

**BLACKSTONE&**

G-Cloud 15 | Lot 3 – Cloud Support Services

AI Factory – AI Solution Build, Delivery & Scale

Service Description

Production AI development service transforming validated prototypes into scalable, enterprise-grade solutions. Following GDS service phases—Alpha, Beta, Live—we architect, build, test, and deploy AI-powered applications including generative AI, large language models, machine learning, and intelligent automation. Delivery ensures secure, resilient, responsible AI solutions with embedded MLOps for continuous improvement.

Service Overview

The AI Factory service bridges the gap between promising AI prototypes and production-grade solutions that deliver value at scale. Many organisations have experienced the frustration of AI pilots that never progress beyond demonstration. Our Factory methodology provides the engineering discipline and delivery practices needed to move from experiment to enterprise.

Building production AI systems is fundamentally different from prototyping. Prototypes can tolerate inconsistency, downtime, and manual intervention; production systems must be reliable, secure, performant, and maintainable. This requires careful attention to architecture, engineering practices, and operational considerations that are often overlooked in the rush from concept to deployment.

Our AI Factory approach applies proven software engineering principles to AI development while accommodating the unique characteristics of AI systems. This includes the iterative nature of model development, the importance of data quality, the need for continuous monitoring, and the challenges of explaining AI decisions.

Key elements of the AI Factory include:

Production Architecture: We design architectures that balance innovation with pragmatism. This includes selecting appropriate cloud services, designing for scalability and resilience, and ensuring security is embedded from the start.

Engineering Excellence: Our delivery teams apply disciplined engineering practices including test-driven development, continuous integration, code review, and documentation. These practices reduce defects, improve maintainability, and accelerate delivery.

MLOps Implementation: AI systems require ongoing attention to model performance, data drift, and retraining. We implement MLOps pipelines that automate these concerns, ensuring AI solutions remain effective over time.

GDS Alignment: We follow GDS service phases (Alpha, Beta, Live) to ensure solutions are developed iteratively with appropriate user research and service assessment preparation.

Responsible AI Governance: We embed responsible AI practices throughout the delivery lifecycle, including bias testing, explainability implementation, and ethical review.

Service Features

- GDS-aligned Alpha, Private Beta, Public Beta, Live delivery
- GenAI and LLM integration: GPT, Claude, Gemini, Llama, Mistral
- RAG architecture, vector databases, embeddings, and knowledge retrieval
- Agentic AI, autonomous workflows, and multi-agent orchestration systems
- MLOps and LLMOps pipelines with CI/CD and model monitoring
- Prompt engineering, fine-tuning, RLHF, and model evaluation frameworks
- Cloud-native deployment on AWS, Azure, and Google Cloud
- Responsible AI governance, explainability, and bias monitoring tooling
- Security-first design aligned to NCSC AI security principles
- AI Factory training building internal AI development capability

Service Benefits

- Proven pathway from validated prototype to production AI
- Reduced delivery risk through phased, iterative development approach
- Enterprise-grade AI built for security, scale, and resilience
- Faster time-to-value leveraging existing prototype investment
- Responsible AI embedded throughout design, build, and deployment
- Seamless integration with existing enterprise systems and data
- Sustainable solutions with MLOps enabling continuous model improvement
- User-centred AI delivering measurable citizen and business outcomes
- Reusable components and patterns accelerating future AI delivery
- Internal delivery capability through AI Factory enablement

Our Approach

AI Factory delivery follows structured phases aligned to GDS service standards:

Alpha Phase: We take validated prototypes and explore how they might be implemented as production solutions. This includes technical architecture design, integration approach validation, and identification of non-functional requirements. Alpha concludes with service assessment and evidence that the solution can be built.

Private Beta: We build a working solution deployed to controlled users. This phase focuses on core functionality, integration with enterprise systems, and validation of the operational model. Continuous user feedback shapes ongoing development.

Public Beta: The solution is deployed to a wider audience while we continue to refine and improve. This phase validates scalability, support processes, and user adoption approaches.

Live: The solution is fully operational with established support and continuous improvement processes. MLOps pipelines ensure model performance is monitored and maintained.

Available through G-Cloud 15 Framework | Lot 3 – Cloud Support Services