Pierre Rognion

Applied AI Architect / Solutions Architect · Systèmes LLM en production

Paris, France · Français / anglais bilingue · pierre@atomly.ai · linkedin.com/in/pierrerognion · atomly.ai · Systèmes en production (détail): atomly.ai/fr/work

Je pars d'un problème métier, je le mène jusqu'à un système d'IA en production, et je le fais adopter. Head of AI de l'une des plus grandes marketplaces freelance de France, où je mène le cadrage, je conçois l'architecture, j'écris le code et je livre, pour tous les métiers qui consomment de l'IA : Sales, Produit, Engineering, Finance, Direction. Avant le code, douze ans face aux clients : une practice IA co-construite de zéro, des PoC d'IA générative menés jusqu'en production chez Bpifrance et Arkéa, et plus de 1 000 professionnels formés à l'IA générative appliquée chez Chanel, Stellantis, Bouygues Telecom, Rakuten, Generali et Institut Pasteur. Je préfère livrer un système dont je sais nommer les modes de défaillance qu'un système qui se contente de bien démontrer.

EXPÉRIENCE

Crème de la Crème · Head of AI / Applied AI Architect — Mai 2025 à aujourd'hui

Responsable technique du portefeuille IA d'une marketplace digital native où plus de 50 000 freelances rencontrent les entreprises qui recrutent, en mission via ma société. Cinq métiers commanditaires : Sales, Produit, Engineering, Finance et Direction, chaque cas d'usage portant son propriétaire, sa mesure de succès et son critère d'arrêt. Problème métier, conception du système, build, production, adoption. Également livrés : un outil de sourcing expert bloqué derrière des licences, ouvert en self-service à toute l'équipe commerciale, une marketplace de skills métier packagées, et un serveur MCP au-dessus du design system. Deux chantiers précoces arrêtés une fois que le CRM en place a sorti l'équivalent.

Atomly AI · Fondateur & AI Engineer — Mai 2024 à aujourd'hui

Ma société (SASU), à travers laquelle je mène mes missions et mes propres builds : agents sur mesure, prototypes internes, et programmes de formation pratiques.

Mozza · GenAI Training Manager — Juin 2024 à juin 2025

Programmes d'IA générative appliquée pour des équipes Produit et Tech : conception de prompts, idéation assistée, automatisation de workflows.

Twelve Consulting · Manager, Innovation & IA (Senior Consultant 2016 à 2021) — Sept. 2016 à avril 2024

Co-construction de zéro de la practice IA du cabinet : conception de l'offre, discovery technique avec les dirigeants clients, cadrage de cas d'usage, et PoC d'IA générative menés jusqu'en pipelines de production, notamment chez Bpifrance. Plusieurs comptes clients portés en parallèle et encadrement des consultants embarqués dessus, une dizaine sur quatre projets simultanés chez Bpifrance. Avant cela, Product Owner d'un moteur de tarification Salesforce.

Avant — 2012 à 2019

Teaching assistant au Wagon (2019) · conseil en intégration CRM, Oracle puis Salesforce · CGI (2015 à 2016) · DISAPI (2014) · stage marketing dans la Bay Area de San Francisco (2012).

COMPÉTENCES TECHNIQUES

Ma façon de construire: Je porte l'architecture et j'écris les systèmes front et back, en TypeScript et en Python, en menant le build avec des agents de code (Codex, Claude Code) et en livrant via la revue de l'équipe engineering du client. Quand un chantier demande plus de bras ou une expertise que je n'ai pas, je fais venir quelqu'un et j'écris qui est propriétaire de quoi ensuite. Socle full-stack issu d'un bootcamp Le Wagon fait en 2019 en parallèle du métier, entretenu depuis.

Langages & plateforme: TypeScript/JavaScript, Python, React, Node, SQL, PostgreSQL avec pgvector, Vercel, PostHog, Git.

Plateforme OpenAI en production: Responses API et sorties structurées validées par schéma, embeddings sous la couche de recherche, en production quotidienne sur le parsing des missions et sur une partie du moteur de matching. Également construit avec : Agent Builder, ChatKit, Codex, génération d'images, et les API Realtime et audio sur des travaux vocaux antérieurs.

Systèmes LLM: Conception de serveurs MCP (OAuth 2.1 + PKCE, autorisation multi-couches, allowlists d'outils, audit), workflows agentiques et sub-agents, skills packagées, sorties structurées validées par schéma, LLM-as-a-judge, harnais d'éval et golden datasets, RAG, contrôle du coût et de la latence. Plusieurs fournisseurs frontier en production côte à côte, choisis selon la tâche plutôt que par défaut. Savoir où chacun gagne, c'est la comparaison que les clients demandent avant de s'engager.

Recherche & delivery: Recherche hybride, kNN vectoriel, BM25, Reciprocal Rank Fusion, text-to-SQL. Conception human-in-the-loop, garde-fous et gouvernance sur données sensibles, arbitrages build-versus-buy, adoption par des équipes non techniques.

FORMATION & CERTIFICATIONS

Formation: Le Wagon, développement web full-stack (2019) · Sciences Po Aix, Master Relations internationales et Business (2010 à 2014) · Truman State University, programme d'échange en Business Administration (2011 à 2012) · Classes préparatoires A/L (2007 à 2010).

Certifications: Introduction to Model Context Protocol · Salesforce Certified AI Associate · Concepteur-Développeur d'Applications Web (RNCP).

SYSTÈMES SÉLECTIONNÉS

Construits dans une marketplace où des milliers de freelances rencontrent les entreprises qui les cherchent, et où une équipe Sales doit faire ce rapprochement vite et sans voir de données qu'elle n'a pas à voir. Le code appartient au client et les dépôts sont privés ; l'architecture, les démos et les résultats mesurés peuvent être présentés en entretien.

Ouvrir le back-office d'une entreprise au langage naturel, sans relâcher une seule permission

Juil. à août 2026 · pilote avant production · MCP · OAuth 2.1 + PKCE · Streamable HTTP · PostgreSQL · 49 outils · seul architecte et développeur

- **Ce que ça débloque.** N'importe qui dans l'entreprise demande en une phrase ce qui demandait un parcours de plusieurs écrans, et l'action est exécutée. Des tâches trop petites pour valoir la peine d'être déléguées peuvent aller à un agent, parce que l'agent hérite des droits de la personne qui demande et non d'un compte de service. La même question posée par deux personnes renvoie deux réponses différentes, et aucune n'apprend ce que l'autre peut voir.
- **Comment, et l'arbitrage.** Trois couches d'autorisation vérifiées côté serveur, des personas dont les périmètres sont calculés à partir des données plutôt que recopiés à la main, une projection du payload champ par champ, et un audit trail requêtable. L'essentiel du travail n'était pas l'API mais de répondre explicitement aux questions de sécurité que l'interface portait implicitement : comment on donne l'accès, ce qu'un salarié qui part perd et quand, ce qui se passe la première fois qu'on demande à un outil une action irréversible. Quand un accès client OAuth manquant a bloqué la campagne d'intégration, j'ai écrit un script qui passe par la vraie boucle d'autorisation plutôt qu'un contournement. Ça a coûté deux jours.

Recherche hybride, scoring et jugement LLM borné

Mai à juin 2026 · en production, quotidien · API OpenAI Responses et embeddings · PostgreSQL/pgvector · RRF · LLM-as-a-judge · TypeScript

- **Ce que ça a changé.** Un brief qui imposait une recherche manuelle rend maintenant une shortlist qu'un commercial peut défendre devant un client. Il rend ce qui correspond puis s’arrête, au lieu de compléter chaque brief jusqu’au même nombre fixe de profils, et il fait remonter des candidats forts que le classement par mots-clés laisse sous la ligne de flottaison. Déployé auprès des Sales avant l'ouverture générale.
- **Comment, et l'arbitrage.** Un rappel à quatre signaux (BM25, recouvrement de compétences, kNN vectoriel, index d'expériences) fusionnés par Reciprocal Rank Fusion, un scoring déterministe pondéré, puis un LLM-as-a-judge borné à un appel par candidat, pour que le coût et la latence par brief restent plats quand le vivier grandit. Le harnais d'éval est venu avant le moteur, qui n'a pas eu le droit de sortir tant qu'il ne battait pas l'existant sur un golden dataset plutôt que sur une démo.

Décision human-in-the-loop sur les candidatures entrantes

Déc. 2025 · en production, quotidien · construit de bout en bout, puis fiabilisé en prod avec une développeuse · sorties structurées validées par schéma · TypeScript

- **Ce que ça a changé.** 90 à 95 % du flux entrant est tranché sans humain, environ sept heures par semaine ont été rendues à une Product Manager, et l'entrée sur la plateforme a cessé de dépendre de la présence d'une seule personne. Validé avant mise en service sur cinquante candidatures déjà tranchées à la main, sans un seul désaccord.
- **Comment, et l'arbitrage.** Des verdicts à trois états scorés axe par axe, des sorties structurées validées par schéma, et un routage par règles de tout doute vers un humain. Trier vaut mieux qu'automatiser à 100 % : garder une personne sur les cas durs est ce qui a rendu l'automatisation acceptable sur une décision aussi sensible. Le schéma est le contrat, pas le modèle, donc le pipeline a traversé des générations de modèles sans changer le format de décision.

Accès en langage courant à une base de production, sous gouvernance

Févr. à mars 2026 · en production, quotidien · serveur MCP base de données et skill packagée · PostgreSQL · TypeScript

- **Ce que ça a changé.** Des extractions qui prenaient de trente minutes à plusieurs heures se font en quelques secondes, en self-service, pour une équipe non technique. Les questions de routine ne font plus la queue derrière un ingénieur, et celles qui comptent en atteignent toujours un.
- **Comment, et l'arbitrage.** Un serveur MCP base de données avec découverte de schéma, plus une skill packagée qui porte le contexte métier que le schéma ne porte pas, et des règles métier câblées dans des templates de requête obligatoires plutôt que laissées à la mémoire de chacun. À l'origine l'agent exécutait seul ; faute d'un environnement sûr où exécuter, je l'ai restreint à générer du SQL qu'un humain relit et exécute. Une décision de gouvernance, pas une limite technique, avec une porte de sortie documentée.