GenAI Safety Telemetry Standards

OpenTelemetry Semantic Conventions Proposal

THE PROBLEM

EU AI Act (Aug 2026) requires automatic logging of AI safety system events. Compliance advisors are telling enterprises to capture "guardrail events" and "blocked/approved cases." **But there is no standard for providers to emit this telemetry.**

OTel GenAI Has:

- Token counts
- Model versions
- Request parameters
- Tool executions

OTel GenAI Missing:

- Safety evaluation signals
- Response modification flags
- Confidence/uncertainty scores
- Guardrail activation events

PROPOSED ATTRIBUTES

gen_ai.safety.evaluation_performed
Safety system processed this request

gen_ai.response.modified
Response was altered post-generation

gen_ai.response.modification_type
safety_filter | pii_redaction | truncation

gen_ai.confidence.score
Provider confidence (0.0-1.0) + method

WHY PROVIDERS SHOULD CARE

Legal Protection	**Enterprise Sales**	**Market Position**
Telemetry is evidence. "We have safety systems" is an assertion. Logs proving safety systems ran on every request is a defense.	"How do we audit your AI?" Currently: "You can't." With this: "OTel-standard telemetry in your existing stack."	First mover defines "responsible AI." Everyone else answers: "Why doesn't your AI have safety telemetry?"

DESIGN PRINCIPLES

- Signal presence, not logic — "evaluation happened" not "why it decided"
- Opaque identifiers OK — audit correlation without exposing internals
- No threshold exposure — "modified" without revealing trigger criteria
- Provider discretion — all attributes Recommended or Opt-In

REGULATORY TIMELINE

Feb 2025: Prohibited AI practices banned

Aug 2025: GPAI model rules active

Aug 2026: Full enforcement (logging required)

Penalty: €35M or 7% global revenue

The question isn't "why would you do this?"

The question is "why wouldn't you, unless you have something to hide?"

Norman Todd • infinitum nihil • github.com/NTinfinitumnihil

Full proposal with technical specification available upon request