April 27, 2026

Mixed observability for GenAI

OTel substrate + GenAI layer

Navigation Roadmap

What this article covers

Scheme → SOTA stack

- Parts 1-2: scheme, architecture.
- Parts 3-4: GenAI traces, OTel/Langfuse.
- Parts 5-6: market and maturity gaps.
- Part 7: SOTA stack and verdict.

Part 1

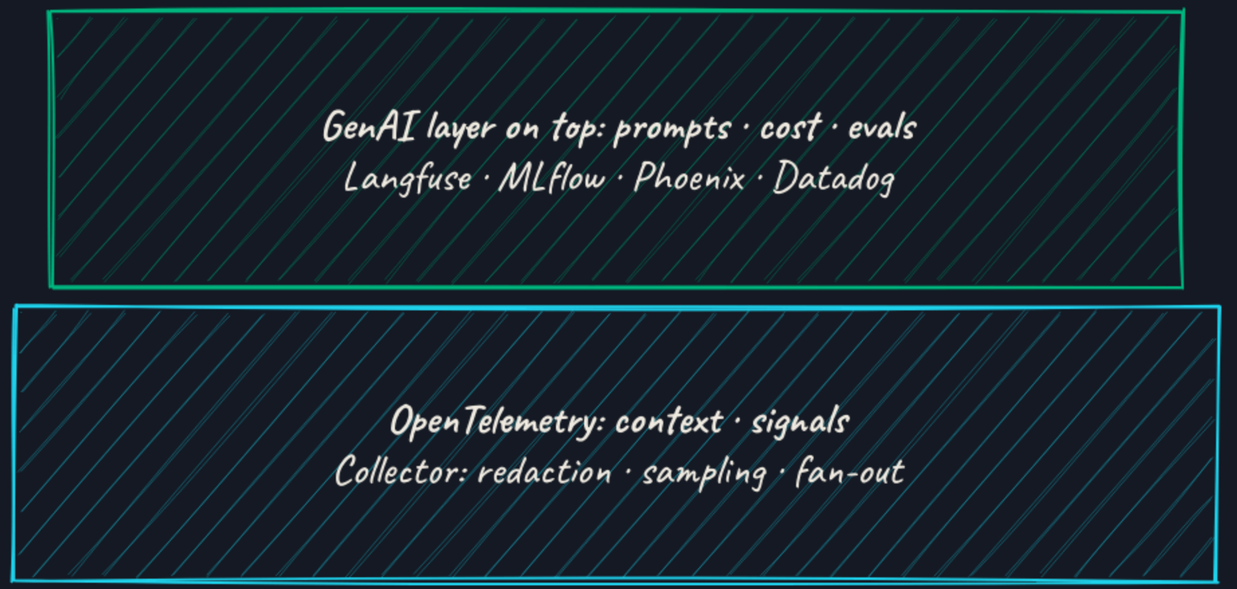
01

Two-layer observability scheme

OTel as the nervous system + specialized GenAI layer on top

OTel substrate + GenAI layer

Two layers Architecture



Part 2

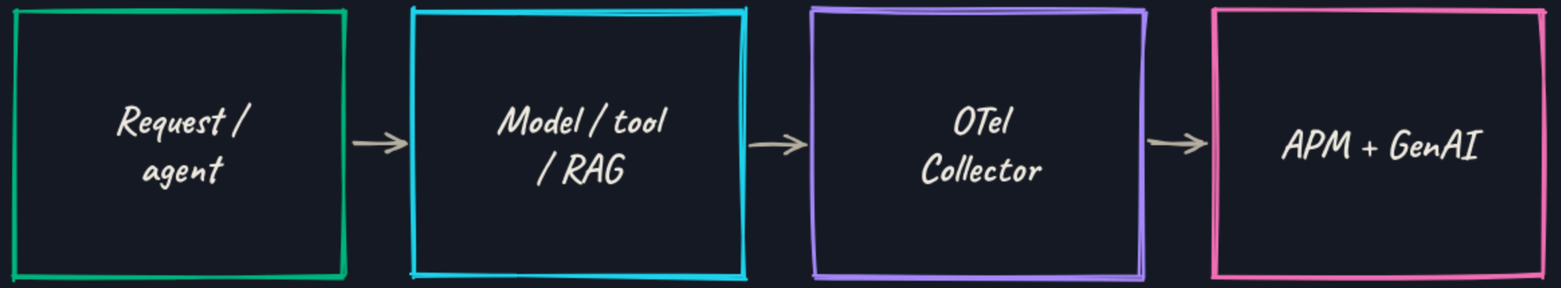
02

Reference architecture

What the chain from user request to AI backend looks like

From request to backend

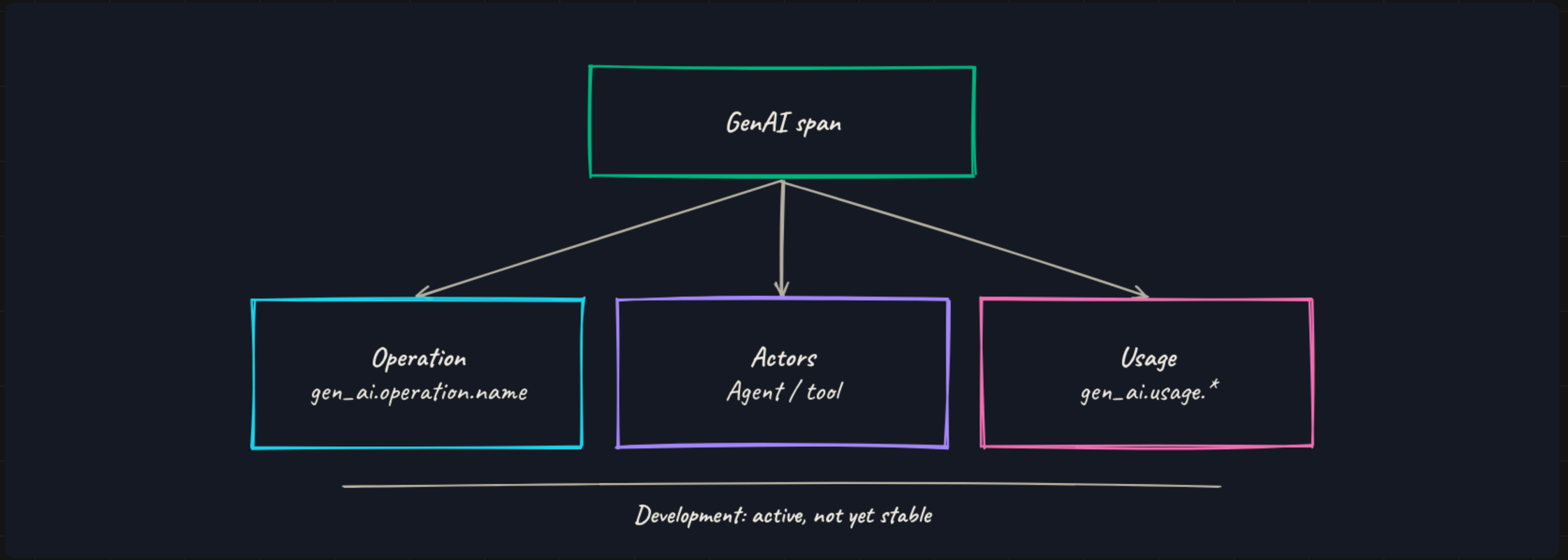
Pipeline Reference



Collector: redaction, sampling, fan-out; context is preserved

OTel semantic conventions for GenAI

Semconv GenAI spans



Part 3

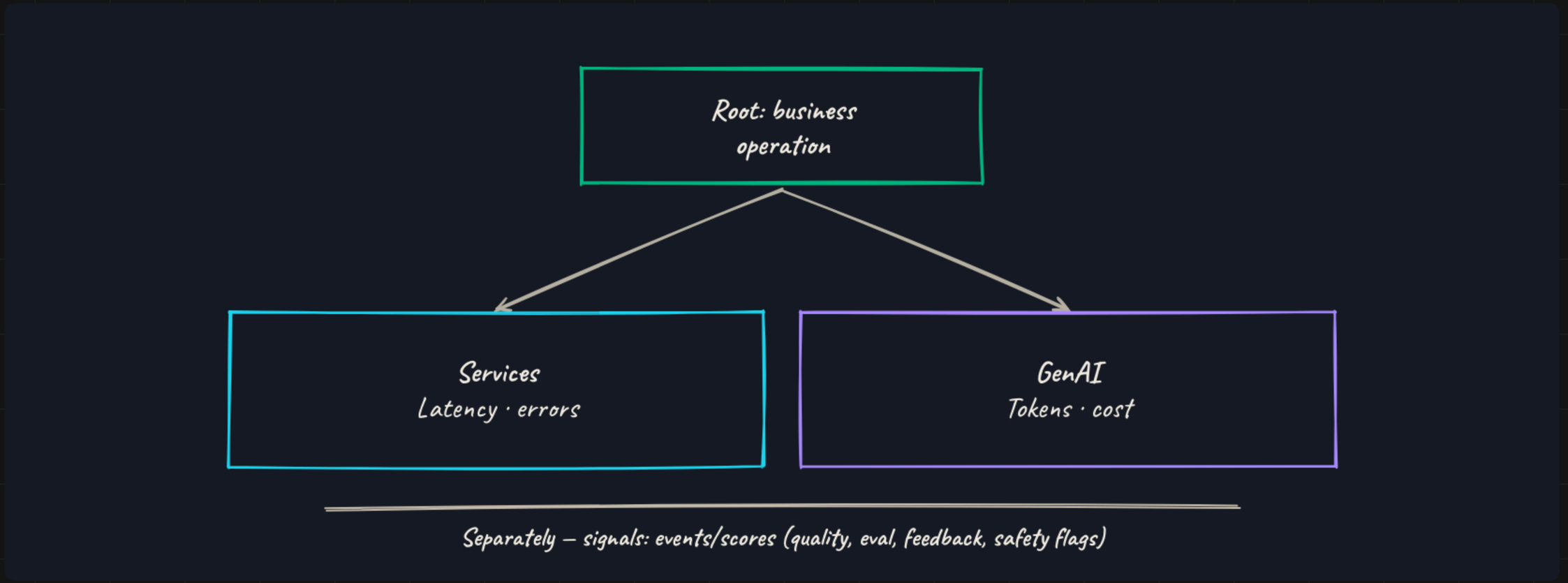
03

GenAI trace best practices

Decision episode: from user request to final response

Structure of a good trace

Decision episode | Best practice



Part 4

04

OTel/Langfuse: how they coexist

OTel-first, Langfuse-as-AI-backend — the most mature pattern

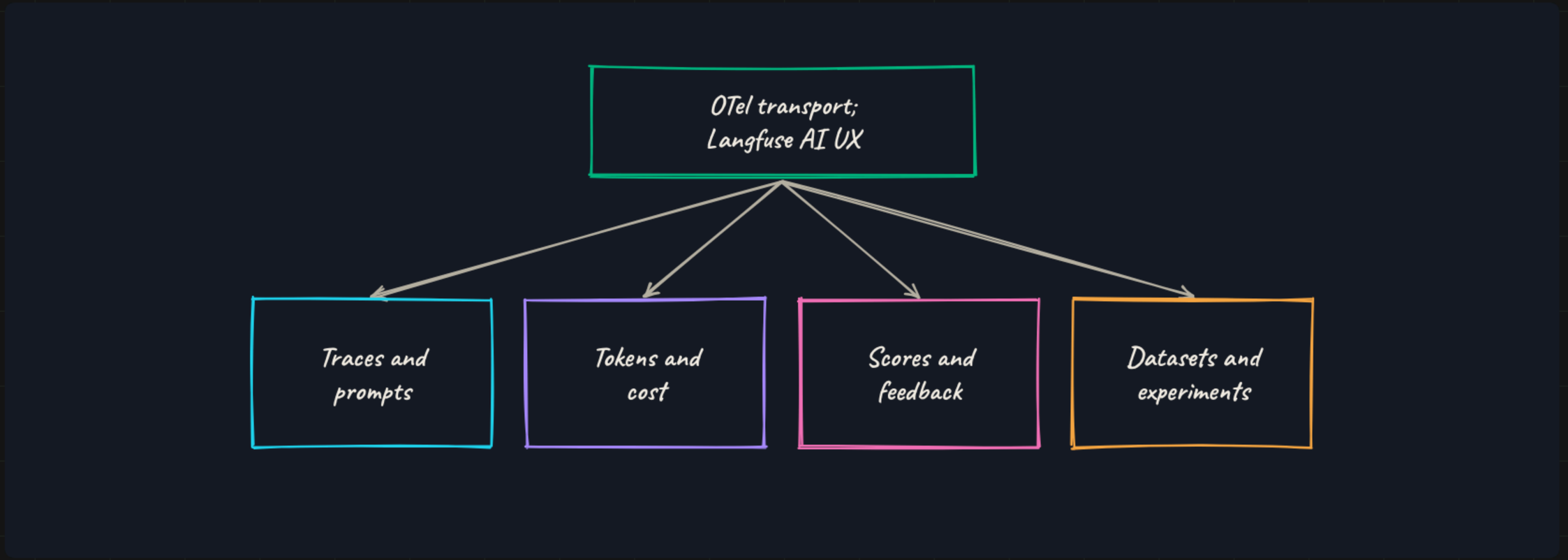
Two integration models

Integration Two models



Langfuse over raw OTel

Added value Langfuse



Part 5

05

Four market approaches

OTel-first, APM-first, AI-platform-first, Auto-instrumentation

4 architectural approaches



Part 6

06

Maturity assessment: 7 pain points

Honest assessment — what's covered, what's partial, what still hurts

What already works well

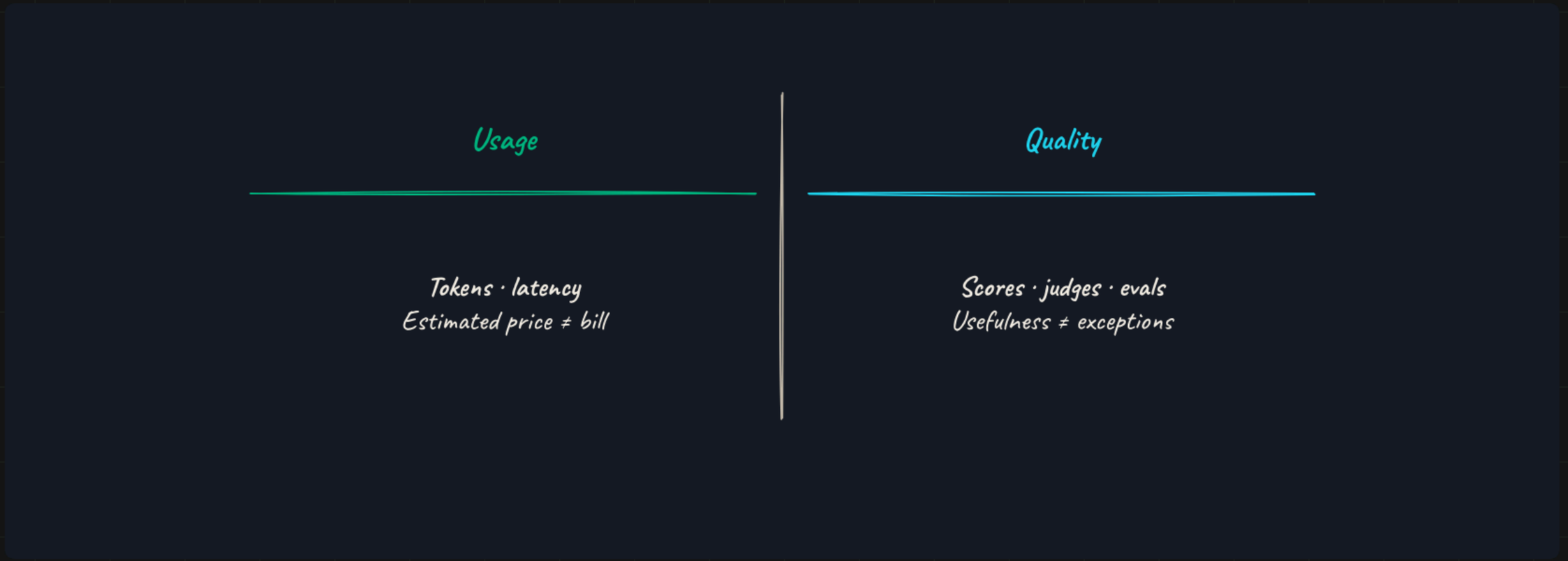
Covered ✓ Good



A shared trace/span ID connects the signals

Token/cost and quality

Partial ~ Partial



What still hurts

Pain ✗ Pain

Privacy	Scale	Filtering	Coexistence
Prompts/outputs: allowlists, redaction	Heavier traces, complex tail sampling	Dropped spans break trace integrity	TracerProvider conflicts between OTel tools

Part 7

07

SOTA stack and verdict

Four principles for the right stack for a mixed prod system

The right SOTA stack

SOTA 4 principles



Verdict

The industry matured, but not fully

- SOTA = OTel protocol/context + AI interpretation/evals.
- End-to-end mixed observability is more standard.
- Quality and privacy/governance remain open zones.

- ↗ tellmeabout.tech · ↗ ai4sdlc-research.space

References and Materials

Sources References

OTel, GenAI telemetry, evals

- OpenTelemetry documentation and GenAI semantic conventions.
- Langfuse, MLflow, Phoenix and OpenInference tracing/evaluation docs.
- Datadog LLM Observability; Azure AI Foundry.

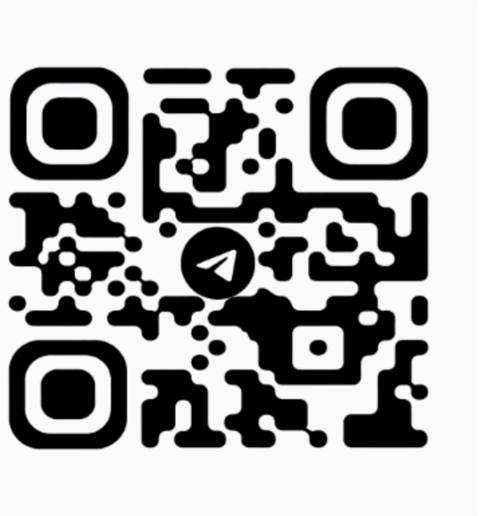
Mixed Observability for GenAI

Thank you!

Materials and links are in the ”Book Cube” channel

Alexander Polomodov

Technical Director & Fellow,
T-Technologies



@Book_Cube