

Productionising Machine Learning (ML) and Generative AI (GenAI) workloads using modern MLOps and LLMOps practices

G-Cloud 15





Table of Contents

1	Service Overview	3
2	Business Need	5
3	Our Approach.....	5
3.1	Experimentation	5
3.2	Reproducibility	6
3.3	Operationalisation.....	6
3.4	Automation	6
4	Buyer Responsibilities	6
5	Service Management	7
6	Protection of Data	7
7	On-boarding and Off-boarding	7
8	Skills and Knowledge Transfer	7
9	Vendor Accreditations/Awards	7
10	Sub-contractors	8
11	Business Continuity and Disaster Recovery	8
12	Pricing	8
13	Ordering and Invoicing	8
14	Termination Terms	8
15	Further Information	9



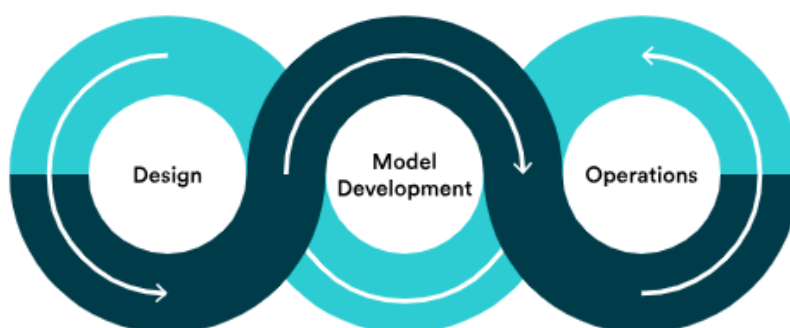
1 Service Overview

Many organisations struggle to extract real value from AI solutions and to integrate them into day-to-day operations. A high proportion of initiatives stall in the Proof of Concept (PoC) phase – the “death valley” between experimentation and production.

MLOps provides the practices, platforms and governance required to move reliably from PoC to production. Built on DevOps principles, it covers the full machine learning lifecycle – from data ingestion and experimentation, through training and evaluation, to deployment, monitoring and continuous improvement. When implemented using cloud-native services such as Azure ML, AWS SageMaker, Google Vertex AI, and open-source tools like MLflow and Kubeflow, MLOps enables faster deployment cycles, scalable workflows and high-quality, well-governed models

The latest **foundation models including Azure OpenAI, AWS Bedrock, and Google Gemini**, Large Language Models (LLMs) are now a core part of digital services. **LLMOps** extends MLOps practices to these foundation models, addressing their specific needs: prompt and system design, retrieval-augmented generation (RAG), safety and guardrails, model evaluation and optimisation, and the orchestration of agentic workflows with support for modern protocols like Model Context Protocol (MCP) and Agent-to-Agent (A2A) communication.

Capgemini combines deep data, AI and cloud engineering expertise with proven accelerators to help public sector bodies industrialise AI across major cloud platforms. Capgemini’s assets and accelerators such as the **Reliable AI Solution Engineering (RAISE)** platform provide modular, battle-tested MLOps and LLMOps blueprints. These are available as reference architectures and implementation templates for Google Cloud, covering data, ML and LLM pipelines, security, monitoring and governance.



Lifecycle of Machine Learning Operations (MLOps)

Capgemini’s services span the full lifecycle of AI productionisation, including :

- **Thought leadership and strategy** – Defining AI and data product strategies, target operating models and platform roadmaps for MLOps and LLMOps, aligned to public sector constraints around cost, security, compliance and sustainability.
- **Pre-built accelerators** – Deployable MLOps / LLMOps foundations, including GitOps-ready CI/CD templates, Terraform-based landing zones, RAG reference architectures, and reusable components for model cataloguing, governance and observability.
- **Model, data and experiment management** – Versioning of datasets, features and traceable lineage across the ML lifecycle; and reproducible experiments that enable collaboration between data science, engineering and operations teams
- **Automated pipelines and orchestration** – End-to-end ML and LLM pipelines built using cloud native CI/CD tools and orchestration platforms, integrating CI/CD/CT for data, models and infrastructure, and leveraging Cloud Composer (managed Airflow) where appropriate.
- **Model deployment and serving** – Secure, scalable deployment of predictive and generative models using cloud native endpoints and container services for real-time workloads, and BigQuery / Dataflow / Vertex AI Batch Predictions for batch inference.

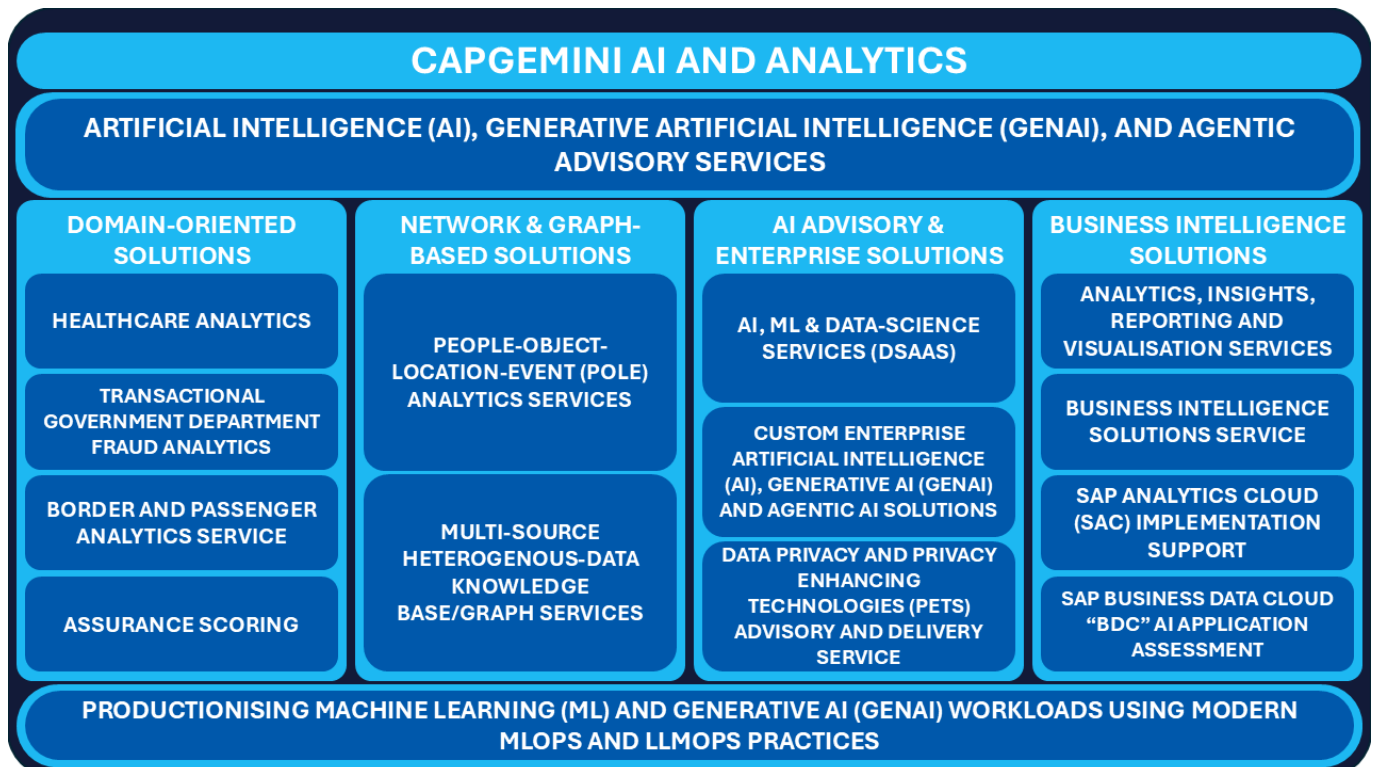


- **Monitoring, drift detection and feedback loops** – Continuous monitoring of data and concept drift, performance, cost and fairness using cloud native monitoring and observability tools; automated re-training triggers; and human-in-the-loop review processes.

Model and AI governance – Operating models, processes and tooling to ensure AI is responsible and compliant, including model catalogues, audit trails, approvals, risk assessments and alignment to public sector standards and “Official” security classification. For LLMOps, Capgemini’s services include:

- **Prompt and system design, fine-tuning and adaptation** of foundation models for specific public sector use cases (e.g. case summarisation, knowledge assistants, citizen-facing virtual agents).
- **Retrieval-Augmented Generation (RAG)** architectures using enterprise data platforms, vector databases, and cloud storage services to ground LLMs in an organisation’s own trusted data, improving accuracy, auditability, privacy and traceability.
- **Agentic AI and workflow orchestration**, using cloud native agent orchestration platforms and frameworks and related services to design multi-step, tool-calling agents that can safely automate complex knowledge and decision-support workflows.
- **Evaluation, safety and guardrails**, including automated and human-in-the-loop evaluation frameworks, safety policies and content filters to manage hallucination, bias, toxicity and data leakage risks in line with government standards.

Capgemini is a strategic partner of Microsoft Azure, AWS and Google Cloud, and has been recognised with multiple partner awards, Capgemini’s global Data & AI practice brings together thousands of specialists, including large numbers of certified consultants with hands-on experience delivering MLOps and LLMOps for major public and private sector clients.



This diagram is for illustration only and does not represent any obligation or responsibility of Capgemini.

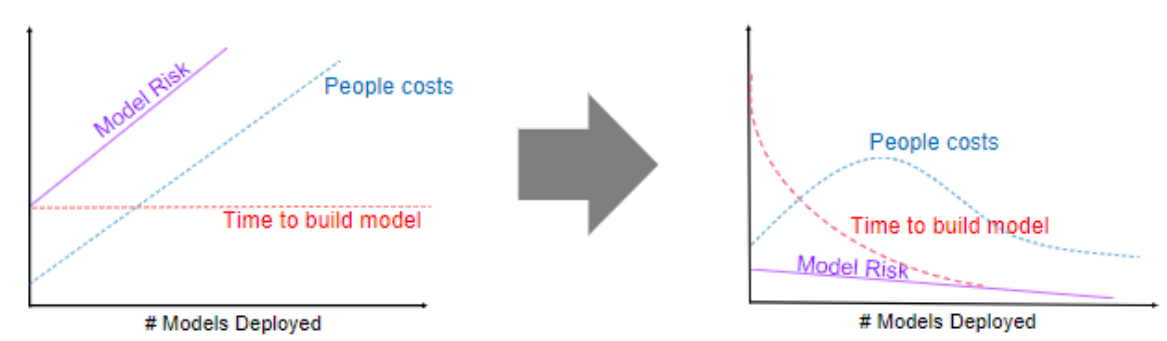
2 Business Need

MLOps is now a critical enabler for organisations seeking to operationalise both predictive models and modern Generative AI solutions at scale. As demand grows for AI-powered data products, the ability to deliver, monitor, and evolve models quickly—while controlling risk and cost—has become a key differentiator. Automation of the ML lifecycle is essential for achieving reliability, reproducibility, and consistent quality across fast-moving AI portfolios.

Modern MLOps streamlines experimentation, training, deployment, and monitoring using unified platforms, reducing manual effort while improving traceability, governance, and explainability. This helps ensure fairness, robustness, and transparency, which are now expected standards for public-facing and internal AI services.

LLMOps extends these principles to foundation models. It addresses challenges unique to large language models: prompt and system design, evaluation of grounded responses, safety and guardrails, and managing rapid changes in model families. Retrieval-Augmented Generation (RAG) and vector search enable organisations to use their private knowledge securely and keep outputs accurate and auditable. Fine-tuning, parameter-efficient adaptation, and model selection must be balanced with cost, privacy, and operational complexity.

Without a mature MLOps and LLMOps capability, investments in data science teams, cloud platforms, and Generative AI technologies cannot be fully realised. An operational AI foundation strengthens collaboration between data scientists, ML engineers, and business stakeholders, ensuring alignment to organisational goals and enabling AI systems that are scalable, governable, and ready for continuous improvement.



Traditional ML practices (left) vs MLOps practices (right)

This diagram is for illustration only and does not represent any obligation or responsibility of Capgemini.

3 Our Approach

Capgemini partners with organisations to select and implement the most appropriate technologies to realise business outcomes from modern MLOps and LLMOps practices. Capgemini's approach builds on deep experience delivering AI for public sector environments and is structured into four delivery phases

3.1 Experimentation

This phase validates the concept and demonstrates value quickly using managed tools. Activities include:

- Requirements and success-criteria definition
- Selection of relevant accelerators and platform components
- Data ingestion, profiling and exploration
- PoC development for ML and LLM use cases (including prompt design or RAG prototypes)
- Findings and recommendations report



3.2 Reproducibility

The goal is to ensure experiments can be repeated reliably and prepared for downstream stages. Activities include:

- Code, data and model versioning
- Dataset and metadata management for lineage and auditability
- Model refinement, evaluation tests, and safety/quality checks (optional)
- Registration of ML and LLM assets for consistent deployment

3.3 Operationalisation

This phase makes the model or Generative AI workflow available to consumers through secure, scalable deployment patterns. Activities include:

- Packaging of ML models or LLM configurations for serving
- Validation, interpretability and safety assessment (where required)
- Deployment to **Vertex AI Endpoints**, **Cloud Run**, or **GKE**
- Integration with RAG infrastructure where needed

3.4 Automation

The final phase establishes a production-grade AI lifecycle with continuous improvement. Activities include:

- Build-out of ML and LLM pipelines
- **CI** – automated testing and integration of model and prompt code
- **CD** – automated deployment to secured production environments
- **CT** – scheduled or event-driven retraining or LLM evaluation workflows
- **CM** – automated monitoring of drift, performance, cost and safety metrics

This framework ensures that both traditional ML and modern Generative AI solutions can be delivered consistently, safely and at scale.

4 Buyer Responsibilities

Please refer to the Supplier Terms listed with this service on the Platform. These may include additional Buyer obligations or costs not explicitly described in this service description.

The Buyer responsibilities for this service are as follows:

- The Buyer must provide an empowered project owner who is available to make timely decisions and act as the main point of contact for the delivery team.
- The Buyer must ensure that any changes to objectives, data access, or scope are documented and communicated promptly to Capgemini.
- The Buyer must provide timely access to relevant data sources, environments, governance requirements, and subject-matter experts needed for AI, MLOps, or LLMOps activities.
- The project owner should attend regular progress and review meetings, including sessions required for model validation, safety assessments, and approval of deployment decisions.

If these responsibilities do not align with the Buyer's expectations, please contact Capgemini so alternative delivery options can be discussed.



5 Service Management

Not applicable.

6 Protection of Data

This service assumes delivery operating at the '**Official**' security classification, with appropriate controls for data access, AI workloads, and model governance. If you require support at a different security classification or need enhanced controls for specific ML or Generative AI use cases, please contact Capgemini to discuss available options.

7 On-boarding and Off-boarding

Prior to execution of the Order, Capgemini and the Buyer will agree the scope and timelines for the on-boarding and exit plan to ensure continuity of service. This will cover requirements for access Cloud environments, data sources, governance processes, and any AI, MLOps or LLMOps assets involved in the delivery.

Capgemini will carry out the agreed on-boarding and off-boarding activities, including the preparation of an exit plan aligned to the Call-Off Contract terms and the handover or decommissioning of relevant AI, data, and pipeline components. These activities will be charged in accordance with the Pricing section for this service.

8 Skills and Knowledge Transfer

Capgemini recognises that effective skills and knowledge transfer is essential to the successful adoption of G-Cloud services, particularly for organisations building capability in AI, MLOps and LLMOps. Knowledge transfer is incorporated into the delivery plan from the outset and tailored to the Buyer's operating model, skills gaps and governance requirements. Our consultants and engineers have extensive experience delivering structured knowledge transfer to public sector teams.

Where appropriate, we apply Capgemini's Assess-Plan-Implement framework to ensure consistency and effectiveness. This approach supports capability-building across the full AI stack, including data engineering, ML/LLM pipelines, model evaluation, and responsible AI practices. It enables teams to gain the skills required to maintain and evolve ML and Generative AI services after the engagement ends, with clear learning objectives, training materials, hands-on sessions and measurable knowledge transfer outcomes.

9 Vendor Accreditations/Awards



For the 13th year in a row, Capgemini has been recognized as one of the World's Most Ethical Companies® by the Ethisphere® Institute. This is an acknowledgment of our ethical culture that makes us an employer of choice and responsible player in the eyes of our clients, shareholders, and the wider community.

Our Data, AI and Analytics practice is consistently recognised by major industry analysts. Capgemini is regularly positioned as a **Leader** in Gartner's evaluations of Data & Analytics service providers, in IDC MarketScape assessments for analytics and AI services, and in Everest Group's PEAK Matrix for Data & Analytics. Our capabilities are also acknowledged in NelsonHall's NEAT analyses across digital, data and regulatory services.



- Diamond Partner with Oracle
- Global Strategic partner Google Cloud
- Gold Partner with Microsoft
- Global Platinum Partner with SAP
- Global Strategic Partner with SAS
- Global Partner with IBM and QlikView
- Data / BI /AI Practice well recognised by industry analysts
 - positioned as a Leader for the third year in a row, in Gartner's Magic Quadrant for Data and Analytics Service Providers, Worldwide
 - named a Leader in IDC's MarketScape Worldwide Business Analytics Consulting and Systems Integration Services 2020 Vendor Assessment
 - positioned as a Leader in Everest Group's PEAK Matrix assessment for Data & Analytics (D&A)
 - identified as a Leader in Nelson Hall's NEAT vendor evaluation for GDPR Services in the overall market segment.

10 Sub-contractors

Capgemini UK may use the following subcontractors to deliver this service:

- Capgemini Technology Services India Limited.

11 Business Continuity and Disaster Recovery

No disaster recovery plan is provided as part of these Services.

12 Pricing

This service is priced in accordance with the SFIA Rate Card attached. Capgemini can also provide offshore resources at reduced rates where appropriate. Projects can be priced either on a Time & Materials or Fixed Price basis.

13 Ordering and Invoicing

Please refer to the Supplier Terms for this service.

We would be pleased to arrange a call or meeting to discuss your requirements of our service in more detail.

14 Termination Terms

Please refer to the Supplier Terms for this service.



15 Further Information

For more information about this or any of our G-Cloud services, please contact our Public Sector Team.

Phone: 0370 904 4858

Email: publicsector.opps.uk@capgemini.com including the following information:

1. The name of this service.
2. The name of your organisation.
3. Your name and contact details.
4. A brief description of your business situation.
5. Your preferred timescales for starting the work.

About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, generative AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2024 global revenues of €22.1 billion.

Make it real. | www.capgemini.com



This presentation contains information that may be privileged or confidential and is the property of the Capgemini Group.

Copyright © 2026 Capgemini. All rights reserved.