

Baptiste Blouin

Data Scientist & Ingénieur IA / NLP, LLM & Full-Stack

Aix-en-Provence, France
✉ BaptisteBlouin@proton.me
🌐 baptisteblouin.fr
in baptiste-blouin
🔊 Baptiste Blouin
Français, Anglais

Résumé Professionnel

Data Scientist et Ingénieur IA avec un Doctorat en Informatique, je combine une expertise en recherche ML/NLP publiée avec une solide pratique d'ingénierie de produits IA en production. De l'architecture à la mise en production, je conçois et réalise des produits IA de bout en bout, en autonomie. Ma plateforme SaaS l'illustre : un système LLM réellement en production, multi-tenant, où j'ai autant soigné la valeur métier que la robustesse d'exploitation — maîtrise des coûts LLM et sécurité de bout en bout. Ma formation par la recherche (7+ publications, conférences internationales, subvention ERC) me donne une rigueur méthodologique que j'applique aussi bien à la conception d'architectures techniques qu'à la résolution de problèmes complexes.

Projets Clés

Plateforme SaaS de Transformation de Données par IA

2025–Présent

Conception & Développement Full-Stack

Projet personnel — [aubier-data.fr](#)

- **Conçu et développé une plateforme SaaS multi-tenant** de bout en bout : extraction documentaire, RAG, text-to-SQL, analyse et prédictions
- **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** convertissant le langage naturel en requêtes SQL avec streaming SSE et raisonnement pas-à-pas
- **Intégré 15+ connecteurs de sources de données** : bases relationnelles (PostgreSQL, MySQL, MongoDB), APIs SaaS (Salesforce, HubSpot, Stripe), 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

HistText - Plateforme d'Analyse de Textes à Grande Échelle

2023–2025

Recherche & Développement Full-Stack

Publié : [JDMDH 2024](#)

- **Développé une plateforme d'analyse de textes** traitant des milliards de tokens de documents historiques
- **Conçu une architecture évolutive** pour gérer des millions de documents aux exigences complexes
- **Implémenté une solution ML full-stack** avec analyses en temps réel et visualisations interactives
- **Déployé la plateforme** pour la communauté internationale de recherche en humanités numériques
- **Stack Technique** : Python, React, Rust, R, PostgreSQL, Apache Solr, Docker

Dataset NER Chinois & Pipeline ML

2023–2024

Ingénierie de Données & ML

Publié : [LREC-COLING 2024](#)

- **Travaillé sur l'annotation** d'un dataset NER chinois historique (1872-1949)
- **Développé des systèmes de contrôle qualité** automatisés pour datasets multilingues
- **Créé un pipeline d'entraînement ML** avec méthodologie d'évaluation
- **Stack Technique** : Python, frameworks d'annotation, PostgreSQL, pipelines de validation

Modèle de Langue Chinoise & Système de Tokenisation

2023

Développement de Modèles ML

Publié : [NLP4DH 2023](#)

- **Analysé les motifs linguistiques** du chinois transitionnel par investigation systématique
- **Développé des modèles de tokenisation** atteignant 83% de précision avec 35% d'amélioration
- **Collaboré avec les chercheurs** d'Academia Sinica sur ce projet international
- **Stack Technique** : PyTorch, TensorFlow, algorithmes de tokenisation, fine-tuning

Simulation d'Erreurs OCR & Robustesse de Modèles

Recherche Expérimentale

2022

Publié : TALN 2022

- Étudié la performance des modèles ML sous conditions de données bruitées
- Développé une approche d'augmentation réduisant l'impact d'erreur de 50%
- Évalué la robustesse des modèles avec un framework de benchmarking
- **Stack Technique** : PyTorch, Transformers, algorithmes de simulation, frameworks d'évaluation

Transfer Learning pour l'Adaptation de Domaine

Développement & Optimisation ML

2021

Publié : NLP4DH 2021

- Appliqué le few-shot learning atteignant 93% de récupération de performance avec données minimales
- Implémenté des architectures neuronales character-aware pour traitement de texte en production
- Évalué les modèles sur plusieurs datasets avec analyse statistique
- **Stack Technique** : BERT, transformers character-aware, techniques d'adaptation

EventExtractionPapers - Liste de Référence Open Source

Veille scientifique & bibliographique

2019–2024

580+ étoiles GitHub

- Créé et entretenu une sélection de référence de la littérature sur l'extraction d'événements (papiers, datasets, modèles, code)
- Plus de 580 étoiles GitHub, ressource utilisée par la communauté de recherche NLP internationale

Expériences professionnelles

Développeur Full-Stack & Ingénieur IA — *Projet Personnel*

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 — *Projet ENP-China, Aix-Marseille Université*

Janv. 2023 – Août 2025

- Chercheur principal et développeur pour la plateforme HistText, gestion de projet indépendante
- Exploration et résolution de défis ML complexes dans le traitement de textes historiques à grande échelle
- Communication de concepts techniques par ateliers de formation dans 4 pays
- Collaboration interdisciplinaire avec historiens, linguistes et informaticiens internationaux
- Entraînement distribué de modèles de langue et NLP à grande échelle sur Jean-Zay (IDRIS/GENCI, supercalculateur national), orchestration SLURM

Doctorant & Scientifique ML — *Laboratoire LIS / IrAsia, Aix-Marseille Université*

Sept. 2019 – Déc. 2022

- Recherche indépendante sur subvention ERC Avancée européenne, apprentissage auto-dirigé
- Publié 5+ articles à comité de lecture, rédaction technique et communication
- Présenté la recherche en conférences internationales dans 7 pays (France, Inde, Vietnam, Japon, Italie, USA, Taïwan), prise de parole publique
- Mentoré et collaboré avec équipes de recherche internationales, leadership interculturel
- Calcul HPC sur Jean-Zay (IDRIS/GENCI) : entraînement de modèles de langue et NLP (NER, segmentation de mots, adaptation de domaine)

Compétences

Langages	Python, Rust, JavaScript / TypeScript, Java, C++, R, SQL
ML & Data Science	PyTorch, TensorFlow, Hugging Face, scikit-learn, FLAML / AutoML, Sentence Transformers, NetworkX, UMAP, Analyse Statistique
NLP & Documents	spaCy, NLTK, NER Multilingue, Extraction d'Événements, Transfer Learning, Adaptation de Domaine, Segmentation de Mots, Annotation de Données, OCR (Docling, Google Cloud Vision), Traitement de Textes Historiques
LLM & IA Générative	litellm, LangChain, LangGraph, Langfuse, Prompt Engineering, Structured Output, Routing Multi-provider & Fallback, RAG, GraphRAG, RAPTOR, Recherche Hybride (BM25 + Vector), Reranking (Cross-Encoder), Embeddings, Vector Search, VLM, Text-to-SQL, Suivi de Coûts LLM
Dév. Full-Stack	React, Vue.js, TailwindCSS, D3.js, Shiny, Node.js, FastAPI, Flask, SQLAlchemy, Pydantic, Async Python, REST API, WebSockets, Auth / RBAC, pytest / TDD, Playwright, Git, CI/CD
Infra & DevOps	PostgreSQL, MySQL, pgvector, Elasticsearch, Apache Solr, Redis, Celery, Alembic, Docker, Nginx, S3 / MinIO, Prometheus / Grafana, OpenTelemetry, GCP, Linux, Jean-Zay / IDRIS, SLURM
Recherche & Communication	Publications Scientifiques, Méthodologie Scientifique, Conception Expérimentale, Rédaction Technique, Présentations Internationales, Revue de Littérature, Mentorat, Collaboration Interculturelle, LaTeX

Formations

Doctorat en Informatique (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. Benoit Favre Projet : Subvention ERC Avancée ENP-China (no. 788476)	
Master Informatique (IA/ML) — Aix-Marseille Université	2019
Licence Informatique — Aix-Marseille Université	2016

Publications

HistText: An Application for Leveraging Large-Scale Historical Textbases	2024
<i>JDMDH — Journal of Data Mining & Digital Humanities</i>	
Outil de fouille de données à grande échelle pour documents historiques chinois : extraction et visualisation sur des milliards de mots.	
B. Blouin, C. Armand, C. Henriot	HAL · GitHub
A Dataset for Named Entity Recognition and Entity Linking in Chinese Historical Newspapers	2024
<i>LREC-COLING, Torino, Italy</i>	
Plus grand corpus chinois historique (1872–1949) annoté manuellement pour la NER, le liage d'entités, la coréférence et les relations.	
B. Blouin, C. Armand, C. Henriot	ACL Anthology
Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts	2023
<i>NLP4DH, Tokyo, Japan</i>	
Modèle de segmentation de mots pour le chinois de transition : plus de 83% de précision, F-score supérieur de 35% aux outils existants.	
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
<i>TALN (JEP-TALN-RECITAL), Avignon, France</i>	
Quantification de l'impact des erreurs OCR sur trois tâches d'extraction d'information ; approche réduisant cet impact de plus de 50%.	
B. Blouin, B. Favre, J. Auguste	ACL Anthology
Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?	2021
<i>NLP4DH, Silchar, India</i>	
Transfert de modèles NER vers le domaine historique : annoter 250 phrases suffit à récupérer 93% des performances pleine-données.	
B. Blouin, B. Favre, J. Auguste, C. Henriot	ACL Anthology
Creating Biographical Networks from Chinese and English Wikipedia	2021
<i>Journal of Historical Network Research, Vol. 5, No. 1</i>	
Réseaux biographiques construits à partir de 338k biographies Wikipédia (chinois et anglais), reliées entre les langues.	
B. Blouin, N. van den Bosch, P. Magistry	HAL
Contextual Characters with Segmentation Representation for Named Entity Recognition in Chinese	2020
<i>PACLIC 34, Hanoi, Vietnam</i>	
Nouvelle représentation des sinogrammes enrichie des frontières de mots pour la NER chinoise, surpassant l'état de l'art.	
B. Blouin, P. Magistry	ACL Anthology