

Ariel Guidi

AI Infrastructure & DevOps Solutions Architect

Paris, France · +33 7 64 20 86 75 · arieleddie1@gmail.com · linkedin.com/in/arieleddieguidi · arielguidi.com

AI infrastructure architect with five years of experience delivering customer-facing AI platforms across AWS, Kubernetes, on-premises GPUs and hybrid edge environments. Combines hands-on Terraform and GitOps delivery with architecture guidance, production deployment and cross-team support.

Selected client experience

SIA Partners · AI & Platform Engineer · Sep 2024 - Present

AlloPneus · infrastructure and DevOps foundation · Mar 2026 - Jun 2026

- Built **development and production infrastructure from scratch** for a customer-facing assistant on the client's commercial portal: reusable Terraform deployment, CI/CD, OIDC federation into AWS, AgentCore Gateway and Observability.
- **Supported the application team through development and deployment** once the foundation was operational.

CACIB, Crédit Agricole CIB · GPU inference and deployment architecture · Nov 2025 - May 2026

- Redesigned the deployment architecture for a C3/C4 banking workflow, **reducing its 50-60 second response time by 50% at equal performance** and serving a fine-tuned small model with vLLM on the client's own GPUs.
- Traced roughly 70% of initial latency to engineering: prompt bloat, unconditional reasoning, cold starts and redundant orchestration. Proposed Kubernetes for the target scale of 10-30k users; **it was not part of the deployed client stack**.

Nhood · Kubernetes AI platform and NVIDIA edge architecture · Mar 2025 - Aug 2025

- Contributed to building an AI platform from scratch on EKS with Terraform and Argo CD, then took **four AI use cases to production** under the client's security and FinOps standards.
- Designed a **hybrid edge architecture with NVIDIA** so GDPR-sensitive video is processed on site and only analytics reach the cloud. Became a reusable internal offering.
- Served as the **main technical point of contact**, coordinating the client's security, infrastructure and use-case teams through delivery.

Sovereign agentic AI platform · project lead · Jul 2026 - Present

- Leading the identity and authorization foundation for a self-hosted platform where **agents act on behalf of users** across C1-C4 data: effective access is the **intersection of user and agent rights**, delegation is revocable mid-task, and decisions are auditable. Deployed under **GitOps on Kubernetes**.

Internal GPU and inference engineering · continuous alongside client work

- Led the design of a **benchmarking CLI** comparing vLLM-heavy serving configurations across latency, throughput and quality, with every run recorded for reproducibility. Operationalised internal GPUs behind identity-based access.

Core technologies

- Kubernetes · EKS · Argo CD
- Terraform · AWS · CI/CD
- vLLM · GPU inference · NVIDIA edge
- SageMaker · MLflow
- Python · Docker · GitOps
- OpenCost · AgentCore Observability

Architecture scope

- Kubernetes platform delivery
- AI inference and deployment
- Cloud, on-premises and edge
- Infrastructure as Code and GitOps
- Security, FinOps and observability
- Customer architecture and handover

Earlier client experience

Artefact

Data & ML Consultant · 2022 - 2024

- Orange: commercial agent validated with sales users; fibre forecasting pipeline.
- Matmut: industrialised an end-to-end pipeline segmenting 3 million members.
- Bayer: production hybrid recommender; coached the team on MLOps.

Zoomer.ai

Machine Learning Engineer · 2021 - 2022

Co-led 3 features of an AI data assistant, halved entity-recognition mismatches and put the model on scheduled retraining.

Languages

French and English, fluent

Education

MSc Applied Mathematics: Modelling and Simulation
Université Paris-Saclay · 2020

Machine Learning Engineer, RNCP 7
CentraleSupélec and OpenClassrooms · 2022