MATTHIEU DUBOIS

Paris, France
duboism@isir.upmc.fr baggerofwords.github.io LinkedIn Google Scholar GitHub

PROFILE

NLP researcher working on how we evaluate what language models write: detecting machine-generated text, and measuring how much of a text's content comes from the model and how much from the person who prompted it. Looking for roles in LLM evaluation, text generation, and understanding how AI was involved in producing a text. Willing to relocate.

RESEARCH EXPERIENCE

PhD Researcher – ISIR, Sorbonne Université & CNRS – Paris, France 2023 – 2026
Supervised by François Yvon (CNRS) and Pablo Piantanida (CentraleSupélec). Thesis: *Identifying Artificial Texts*, defended 13 November 2026.

- **MOSAIC: detecting AI text with several models at once** (Findings of ACL 2025). The performance of zero-shot detectors depends heavily on the fixed language models they use to score text. MOSAIC combines several models, reweighting them at every token, which makes detection robust across generators and domains.
- **How decoding choices affect detectors** (Findings of EMNLP 2025). A systematic study of 37 decoding configurations (temperature, top-k, top-p, repetition penalty, typical and η sampling). Small changes to these settings can drop detector AUROC from near-perfect to chance; the failures are linked to divergences between the generating and scoring distributions. Dataset and code released.
- **Self-Conditioning: how much was the model told?** (AI for Meta-Science workshop @ NeurIPS 2026, non-archival; extended version in preparation). An unsupervised estimator of the outside information a text carries, comparing its likelihood with and without hints drawn from the text itself. Separates fully AI-delegated peer reviews from AI-polished ones with AUC up to 1.0, is robust to rewording, and also applies to student essays and reference letters.

Research Intern, Semantic Communications – Orange Labs – Grenoble, France 2023
Supervised by Louis-Adrien Dufrene, Quentin Lampin and Stéphane Coutant.

- Tested whether sending sentence embeddings over a channel, instead of bits (the current paradigm), makes communication more efficient.
- Measured how robust embeddings are to channel noise, and quantized them heavily to minimise their size while preserving their meaning.

TEACHING

Teaching Assistant – Sorbonne Université – Paris, France Spring 2026
RITAL: Information Retrieval and Natural Language Processing, Master MIND (M1). Led practical sessions, graded student projects and ran oral examinations.

EDUCATION

Sorbonne Université & CNRS – Paris, France 2023 – 2026
PhD in Computer Science

Institut Polytechnique de Paris – Palaiseau, France 2022 – 2023
Master Year 2 – Applied Mathematics and Data Science
Relevant courses: Natural Language Processing; Advanced Learning for Text and Graph Data; Deep Learning; Optimization for Data Science; Reinforcement Learning; Structured Data: Learning and Prediction; Law and Ethics of Artificial Intelligence

Télécom SudParis – Évry, France 2020 – 2023
Diplôme d'ingénieur (French engineering degree), Statistics major

Relevant courses: Machine Learning, Classification and Data Mining; Applied Statistics; Probability and Statistics; Stochastic Processes; Bayesian Estimation in Markovian Models; Optimization; Scientific Computing; Advanced Signal Processing

Sorbonne Université – Paris, France 2020 – 2021
Language Sciences, Computer Science option

Lycée Louis-le-Grand – Paris, France 2018 – 2020
Classe préparatoire MPSI/MP*: two-year intensive programme in mathematics and physics preparing for the competitive entrance exams of the French *Grandes Écoles* (the star marks the most selective track)

École Centrale de Pékin – Beijing, China 2016 – 2018
First two years of the general engineering cycle (Centrale-Supélec group and Beihang University)

PUBLICATIONS

Matthieu Dubois, Pablo Piantanida, François Yvon. 2026.
How Much Were You Told? Measuring External Information in Peer Reviews. AI for Meta-Science Workshop @ NeurIPS 2026 (non-archival). [arXiv] [code]

Matthieu Dubois, François Yvon, Pablo Piantanida. 2025.
How Sampling Affects the Detectability of Machine-written Texts: A Comprehensive Study. Findings of EMNLP 2025. [paper] [code]

Matthieu Dubois, François Yvon, Pablo Piantanida. 2025.
MOSAIC: Multiple Observers Spotting AI Content. Findings of ACL 2025. [paper] [code]

Matthieu Dubois, Pablo Piantanida, François Yvon. 2025.
MOSAIC : Mélange d'experts pour la détection de textes artificiels. TALN 2025 (in French). [paper] [code]

Matthieu Dubois, François Yvon, Pablo Piantanida. 2025.
MOSAIC at GenAI Detection Task 3: Zero-Shot Detection Using an Ensemble of Models. GenAIDetect Workshop, COLING 2025. [paper] [code]

TALKS

- AI for Meta-Science Workshop, NeurIPS – Paris, France Dec 2026
  How Much Were You Told? Measuring External Information in Peer Reviews
- PhD defense, Sorbonne Université – Paris, France 2026
  Identifying Artificial Texts
- DIRO, Université de Montréal – Montréal, Canada 2025
  Research presentation
- Mila NLP Reading Group – Montréal, Canada 2025
  Research presentation
- TALN 2025 – Marseille, France 2025
  Oral presentation: *MOSAIC : Mélange d'experts pour la détection de textes artificiels*
- *Internet Exploreuses* (Origami), S02E02 – podcast/video, in French 2024
  Interview on detecting AI-generated text

SKILLS

Programming: Python, PyTorch, Hugging Face Transformers, large-scale experiments on a Slurm GPU cluster

Research: LLM evaluation and benchmarking, machine-generated text detection, information-theoretic methods, large-scale generation experiments

Languages: French (native), English (C2, IELTS 8.5/9), Chinese (C1, HSK 5)

Miscellaneous: Bouldering – 7a+