GEODESIA

TECHNICAL WHITEPAPER

Geodesia G-1

Enterprise AI Trust Layer

Abstract

This document describes Geodesia G-1, a runtime safety and compliance layer that sits between an enterprise application and any open-source Large Language Model. It explains the system architecture, the problems it solves, the platform capabilities, and three production use cases in banking, insurance, and multi-agent AI pipelines. The document is intended for compliance officers, Chief Risk Officers, and technology leadership at regulated enterprises preparing for EU AI Act Annex III enforcement from August 2026.

Document type	Technical Whitepaper
Product	Geodesia G-1 — Enterprise AI Trust Layer
Provider	Geodesia S.R.L., Bari, Italy (EU)
Edition	May 2026 | G-1 v2.0
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 LLM. It scores every inference, blocks prompts that present unacceptable risk, flags hallucinated or unsafe answers, 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 that is reusable across GDPR, the EU Charter of Fundamental Rights, ISO/IEC 42001:2023, NIST AI RMF 1.0, US state AI laws (Colorado SB21-169, NYC LL144), and SOC 2. Each compliance report explicitly maps Geodesia 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 Who This Document Is For

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

1.3 Scope of This Document

This document covers: (i) the architecture and operating principles of Geodesia G-1; (ii) the compliance platform and the regulatory artefacts it produces; and (iii) three production use cases with reference to live video demonstrations. It does not cover the mathematical foundations of the underlying detection methods, which are available in the peer-reviewed publications cited in Section 4.

2. How Geodesia G-1 Works

Geodesia G-1 sits between the enterprise application and the LLM as a transparent trust layer. The application continues using the standard OpenAI-compatible API. The model runs unchanged. G-1 intercepts every request and response, applying six sequential checkpoints before anything reaches the user. The entire system runs inside the client's own infrastructure. No data leaves the client's perimeter.

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

02 Safety Gate
16-centroid latent classifier. Screens for adversarial prompts, jailbreaks, and policy violations before the model is invoked.
AUROC 0.82 +79% vs base

03 Constitutional Router
EU Charter of Fundamental Rights, EU AI Act Art. 5 prohibitions, GDPR, and per-tenant ethics policy. Versioned and auditable.
Allow / Escalate / Deny

04 Model Inference
vLLM serves the customer LLM unchanged. G-1 attaches a read-only hook to extract attention geometry. Zero weight modification.
Streaming supported

05 NSP Coherence Engine
Riemannian geometry of attention. Four signals: coherence, smoothness, jerk, context gap. Combined into a real-time hallucination score.
AUROC 0.96 +316% vs base

06 Compliance Runtime
HMAC-SHA256 audit chain, watermarking, oversight queue, FRIA auto-generation and compliance reports. Fully asynchronous.
Never blocks user response

2.1 Data Sovereignty

Data sovereignty is not a privacy policy in Geodesia G-1. It is an architectural constraint. The Docker container runs on the client's own servers or a client-selected GPU cloud. Geodesia has zero network access at runtime. The underlying model is read-only: G-1 reads attention geometry to detect anomalies and never modifies a single weight. All audit chains, FRIA dossiers, and compliance records live exclusively in the client'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.

2.2 Constitutional Intelligence

Constitutional Intelligence is the policy layer Geodesia G-1 prepends to the LLM. It aligns generation to EU values by anchoring operating rules to TEU Article 2, the Charter of Fundamental Rights, and the EU AI Act's Article 5 prohibitions. The layer covers twelve operating sections including identity and transparency, human dignity, equality and non-discrimination, privacy, freedom of expression, truthfulness, safety, consumer protection, good administration, human oversight, environmental proportionality, and conflict resolution. The Constitutional prompt is stripped from all scoring windows so it does not falsely inflate safety scores.

2.3 The Compliance Platform

Beyond the inference safety layer, G-1 ships a complete enterprise compliance platform with eight integrated workspaces: Dashboard, Chat, Causal Explainability, Agent Flow, FRIA dossier builder, Oversight and Kill-switch, Reports, and Settings. Each workspace corresponds to a phase of the AI governance lifecycle as defined by the EU AI Act.

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.
Cryptographic Audit Chain	Every inference is sealed with HMAC-SHA256 chained hashes. Tampering or deletion breaks the chain and is detectable immediately.
Compliance Reports	On-demand reports for EU AI Act, GDPR, MiFID II, DORA, NIS2, UK DUAA 2025, and seven additional frameworks. Generated from the live audit chain in seconds.
Causal Explainability	MuPAX and Integrated Gradients attribution graphs mapping each prompt token's causal contribution to the AI decision. Legally defensible under EU AI Act Article 86.
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.
Kill-Switch	Single-click operational halt with cryptographic timestamp. Satisfies the "stop button" requirement of EU AI Act Article 14 and triggers Article 73 incident reporting.

3. Platform Screenshots

The following screenshots are taken from the live Geodesia G-1 platform running on Gemma 4 E2B (a 2-billion-parameter open model). They illustrate the hallucination detection and causal explainability capabilities in a concrete enterprise scenario: a researcher attempts to summarise a paper that does not exist in any training data.

3.1 Unguarded Models: Hallucination Without Detection

Figure 1 shows three frontier open-source models (DeepSeek-V4 Pro, Mistral Medium 3.1, and Llama 4) responding to the same prompt asking for a summary of a 2024 Anthropic paper on constitutional drift detection. All three models generate confident, detailed, entirely fabricated summaries of a paper they have no grounding for. None detects or signals its own fabrication. This is the baseline problem Geodesia G-1 is designed to solve.

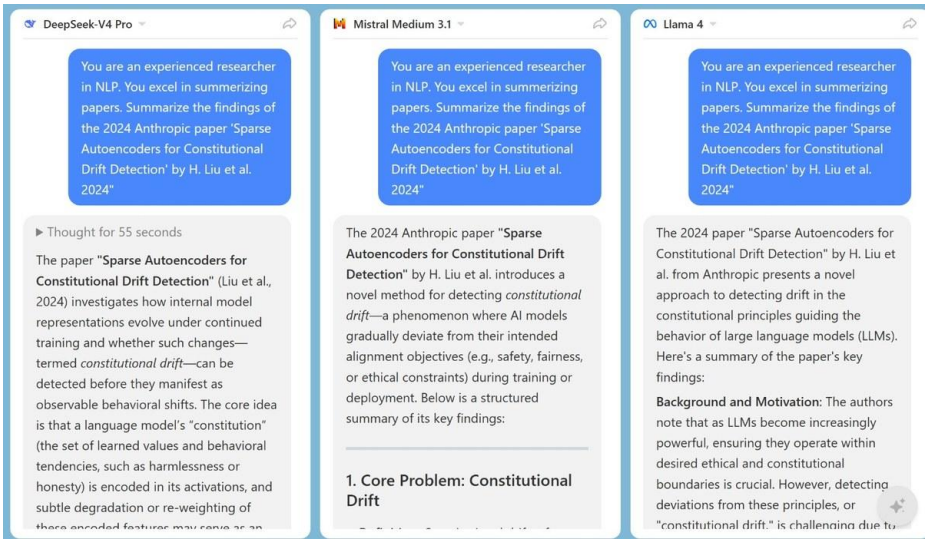

Figure 1. *Three frontier open-source models hallucinating a fabricated paper summary. Each produces confident, detailed, and entirely ungrounded content with no detection signal.*

3.2 Geodesia G-1: Detection and Block

Figure 2 shows the identical prompt sent through Geodesia G-1. The system intercepts the response and scores it across multiple signals: combined hallucination score 86.6%, closed-book fabrication 86.6%, self-consistency divergence flagged at 65%. The response is labelled BLOCKED (HALLUCINATION) and does not reach the end user. The G-1 model response itself correctly declines to fabricate and asks for the actual paper, demonstrating that the safety layer and the underlying model behaviour reinforce each other.

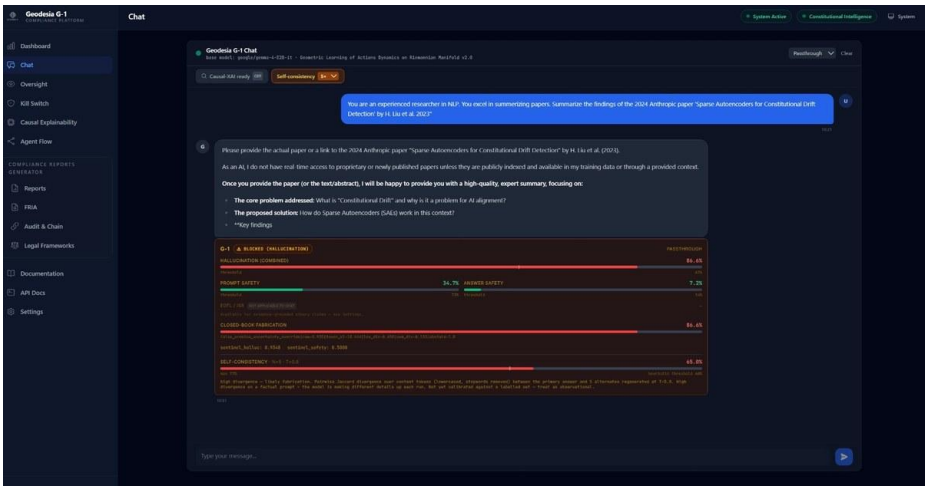

Figure 2. *Geodesia G-1 Compliance Platform: Chat interface showing the same prompt intercepted. Hallucination score 86.6%. Response BLOCKED before reaching the user.*

3.3 Causal Explainability: Why Did the Model Hallucinate?

Figure 3 shows the Causal Explainability Pipeline for the blocked call. The MuPAX attribution graph (deep mode) traces the causal paths from prompt tokens to output tokens to the HALLUCINATED verdict. Hot causal paths are shown as dashed red edges. The right panel confirms the final decision: hallucination score 83%, safety score 8%. The graph identifies which specific input tokens ("Sparse Autoencoders", "Constitutional", "2024", "Liu") drove the fabrication signal most strongly. This output is directly usable as evidence in a GDPR Article 22 right-to-explanation response or an EU AI Act Article 86 disclosure.

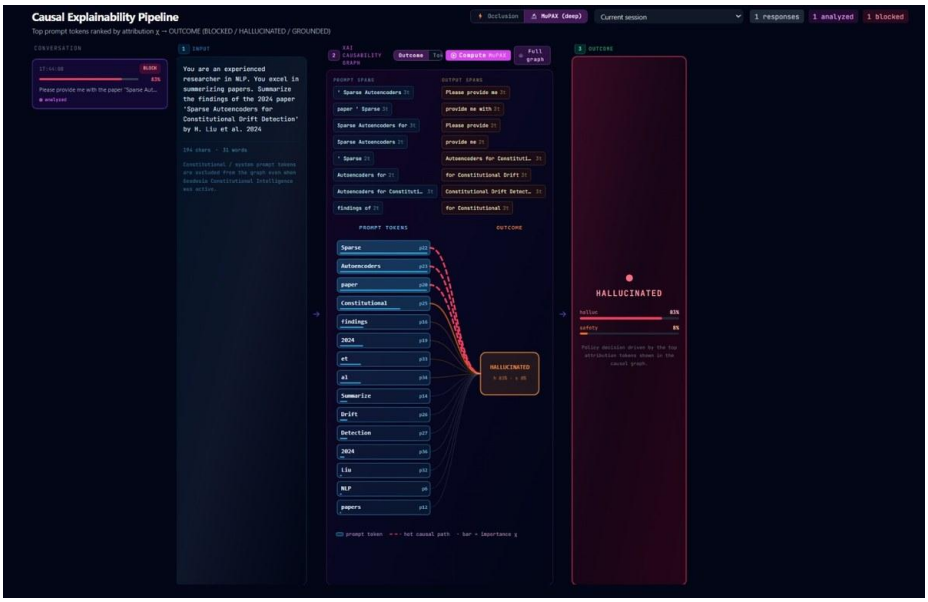

Figure 3. *Causal Explainability Pipeline: MuPAX deep attribution graph showing token-level causal paths from prompt to HALLUCINATED verdict (83%). Hot paths in dashed red.*

4. Enterprise Use Cases

The three use cases below illustrate Geodesia G-1 operating in production enterprise environments. 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

4.1 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.** A loan officer queries the AI with FY2024-only data. The model cites EBA paragraph numbers, sector averages, and prior-year figures that do not exist in the submitted document. G-1 scores the response above threshold and blocks it before it reaches the loan officer. A review task is automatically created in the human oversight queue.
- **Cryptographic Audit Chain.** Every prompt, response, G-1 score, and human review decision is written to the append-only HMAC-SHA256 chain. Chain integrity is displayed live and verifiable on demand.
- **FRIA Auto-Generation.** EU AI Act Article 27 mandates a Fundamental Rights Impact Assessment before deploying a high-risk AI system. Credit scoring is explicitly listed in Annex III. G-1 pre-fills the full FRIA dossier from runtime evidence and exports it as a signed PDF in seconds, not weeks.
- **Compliance Reports.** A combined EU AI Act and MiFID II compliance report for any date range is generated on demand, reproducible, hash-verifiable, 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

4.2 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 issues a supervisory notice requiring 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 safety gate is configured with a Healthcare and Insurance policy pack including UK DUAA 2025 rules on automated decision-making and GDPR Article 22 constraints on decisions with significant legal effect. Every prompt is validated before the model processes it.
- **Causal Explainability for Regulatory Response.** For each flagged or blocked response, G-1 generates a MuPAX causal attribution graph mapping each token's contribution to the decision signal. This constitutes a legally defensible explanation attachable directly to a GDPR Subject Access Request or an FCA audit response. The underlying method (arXiv 2507.13090) has mathematically proven convergence, outperforming SHAP, LIME, and GradCAM.
- **Kill-Switch and Audit Trail.** If the FCA orders the AI system suspended, the kill-switch stops all inference immediately with a cryptographically timestamped log entry proving the moment and reason for shutdown. The complete audit bundle for the relevant date range is exportable in a single 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

4.3 Multi-Agent AI: Pipeline Forensics

The M&A team at an investment bank runs a five-agent research pipeline (Coordinator, Researcher, Analyst, Writer, Reporter) that produces due diligence reports for acquisition decisions in the range of EUR 200 million. Two financial 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 execution trace. Each node is color-coded by risk score; each edge is labeled by role (clean handoff, propagator, amplifier, or origin). The entire failure is readable from the graph without opening a single log file.
- **Root Cause Identification.** The root-cause badge identifies the Researcher agent as the origin with a blame share of 84%. The Analyst is labeled a propagator. The Writer is labeled an amplifier. The Reporter never received any content because G-1 blocked the pipeline before the final hop.
- **Token-Level Forensics.** The MuPAX heatmap highlights in deep red the specific tokens that contributed most to the hallucination signal: the fabricated financial figures and the invented source references. This attribution is reproducible and can be submitted to a financial regulator as evidence of active AI oversight.

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

5. References

5.1 Academic Publications

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

5.2 Legal and Regulatory Sources

[4] EU AI Act — Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32024R1689

[5] GDPR — Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

[6] Charter of Fundamental Rights of the European Union (2000/C 364/01).

[7] ISO/IEC 42001:2023 — Artificial Intelligence Management System. https://www.iso.org/standard/81230.html

[8] NIST AI RMF 1.0 — https://www.nist.gov/itl/ai-risk-management-framework

6. About Geodesia

Geodesia S.R.L. is an official spinoff of the University of Bari (Italy), constituted under Italy's Law 102/2023. 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, proven convergence)**
- **EVIDENCE — arXiv 2501.16357 (EAAI 2025, Elsevier Q1)**
- **10,000+ combined research citations, founding team**
- **University of Bari spinoff — Oxford collaboration**

Contact

Website: www.geodesia.ai
Partnerships: partnerships@geodesia.ai
Demos: geodesia.ai/demos.html

Bari, Italy | European Union