



# Solène Tarride

Applied AI Researcher

April 29, 1995

Nantes (France)

solene [dot] tarride [at] gmail [dot] com

Website

LinkedIn

Google Scholar

## About me

PhD researcher specializing in document understanding and multimodal AI. 7+ years developing production ML systems for handwriting recognition, layout analysis, and information extraction. Contribute to interdisciplinary research projects and open-source tools.

## Languages

French

English

## Work experience

- 2022 – **AI Researcher** TEKLIA, Paris (remote), France  
Present Research and development on Document AI (handwritten/printed).
- Conducted applied research on document AI : image and text clustering/classification, object detection, layout analysis, reading-order detection, text recognition (HTR/OCR), named entity recognition and linking, text translation, image captioning, VQA...
  - Drove interdisciplinary research projects bridging computer science and social sciences and humanities.
  - Strong production experience across the full ML lifecycle : data acquisition and curation, model selection, efficient training/fine-tuning strategies, reproducible experiments orchestration, optimized inference, evaluation metrics, deployment.
  - Contributed to open-source ML tools (PyLaia, DAN, nerval, IE-eval), open-source models and datasets ([HuggingFace](#), [Zenodo](#)).
  - Wrote scientific articles, presentations, and technical reports.
- 2019 – **Industrial PhD Student** IRISA/Doptim, Rennes, France  
2022 Research on information extraction from French handwritten historical parish records with few labeled samples.  
Supervisors : Bertrand Coüasnon, Aurélie Lemaitre.
- Developed a full document understanding pipeline (classification, layout analysis, HTR, NER).
  - Dataset curation with synthetic data generation methods.
  - Dissemination through publications in international conferences and journals.
- 2018 – **Research Engineer** IRISA, Rennes, France  
2019 Automatic information extraction models for scanned administrative forms. Results were integrated to the [GéoBretagne](#) public platform.
- 2018 **Research Intern** IRISA, Rennes, France  
Multiple Sclerosis lesion segmentation on 3D MRI images using CNNs.
- 2017 **Research Intern** Jožef Stefan Institute, Ljubljana, Slovenia  
Statistical modeling of nuclear fission yields

## Education

- 2019– **Doctoral training** INSA/IRISA, Rennes, France  
2022 Specializing in Automatic Document Understanding. Coursework and research in Deep Learning (NLP, Computer Vision), Statistical Modeling, Scientific Communication, English, and Research Ethics.
- 2018 **Master's degree** Université Paris-Saclay, Paris, France  
Research-oriented Master in Data Science and Image Processing, graduated with honours.
- 2018 **Engineering degree** ENSIIE, Evry, France  
Specialized in Computer Science, Applied Mathematics, and Statistics.

## Scientific activities

- **15+ peer-reviewed publications** with +130 citations at international journals and conferences (ICDAR, IJDAR, DH).
- **Executive board member** of the GRCE French-speaking research group on document understanding.
- **PhD co-supervision** of Natalia Bottaioli (ENS/Centre Borelli - ongoing).
- **Conference organization** : Co-organizer (SIFED 2024), Session Chair (SIFED 2023), Award Chair (SIFED 2025) for the French conference on Automatic Document Understanding.
- **Reviewer** for international journals and conferences (ICDAR, IJDAR, ICLR, ICIAP, NeurIPS).
- **Mentorship** : 5+ MSc interns on LLM/CV/NLP projects, resulting in 2 publications.
- **Public outreach** : Science communication and education through creative coding workshops (L codent L créent), science festival co-organization (Pint of Science), and invited talks (TESE days, Breizh Data Club, Science Festival).

## Key research projects

- **Finlam (2024 - Present)** | Research lead | [website](#)  
Developed multimodal models for historical newspaper recognition, focusing on reading order detection and logical organization (section, article, advertisements) using multiple modalities (image, text, type, position, separators). Benchmarked 20+ OCR models on complex newspaper layouts and developed DLA models for zone detection.
- **DAI-CRetDHI (2024 - Present)** | Research lead | [website](#), [ressources](#)  
Leading ML development for 16th-19th century French population records. Benchmarked 20+ HTR models and fine-tuned VLMs (QWEN-VL LoRA) using hybrid training : 100k+ noisy samples with 5k gold annotations. Open-sourced models on HuggingFace. Currently developing NER models robust to noisy transcriptions.
- **Socface (2022 - 2025)** | Research partner | [website](#), [article](#)  
Contributed to France's largest census digitization project (15M pages, 700M individuals). Developed joint HTR-NER model achieving 95% accuracy on first names, 85% on surnames, enabling century-scale demographic research across 20 censuses.
- **Hugin-Munin (2022 - 2024)** | Research lead | [website](#), [article](#), [ressources](#)  
Led HTR development for Norwegian archives using supervised, active, unsupervised, transfer, and zero-shot learning. Benchmarked CRNNs vs Transformers and line vs page-level architectures. Developed statistical LMs improving accuracy and attention-based word detection. Published datasets/models on HuggingFace and enhanced PyLaia/DAN open-source frameworks.
- **HikarIA (2023 - 2025)** | Research partner | [website](#), [demo](#)  
Supervised intern working on image captioning (5-model benchmark), clustering, CLIP encoding, and object detection for photographic collections.
- **Balsac (2019 - 2023)** | Research partner | [website](#), [article](#)  
Enhanced HTR pipeline for 6M Quebec genealogical records, achieving 6% CER through architecture optimization and custom statistical LMs for French/English name recognition. Benchmarked NER models for entity extraction from historical records.

## Technologies

- **ML Frameworks** : PyTorch, HuggingFace Transformers, TensorFlow, ONNX Runtime
- **LLM Libraries** : transformers, datasets, peft, trl, accelerate, bitsandbytes, flash-attn, vLLM, pydantic, outlines
- **MLOps & Tooling** : Weights & Biases, MLflow, TensorBoard, Docker, Gradio, HuggingFace Spaces
- **Development** : Python, Git, CI/CD, unit testing, documentation
- **Document AI** : Donut, PyLaia, Kraken, TrOCR, DAN, LayoutLM, LayoutReader, DocFormer, Doc-UFCN, Churro
- **Computer Vision** : YOLO, SAM, DETR, Swin Transformer, ViT
- **Language & Multimodal Models** : QWEN(-VL), Gemma, Florence, DeepSeek(-VL), CLIP, SigLIP, LLaVA
- **APIs** : OpenAI (GPT), Anthropic (Claude), Google AI (Gemini), Mistral AI

**i** For a more comprehensive view of my research publications and projects, find my [Google Scholar page](#) and my [personnal website](#).