Anas Belfathi

LLM Research Engineer (PhD C.) — ML Systems, Retrieval, Evaluation @ Scale

+33 7 45 47 90 70 | anasbelfathi247@gmail.com | linkedin.com/in/anas-belfathi | anas-belfathi.vercel.app/

Skills

Languages: Python, C/C++, Java, JavaScript, TypeScript, SQL

Technologies & Tools: Hugging Face (Transformers/Datasets/PEFT), PyTorch, FAISS, Docker, vLLM, Weights & Biases, Git, MS Azure, Amazon AWS, Google Cloud

Work Experience

Nantes Laboratory of Digital Sciences (LS2N) — Oct 2023 – Present

LLM Research Engineer (Ph.D. Candidate) — *France*

- Designed a hybrid **BERT** $\leftrightarrow$ **LLM** system combining **BERT** contextual encoders with generative **LLMs** for label verification, reranking, and calibration, achieving up to **+15% macro-F1** improvement over BERT-only baselines on long judicial documents.
- Built the first large-scale annotated dataset of **U.S. Supreme Court** opinions with three levels of granularity (**Category**, **Rhetorical Function**, and **Steps**) — over **30K** opinions and **1.2M** sentences, achieving inter-annotator agreement $\kappa = \mathbf{0.72}$. Released **baseline models** and **training scripts**.
- Conducted **text summarization** experiments using **LLMs** with controlled sampling and constrained decoding, including **factuality** and **faithfulness** analyses (**citation precision/recall**, **evidence-free rate**). Integrated **RAG**-based retrieval and constraints, yielding **+8.9** factuality points and **-22%** hallucination rate on internal benchmarks.

TALN Research Labs – Legal NLP Group — Feb 2023 – Oct 2023

Data Scientist Intern — *France*

- Collected and centralized the complete history of **SCOTUS** legal opinions from **CourtListener** and French court decisions from **Légifrance**; applied **topic modeling** and **clustering** to build a structured, searchable foundation for future **LLM**-based applications.
- Delivered the first **state-of-the-art** contribution to **Rhetorical Role Labeling (RRL)** with an enhanced **hierarchical architecture** based on **BERT**-type models, capable of processing full-length legal opinions efficiently and achieving superior performance at reduced computational cost.
- Deployed the developed system on a **law school learning platform** to assist students in exploring **argumentation structures** in court opinions; conducted **training sessions** and **workshops** with legal scholars and engineers to foster cross-disciplinary adoption of **GenAI** tools.

Education

Nantes University — Expected Sep 2026

Ph.D. Candidate in Artificial Intelligence — *France*

- Research focus on LLMs for long legal and medical documents: classification, summarization, NER, retrieval-augmented generation, evaluation and serving
- **7+** publications on LLMs with **4** under review; **2** Best Paper Awards; recent submissions to TALN, ICAIL and ACL

Sorbonne Paris Nord University — 2023

M.Sc. in Data Science (Artificial Intelligence), Valedictorian — *France*

- Program recognized among France's leading AI/Data Science tracks
- Core coursework: Machine Learning, Deep Learning, Natural Language Processing, Probabilistic Modeling, Optimization

Mediterranean Machine Learning Summer School (by DeepMind) — 2025

Participant (selective summer school) — *Croatie*

- Team project on reproducible ML and responsible AI practices; poster/demo at the end of the school.
- Selected participant in a competitive international cohort; collaborated with peers from diverse research backgrounds.
- Completed intensive modules in deep learning, representation learning, and RL; hands-on sessions with research mentors.

Publications & Awards

[1] Coupling Local Context and Global Semantic Prototypes via a Hierarchical Architecture for Rhetorical Role Labeling. **Anas Belfathi**, Nicolas Hernandez, Laura Monceaux, Mary C. Lavissière, Warren Bonnard, Christine Jacquin, Richard Dufour. **EACL 2026 (Under Review)**.

[2] A Simple but Effective Context Retrieval for Sequential Sentence Classification in Long Legal Documents. **Anas Belfathi**, Nicolas Hernandez, Laura Monceaux, Richard Dufour. **ACL 2025**.

[3] Is Selective Masking a Key to Improving Domain Adaptation for Masked Language Models? **Anas Belfathi**, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour. **ICAIL 2025**.

[4] The Role of Context in Sequential Tasks for Long Documents. **Anas Belfathi**, Nicolas Hernandez, Laura Monceaux, Richard Dufour. **TALN 2025**, **Best Paper Award**.

[5] Adapting Language Models to Specialized Domains via Genre- and Topic-Based Selective Masking. **Anas Belfathi**, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour. **TALN 2024**.

[6] Harnessing GPT-3.5-Turbo for Rhetorical Role Prediction in Legal Cases. **Anas Belfathi**, Nicolas Hernandez, Laura Monceaux, Richard Dufour. **JURIX 2023**, **Best Paper Award**.

[7] Enhancing Pre-trained Language Models with Sentence Position Embeddings for Rhetorical Role Recognition in Legal Opinions. **Anas Belfathi**, Nicolas Hernandez, Laura Monceaux, Richard Dufour. **ICAIL 2023**.

Side Projects

Arabic Dialect Translation and Evaluation Benchmark — 2024 – Present

With MBZUAI

- Focuses on reproducibility and evaluation: standardized datasets and metrics (BLEU, COMET, human adequacy) are used to benchmark cross-dialect and cross-domain performance (legal, medical).
- Delivers cleaned and aligned datasets, modular training/evaluation scripts, and a benchmark suite to advance research on Arabic dialect translation.
- Explores how large multilingual language models can adapt to Arabic dialectal variation (MSA, Moroccan, Egyptian, Gulf, Levantine, Tunisian) through corpus curation, retrieval-augmented context selection, and adapter-based fine-tuning.

Mixtral Constrained Decoding for Regulated Domains — 2023 – 2024

With Mistral.AI

- Implemented regex/grammar/JSON-schema constraints and ontology-backed function calls to reduce hallucinations and enforce structured outputs.
- Benchmarked Mixtral 8x22B vs Mistral Large on curated QA and summarization sets, plus adversarial stress tests (ambiguous prompts, out-of-ontology entities, long-context, code-switching). Measured factuality (citation precision/recall, evidence-free rate), faithfulness (answer-to-evidence alignment).
- Shipped a reference repo and evaluation scripts with guidance on quality/latency/cost trade-offs.

FrenchLegalRAG-Bench — 2023 – 2024

With Hugging Face

- Built a reproducible legal RAG benchmark over case law, statutes, and contracts: ingestion with Hugging Face Datasets, document chunking and metadata normalization.
- Implemented modular pipelines using Transformers and FAISS (BM25 plus dense retrievers with Legal-BERT and E5), bge-reranker for passage ranking, and generation on Text Generation Inference (TGI) or Inference Endpoints; enforced JSON Schema outputs and citation tags for paragraph-level grounding.