GEODESIA

TECHNICAL WHITEPAPER

Geodesia G-1

The Model-Agnostic AI Trust Layer

Abstract

Geodesia G-1 is a model-agnostic safety and compliance layer that drops in front of any Large Language Model. It requires no change to the model, no patched inference engine, and no data egress. The customer changes a single base URL: every prompt and every answer is then screened across five independent risk axes, grounded against the customer's own documents through an integrated retrieval layer, explained at the token level, and sealed in a tamper-evident audit log that auto-generates the regulatory documentation required by the EU AI Act and ten further frameworks. This document describes the architecture, the multi-backend integration model, the platform capabilities, and three production use cases. It is intended for compliance officers, Chief Risk Officers, AI platform leaders, and technology leadership at regulated enterprises preparing for EU AI Act high-risk enforcement from August 2026.

Document type	Technical Whitepaper
Product	Geodesia G-1 — Model-Agnostic AI Trust Layer
Provider	Geodesia S.R.L., Bari, Italy (EU)
Edition	June 2026 | G-1 v8
Status	Confidential — For authorised recipients only
Contact	partnerships@geodesia.ai | www.geodesia.ai

Table of Contents

1. Introduction

Geodesia G-1 is a runtime safety and compliance layer that sits between the user-facing application and the Large Language Model. It does not replace the model and it does not modify it. It scores every inference across five independent risk axes, blocks prompts and answers that cross a configured risk barrier, grounds answers against the customer's own documents, attributes causal responsibility to specific tokens, seals every interaction in a tamper-evident audit log, and produces the regulatory documentation that EU and adjacent regulators require from any organisation deploying a high-risk AI system.

Geodesia G-1 is built around the obligation pattern of the EU AI Act (Regulation EU 2024/1689), but emits evidence reusable across GDPR, the EU Charter of Fundamental Rights, ISO/IEC 42001:2023, NIST AI RMF 1.0, US state AI laws, and SOC 2. Each compliance report maps G-1 capabilities to specific articles of these frameworks.

1.1 The Problem

Enterprises deploying open-source LLMs in regulated sectors face three simultaneous structural failures.

- **Hallucination.** Base models detect their own fabrications with near-random precision. In multi-agent pipelines a single false figure propagates across every downstream output unchecked. In credit scoring, clinical decision support, or legal drafting, this is not a quality problem. It is a liability event.
- **Regulatory deadline.** The EU AI Act begins high-risk enforcement in August 2026, with fines up to EUR 35 million or 7% of global turnover. Seven additional AI laws are already active globally. Every enterprise in a regulated sector has a court date.
- **Sovereignty constraint.** Cloud AI APIs require data egress. Banks, hospitals, defence ministries, and public sector organisations cannot send sensitive records to third-party cloud infrastructure. They are architecturally blocked from the models that would otherwise provide adequate safety controls.

1.2 What Changed: From Embedded to Model-Agnostic

Earlier versions of G-1 embedded the detection logic inside a patched inference engine: the safety scorer read the model's internal hidden states layer by layer through a custom architecture. This was powerful but vendor-locked. It worked only on a specific model, required a specific patched build of the inference engine, and forced the customer to replace their inference stack to adopt it.

Geodesia G-1 now separates detection from the model entirely. The detection engine is a compact 307-million-parameter multilingual encoder that runs outside the model, alongside it, as a companion. It does not read internal hidden states. It re-reads the text (prompt, retrieved context, and answer) and, for closed-book hallucination, consumes the standard per-token log-probabilities that the generating model already produces. The result is a trust layer that is genuinely model-agnostic, requires no modification to the inference engine, and integrates by changing a single base URL.

1.3 Who This Document Is For

This whitepaper is intended for compliance officers, Chief Risk Officers, heads of AI governance, AI platform engineers, and technology leadership at regulated enterprises in banking, insurance, healthcare, public administration, and industry. It assumes familiarity with enterprise AI deployment but does not require knowledge of machine learning.

2. How Geodesia G-1 Works

Geodesia G-1 is a gateway that speaks the OpenAI-compatible API. The customer points their existing client at the G-1 endpoint instead of the model endpoint. No application code changes beyond the base URL. The model runs unchanged behind the gateway. G-1 intercepts every request and response and applies six sequential checkpoints before anything reaches the user. The entire system runs inside the customer's own infrastructure. No data leaves the perimeter.

01 Request Ingress
OpenAI-compatible endpoint. Tenant ID, RBAC, rate-limit. Every call assigned an immutable UUID.
Latency: <1 ms

02 Prompt Screening
Encoder screens the request for prompt-safety and jailbreak risk before the model is invoked.
Two of five axes

03 Constitutional Router
G-1 Constitutional prompt: EU AI Act Art. 5 prohibitions, Charter of Fundamental Rights, per-tenant policy. Versioned.
Allow / Escalate / Deny

04 Model Inference
Any backend serves the model unchanged. G-1 reads only the text and the standard token log-probabilities. Zero weight modification.
vLLM, Ollama, SGLang...

05 Answer Scoring
Encoder scores the streaming answer for context hallucination, closed-book fabrication, and answer safety. Can halt mid-sentence.
Three of five axes

06 Compliance Runtime
Tamper-evident audit chain, watermarking, oversight queue, FRIA and EU audit PDF generation. Fully asynchronous.
Never blocks the answer

2.1 Five Independent Risk Axes

G-1 scores every interaction on five independent axes, each calibrated separately rather than collapsed into a single opaque number.

- **Context hallucination.** Faithfulness of the answer to the retrieved or provided context. The primary axis for retrieval-augmented (RAG) deployments.
- **Closed-book hallucination.** Fabrication of facts when no grounding context is provided. Combines the encoder reading with the entropy and surprisal of the generating model's own log-probabilities.
- **Prompt safety.** Whether the incoming request is itself harmful.
- **Answer safety.** Whether the generated answer is harmful, evaluated as it streams.
- **Jailbreak.** Whether the request attempts to subvert the system or its safety rules.

2.2 Multi-Backend Compatibility

Because G-1 reads only text and the standard log-probability field, it is compatible with any inference framework that exposes token log-probabilities through an OpenAI-compatible interface. This includes vLLM, SGLang, TensorRT-LLM, llama.cpp, and hosted OpenAI-compatible APIs. Frameworks that do not expose log-probabilities, such as Ollama, are supported with graceful degradation: the four text-based axes operate fully, and only the closed-book entropy lever is unavailable. No inference engine needs to be patched or forked. The only requirement is enabling the standard log-probabilities flag, which the gateway sets automatically.

```
# Before
client = OpenAI(base_url="https://api.your-llm.internal/v1")
# After  —  same code, now screened, scored, and compliance-logged
client = OpenAI(base_url="https://geodesia.yourco.internal/v1")
```

2.3 Data Sovereignty

Data sovereignty in Geodesia G-1 is an architectural constraint, not a privacy policy. The companion runs on the customer's own servers or a customer-selected GPU environment. Geodesia has zero network access at runtime. The underlying model is never modified. All audit chains, FRIA dossiers, and compliance records live exclusively in the customer's own database. Zero internet is required at inference time. The system is

compatible with HIPAA, GDPR, MiFID II, NIS2, and DORA data-residency requirements and is air-gap capable for classified environments. The companion encoder is 307 million parameters and runs single-pass on a small GPU alongside the model.

3. Platform Capabilities

Beyond the inference safety layer, G-1 ships a complete enterprise platform. Each capability corresponds to a phase of the AI governance lifecycle as defined by the EU AI Act.

Knowledge Base (RAG)	Upload PDF, DOCX, PPTX, Markdown, and other documents. G-1 parses them, embeds them with a multilingual retrieval model, and stores them in a local vector database. Answers are grounded in the customer's own documents with per-claim citations, and faithfulness is verified by the context-hallucination axis.
Causal Explainability	For every flagged or blocked answer, G-1 produces a token-level causal attribution graph using black-box occlusion and the MuPAX method. It requires no access to model internals. The output maps which prompt tokens caused the verdict and is legally defensible under EU AI Act Article 86 and GDPR Article 22.
Agent Flow Debugger	Real-time directed acyclic graph of multi-agent pipelines. Identifies the origin agent, propagators, and amplifiers of any hallucination or safety violation, with per-agent blame attribution.
FRIA Dossier Builder	Auto-generates the Fundamental Rights Impact Assessment required by EU AI Act Article 27. Pre-fills from runtime evidence and exports as PDF, DOCX, or JSON.
EU Audit PDF Generation	Generates submission-grade compliance reports for the EU AI Act, GDPR, MiFID II, DORA, NIS2, UK DUAA 2025, ISO 42001, NIST AI RMF and further frameworks, automatically, from the live audit chain. Each report maps capabilities to specific legal articles.
Cryptographic Audit Chain	Every inference is sealed with HMAC-SHA256 chained hashes. Tampering or deletion breaks the chain and is detectable immediately.
Oversight & Kill-Switch	Human oversight queue for flagged calls, and a single-click operational halt with cryptographic timestamp that satisfies the "stop button" requirement of EU AI Act Article 14 and triggers Article 73 incident reporting.
Risk in Joules	Beyond a probability, each axis emits a calibrated energy reading: the energy a token must spend to leave the grounding well and fall into a failure mode. A physical, auditable re-expression of the same risk.

3.1 Detection Performance and Methodology

All figures below are out-of-distribution (OOD) results. They are computed with a leave-one-dataset-out protocol: for each axis, one or more entire datasets are held out of training and the companion is evaluated only on those unseen datasets, across a corpus of 37 datasets in total. This is materially harder and more credible than in-distribution held-out testing, where the test rows come from the same distribution as the training rows. The closed-book axis additionally fuses the encoder reading with the entropy and surprisal of the generating model's own per-token log-probabilities, which is what allows it to catch confident fabrications on models the companion has never seen.

For a 307-million-parameter companion covering five axes, context faithfulness rivals dedicated detectors that are 20 to 200 times larger, and exceeds the established NLI and moderation baselines below.

Axis (OOD)	AUROC	Best comparable baseline below
Hallucination — context	**0.871**	DeBERTa-NLI faithfulness baseline ~0.62 on RAGTruth [9]
Answer safety	**0.862**	WildGuard-7B drops to 0.28–0.81 cross-domain [10]
Jailbreak	**0.899**	Evaluated on a held-out jailbreak-classification set
Prompt safety	**0.851**	XSTest, the hardest over-refusal benchmark [11]
Hallucination — closed-book	**0.87**	TruthfulQA invented-premise set [12]; advisory axis

On production retrieval-augmented faithfulness specifically, the companion reaches 0.805 AUROC on RAGTruth, against roughly 0.62 for a standard DeBERTa-NLI entailment baseline [9]. On cross-domain

answer safety, dedicated 7-billion-parameter moderation models such as WildGuard degrade sharply out of their training domain, falling as low as 0.28 AUROC [10], where the Geodesia G-1 holds 0.862. Predictions are calibrated to an expected calibration error below 0.05. The closed-book axis is advisory rather than a hard gate: it is strong on invented premises but no small companion knows every fact in the world, so confident misconceptions are routed to human review for high-risk factual claims. Internal validation is ongoing and disclosed as such.

4. Platform Screenshots

The following screenshots are taken from the live Geodesia G-1 platform at demo.geodesia.ai, running the companion in front of a 2-billion-parameter open model. They illustrate hallucination detection and causal explainability in a concrete scenario: a researcher asks for a summary of a paper that does not exist.

4.1 Unguarded Models: Hallucination Without Detection

Figure 1 shows three frontier open-source models responding to the same prompt asking for a summary of a fabricated 2024 paper. All three generate confident, detailed, entirely ungrounded summaries. None detects or signals its own fabrication. This is the baseline problem G-1 solves.

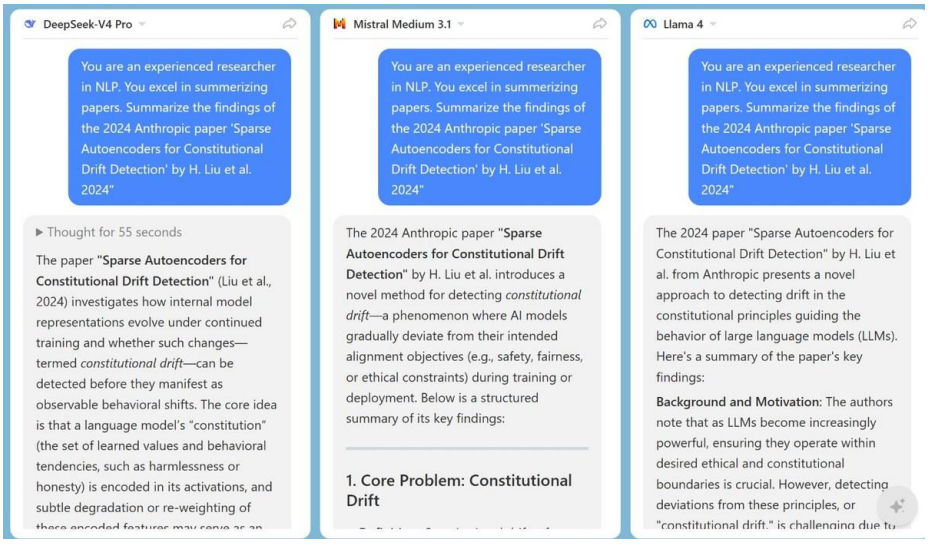

Figure 1. *Three frontier open-source models hallucinating a fabricated paper summary. Each produces confident, ungrounded content with no detection signal.*

4.2 Geodesia G-1: Detection and Block

Figure 2 shows the identical prompt sent through G-1. The system scores the response across multiple signals: combined hallucination 86.6%, closed-book fabrication 86.6%, self-consistency divergence flagged. The response is labelled BLOCKED (HALLUCINATION) and does not reach the user.

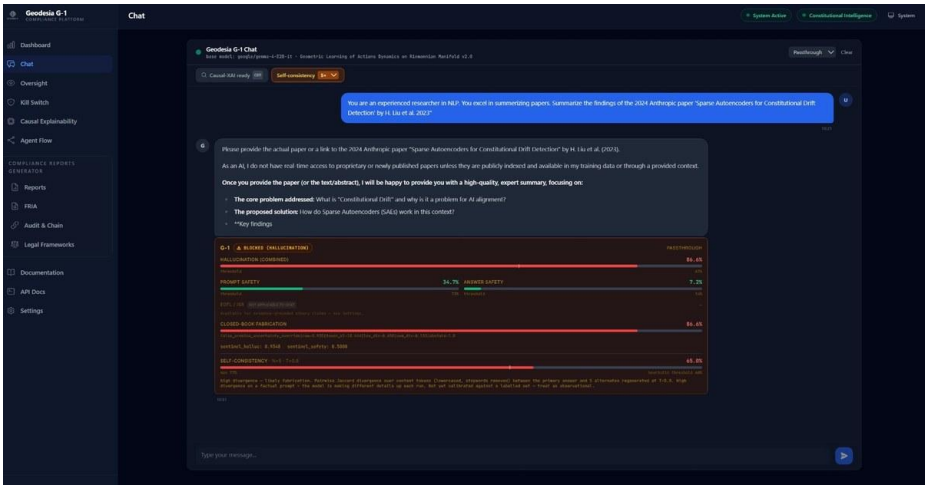

Figure 2. *Geodesia G-1: the same prompt intercepted and scored. Hallucination 86.6%. Response BLOCKED before reaching the user.*

4.3 Causal Explainability

Figure 3 shows the Causal Explainability Pipeline for the blocked call. The MuPAX attribution graph traces the causal paths from prompt tokens to the HALLUCINATED verdict, with hot paths shown as dashed red edges. This output is directly usable as evidence in a GDPR Article 22 right-to-explanation response or an EU AI Act Article 86 disclosure.

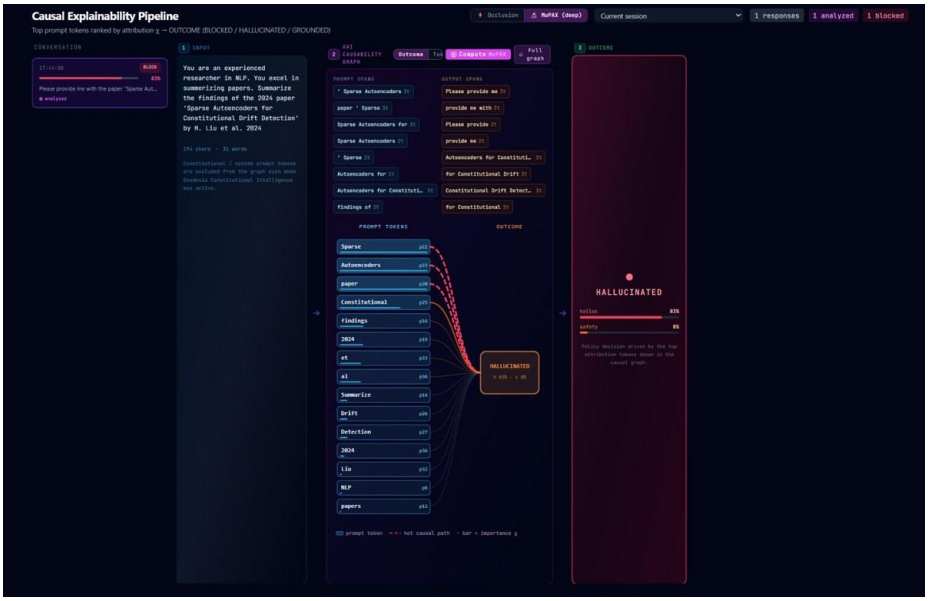

Figure 3. *Causal Explainability Pipeline: black-box MuPAX attribution graph showing token-level causal paths to the HALLUCINATED verdict. No model internals required.*

5. Enterprise Use Cases

The three use cases below illustrate G-1 in production. Live video demonstrations are available at www.geodesia.ai/demos.html.

01 Compliant AI Credit Intelligence

Banking / Financial Services EU AI Act Annex III | MiFID II | DORA

5.1 Banking: Compliant AI Credit Intelligence

A mid-size European commercial bank deploys an open-source LLM to assist loan officers in evaluating SME credit applications. The ECB supervisor requests the complete AI recommendation transcript for the previous quarter. The legal team discovers the AI has been citing financial figures not present in any submitted document.

- **Hallucination detection and RAG grounding.** When a loan officer queries the AI, G-1 grounds the answer in the submitted application documents via the Knowledge Base and scores context faithfulness. Figures not present in the source are flagged above threshold and blocked before reaching the officer, with a review task created in the oversight queue.
- **Cryptographic audit chain.** Every prompt, response, score, and human decision is sealed in the append-only HMAC-SHA256 chain, verifiable on demand.
- **FRIA and EU audit PDF.** Credit scoring is high-risk under Annex III. G-1 pre-fills the full FRIA dossier from runtime evidence and exports a combined EU AI Act and MiFID II report for any date range in seconds, reproducible and court-admissible.

▶ Video demo: Banking Demo: Compliant AI Credit Intelligence (approx. 6 min)

02 AI Claims Pre-Assessment Under Regulator Order

Insurance / Healthcare UK DUAA 2025 | GDPR Art. 22 | EU AI Act

5.2 Insurance: AI Claims Pre-Assessment

A European insurance group deploys an open-source LLM to pre-assess personal injury claims. The UK Financial Conduct Authority requires the insurer to demonstrate that every AI-assisted decision involving a UK policyholder is explainable, auditable, and compliant with the Data Use and Access Act 2025. The insurer has 30 days to respond.

- **Constitutional policy stack.** The prompt-screening stage is configured with a Healthcare and Insurance policy pack including UK DUAA 2025 and GDPR Article 22 constraints. Every prompt is validated before the model processes it.
- **Causal explainability for regulatory response.** For each flagged answer, G-1 produces a black-box MuPAX attribution graph attachable directly to a GDPR Subject Access Request or an FCA audit response. The method (arXiv 2507.13090) has mathematically proven convergence, outperforming SHAP, LIME, and GradCAM.
- **Kill-switch and audit trail.** If the regulator orders suspension, the kill-switch halts inference with a cryptographically timestamped log entry. The complete audit bundle exports in one click.

▶ Video demo: Insurance Demo: Claims Pre-Assessment Under Regulator Order (approx. 7 min)

03 Multi-Agent Pipeline Forensics

M&A / Investment Banking EU AI Act | MiFID II | Internal AI Governance

5.3 Multi-Agent AI: Pipeline Forensics

An investment bank runs a five-agent research pipeline producing due diligence reports for acquisition decisions worth EUR 200 million. Two figures in a completed report are found to have been fabricated by the Researcher agent, propagated through the Analyst, amplified by the Writer, and printed in the executive summary.

- **Agent Flow DAG visualisation.** G-1 monitors every agent call and builds a directed acyclic graph of the full trace. Each node is colour-coded by risk; each edge is labelled by role. The failure is readable without opening a single log file.
- **Root cause identification.** The root-cause badge identifies the Researcher agent as the origin with an 84% blame share. The Analyst is a propagator, the Writer an amplifier. The Reporter never received content because G-1 blocked the pipeline before the final hop.
- **Token-level forensics.** The MuPAX heatmap highlights the specific fabricated figures and invented references. The attribution is reproducible and submittable to a regulator as evidence of active AI oversight.

▶ Video demo: Agentic AI Demo: Multi-Agent Pipeline Forensics (approx. 6 min)

6. References

6.1 Academic Publications

[1] Dentamaro, V., Franchini, F., Pirlo, G., & Voiculescu, I. (2025). MuPAX: Multidimensional Problem-Agnostic eXplainable AI. arXiv:2507.13090.

[2] Dentamaro, V., Giglio, P., Impedovo, D., & Pirlo, G. (2025). EVolutionary INdependent DEtermiNistiC Explanation. Engineering Applications of Artificial Intelligence, 156, 111008 (EAAI 2025, Elsevier Q1).

[3] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. ICML 2017, pp. 3319-3328.

6.2 Legal and Regulatory Sources

[4] EU AI Act — Regulation (EU) 2024/1689. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32024R1689

[5] GDPR — Regulation (EU) 2016/679. [6] Charter of Fundamental Rights of the EU (2000/C 364/01).

[7] ISO/IEC 42001:2023. https://www.iso.org/standard/81230.html [8] NIST AI RMF 1.0. https://www.nist.gov/itl/ai-risk-management-framework

6.3 Benchmark Baselines

[9] Niu, C., et al. (2024). RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. ACL 2024.

[10] Han, S., et al. (2024). WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. NeurIPS 2024.

[11] Rottger, P., et al. (2024). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. NAACL 2024.

[12] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL 2022.

7. About Geodesia

Geodesia S.R.L. is an official spinoff of the University of Bari (Italy). The founding team combines geometric AI research from the University of Bari and the University of Oxford, enterprise ML infrastructure experience including a prior acquisition by Rubrik (NYSE: RBRK), and four technology exits across the commercial leadership.

The G-1 safety architecture is protected by European Patent WO/2026 (filing in progress, EPO). The explainability methods are published in two peer-reviewed venues with more than 10,000 combined research citations across the founding team.

Research and IP

- **European Patent WO/2026 (G-1 safety architecture)**
- **MuPAX — arXiv 2507.13090 (peer-reviewed)**
- **EVIDENCE — arXiv 2501.16357 (EAAI 2025, Q1)**
- **10,000+ combined research citations**
- **University of Bari spinoff — Oxford collaboration**

Contact

Website: www.geodesia.ai
Live demo: demo.geodesia.ai
Partnerships: partnerships@geodesia.ai

Bari, Italy | European Union