exercise generation, graph reasoning).
- **Collaboration:** Developed a mathematical reasoning solver for the University of Arizona, integrated into their learning platform.
- **Tech Stack:** Python, Go, TypeScript, React, Elasticsearch, Qdrant, Milvus, PyTorch, TensorFlow, Docker, Kubernetes, AWS / GCP / Scaleway.

### Technicolor R&D — Rennes, France

**Deep Learning Research Intern**
*Apr 2018 – Sep 2018*
Researched deep neural network architectures for style-transfer in images and audio (Gatys, Fader Networks, Adversarial Autoencoders).
Technologies: PyTorch, TensorFlow, Audio feature extraction, WaveNet, NSynth.

---

### BioSerenity — Paris, France

**Java Backend Developer Intern**
*Jun 2017 – Sep 2017*
Developed real-time signal-processing pipelines for a medical IoT system, 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)

---

## TECHNICAL SKILLS

**AI & ML Systems:** Graph-RAG, Retrieval-Augmented Generation, Agentic AI, LLM fine-tuning, Text Embeddings, Vector Search, NLP, Information Retrieval
**Frameworks:** PyTorch, TensorFlow, Hugging Face, OpenAI Agents SDK, SmolAgents
**Infrastructure & Orchestration:** Docker, Kubernetes, Prefect, Airflow
**Databases & Search Engines:** PostgreSQL, Elasticsearch, Qdrant, Milvus
**Cloud & DevOps:** AWS (EC2, RDS, S3), GCP, Scaleway, Azure, Terraform
**Languages:** Python, Go, Java, TypeScript, Node.js
**Frontend:** React, Zustand, TanStack React Query
**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