

Associate Director - AI Infrastructure Architect (AI Factory)

Job ID

REQ-10088422

Sep 22, 2026

LOC_IN

About the Role

Key Responsibilities:

- Own the AI infrastructure reference architecture and the standards that govern training, fine-tuning, inference and retrieval-augmented patterns across public cloud and private AI hosting.
- Define the standard AI workload intake pattern, including architecture review criteria, data classification handling, environment selection and the conditions under which a use case may go to production.
- Set guardrails for sensitive data, model access, prompt and inference logging, monitoring, resilience, retention and cost, in partnership with Security, Data Privacy and Quality.
- Design GPU and accelerator capacity patterns - shared pools, reservations, scheduling, quotas and burst to cloud - and support the associated capacity and cost models.
- Support qualification of AI platforms, regions and services for enterprise consumption together with AI & HPC Engineering and the public cloud architects.
- Define the MLOps and LLMOps reference chain: model registry, versioning, evaluation, deployment, drift and performance monitoring, and decommissioning.
- Maintain architecture decision records for AI platform choices, linking each to business value, risk posture and total cost of ownership.
- Advise business, research and Infrastructure Solution Delivery stakeholders on feasible AI infrastructure options, and translate scientific and commercial requirements into technical designs.
- Track the AI infrastructure market, evaluate emerging accelerators, model-serving stacks and platform services, and bring credible innovation into the roadmap.
- Coach engineers and architects on AI infrastructure practice and act as a role model for the Novartis Values and Behaviours.

Essential Requirement :

- 8+ years' experience in infrastructure, cloud, or platform architecture, with a strong track record of designing and supporting enterprise-scale technology solutions.
- Demonstrated expertise in building, deploying, and operating AI and machine learning infrastructure within production environments.
- Experience enabling AI solutions to transition from proof-of-concept stages into scalable, governed, and business-critical services.
- Strong knowledge of GPU and accelerator technologies, Kubernetes, container orchestration, and cloud-based AI platforms such as Azure and AWS.
- Hands-on experience with model serving, inference optimization, MLOps/LLMOps practices, and modern data and vector pipeline architectures.
- Solid understanding of identity and access management, observability, cost optimization (FinOps), and secure AI platform operations.

- Familiarity with Responsible AI principles, including explainability, auditability, logging, human oversight, and risk management controls.
- Excellent stakeholder management and communication skills, with the ability to explain complex AI infrastructure concepts to technical and non-technical audiences while driving collaboration across engineering, security, privacy, and business teams.

Desirable Requirement:

- Fluent English, written and spoken. Additional languages are an asset.

You'll receive: You can find everything you need to know about our benefits and rewards in the Novartis Life Handbook.

<https://www.novartis.com/careers/benefits-rewards>

Commitment to Diversity and Inclusion:

Novartis is committed to building an outstanding, inclusive work environment and diverse teams' representative of the patients and communities we serve.

Join our Novartis Network: If this role is not suitable to your experience or career goals but you wish to stay connected to hear more about Novartis and our career opportunities, join the Novartis Network here:

<https://talentnetwork.novartis.com/network>

Role Requirements

Division

DIV_TO

Business Unit

Information Technology

Location

LOC_IN

Site

Hyderabad (Office)

Company / Legal Entity

IN10 (FCRS = IN010) Novartis Healthcare Private Limited

Functional Area

FCT_IT

Job Type

Full time

Employment Type

Regular

Shift Work

No

[Apply to Job](#)

Job ID

REQ-10088422

Associate Director - AI Infrastructure Architect (AI Factory)

Source URL: <https://jobapi.novartis.com/req-10088422-associate-director-ai-infrastructure-architect-ai-factory>

List of links present in page

1. <https://jobapi.novartis.com/req-10088422-associate-director-ai-infrastructure-architect-ai-factory>
2. <https://www.novartis.com/careers/benefits-rewards>
3. <https://talentnetwork.novartis.com/network>
4. https://novartis.wd3.myworkdayjobs.com/en-US/Novartis_Careers/job/Hyderabad-Office/Associate-Director---AI-Infrastructure-Architect--AI-Factory-_REQ-10088422
5. https://novartis.wd3.myworkdayjobs.com/en-US/Novartis_Careers/job/Hyderabad-Office/Associate-Director---AI-Infrastructure-Architect--AI-Factory-_REQ-10088422