# Rian Touchent

riantouchent@gmail.com | linkedin.com/in/rian-t | github.com/rian-t | Google Scholar | Hugging Face

*Research Scientist | LLM Pretraining | Data Curation at Scale | Domain Adaptation*

## Education

**Inria Paris - Sorbonne Université** — Paris, France
*Ph.D. in Natural Language Processing - ALMAnaCH Team* — *2023 – June 2026 (expected)*

- **Thesis**: Public-Data Pretraining for Clinical Information Extraction
- **Advisors**: Laurent Romary and Éric de la Clergerie
- Trained 7B decoder (128 MI250X, 8.7k GPU-hours) and encoder matching English clinical SOTA with **2.5x fewer tokens**, using neural data filters
- Core contributor to GAPeron (1.5B-24B open LLMs): designed synthetic quality annotations (500k docs) and benchmark contamination study (BIaHs)

**CentraleSupélec - Université Paris-Saclay** — Gif-sur-Yvette, France
*MSc in Artificial Intelligence* — *2021 – 2022*

**ECE Paris School of Engineering** — Paris, France
*Master's in Computer Science (Diplôme d'Ingénieur)* — *2017 – 2022*

## Publications

*2 first-author papers, 53 citations. Full list: Google Scholar*

**CamemBERT-bio: Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data** — *LREC-COLING 2024*

- R. Touchent, L. Romary, É. de la Clergerie
- Collected 413M-word French biomedical corpus from diverse sources
- Trained SOTA French biomedical encoder (100k+ downloads) with 32x less compute than from-scratch

**Biomed-Enriched: Data-Efficient Biomedical Pretraining via Paragraph-Level Annotation** — *Submitted to ACL 2026*

- R. Touchent, N. Godey, É. de la Clergerie
- Encoder matches English clinical SOTA (BioClinical-ModernBERT) with 2.5x fewer tokens using only public data. 7B decoder reaches baseline with 3x fewer tokens

**GAPeron: A Peppered English-French Generative LLM Suite** — *Submitted to ACL 2026*

- N. Godey, W. Antoun, R. Touchent, et al.
- Contributions: data curation, synthetic quality annotations, BIaHs study (showed quality filters amplify benchmark leakage 20x)

**CamemBERT 2.0: A Smarter French Language Model Aged to Perfection** — *arXiv 2024*

- W. Antoun, F. Kulumba, R. Touchent, et al.
- SOTA French encoder (DeBERTaV3 + RoBERTa). Contribution: biomedical evaluation

## Research Experience

**R&D Consultant - LLM Infrastructure** — June 2023 – June 2024
*Praxysanté – Healthcare AI startup (Freelance)* — *Paris, France*

- Deployed open-source LLMs (7B-70B) for healthcare applications with quantization (GPTQ, AWQ) on high-end GPUs
- Built production inference pipeline with vLLM and FastAPI

**R&D Consultant - NLP/NER** — Spring 2025
*ViaDialog (Freelance)* — *Paris, France*

- Built French text anonymization system using NER for customer data

- Designed LLM annotation pipeline (vLLM + constrained decoding) to generate synthetic training data; distilled to production NER model (0.82 F1)

**Pre-doctoral Researcher - Biomedical NLP** June 2022 – November 2022
*Inria Paris - ALMAnaCH Team* *Paris, France*

- Built CamemBERT-bio (see Publications). Work led to PhD position

**Research Intern - Multimodal Generation** May 2021 – August 2021
*Inria Sophia-Antipolis - STARS Team* *Sophia-Antipolis, France*

- Built audio-to-video pipeline: synthesized realistic talking faces from speech input (CNNs, GANs)
- Collected multimodal dataset (audio + video), designed cross-modal lip-sync architecture

## Teaching Experience

**Lecturer - Advanced NLP** Fall 2025
*EPITA School of Engineering* *Paris, France*

- Taught Advanced NLP course (ANLP) to final-year engineering students
- Topics: Domain-specific NLP, Multilingual NLP, Advanced NLP Tasks

**Teaching Assistant - Computer Science** 2023 – 2024
*CentraleSupélec* *Gif-sur-Yvette, France*

- Taught introductory CS to first-year engineering students (Python, Git, SQL)

## Technical Skills

**Research**: LLM Pretraining, Domain Adaptation, Data Curation, Biomedical NLP, NER, Multilingual NLP
**Deep Learning**: PyTorch, Hugging Face, vLLM, DeepSpeed, FSDP, FlashAttention
**Data Processing**: Datatrove, Multinode CPU/GPU pipelines, Corpus preprocessing
**Infrastructure**: Distributed Training, AMD MI250x, NVIDIA H100/A100, Slurm
**Programming**: Python, Bash, SQL
**Tools**: Git, Weights & Biases, Docker, Linux, LaTeX

## Selected Honors

**PAISS 2021** - Paris AI Summer School (lectures by Y. LeCun, C. Schmid, and others)
**Merit Scholarship (2019)** - Ranked 1st/379 at ECE Paris
**Prologin Semi-finalist (2018)** - French National Computer Science Contest