Francis **Kulumba**

Ph.D. Candidate in NLP · Research Scientist

Paris, France

franciskulumbacs@gmail.com | franciskulumba.fr | Madjakul | francis-k | Francis Kulumba

Summary

Ph.D. candidate in natural language processing at Inria Paris (ALMAnaCH), **defending in late 2026**. My research focuses on authorship attribution through learned representations of writing style, combining contrastive learning, late-interaction retrieval, and mechanistic interpretability. I built and released HALvest, a 17-billion-token multilingual scholarly corpus, and its contrastive derivative HALvest-Contrastive; trained embedding models that outperform single-vector baselines by a factor of four on stylometric retrieval; characterized where authorship signal emerges in encoder-based language models using mechanistic interpretability and traced the internal circuits of an 8B-parameter language model to explain how a planted backdoor trigger reroutes its output. Currently leading the research and deployment of a domain-specific French embedding model at the French Ministry of Defense.

Education

Sorbonne University

Paris, France

Ph.D in Computer and Information Sciences — *Nov. 2022 – Sep. 2026*

- **Subject:** Language Models and Authorship Attribution: Divergences and Insights from Semantic Retrieval.
- **Lab.:** ALMAnaCH, Inria Paris.
- **Adviser:** Laurent Romary.

École Supérieure d'Ingénieurs Léonard de Vinci (ESILV)

Paris La Défense, France

Engineer's Degree in Mathematics and Computer Science — *Sep. 2017 – Jul. 2022*

- Coursework: probability, linear algebra, optimization, algorithms and data structures, machine learning, deep learning, NLP.

Chung-Ang University

Seoul, South Korea

Undergraduate Exchange in Mathematics and Computer Science — *Sep. 2019 – Dec. 2019*

- Coursework: probability, statistics, Linux systems, object-oriented programming (C++).

Work Experience

French Ministry of Defense

Paris, France

AI Research Scientist — *May. 2026 – Today*

- Advise on best practices for efficient dense retrieval in restricted, closed-domain settings.
- Co-lead the end-to-end research and deployment of a French embedding model for search and retrieval over restricted, heterogeneous administrative corpora.
- Adapted the self-instruct methodology to generate synthetic contrastive training pairs against domain-specific administrative documents.
- Shipped a working prototype adopted by another office, improving fine-tuning efficiency.

Inria – ALMAnaCH Research Team

Paris, France

Ph.D Student — *Nov. 2022 – Apr. 2026*

- Designed and built HALvest, a 17-billion-token multilingual corpus of 778k open-access scholarly papers across 56 languages and 16 domains, released under an open license.
- Trained encoder-based embedding models with InfoNCE and multi-GPU full-gather (4×H100, 4×GB200), achieving a four-fold improvement by replacing mean pooling with late interaction scoring.
- Introduced Patch-Level Late Interaction, a compression scheme that groups tokens into patches before matching, and characterized an empirical square-root fit for optimal patch size across sequence lengths.
- Applied mechanistic interpretability (activation patching, linear probing, training dynamics) to explain the four-fold performance gap between scoring mechanisms.
- Mapped the internal circuit of a language-switching backdoor in an 8B-parameter LLM, identifying a three-phase pipeline.
- Released all code, corpus construction scripts, and trained models under open licenses.

French Ministry of Defense

Paris, France

Data Scientist Apprentice — *Sep. 2020 – Aug. 2022*

- Researched named entity recognition (NER) methods in low-resource settings, focusing on weak supervision (self-learning, model distillation, and bootstrapping).
- Implemented and scaled a custom search engine for the Ministry's legislative teams.

Teaching

Tutoring
Paris, France

Private tutor — *Sep. 2023 –*

- Support French middle and high-school students in mathematics, physics and computer science.
- Help students and parents with academic guidance.

EPITA
Paris, France

Lecturer — *Sep. 2023 – Dec. 2025*

- Co-designed and taught an Advanced NLP graduate course. Fall 2023, 2024, 2025.
- Wrote and delivered lectures on tokenization, language modeling, efficient NLP with limited resources, and modern interpretability.
- Supervised student team projects (groups of 3–5) on applied NLP, from proposal review through final oral presentations.

University of Paris I Panthéon-Sorbonne
Paris, France

Teaching Assistant — *Sep. 2023 – Dec. 2024*

- Led Python and algorithm tutorial sections for undergraduates (Spring 2024).
- Led relational database tutorial sections for undergraduates (Fall 2023, 2024).

Skills and Tools

Research areas	Natural Language Processing, Representation Learning, Contrastive Learning, Information Retrieval, Stylometry, Mechanistic Interpretability, Language Modeling, Embedding Models
Programming and scripting	Python, C++, Bash, LaTeX
ML / DL Frameworks	PyTorch, PyTorch Lightning, HuggingFace Transformers, HuggingFace Datasets, nnsight
Infrastructure	SLURM, multi-GPU distributed training (DDP, full-gather), Git, GNU/Linux (Ubuntu), Elasticsearch
Tooling	Neovim, Tmux, Zsh, conda, pip
Languages	*French* (Native), *English* (Fluent), *Spanish* (Intermediate), *Korean* (Notions)

Publications

Other Publications

Language-Switching Triggers Take a Latent Detour Through Language Models
Francis Kulumba, Wissam Antoun, Théo Lasnier, Benoît Sagot, Djamé Seddah
2026

HALvest-Contrastive: Retrieval-Like Authorship Attribution with Patch-Level Late Interaction
Francis Kulumba, Wissam Antoun, Guillaume Vimont, Laurent Romary, Florian Cafiero
2026

Where Does Authorship Signal Emerge in Encoder-Based Language Models?
Francis Kulumba, Guillaume Vimont, Laurent Romary, Florian Cafiero
2026

Triggers Hijack Language Circuits: A Mechanistic Analysis of Backdoor Behaviors in Large Language Models
Théo Lasnier, Wissam Antoun, Francis Kulumba, Djamé Seddah
2026

CamemBERT 2.0: A Smarter French Language Model Aged to Perfection
Wissam Antoun, Francis Kulumba, Rian Touchent, Éric Clergerie, Benoît Sagot, Djamé Seddah
2024