

CONFIDENTIAL



Executive committee extract

AI cost and reliability discussion – 3 March 2026

Attendees: CEO, CTO, CFO, COO, Head of infrastructure, Chief risk officer

CTO

Stratum Retail runs steady AI workloads on owned infrastructure. That gives baseline cost control.

For seasonal demand and new features, we burst to cloud.

There is no formal policy defining when a workload must use owned capacity and when cloud is acceptable.

Teams decide case by case.

CFO

AI-related cloud spend increased 34 per cent last quarter.

We cannot clearly attribute that increase to specific features.

Cost is tracked by environment. It is not tracked by AI service or product capability.

Without feature-level visibility, we cannot measure margin impact.

Head of infrastructure

Owned GPU clusters are near steady utilisation.

Batch scheduling is informal. Some teams prioritise speed over efficiency.

There is no documented lifecycle for AI endpoints once deployed. Some remain active without review.

COO

During peak events, performance matters more than experimentation.

We do not have defined cost ceilings for AI auto-scaling during peak windows.

There is no clear rule for when AI services should degrade or pause under load.

Chief risk officer

Data transfer between on-premise systems and cloud is increasing.
Large transfers and third-party model calls are not centrally tracked.
This is both a cost and compliance exposure.

CEO

Stratum Retail cannot afford uncontrolled variability.

AI must improve margin, not threaten it.

We need:

- clear lifecycle ownership
- defined workload placement rules
- peak protection guardrails
- visibility of AI cost at feature level