

Baptiste Blouin

AI Engineer / LLM, Agents, Full-Stack

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

Professional Summary

AI Engineer with a PhD in Computer Science (NLP), I design, ship and operate LLM systems in production, single-handedly and end to end: architecture, backend, frontend, infrastructure. Aubier, my multi-tenant SaaS platform, makes company data queryable in natural language (hybrid RAG, GraphRAG, text-to-SQL), with particular care for security and LLM cost control. My portfolio's public agent combines LangGraph, multi-provider fallback, an MCP server, prompt-injection guardrails and an evaluation bench; two private MCP servers, protected by OAuth, complement the public one. Seven years of NLP, including a PhD and 7 peer-reviewed publications, give me the rigor to measure what a model actually does.

Key Projects

Aubier, Data & AI SaaS Platform

Full-Stack Design & Development

2025–Present

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

Production AI agent & public MCP server

2026–Present

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

4 applications in production & 2 private MCP servers

2026–Present

- **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

- **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

LLM & Agents	AI Agents, LangGraph, LangChain, MCP, PydanticAI, litellm, Claude, Mistral, Gemini, multi-provider routing and fallback, Prompt Caching, Structured Output, inference cost tracking
RAG	Hybrid search (pgvector + BM25), GraphRAG, RAPTOR, Reranking (Cross-Encoder), Embeddings, fastembed, Text-to-SQL, VLM, Docling
Quality & Security	LLM and RAG evaluation (LLM-as-judge), guardrails (injection detection, ONNX Runtime), monitoring (Langfuse, OpenTelemetry), pytest, Vitest, ruff / mypy, supply-chain security
Backend	Python, FastAPI (async), SQLAlchemy 2.0, Pydantic, Celery, Redis, PostgreSQL, pgvector, Alembic, MySQL, MongoDB
Frontend	React 18, TypeScript, TailwindCSS, D3.js, Astro, Vega-Lite, React Query, Zustand
Architecture & Infra	SaaS multi-tenant, RBAC, JWT RS256, SSO (OIDC), 2FA (TOTP), GDPR, Stripe, Docker, Cloudflare (Workers, D1, KV, Workers AI), GCP, Linux, n8n, GitHub Actions, Prometheus
NLP & ML	PyTorch, Hugging Face, spaCy, scikit-learn, pandas, Multilingual NER, Transfer Learning, OCR

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