

Baptiste Blouin

Data Scientist / NLP, Machine Learning

Aix-en-Provence, France
✉ BaptisteBlouin@proton.me
🌐 baptisteblouin.fr
in [baptiste-blouin](#)
🔗 [BaptisteBlouin](#)
French, English

Professional Summary

Data Scientist with a PhD in Computer Science (NLP / Machine Learning), seven years of experience from raw data to production models. I build annotated datasets, train and adapt language models (PyTorch, Hugging Face), including on the Jean-Zay supercomputer, and design the evaluation protocols that judge them. My results are published: +35 F-score points over Jieba in Chinese word segmentation, -50% impact of OCR errors, 93% of performance recovered with 250 annotated sentences. I now bring these methods to production in Aubier, an AI-powered data analytics SaaS platform (text-to-SQL, RAG, AutoML predictions).

Key Projects

Chinese NER Dataset & ML Pipeline

2023–2024

Published: [LREC-COLING 2024](#)

Data Engineering & ML

- **Co-built the largest NER dataset** of manually annotated historical Chinese newspapers (1872-1949): 196 documents, 6,491 entities, entity linking and coreference
- **Built quality-control systems** automated for multilingual datasets
- **Created an ML training pipeline** with an evaluation methodology
- **Tech Stack:** Python, annotation frameworks, PostgreSQL, validation pipelines

Chinese Language Model & Tokenization System

2023

Published: [NLP4DH 2023](#)

ML Model Development

- **Analyzed the linguistic patterns** of transitional Chinese through systematic investigation
- **Developed tokenization models** reaching an F-score above 83%, +35 points over Jieba
- **Collaborated with researchers** at Academia Sinica on this international project
- **Tech Stack:** PyTorch, TensorFlow, tokenization algorithms, fine-tuning

OCR Error Simulation & Model Robustness

2022

Published: [TALN 2022](#)

Experimental Research

- **Studied ML model performance** under noisy-data conditions
- **Developed an augmentation approach** reducing error impact by 50%
- **Evaluated model robustness** with a benchmarking framework
- **Tech Stack:** PyTorch, Transformers, simulation algorithms, evaluation frameworks

Transfer Learning for Domain Adaptation

2021

Published: [NLP4DH 2021](#)

ML Development & Optimization

- **Applied few-shot learning** reaching 93% performance recovery with minimal data
- **Implemented neural architectures** character-aware for production text processing
- **Evaluated the models** across several datasets with statistical analysis
- **Tech Stack:** BERT, character-aware transformers, adaptation techniques

Aubier, Data & AI SaaS Platform

2025–Present

Independent project, [aubier-data.fr](#)

Full-Stack Design & Development

- **Multi-tenant SaaS platform** end-to-end: document extraction, RAG, text-to-SQL, analytics and predictions
- **Advanced RAG engine** (hybrid search pgvector + BM25, GraphRAG, RAPTOR) and **text-to-SQL** with SSE streaming
- **Full security:** JWT RS256, RBAC, tenant isolation, GDPR, Stripe billing
- **Stack:** FastAPI, PostgreSQL/pgvector, Celery, React 18, TypeScript, LangChain/LangGraph, Docker

- **Text analysis platform** over 15 million historical documents (20 billion words), 200+ users
- **Stack:** Python, React, Rust, R, PostgreSQL, Apache Solr, Docker

Professional Experience

AI & Full-Stack Engineer, Aubier (independent)

Sep. 2025 – Present

- **Full product ownership:** from architecture design to production deployment, fully autonomous
- **Managing technical complexity:** data modeling, APIs, async workers, security, frontend

Data Scientist & ML Engineer, Protisvalor, ENP-China Project, Aix-Marseille Université

Jan. 2023 – Aug. 2025

- **Technical lead and lead developer** of the HistText platform, independent project management; helped secure an ERC Proof of Concept grant to bring it to market
- **Explored and solved** complex ML challenges in large-scale historical text processing
- **Communicated technical concepts** through training workshops and international conferences (NLP4DH 2023, LREC-COLING 2024)
- **Interdisciplinary collaboration** with international historians, linguists and computer scientists
- **Distributed training** of large-scale language models and NLP on Jean-Zay (IDRIS/GENCI national HPC cluster), SLURM orchestration

PhD Candidate & ML Scientist, Protisvalor, LIS / IrAsia, Aix-Marseille Université

Sep. 2019 – Dec. 2022

- **Independent research** on a European ERC Advanced grant, self-directed learning
- **Published 4 peer-reviewed papers during the PhD (7 in total)**, technical writing and communication
- **Presented research** at international conferences (PACLIC, NLP4DH, TALN), public speaking
- **Mentored and collaborated** with international research teams, cross-cultural leadership
- **HPC computing on Jean-Zay** (IDRIS/GENCI): language model and NLP training (NER, word segmentation, domain adaptation)

Skills

ML & Data Science	PyTorch, TensorFlow, Hugging Face, scikit-learn, pandas, FLAML / AutoML, Sentence Transformers, NetworkX, UMAP, DuckDB, Statistical Analysis
NLP	spaCy, NLTK, Multilingual NER, Event Extraction, Transfer Learning, Domain Adaptation, Word Segmentation, Data Annotation, OCR (Docling, Google Cloud Vision)
Data	SQL, PostgreSQL, pgvector, MySQL, MongoDB, Elasticsearch, Apache Solr, Redis
LLM	RAG hybrid, GraphRAG, Text-to-SQL, Embeddings, Reranking, LangChain, LangGraph, AI Agents, LLM and RAG Evaluation, Langfuse
Compute & Tools	Jean-Zay / IDRIS (HPC), SLURM, CC-IN2P3, Docker, Linux, Git, CI/CD, FastAPI
Languages	Python, R, SQL, Rust, JavaScript / TypeScript, Java, C++

Education

PhD in Computer Science (NLP / Machine Learning), Aix-Marseille Université	2022
Thesis: Event extraction from facsimiles of historical documents for history studies	
Research skills: Independent investigation, hypothesis testing, experimental design, technical writing	
Advisor: Prof. Benoît Favre Project: ERC Advanced Grant ENP-China (no. 788476)	
MSc in Computer Science (AI/ML), Aix-Marseille Université	2019
BSc in Computer Science, Aix-Marseille Université	2016