

Baptiste Blouin

Data Scientist & AI Engineer / NLP, LLM & Full-Stack

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

Professional Summary

Data Scientist and AI Engineer with a PhD in Computer Science, I combine published ML/NLP research expertise with strong hands-on engineering of AI products in production. From architecture to production, I design and build end-to-end AI products, autonomously. My SaaS platform illustrates this: a genuinely production-grade, multi-tenant LLM system where I cared as much about business value as operational robustness (LLM cost control and end-to-end security). My research background (7+ publications, international conferences, ERC grant) gives me a methodological rigor that I apply equally to designing technical architectures and solving complex problems.

Key Projects

Aubier, Data & AI SaaS Platform

2025–Present

Full-Stack Design & Development

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

- **Single-handedly designed and built an end-to-end multi-tenant SaaS platform:** document extraction, RAG, text-to-SQL, analytics and predictions, covered by over 6,600 automated tests
- **Implemented an advanced RAG engine** with hybrid search (pgvector + BM25), GraphRAG and RAPTOR over unstructured documents (PDF, OCR, VLM)
- **Built a text-to-SQL engine** converting natural language into SQL queries with SSE streaming and step-by-step reasoning
- **Integrated 22 data source connectors:** databases (PostgreSQL, MySQL, Oracle, MongoDB), ERP and accounting (Sage, Cegid, Pennylane), CRM (Salesforce, HubSpot), S3
- **Set up a complete security architecture:** JWT RS256, RBAC, tenant isolation, GDPR, audit logging, rate limiting, Fernet encryption
- **Stack:** FastAPI, PostgreSQL/pgvector, SQLAlchemy 2.0, Celery, Redis, React 18, TypeScript, LangChain, LangGraph, Langfuse, litellm, fastembed, Stripe, Docker

baptisteblouin.fr: Portfolio & AI Agent

2026–Present

Production AI agent & public MCP server

baptisteblouin.fr

- **Designed and operate a LangGraph agent** that answers recruiters about my background, assesses job-offer fit and drafts a contact message
- **Reliability and cost:** multi-provider LLM chain with automatic fallback (Claude, Gemini, Mistral), prompt caching, inference cost under half a euro cent per answer, Langfuse and OpenTelemetry monitoring
- **RAG evaluation:** LLM-as-judge bench (grounding, completeness, no invention) on 38 gold cases, 35 passed, rerun on every model change
- **Security:** prompt-injection classifier (ONNX Runtime, 24 ms) picked from 7 benchmarked models, recall from 10% to 73% on a held-out set, no real question blocked; 20 documented architecture decisions (ADRs)
- **Published a public MCP server** on a Cloudflare Worker (Hono, D1, KV, Workers AI), alongside a daily AI news digest and a radar of AI-ecosystem signals
- **Tech Stack:** LangGraph, FastAPI, PostgreSQL/pgvector, MCP, Cloudflare Workers, Astro, Langfuse, Docker

Private Apps, Agents & Authenticated MCP Servers

2026–Present

4 applications in production & 2 private MCP servers

- **Two private MCP servers protected by OAuth** (RFC 9728/8414, dynamic client registration), every write subject to human confirmation
- **Framework-free tool-using agents** (Claude Haiku, Gemini fallback, bounded turns) in a stock screener (DuckDB) and a business application encrypted at rest (SQLCipher)
- **Operations:** monitor for the site's agent and an infrastructure dashboard with no Docker socket and no SSH, over 3,000 tests
- **Tech Stack:** FastAPI, HTMX, MCP, OAuth, DuckDB, SQLCipher, n8n, Docker

HistText: Large-Scale Text Analysis Platform

Full-Stack Research & Development

2023–2025

Published: JDMDH 2024

- **Built a text analysis platform** querying over 15 million multilingual historical documents, more than 20 billion words
- **Designed a scalable architecture** (Rust backend, Apache Solr): search, concordances, NER and word segmentation applied to whole corpora
- **Implemented a full-stack ML solution** with real-time analytics and interactive visualizations
- **Deployed the platform** to over 200 users in the international digital humanities research community
- **Tech Stack:** Python, React, Rust, R, PostgreSQL, Apache Solr, Docker

Chinese NER Dataset & ML Pipeline

Data Engineering & ML

2023–2024

Published: LREC-COLING 2024

- **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

ML Model Development

2023

Published: NLP4DH 2023

- **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

Experimental Research

2022

Published: TALN 2022

- **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

ML Development & Optimization

2021

Published: NLP4DH 2021

- **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

EventExtractionPapers: Open Source Reference List

Scientific literature watch

2019–2024

580 GitHub stars (09/2026)

- **Created and maintained** a reference selection of event-extraction literature (papers, datasets, models, code)
- **A resource used** by the international NLP research community

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

Languages	Python, Rust, JavaScript / TypeScript, Java, C++, R, SQL
ML & Data Science	PyTorch, TensorFlow, Hugging Face, scikit-learn, FLAML / AutoML, Sentence Transformers, NetworkX, UMAP, pandas, DuckDB, ONNX Runtime, Statistical Analysis
NLP & Documents	spaCy, NLTK, Multilingual NER, Event Extraction, Transfer Learning, Domain Adaptation, Word Segmentation, Data Annotation, OCR (Docling, Google Cloud Vision), Historical Text Processing
LLM & Generative AI	litellm, LangChain, LangGraph, Langfuse, Prompt Engineering, Structured Output, Multi-provider Routing & Fallback, RAG, GraphRAG, RAPTOR, Hybrid Search (BM25 + Vector), Reranking (Cross-Encoder), Embeddings, Vector Search, VLM, Text-to-SQL, LLM Cost Tracking, AI Agents, MCP (Model Context Protocol), Claude (Anthropic), Mistral, Gemini, OpenRouter, PydanticAI, Prompt Caching, LLM Evaluation, LLM Guardrails (Injection Detection)
Full-Stack Dev.	React, Vue.js, TailwindCSS, D3.js, Shiny, Node.js, FastAPI, Flask, SQLAlchemy, Pydantic, Async Python, REST API, WebSockets, Auth / RBAC, OAuth, pytest / TDD, Vitest, ruff / mypy, Playwright, Git, CI/CD, Astro, Hono, HTMX, Streamlit
Infra & DevOps	PostgreSQL, MySQL, pgvector, Elasticsearch, Apache Solr, Redis, Celery, Alembic, Docker, Nginx, S3 / MinIO, SQLCipher, Prometheus / Grafana, OpenTelemetry, Loki / Jaeger, OpenSearch, Cloudflare (Workers, Pages, D1, KV, Workers AI, Tunnel), AWS (CDK, Lambda, Bedrock Guardrails), GCP, Linux, n8n, GitHub Actions, Supply-Chain Security (SBOM, Trivy), Jean-Zay / IDRIS, SLURM
Research & Communication	Scientific Publications, Scientific Methodology, Experimental Design, Technical Writing, International Presentations, Literature Review, Mentoring, Cross-cultural Collaboration, ADRs, LaTeX

Education

PhD in Computer Science (NLP / Machine Learning) , <i>Aix-Marseille Université</i> 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)	2022
MSc in Computer Science (AI/ML) , <i>Aix-Marseille Université</i>	2019
BSc in Computer Science , <i>Aix-Marseille Université</i>	2016

Publications

HistText: An Application for Leveraging Large-Scale Historical Textbases <i>JDMDH, Journal of Data Mining & Digital Humanities</i> Large-scale data-mining tool for Chinese historical documents: extraction and visualization across billions of words. <i>B. Blouin, C. Armand, C. Henriot</i>	2024
A Dataset for Named Entity Recognition and Entity Linking in Chinese Historical Newspapers <i>LREC-COLING, Turin, Italy</i> Largest manually annotated historical Chinese corpus (1872–1949) for NER, entity linking, coreference and relations. <i>B. Blouin, C. Armand, C. Henriot</i>	2024

HAL · GitHub

ACL Anthology

- Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts** 2023
NLP4DH, Tokyo, Japan
 Word-segmentation model for transitional Chinese: F-score above 83%, +35 points over the Jieba tokenizer.
B. Blouin, H.-H. Huang, C. Henriot, C. Armand ACL Anthology
- Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques** 2022
TALN (JEP-TALN-RECITAL), Avignon, France
 Quantifying the impact of OCR errors on three information-extraction tasks; an approach reducing that impact by over 50%.
B. Blouin, B. Favre, J. Auguste ACL Anthology
- Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?** 2021
NLP4DH, Silchar, India
 Transferring NER models to the historical domain: annotating 250 sentences is enough to recover 93% of full-data performance.
B. Blouin, B. Favre, J. Auguste, C. Henriot ACL Anthology
- Creating Biographical Networks from Chinese and English Wikipedia** 2021
Journal of Historical Network Research, Vol. 5, No. 1
 Biographical networks built from 338k Wikipedia biographies (Chinese and English), linked across languages.
B. Blouin, N. van den Bosch, P. Magistry HAL
- Contextual Characters with Segmentation Representation for Named Entity Recognition in Chinese** 2020
PACLIC 34, Hanoi, Vietnam
 A new sinogram representation enriched with word-boundary information for Chinese NER, outperforming the state of the art.
B. Blouin, P. Magistry ACL Anthology