

Baptiste Blouin

Ingénieur IA / LLM, agents, full-stack

Aix-en-Provence, France
✉ BaptisteBlouin@proton.me
🌐 baptisteblouin.fr
in [baptiste-blouin](#)
🐙 [BaptisteBlouin](#)
Français, anglais

Résumé professionnel

Ingénieur IA et docteur en informatique (NLP), je conçois, livre et exploite des systèmes LLM en production, seul et de bout en bout : architecture, backend, frontend, infrastructure. Aubier, ma plateforme SaaS multi-tenant, rend la donnée d'entreprise interrogeable en langage naturel (RAG hybride, GraphRAG, text-to-SQL), avec une attention particulière à la sécurité et à la maîtrise des coûts LLM. L'agent public de mon portfolio associe LangGraph, repli multi-fournisseurs, serveur MCP, garde-fous contre l'injection de prompt et banc d'évaluation ; deux serveurs MCP privés, protégés par OAuth, complètent le public. Sept ans de NLP, dont une thèse et 7 publications à comité de lecture, me donnent la rigueur de mesurer ce qu'un modèle fait vraiment.

Projets clés

Aubier, plateforme SaaS Data & IA

2025–présent

Conception & développement full-stack

Projet indépendant, aubier-data.fr

- **Conçu et développé seul une plateforme SaaS multi-tenant** de bout en bout : extraction documentaire, RAG, text-to-SQL, analyse et prédictions, couverte par plus de 6 600 tests automatisés
- **Implémenté un moteur RAG avancé** avec recherche hybride (pgvector + BM25), GraphRAG et RAPTOR sur des documents non-structurés (PDF, OCR, VLM)
- **Développé un moteur text-to-SQL** du langage naturel vers SQL, avec streaming SSE et raisonnement pas-à-pas
- **Intégré 22 connecteurs de sources de données** : bases (PostgreSQL, MySQL, Oracle, MongoDB), ERP et comptabilité (Sage, Cegid, Pennylane), CRM (Salesforce, HubSpot), S3
- **Mis en place une architecture de sécurité complète** : JWT RS256, RBAC, isolation tenant, GDPR, audit logging, rate limiting, chiffrement Fernet
- **Stack** : FastAPI, PostgreSQL/pgvector, SQLAlchemy 2.0, Celery, Redis, React 18, TypeScript, LangChain, LangGraph, Langfuse, litellm, fastembed, Stripe, Docker

baptisteblouin.fr : portfolio & agent IA

2026–présent

Agent IA en production & serveur MCP public

baptisteblouin.fr

- **Conçu et exploité un agent LangGraph** qui répond aux recruteurs sur mon parcours, évalue l'adéquation d'une offre et prépare un message de contact
- **Fiabilité et coûts** : chaîne de fournisseurs LLM à repli automatique (Claude, Gemini, Mistral), cache de prompt, coût d'inférence sous le demi-centime par réponse, monitoring Langfuse et OpenTelemetry
- **Évaluation RAG** : banc LLM-as-judge (ancrage, complétude, absence d'invention) sur 38 cas de référence, 35 réussis, rejoué à chaque changement de modèle
- **Sécurité** : classifieur d'injection de prompt (ONNX Runtime, 24 ms) retenu parmi 7 modèles comparés, rappel de 10 % à 73 % sur un jeu tenu à part, aucune vraie question bloquée ; 20 décisions d'architecture documentées (ADR)
- **Publié un serveur MCP public** sur un Worker Cloudflare (Hono, D1, KV, Workers AI), avec une veille IA quotidienne et un radar des signaux de l'écosystème IA
- **Stack technique** : LangGraph, FastAPI, PostgreSQL/pgvector, MCP, Cloudflare Workers, Astro, Langfuse, Docker

Applications privées, agents et serveurs MCP authentifiés

2026–présent

4 applications en production & 2 serveurs MCP privés

- **Deux serveurs MCP privés protégés par OAuth** (RFC 9728/8414, enregistrement dynamique des clients), toute écriture soumise à confirmation humaine
- **Agents à outils sans framework** (Claude Haiku, repli Gemini, tours bornés) dans un screener boursier (DuckDB) et une application métier chiffrée au repos (SQLCipher)
- **Exploitation** : moniteur de l'agent du site et tableau de bord d'infrastructure sans socket Docker ni SSH, plus de 3 000 tests
- **Stack technique** : FastAPI, HTMX, MCP, OAuth, DuckDB, SQLCipher, n8n, Docker

- **Plateforme d'analyse de textes** sur 15 millions de documents historiques (20 milliards de mots), 200+ utilisateurs
- **Stack** : Python, React, Rust, R, PostgreSQL, Apache Solr, Docker

Expériences professionnelles

Ingénieur IA & full-stack, Aubier (indépendant) **Sept. 2025 – présent**

- **Ownership produit complet** : de la conception de l'architecture à la mise en production, en autonomie totale
- **Gestion de la complexité technique** : modélisation de données, API, workers async, sécurité, frontend

Data scientist & ingénieur ML, Protisvalor, projet ENP-China, Aix-Marseille Université **Janv. 2023 – août 2025**

- **Responsable technique et développeur principal** de la plateforme HistText, gestion de projet indépendante; contribution à l'obtention d'une bourse ERC Proof of Concept pour sa valorisation
- **Exploration et résolution** de défis ML complexes dans le traitement de textes historiques à grande échelle
- **Communication de concepts techniques** en ateliers et conférences (NLP4DH 2023, LREC-COLING 2024)
- **Collaboration interdisciplinaire** avec historiens, linguistes et informaticiens internationaux
- **Entraînement distribué** de modèles de langue sur le supercalculateur Jean-Zay (IDRIS/GENCI), SLURM

Doctorant & scientifique ML, Protisvalor, LIS / IrAsia, Aix-Marseille Université **Sept. 2019 – déc. 2022**

- **Recherche indépendante** sur subvention ERC avancée européenne, apprentissage auto-dirigé
- **Publié 4 articles à comité de lecture pendant la thèse (7 au total)**, rédaction technique et communication
- **Présenté la recherche** en conférences internationales (PACLIC, NLP4DH, TALN), prise de parole publique
- **Mentoré et collaboré** avec équipes de recherche internationales, leadership interculturel
- **Calcul HPC sur Jean-Zay** (IDRIS/GENCI) : entraînement de modèles NLP (NER, segmentation, transfert)

Compétences

LLM & agents	Agents IA, LangGraph, LangChain, MCP, PydanticAI, litellm, Claude, Mistral, Gemini, routage multi-fournisseurs et repli, Prompt caching, structured output, suivi du coût d'inférence
RAG	Recherche hybride (pgvector + BM25), GraphRAG, RAPTOR, Reranking (cross-encoder), Embeddings, fastembed, Text-to-SQL, VLM, Docling
Qualité & sécurité	Évaluation LLM et RAG (LLM-as-judge), garde-fous (détection d'injection, ONNX Runtime), monitoring (Langfuse, OpenTelemetry), pytest, Vitest, ruff / mypy, sécurité supply chain
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, NER multilingue, Transfer learning, OCR

Formations

Doctorat en informatique (TAL / machine learning), Aix-Marseille Université **2022**

Thèse : Extraction d'événements à partir de fac-similés de documents anciens pour les études en histoire

Compétences de recherche : Investigation indépendante, test d'hypothèses, conception expérimentale, rédaction technique

Directeur : Prof. Benoît Favre | **Projet** : Subvention ERC avancée ENP-China (n° 788476)

Master Informatique (IA/ML), Aix-Marseille Université **2019**

Licence Informatique, Aix-Marseille Université **2016**