GenAI Safety Telemetry Standards

A Proposal for OpenTelemetry Semantic Conventions

Enabling Compliance, Trust, and Human Oversight at Scale

Prepared by:

Norman Todd

infinitum nihil

January 28, 2026

Version 1.0 - Public Draft

Executive Summary

The EU AI Act (Article 12) requires high-risk AI systems to maintain automatic logging of events throughout their lifecycle. Article 14 mandates human oversight capabilities. Compliance advisors are already instructing enterprises to capture "guardrail events" and demonstrate "blocked and approved cases" of output moderation.

There is no standard for AI providers to emit this telemetry.

The OpenTelemetry GenAI semantic conventions (currently in development status) comprehensively cover operational telemetry—token counts, latency, model versions, tool executions—but contain no provisions for safety system engagement signals.

This creates an untenable situation: enterprises are legally required to prove safety system coverage using data that providers do not emit and for which no standard exists.

This proposal defines semantic conventions for GenAI safety and confidence telemetry, enabling:

- Regulatory compliance with EU AI Act logging requirements
- Enterprise audit trails for safety system engagement
- Human-in-the-loop trust calibration
- Provider differentiation through transparency
- Interoperability across multi-provider agent architectures

First-mover advantage accrues to the provider who implements this standard. Liability accrues to those who do not.

The Regulatory Mandate

EU AI Act Requirements

The EU AI Act establishes specific logging requirements for high-risk AI systems:

Article 12 (Record-keeping) [1]:

> *"High-risk AI systems shall technically allow for the automatic recording of events (logs) over the lifetime of the system."*

Logs must enable: (a) identifying situations that may result in risk, (b) facilitating post-market monitoring, and (c) monitoring the operation of high-risk AI systems.

Article 14 (Human Oversight) [1]:

> *"High-risk AI systems shall be designed and developed in such a way...that they can be effectively overseen by natural persons."*

Enforcement Timeline

Date	**Requirement**
Feb 2, 2025	Prohibited AI practices banned
Aug 2, 2025	General-purpose AI model rules take effect
Aug 2, 2026	**Full enforcement begins - most obligations including logging**
Aug 2, 2027	High-risk AI in regulated products deadline

Penalties [6]: Up to €35 million or 7% of global annual turnover, whichever is higher.

What Compliance Advisors Are Telling Enterprises

Industry guidance already assumes this telemetry should exist:

> *"Evidence pack: capture prompts, model/router versions, human-in-the-loop actions, and guardrail events to feed technical documentation and post-market monitoring."*
>
> — ABV EU AI Act Compliance Checklist [4]
>
> *"Safety: Demonstrate both blocked and approved cases of output moderation or escalation."* [5]
>
> — Sombra AI Regulations Guide 2026
>
> *"In the 2026 compliance environment, screenshots and declarations are no longer sufficient—only operational evidence counts."* [5]
>
> — Sombra AI Regulations Guide 2026

The Gap: What Exists vs. What's Needed

Current OTel GenAI Semantic Conventions

The OpenTelemetry GenAI semantic conventions [2][3] (Development status) provide comprehensive operational telemetry:

- Token counts (input, output, cached)

- Model identifiers and versions
- Request parameters (temperature, top_p, penalties)
- Tool definitions and executions
- Conversation threading
- Error types and finish reasons
- Input/output message content (opt-in)

What's Missing

Zero provisions for:

- Safety classifier evaluation signals
- Response modification/steering events
- Guardrail activation indicators
- Confidence/uncertainty signals
- Regeneration events
- Content filtering actions

> **The industry standard for GenAI observability doesn't have a vocabulary for safety system engagement. The spec doesn't contemplate it.**

Proposed Specification

GenAI Safety and Confidence Attributes

Status: Proposal (Development)

These attributes provide observability into AI safety system interactions and model confidence signals, enabling appropriate trust calibration, compliance auditing, and human oversight of autonomous AI systems.

These attributes do not expose safety system logic, thresholds, or decision criteria. They signal that an evaluation occurred, not why it reached a particular conclusion.

Attribute	Type	Description
gen_ai.safety.evaluation_performed	boolean	Safety evaluation was performed on this request/response
gen_ai.safety.evaluation_ids	string[]	Identifiers for safety evaluations performed (opaque)
gen_ai.response.modified	boolean	Response was modified after initial generation
gen_ai.response.modification_type	string	Category: safety_filter, pii_redaction, truncation, etc.

gen_ai.response.generation_attempts	int	Number of generation attempts before final response
gen_ai.confidence.score	float	Provider-determined confidence score (0.0-1.0)
gen_ai.confidence.method	string	Method: logprob_derived, self_evaluation, ensemble, etc.
gen_ai.confidence.abstention_recommended	boolean	Provider signal that human review may be appropriate

Design Principles

1. Signal Presence, Not Logic

Attributes indicate *that* evaluation occurred, not *how* decisions were made. This preserves safety system integrity while enabling compliance.

2. Opaque Identifiers Permitted

evaluation_ids may be opaque strings that support audit correlation without exposing internal system names or architectures.

3. Provider Discretion on Confidence

Confidence scores are provider-computed using provider-determined methods. The specification requires disclosure of method, not standardization of computation.

4. No Threshold Exposure

The modified flag signals outcome without revealing what triggered modification. Providers retain complete control over their safety criteria.

The Strategic Case for Providers

Telemetry Is Evidence, Not Exposure

When something goes wrong with AI—a harmful output, a manipulated system, an unexpected behavior—the provider's current position is: "We have safety systems."

That's not a defense. That's an assertion.

When the regulatory inquiry comes:

> **Regulator:** "Can you prove your safety systems were active when this output was generated?"
>
> **Provider:** "We have safety systems."
>
> **Regulator:** "But were they running? On this specific request? Show me the logs."

Providers who emit gen_ai.safety.evaluation_performed: true on every request have proof they did their job. Providers who don't have assertions.

First Mover Advantage

The provider who adopts this first gets to say:

> *"We're so confident in our safety systems that we instrument them for external verification. Our competitors are still asking you to trust them."*

That's not defensive. That's a competitive position.

Enterprise Sales Enablement

Fortune 500 compliance officers are asking: "How do we audit this AI system?"

Right now the answer is: "You can't."

The provider who can say "We emit OTel-standard safety telemetry that integrates with your existing observability stack" wins those contracts.

Regulatory Arbitrage

EU AI Act requires "human oversight" for high-risk systems. What does that mean in practice? The interpretation is still being written.

The provider who defines what auditable AI looks like—by shipping the telemetry standard—shapes that interpretation. They're not responding to regulation. They're writing the playbook that becomes regulation.

Use Cases

Compliance Auditing

> **Query:** Show all responses where safety evaluation was performed but response was not modified
>
> **Purpose:** Validate safety system coverage without false positive burden

Human Review Routing

> if **gen_ai.response.modified == true** OR **gen_ai.confidence.abstention_recommended == true**:
>
> route_to_human_review()

Agent Debugging

> **Query:** Trace where gen_ai.response.generation_attempts > 2
>
> **Purpose:** Identify prompts causing generation difficulty

Trust Calibration

```
if gen_ai.confidence.score < 0.7 AND gen_ai.confidence.method == "ensemble":
display_uncertainty_indicator_to_user()
```

Multi-Provider Agent Observability

When agents span multiple providers (OpenAI tool calls Anthropic which queries Bedrock), a response that was "steered" at one layer affects downstream reasoning. Without standard signals, multi-provider agent debugging is impossible.

Privacy and Security Considerations

These attributes are designed to support observability without compromising safety system integrity:

1. **No logic exposure:** Attributes indicate that evaluation occurred, not how decisions were made
2. **No threshold exposure:** Pass/fail states are not directly indicated
3. **Opaque identifiers permitted:** evaluation_ids may be opaque, supporting audit correlation without exposing system internals
4. **Provider discretion:** All safety-related attributes are Recommended or Opt-In, allowing providers to implement based on their security posture

The Alternative: Adversarial Workarounds

Without standard telemetry, the industry will build workarounds:

- Prompt injection detection
- Output diffing against expected responses
- Behavioral fingerprinting
- Cross-model comparison to detect guardrail activation

These approaches are adversarial by nature. They treat providers as opaque systems to be probed and reverse-engineered.

A standard telemetry signal is cooperative. It aligns provider and enterprise incentives around transparency and trust.

Call to Action

For OpenTelemetry Working Group

Consider this proposal for inclusion in the GenAI semantic conventions. The regulatory deadline of August 2026 creates urgency for standardization.

For AI Providers

First mover advantage is real. The provider who ships this telemetry first defines what "responsible AI" looks like in practice. Everyone else plays catch-up—and faces the question: "Why doesn't your AI have safety telemetry?"

For Enterprises

Ask your AI providers for this telemetry. Include it in RFPs. Make it a procurement requirement. Market demand accelerates standardization.

Conclusion

The demand for AI safety telemetry exists. The regulatory requirement is in law. Compliance advisors are already telling enterprises to capture data that providers don't emit.

This proposal closes the gap between what regulators require and what the industry can deliver.

We're not asking providers to expose safety system logic. We're asking them to prove their safety systems exist—at scale, in a way that protects them legally, wins enterprise deals, and positions them as responsible AI leaders.

> ***The question isn't "why would you do this?" The question is "why wouldn't you, unless you have something to hide?"***

Contact

Norman Todd

infinitum nihil

github.com/NTinfinitumnihil

Appendix A: Full Attribute Specification

The following specification follows OpenTelemetry semantic conventions format and is intended for inclusion in the GenAI semantic conventions.

gen_ai.safety.evaluation_performed

Type	boolean
Requirement	Recommended
Description	Indicates that one or more safety evaluation systems processed this request or response.
Example	true

Rationale: Enables operators to distinguish between environments/requests where safety systems are active versus inactive, supporting compliance documentation and coverage analysis.

gen_ai.safety.evaluation_ids

Type	string[]
Requirement	Opt-In

Description	Provider-defined identifiers for safety evaluations performed. Identifiers SHOULD be stable across requests but need not expose internal system names. Providers MAY use opaque identifiers.
Example	["content_policy_v2", "pii_detection"]

Rationale: Supports debugging and audit trails when operators need to correlate behaviors across requests. Does not indicate pass/fail status.

gen_ai.response.modified

Type	boolean
Requirement	Recommended
Description	Indicates the final response differs from initial generation due to post-generation processing. This includes safety filtering, content adjustment, or policy application.
Example	true

Rationale: Critical for trust calibration. A modified response may warrant different downstream handling than an unmodified response. Enables human-in-the-loop systems to allocate review attention appropriately.

gen_ai.response.modification_type

Type	string
Requirement	Conditionally Required if gen_ai.response.modified is true
Description	Categorizes the type of modification applied to the response.

Well-known values:

Value	**Description**
safety_filter	Content modified due to safety policy
format_adjustment	Structure/format changed (not content)
truncation	Response truncated (length, time, tokens)
pii_redaction	Personally identifiable information removed
citation_injection	Citations or attributions added
_OTHER	Other modification type

Rationale: Different modification types have different implications for downstream use. A truncated response may be continued; a safety-filtered response should not be prompt-engineered around.

gen_ai.response.generation_attempts

Type	int
Requirement	Opt-In
Description	Count of generation attempts before producing final response. A value greater than 1 indicates regeneration occurred.
Example	3

Rationale: Multiple generation attempts may indicate the model encountered difficulty with the request. Useful for debugging, cost attribution, and latency analysis.

gen_ai.confidence.score

Type	float
Requirement	Opt-In
Description	Provider-computed confidence score representing the provider's assessment of response reliability. Scale of 0.0 (no confidence) to 1.0 (full confidence). Providers SHOULD document their confidence computation methodology.
Example	0.85

Rationale: Enables downstream systems to implement confidence-based routing, human review thresholds, and appropriate uncertainty communication to end users.

gen_ai.confidence.method

Type	string
Requirement	Conditionally Required if gen_ai.confidence.score is provided
Description	Documents the method used to compute gen_ai.confidence.score.

Well-known values:

Value	**Description**
logprob_derived	Derived from token log probabilities
self_evaluation	Model self-assessment
ensemble	Agreement across multiple generations
classifier	External classifier assessment

calibrated_hybrid	Combination of methods with calibration
_OTHER	Other method

Rationale: Confidence scores computed via different methods have different reliability characteristics. Downstream systems may weight or interpret scores differently based on method.

gen_ai.confidence.abstention_recommended

Type	boolean
Requirement	Opt-In
Description	Provider signal indicating this response may benefit from human review, additional verification, or abstention from autonomous action. This is a provider recommendation, not a mandate.
Example	true

Rationale: Enables providers to communicate "soft" uncertainty that may not be captured in a numeric score. Supports human-in-the-loop architectures without requiring providers to expose detailed reasoning.

Appendix B: Open Questions for Working Group

5. Should **gen_ai.response.modified** distinguish between "modified and served" versus "blocked entirely"? Current proposal treats refusal as a modification type.

6. Should confidence scores be standardized across providers, or is provider-specific calibration acceptable with method disclosure?

7. Is there value in a **gen_ai.safety.evaluation_latency** attribute for performance analysis of safety systems?

8. Should **modification_type** support multiple values (array) for responses with multiple modifications?

References

[**1**] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union, August 1, 2024. Article 12 (Record-keeping), Article 14 (Human oversight). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[**2**] OpenTelemetry Authors. *Semantic Conventions for GenAI operations.* OpenTelemetry Specification, Development status. https://opentelemetry.io/docs/specs/semconv/gen-ai/

[**3**] OpenTelemetry Authors. *Semantic Conventions for Gen AI Agent operations.* OpenTelemetry Specification, Development status. https://opentelemetry.io/docs/specs/semconv/gen-ai/gen-ai-agents/

[**4**] ABV Group. *EU AI Act Compliance Checklist for High-Risk AI Systems.* 2025. https://abv-group.com/eu-ai-act-compliance-checklist/

[**5**] Sombra. *AI Regulations Guide 2026: Enterprise Compliance Requirements.* 2025. https://sombra.ai/regulations-guide/

[**6**] European Commission. *AI Act Implementation Timeline and Penalties.* Digital Strategy, 2024. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[**7**] Xiong, M., Hu, Z., Lu, X., et al. *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs.* ICLR 2024. https://arxiv.org/abs/2306.13063

[**8**] Kadavath, S., Conerly, T., Askell, A., et al. *Language Models (Mostly) Know What They Know.* arXiv:2207.05221, 2022. https://arxiv.org/abs/2207.05221

[**9**] Kuhn, L., Gal, Y., Farquhar, S. *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation.* ICLR 2023. https://arxiv.org/abs/2302.09664