

# The GenAI POCs worked... So why is production a nightmare?

Go from sleepless nights to real peace of mind with these tweaks.

Written by



In partnership with



# Why read this guide



The internet and LinkedIn are awash with GenAI advice. This guide focuses on a specific and common hurdle in the journey to get value:

The struggle to effectively convert successful GenAI POCs into production-grade solutions that run smoothly

From spiralling costs to hallucinated results and slow response times, why is it so hard to scale reliable, cost-effective GenAI solutions?

Your GenAI dream doesn't have to become a production nightmare

A few smart tweaks can make a huge difference, so let's look at:

- Root causes
- Straightforward solutions
- Case studies

# Table Of Contents

Discover why this guide is essential, understand the challenges of converting successful GenAI POCs into production-grade solutions, and learn about the importance of Managed AI. Get an overview of what to expect in the book, covering the full lifecycle of managed AI.

## 01. Why GenAI solutions fail in production

Explore the reasons why GenAI solutions often fail in production, including cost uncertainty, inconsistent and unreliable outputs, degrading performance, and slow response times.

## 02. Why these failures happen

Learn why these failures occur and understand the need for Managed AI to cover the AI layers of your application.

## 03. GenAI managed services requirements

Discover the requirements for GenAI managed services, including strategies for cost and performance optimisation, ensuring consistent and reliable outputs, and sustaining performance in line with evolving business needs.

## 04. Managed AI in action

See Managed AI in action through real-world case studies that demonstrate cost control, consistent and reliable outputs, sustained performance and model suitability, fast response times at scale, and effective integration.

## 05. Conclusion

Understand the benefits of adding Managed AI to your managed services stack, including how it helps control costs, ensure consistent and reliable outputs, sustain performance, and maintain seamless integration.

## Chapter 1

# Why GenAI solutions fail in production

### Cost uncertainty

#### How can we predict costs and stick to budgets?

Costs can vary widely based on request number and complexity, task complexity, user load, data quality and quantity, computational power, infrastructure requirements, model choice and the time it takes to fine-tune models for specific needs.

### Inconsistent and unreliable outputs

#### Can you trust what the app says?

Variability in user input leads to variability in model responses. Prompt quality affects response consistency. Response accuracy can vary depending on the model's parameter settings. It could be trained using poor-quality, biased or noisy data. There may not be sufficient or concrete context for the model to draw on. Overly strict safety filters and guardrails may cause cautious or inconsistent responses.

### Degrading performance

#### The app's not exactly future-proof, is it?

Generative models – particularly in fast-moving fields like finance, healthcare and e-commerce – may lose relevance if the context they were trained on changes. Some models may not be updated frequently enough to reflect new data or emerging trends, and without regular fine-tuning, they may be less effective as they become outdated.

### Integration issues

#### Why won't the app work seamlessly as part of workflows?

Failure of third-party APIs, cloud services or data sources can lead to unexpected interruptions as well as low adoption (if the app isn't well integrated into existing tools and systems). Bugs in the underlying software can also cause prediction errors or lead the model to behave unpredictably.

### Slow response times

#### Why can't we get responses fast enough as we scale to more users?

It wasn't necessarily apparent during build or POC, but when you start scaling, the app isn't working at the speed users need. This can be due to throughput limits, token quotas, the time it takes to access large datasets and inefficient resource allocation. How can you optimise workloads to reduce the number of requests, optimise prompts so they use fewer tokens, and configure RAG solutions to return the optimal amount of data?

*"The gap between AI research and AI in production is growing. The challenge is not just technical, but also organizational."*

- Andrew Ng, Founder of Landing AI



## Chapter 2

# Why these failures happen

There's an underlying reason behind these challenges: managed services haven't evolved to cover the AI layers of the app:

- The traditional app managed services stack covers the infrastructure (cloud), app and data pipeline layers.
- With a GenAI app, you need extra management layers covering the AI model(s). This is called Managed AI.

To productise GenAI apps, you need AI model management alongside data, application and cloud managed services. This specific managed services element, Managed AI, focuses on maintaining and optimising GenAI solutions in the production phase – ensuring they perform and scale as needed (while minimising cost uncertainty and proactively driving cost optimisation).



## Chapter 3

# GenAI managed services requirements (above classic managed services)

There's a similar principle behind Managed AI and classic cloud and application managed services: optimise and manage in the background so everyone has peace of mind on performance, security and costs.

However, elements specific to AI workloads need dedicated support to increase POC to production conversion rates and enable effective scaling. When you have these in place, you can eliminate the 5 common failure points we discussed in the previous chapter.

---

### To control costs

#### Cost and performance optimisation

- Data pipelines – Finding opportunities to enhance pipeline throughput and reliability, as well as optimising pipelines to reduce operational costs without compromising performance.
- AI application – Identifying ways to improve responsiveness and accuracy, together with cost-saving opportunities such as reducing token consumption with prompt engineering.
- AI model – Optimising performance through regular evaluations and adjustments, as well as selecting appropriate models that minimise costs while maintaining sufficient levels of accuracy and performance.

---

### To ensure consistent and reliable outputs

#### Ongoing evaluation

- AI application – Ongoing evaluation and regular reviews, looking at elements including accuracy, toxicity, bias, completeness, hallucination, grounding, relevance, context adherence, chunk attribution, chunk utilisation and chunk relevance.
- AI model – Ongoing evaluation and regular assessments to maintain accuracy and relevance.

#### Prompt engineering

- AI application – Designing and refining prompts to ensure the AI model generates relevant, accurate responses that meet application requirements.
- AI model – Adjusting and customising prompts to specific models to optimise their performance, accuracy and consistency in generating desired outputs.

### To ensure sustained performance in line with evolving business needs

#### Business alignment reviews

- AI application – Looking at accuracy and efficiency, with recommendations in areas such as prompt improvements or RAG setting adjustments.
- AI model – Reviewing accuracy and relevance, including highlighting opportunities for model updates or fine-tuning.

#### Model fine-tuning and retraining

- AI model – Optimising and sustaining performance by customising pre-trained models with targeted datasets and regularly updating them with new data. Fine-tuning involves customising pre-trained AI models with specific datasets to optimise performance for specific tasks or domains. Re-training involves regularly updating models with new data to maintain their accuracy and relevance.

---

### To maintain the speed, quality and seamless workflow integration users need

#### Real-time monitoring and automated alerting

- Data pipelines – Detecting and alerting on anomalies and performance issues.
- AI application – Identifying and resolving issues fast.
- AI model – Ensuring sustained accuracy and efficiency.

#### Incident and problem management

- Data pipelines – Proactive detection, resolution and root cause analysis to minimise downtime and data flow disruptions.
- AI application – Addressing issues quickly to ensure consistent performance and availability.
- AI model – Mitigating model-related incidents to maintain accuracy and performance.

#### Service request and change management

- Data pipelines – Addressing user requests for pipeline configuration adjustments and implementing changes to data pipelines with minimal disruption to operations.
- AI application – Managing user requests related to configuration and prompt engineering support, as well as ensuring controlled and documented changes to maintain operational stability.
- AI model – Adjusting AI model settings and configurations, deploying additional AI models and managing model updates/changes so they remain current and effective.



Chapter 4

# Managed AI in action

By integrating Managed AI into your managed services stack, you can detect, resolve and even prevent issues that hold back successful GenAI scaling.

These case studies show you how.

## Cost control

# Spiralling costs for AI-generated reports [Consulting firm]

By integrating Managed AI into your managed services stack, you can detect, resolve and even prevent issues that hold back successful GenAI scaling. These case studies show you how.

### Problem

The firm noticed that the cloud expenses for running the application were soaring, significantly impacting their budget.

### Analysis

We conducted a cost analysis and identified the specific configuration settings contributing to high token usage. We discovered that the RAG solution was retrieving long data chunks and processing more chunks than needed to generate the reports. This over-retrieval resulted in excessive token usage, inflating the costs for cloud infrastructure and AI model APIs.

### Solution

Our engineers analysed the RAG configuration settings and recommended adjustments to reduce chunk sizes and limit the number of chunks retrieved.

### Outcome

Token usage reduced significantly, resulting in lower costs for model API calls. The system now retrieves only the necessary amount of data to generate accurate reports, improving relevance while minimising costs.

**3.5x**

According to a 2024 report by McKinsey, organizations that adopt AI at scale are 3.5 times more likely to report significant improvements in cost optimization.

**65%**

A 2024 survey by Deloitte found that 65% of organizations that have adopted AI at scale report significant reductions in operational costs.

**2.8x**

According to a 2024 report by PwC, organizations that adopt AI at scale are 2.8 times more likely to report significant improvements in cost-effectiveness.

## Consistent and reliable outputs

# Issues with accuracy of call centre transcript analyses [Retailer]

A large retail company uses an AI-driven application to monitor and assess call centre interactions with customers. The application analyses random transcript samples daily to ensure comprehensive evaluation across different shifts and call types. It looks at metrics like customer sentiment, agent response efficiency and compliance with standard procedures.

The company needs to ensure the results are trustworthy and accurate so they can feed into customer service performance reviews, continuous improvement and agent training programmes. Real-time monitoring and automated alerting provide this.

### Metric

Accuracy with which the AI model detects and categorises customer sentiment during calls.

### Automatic alert trigger

Accuracy falls below 90% for more than 10% of analysed calls within a day.

### Resolution process

A ticket is automatically created to track the issue. Relevant team members are assigned to investigate the root cause (like model drift, data inconsistencies or processing delays). The team takes corrective action, which might include switching to a different model, improving the code or adjusting system resources.

**75%**

A 2024 survey by Gartner found that 75% of organizations that have adopted AI at scale report significant improvements in the accuracy of their AI outputs.

**3.2x**

According to a 2024 report by McKinsey, organizations that adopt AI at scale are 3.2 times more likely to report significant improvements in the reliability of their AI models.

**70%**

A 2024 survey by IDC found that 70% of organizations that have adopted AI at scale report significant improvements in the consistency of their AI outputs.

Consistent and reliable outputs

# Customer service chatbot replying with errors [E-commerce]

An e-commerce company uses an AI-powered customer chatbot to help with customer enquiries.

## Problem

Over a few weeks, the chatbot intermittently responded erroneously during conversations, disrupting communications and frustrating users. The monitoring systems detected these issues and automatically flagged them as recurring incidents.

## Analysis

Reviewing the logs identified a pattern correlating errors with high request volumes. Detailed examination of AI model request metrics revealed that the errors predominantly occurred when the number of requests approached or exceeded the quota limits set by the AI model provider. Further analysis determined that quota limits weren't adequately scaled to accommodate peak traffic or the company's growing user base.

## Solution

An internal improvement ticket was created, and a team member was assigned to investigate and propose long-term solutions. Recommendations included increasing quota limits for AI model requests and implementing load balancing strategies to prevent bottlenecks during peak usage. It was ultimately decided to change the AI model configuration to accommodate higher request volumes and optimise resource allocation.

## Outcome

The issue of random chat errors due to quota limits was resolved, making the chatbot more stable and reliable.

**2.7x**

According to a 2024 report by PwC, organizations that adopt AI at scale are 2.7 times more likely to report significant improvements in the quality of their AI outputs.

**75%**

A 2023 survey by Gartner found that 75% of organizations that have adopted AI at scale report significant improvements in customer satisfaction.

**2.5x**

According to a 2023 report by PwC, organizations that adopt AI at scale are 2.5 times more likely to report significant improvements in the integration of their AI models into broader business processes and workflows.

**80%**

A 2024 survey by Deloitte found that 80% of organizations that have adopted AI at scale report significant improvements in the predictability of their AI outputs.

Consistent and reliable outputs

# Diagnostic app was returning different results for similar symptoms [Healthcare]

A healthcare provider used an AI-powered application to help medical professionals diagnose conditions based on the symptoms patients described during consultations. The application relied heavily on prompts to generate accurate and contextually relevant diagnostic suggestions.

## Problem

Clinicians reported instances where the AI application provided inconsistent diagnostic suggestions for similar patient symptoms. For example, when 2 patients described nearly identical symptoms, the model gave significantly different responses, some of which contained irrelevant or incorrect diagnostic information.

## Analysis

Monitoring and logs revealed variability in the AI model's responses to similar prompts. Our prompt engineering team discovered that the prompts lacked specificity and didn't capture the context needed for accurate diagnosis.

## Solution

Our experts created more specific and context-rich prompts by incorporating domain-specific language and detailed symptom descriptions. The newly designed prompts were subjected to rigorous testing with various patient symptom scenarios to evaluate their effectiveness and consistency. Based on test results, prompts were further refined to enhance their specificity and contextual understanding. This iterative testing process ensured the prompts consistently yielded accurate and reliable suggestions. Optimised prompts were deployed in the AI application. Ongoing evaluation monitors their performance and consistency.

## Outcome

The issue was resolved. Medical professionals now receive accurate and contextually relevant responses from the AI application, improving their efficiency and supporting better patient care.

**85%**

A 2024 survey by Gartner found that 85% of organizations that have adopted AI at scale report significant improvements in the maintainability of their AI models.

**3.3x**

According to a 2024 report by PwC, organizations that adopt AI at scale are 3.3 times more likely to report significant improvements in the interpretability of their AI models.

**75%**

A 2024 survey by IDC found that 75% of organizations that have adopted AI at scale report significant improvements in the transparency of their AI models.

Sustained performance and model suitability

# Providing accurate insurance premiums despite market changes [Insurance company]

An insurance company uses an AI-powered application to assess potential customers' risk. The application evaluates factors such as age, employment history, health records and previous claims to determine the risk level and generate a tailored premium.

## Problem

The AI model was showing signs of performance degradation.

## Objectives

We helped them define objectives for the model, aiming to improve risk assessment accuracy and relevance, particularly in evaluating health records and employment history.

## Dataset creation

We selected a pre-trained model known for its effectiveness in risk assessment tasks. Then, we created a dataset by compiling recent application data, including updated health statistics, employment trends and claims records.

## Training

The model was trained on this tailored dataset. Rigorous evaluation and testing were conducted to validate the model's accuracy, consistency and reliability in assessing various risk scenarios.

## Deployment

Once the fine-tuned model achieved the desired accuracy, it was deployed in the AI application. Ongoing monitoring tracks the model's performance and ensures consistency and effectiveness.

## Retraining

As time passes, societal trends, health data and employment statistics evolve. Therefore, the model requires periodic retraining to maintain its accuracy and relevance. We regularly update the dataset with new application data and relevant statistics, which are used to retrain the model.

## Outcome

By fine-tuning and periodically retraining the AI model, the insurance company can consistently assess risk accurately as it adapts swiftly to market changes.

Fast response times at scale

# Reports were taking too long to generate [Marketing agency]

A marketing agency uses an AI-powered application to create campaign reports for clients. A user noticed that the application was taking longer than expected to respond when asked to generate a report.

There is ongoing tracking of response times for each query the chatbot handles. This ensured the latency issue was detected and addressed early, limiting the impact and preventing more users from being affected.

**3.8x**

According to a 2024 report by McKinsey, organizations that adopt AI at scale are 3.8 times more likely to report significant improvements in response times.

**82%**

A 2024 survey by Gartner found that 82% of organizations that have adopted AI at scale report significant improvements in the speed of their AI models.

**2.9x**

According to a 2024 report by PwC, organizations that adopt AI at scale are 2.9 times more likely to report significant improvements in the latency of their AI models.

## Metric

AI model response time when generating replies to user queries.

## Automatic alert trigger

Latency exceeds 5 seconds for more than 5% of queries within 1 hour.

## Resolution process

A ticket was automatically created to track the issue. The incident was investigated to identify the root cause of the latency (resource constraints, network issues). Appropriate team members were assigned to adjust infrastructure to reduce the latency in line with the target threshold.

Effective integration

# Upgrading a preventative maintenance app with new sensor data [Manufacturing]

A manufacturing company relies on an AI-powered application to predict equipment failures and proactively schedule maintenance tasks. The application processes data from IoT sensors attached to machinery, analysing performance metrics and identifying signs of wear and tear.

## Problem

The company upgraded its IoT sensors to capture more detailed and higher frequency data. This required changes to the data ingestion pipeline, AI model configurations and cloud infrastructure.

## Analysis

We assessed the impact of integrating the new IoT sensors on the AI application. We outlined the necessary modifications to data pipelines, model configurations and cloud resources, complete with effort estimates.

## Solution

Data pipeline modifications were implemented, enabling the ingestion and processing of higher frequency data from the upgraded sensors. AI model configurations were adjusted to handle the increased data volume and complexity more efficiently. Cloud infrastructure was scaled to support the enhanced data processing requirements. Automated monitoring and alerting systems were updated to ensure seamless detection of any anomalies during the transition. Rigorous testing was conducted to validate the changes, ensuring the updated AI application functioned correctly with the new sensors. Performance metrics were evaluated to confirm improvements in efficiency and accuracy. After successful testing and validation, the changes were deployed during the scheduled window. Ongoing monitoring checks the application's performance to rapidly address any issues if they arise.

## Outcome

The structured change management process ensured minimal disruption. The AI application has adapted to the upgraded IoT sensors, enhancing predictive maintenance.

### 4.5x

According to a 2024 report by McKinsey, organizations that adopt AI at scale are 4.5 times more likely to report significant improvements in the integration of their AI models into broader business processes and workflows.

### 83%

A 2024 survey by Gartner found that 83% of organizations that have adopted AI at scale report significant improvements in the interoperability of their AI models with other systems and applications.

### 3.1x

According to a 2024 report by PwC, organizations that adopt AI at scale are 3.1 times more likely to report significant improvements in the alignment of their AI models with business objectives and goals.



## Conclusion

# Go from sleepless nights to peace of mind

We've seen how adding Managed AI to your managed services stack helps address key challenges with scaling GenAI POCs into production-ready solutions.

So what exactly should you be aware of to escape GenAI production nightmares and achieve that vital peace of mind?

### Remember the root cause of your production nightmares

GenAI applications behave differently to traditional applications. If you don't have AI model management alongside data, application and cloud managed services, you run into a host of cost, performance and scaling issues.

### Managed AI takes care of that root cause

Managed AI maintains and optimises GenAI solutions in the production phase, helping you increase POC to production conversion rates and:

- Control costs
- Ensure consistent and reliable outputs
- Sustain performance in line with evolving business needs
- Maintain the speed, quality and seamless workflow integration users need

### What capabilities do you need?

At a minimum, look for these:

- Cost optimisation advisory
- Monitoring and alerting
- Incident management
- Service request management
- Ongoing evaluation

If you have more complex applications or are at a point where AI is starting to be an integrated part of at least one core business process/function, it's also useful to have:

- Performance optimisation
- Problem management
- Prompt engineering
- Quarterly reviews

### Don't let a simple managed services gap lead to so many problems

Adding Managed AI to your stack is easy – and crucial to realising investment from your GenAI investment and ensuring your AI maturity doesn't stall unnecessarily.

*"The key to successful AI is not just the technology, but the people and processes that support it."*

- Fei-Fei Li, Co-Director of the Stanford Institute for Human-Centered AI



## What GenAI app nightmares keep you up at night?

Let's chat about how we can help solve them with streamlined, effective management of your AI workloads. Our Managed AI experts help you save money, future-proof solutions and give you peace of mind on security, performance and governance.

[Contact us](#)

### About Firemind

Firemind Group helps organisations accelerate the path to GenAI maturity, providing the consultancy, software and services needed to go from proof-of-concept to production – scaling GenAI to create new value.

We're at the cutting edge of the emerging managed AI space and recognised as one of the most accredited AWS partners in the AWS Partner Network. We're an AWS Premier Tier Services Partner, won AWS Rising Star Partner of the Year 2023, were an AWS GenAI Tools Partner of the Year 2024 finalist and have helped more than 200 organisations launch successful GenAI projects in the cloud.