

Nicolas Ha

Ingénieur logiciel freelance senior

France | nicolas@klar.im | nicolas-ha.com | github.com/nha | stackoverflow.com/u/1327651 | linkedin.com/in/nicolasha
Droit de travailler : UE (française), Royaume-Uni (britannique)

PROFIL

Ingénieur logiciel freelance senior, 12+ ans d'expérience, aujourd'hui sur les systèmes d'IA : petits modèles spécialisés, orchestration et workflows agentiques, RAG, outillage MCP, mémoire d'agents.
Références : Griffin Bank, MailOnline.

J'ai dirigé une équipe backend sur un pipeline servant 1,3 milliard de vues par mois, participé à la création de nouvelles équipes dans une banque pendant l'obtention de sa licence bancaire. J'ai livré plusieurs produits de 0 à 1. Dernier en date : Klar, filtre anti-spam IA pour Apple Mail.

EXPÉRIENCE PROFESSIONNELLE

Fondateur & Lead AI Engineer

Klar (klar.im), Télétravail, France | 01/2026 – aujourd'hui

En production sur l'App Store : filtre anti-spam pour Apple Mail qui tourne entièrement sur la machine, sans rien transmettre.

- Entraînement du classifieur en Python : une tête à quatre classes sur un encodeur XLM-RoBERTa gelé, quantifié pour tourner sur le GPU du Mac en moins de 50 ms par message, sans accès réseau.
- Développement du moteur d'inférence C++17 (llama.cpp/GGML), exposé via une API C stable, avec bindings Node et Swift, un milter Postfix et des tests end-to-end Docker.
- Outillage IA : mémoire et recherche sur mesure (MCP, LangChain, RAG, hooks et skills d'agents).

Développeur Staff

Narval (GM NFT), Paris, France | 06/2025 – 12/2025

Rejoint une société de 4 personnes avant la création du produit.

- Intégration de custodians institutionnels (BitGo, Zodia) pour l'accès à la DeFi (staking Aave, transactions TON), avec authentification et autorisation GNAP/OAuth sur un moteur de règles OPA.
- Développement de l'infrastructure de décodage et de proxy pour la validation des transactions on-chain (EIP-7730).

Ingénieur logiciel senior

Griffin Bank (BaaS), Londres, Royaume-Uni | 04/2022 – 06/2025

Composants de plateforme dont dépendait l'obtention de la licence bancaire britannique, en environnement fortement réglementé.

- Pilotage de plusieurs nouveaux chantiers de plateforme de bout en bout, et participation à la création de nouvelles équipes d'ingénierie à mesure que l'entreprise se structurait pour obtenir sa licence.
- Implémentation de l'autorisation admin sécurisée, d'une architecture de webhooks et de l'authentification multi-facteurs par passkeys.
- Conception des systèmes de queuing et de traitement, du logging d'audit, de l'infrastructure de jobs planifiés et de la signature HTTP des API.

Développeur indépendant

TurtleQueue (SaaS), Londres, Royaume-Uni | 09/2019 – 04/2022

Direction du développement de TurtleQueue, plateforme SaaS de message broker compatible Apache Pulsar et Apache Kafka.

- Produit mené de 0 à 1 : conception et construction de toute la stack technologique from scratch ; obtention de crédits d'impôt R&D.
- Développement d'un proxy parlant les protocoles réseau d'Apache Kafka et d'Apache Pulsar : les applications clientes existantes se connectent sans modification.
- Recrutement et encadrement d'une petite équipe de prestataires pour accélérer la livraison.
- En parallèle, prestations d'ingénierie freelance pour MailOnline.

Responsable d'équipe backend

MailOnline (Média), Londres, Royaume-Uni | 08/2016 – 07/2019

Progression d'ingénieur senior à responsable d'équipe backend, encadrant des développeurs expérimentés.

- Architecture et montée en charge du pipeline de publication jusqu'à 1,3 milliard de vues par mois.
- Implémentation d'un ad placement augmentant les revenus ; analytics temps réel réduisant les cycles de feedback.

Ingénieur logiciel

Go Find It (Startup mobile apps), Paris, France | 05/2015 – 04/2016

Développement d'un CMS mobile full-stack pour hôtels de luxe.

- Pilotage du pipeline complet, du développement au DevOps et à la publication App Store, pour une équipe de 10 développeurs mobiles, avec un outillage de déploiement SSH en Python.

Ingénieur logiciel

Apizee (Startup web telecommunications), Lannion, France | 03/2014 – 05/2015

Premier ingénieur backend recruté.

- Prototype mené de 0 à 1 : connecteurs entre clients WebRTC et systèmes téléphoniques existants (Alcatel-Lucent SIP IMS, RCS), et l'outillage de mesure pour Orange Labs à l'origine d'un contrat pluriannuel avec Orange.
- Technologie démontrée au Mobile World Congress de Barcelone, sur le stand de HP.

Stagiaire

Siléane (PME Robotique), France | 2012

- Prototype mené de 0 à 1 : contrôleur de bras robotique industriel Fanuc piloté par la vision Microsoft Kinect, permettant à des utilisateurs non techniques de programmer des trajectoires par gestes.

Stagiaire

Dassault Aviation (division équipements), France | 2010

- Développement d'un banc de test automatisé pour des cartes de contrôle de vol avionique militaire, permettant des cycles de test de 24 h et réduisant le temps de test de 50 %.

COMPÉTENCES

IA AGENTIQUE & SYSTÈMES LLM	LangChain / LangGraph | Serveurs et outils MCP | RAG et recherche hybride | Mémoire d'agents | Workflows multi-agents | Évaluation et benchmarking d'agents | Embeddings et inférence locale (ONNX, llama.cpp) | Recherche vectorielle
LEADERSHIP	Leadership technique | Encadrement et mentorat d'ingénieurs | Création de nouvelles équipes | Recrutement et pilotage de prestataires | Responsabilité de l'architecture
LANGAGES DE PROGRAMMATION	Python | C++ | Clojure | TypeScript | JavaScript
PLATEFORMES & INFRASTRUCTURE	Node.js | React | Docker | Kubernetes | AWS | CI/CD | Git/GitLab
DONNÉES & MESSAGERIE	PostgreSQL | Redis | SQLite | FoundationDB | ElasticSearch | MongoDB | Kafka | Pulsar | RabbitMQ
MA FAÇON DE TRAVAILLER	Itération pilotée par les benchmarks | Systèmes de production scalables et fiables | Tests (unitaires, intégration, e2e, génératifs)

FORMATION

Diplôme d'ingénieur en informatique

ENSSAT, Lannion, France | 2011 – 2014

Représentant étudiant. Échange Erasmus en Allemagne, cours dispensés en allemand.

Bachelor of Engineering in Electronic and Computer Engineering

Edinburgh Napier University, Édimbourg, Royaume-Uni | 2010 – 2011

Cours de C, conception de systèmes embarqués et temps réel, machines à états.

LANGUES

Français (langue maternelle) | Anglais