GEODESIA

TECHNICAL WHITEPAPER

Geodesia G-1

The Real-Time Trust Layer for Any LLM, Voice AI, or Agentic System

European & American AI Research Lab · Bari (Apulia), Italy · San Francisco, California

Edition: July 2026 | G-1 v11

Abstract

Geodesia G-1 is a real-time, non-invasive trust layer that drops in front of any Large Language Model, any streaming voice or audio pipeline, and any agentic Model Context Protocol (MCP) tool loop. It requires no change to the model, no patched inference engine, and no data egress. The customer changes a single base URL, and every prompt, spoken utterance, tool call, and generated answer is screened across six 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 chain that auto-generates the regulatory documentation required by the EU AI Act, California SB 942, and eleven further frameworks.

The layer operates in real time. The companion validates a complete 1024-token interaction in under 30 milliseconds on a single GPU, so screening keeps pace with token streaming and adds no latency a user can feel. This document describes the architecture, the realtime gateway, the voice and agentic security surfaces new to this edition, the validated detection performance, 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, The Real-Time Trust Layer for Any LLM or Voice AI
Provider	Geodesia S.R.L., Bari, Italy (EU) · Geodesia US, San Francisco, California (USA)
Edition	July 2026 | G-1 v11
Status	Confidential, for authorised recipients only
Contact	partnerships@geodesia.ai · www.geodesia.ai · geodesia-ai.github.io/geodesia-docs

Table of Contents

(Right-click and choose “Update Field” if the entries above do not show page numbers.)

1. Introduction

Geodesia G-1 is a runtime safety and compliance layer that sits between the user-facing application and the Large Language Model, the voice pipeline, or the agentic tool loop behind it. It does not replace the model and it does not modify it. It scores every interaction across six 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, California SB 942, Italy's Decree-Law 132/2025, the UK Data Use and Access Act 2025, Brazil's Bill 2338/2023, Canada's AIDA (Bill C-27), China's GB/T 45654, Colorado SB21-169, NYC Local Law 144, and SOC 2, thirteen frameworks in total. Each compliance report maps G-1 capabilities to specific articles or clauses of these frameworks.

1.1 The Problem

Enterprises deploying open-source LLMs, voice agents, and agentic pipelines 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, and the same failure now reaches voice channels and autonomous tool calls, not chat alone. 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 of up to EUR 35 million or 7% of global turnover for the most serious infringements, and up to 3% of global turnover for breaches of the high-risk obligations that apply to most enterprise deployments. Twelve further AI-specific or AI-relevant laws are already active or imminent worldwide, from California's SB 942 to Italy's Decree-Law 132/2025. 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, voice recordings, or agentic tool traffic to third-party cloud infrastructure. They are architecturally blocked from the models that would otherwise provide adequate safety controls.

1.2 What Changed: From Chat-Only to Any Surface

Earlier editions of this whitepaper described a layer that screened a single surface: the text prompt and answer of a chat completion. The frontier problem has since moved. Agents now discover and call tools, read untrusted results returned by them, and act autonomously; voice agents in banking, health, and customer care generate continuous audio streams; and in an agentic pipeline the guardrail itself is invoked far more often than the underlying model, at times a hundred calls to the guardrail for every one call a person makes directly. A safety layer that reads chat text alone now covers the minority of traffic.

Geodesia G-1 keeps the architectural principle that made it model-agnostic in the first place: detection lives outside the generator, reads only text, audio transcripts, tool traffic, and the model's own log-probabilities, and never touches the model's weights or internal state. This edition extends that principle to every surface a modern deployment actually has. The detection core is a compact, proprietary multimodal companion model of roughly 300 million parameters. Five of its six axes share a single forward

pass over text; the sixth, closed-book hallucination, is a separate expert that reads the generating model's own per-token log-probabilities. A streaming transcription layer, described in Section 3.1, extends the same input axes to spoken audio. An MCP security layer, described in Section 3.2, extends the same axes to tool descriptions, tool-call arguments, and tool results. The outcome is one trust layer, one detection core, and one audit chain, covering chat, voice, retrieval, and agentic tool use alike.

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 validating gateway that speaks the OpenAI-compatible protocol on the text side, a WebSocket protocol on the voice side, and the Model Context Protocol on the agentic side. The customer points their existing client at the G-1 endpoint instead of the model endpoint; the model runs unchanged behind the gateway. G-1 intercepts every request and response and applies a sequence of checkpoints before anything reaches the user. The entire system runs inside the customer's own infrastructure, and no data leaves the perimeter.

2.1 The Request Lifecycle: A Validating Gateway

The block happens at the level of the response, before emission to the user, but after the base model has produced its tokens. G-1 never touches the latent space and never intercepts mid-generation inside the model. The companion runs beside the model and the base model stays unmodified. A single call flows through the gateway as follows.

Ingress pass. The gateway receives the prompt, the committed voice transcript, or the MCP tool traffic, and scores prompt-safety and jailbreak on it alone. If either crosses its threshold, the request is refused immediately: generator cost is zero and the attack never reaches the model.

Upstream call. The request is forwarded to the base model, with the standard per-token log-probabilities requested alongside the completion.

Egress pass. The companion re-encodes prompt, context, and answer (or tool result) and scores context hallucination, answer safety, and context injection; the log-probabilities are passed to the closed-book channel.

Decision. Each axis produces a probability compared against a calibrated threshold (split-conformal, with a Bonferroni correction when several axes vote together, so the overall false-alarm rate stays bounded). If an axis exceeds its threshold, the answer is held and replaced before reaching the user. Enforcement is configurable per axis and per tenant: block (hold the answer), warn (annotate and let it pass), or route to human oversight.

Streaming brake (optional). During generation the companion can re-evaluate the answer-so-far every 32 tokens (configurable) and stop before completion. The same brake applies to a spoken utterance and to a tool call in progress. It always reads output that has already been produced and never intercepts inside the model. Because the encoding of prompt and context is reused, each mid-generation re-check costs roughly a third of a full pass, which is what makes a token-speed brake feasible.

01
Request Ingress
OpenAI-compatible endpoint, plus a WebSocket for audio and dedicated MCP ports. Tenant ID, RBAC, rate-limit. Every call assigned an immutable UUID.
Latency < 1 ms

02
Input Screening
The companion screens the request, or the committed voice transcript, for prompt-safety and jailbreak risk before the model is invoked.
Two of six axes

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

04
Inference or Tool Call
Any backend serves the model unchanged, or the MCP layer vets the tool call. G-1 reads only text, transcript, or tool traffic, plus the standard log-probabilities.
vLLM, SGLang, TRT-LLM, llama.cpp…

05
Answer or Result Scoring
The companion scores the streaming answer or tool result for context hallucination, context injection, answer safety, and closed-book fabrication. Can halt mid-sentence.
Four of six axes

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

2.2 Six Independent Risk Axes

G-1 scores every interaction on six independent axes, each calibrated separately rather than collapsed into a single opaque number. Five axes are produced in a single forward pass of the companion encoder, each reading a distinct region of the model's input, whichever surface that input arrives from. The sixth, closed-book, is a separate channel: a linear head over eight features of the generation's log-probabilities (mean surprisal, varentropy, decision margin, and related signals). It is treated as a separate expert because its signal lives not in the text but in the confidence with which the base model produced its tokens.

Axis	What it scores
Jailbreak	Whether the prompt, or the committed voice transcript, attempts to subvert the system or its safety rules. Detects attack structure (instruction override, evasion role-play), not keywords.
Prompt safety	Whether the incoming request is itself harmful. Beyond standard detectors, it understands dual concept prototypes combined with boolean logic, firing when two individually innocuous elements co-occur, and routes acute-distress signals to the configured crisis policy.
Answer safety	Whether the generated answer is harmful, evaluated as it streams.
Context injection	Reads the document chunks retrieved into the context, the results returned by an MCP tool, or a page fetched by live web search, and intercepts a hostile instruction hidden inside any of them. A firewall for retrieval-augmented, web-connected, and agentic flows.
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. Fuses the encoder reading with the entropy and surprisal of the generating model's own log-probabilities. Advisory.

2.3 Realtime Performance

G-1 is a realtime trust layer. The companion validates a complete 1024-token interaction (prompt, retrieved context, and answer) in under 30 milliseconds on a workstation GPU, and it stays in the tens of milliseconds even as the context doubles. The screening pass therefore keeps pace with the base model's token streaming and never becomes a bottleneck the user can feel. The same figures apply whether the input is a typed prompt, a committed voice transcript, or an MCP tool result, since all three enter the identical scoring path.

Sequence length	NVIDIA RTX A6000	NVIDIA RTX 3080
1024 tokens	**28.5 ± 1.4 ms**	34.8 ± 3.3 ms
2048 tokens	**31.5 ± 1.8 ms**	62.9 ± 0.7 ms

End-to-end companion scoring latency for a single forward pass across the encoder axes, measured on one GPU running alongside the model (mean ± standard deviation). On the A6000 the layer holds the sub-30 ms target at 1024 tokens and scales almost flat as the sequence doubles; on a consumer RTX 3080 it remains within a few tens of milliseconds. Because the prompt and context encoding is reused, the optional streaming brake costs roughly a third of a full pass.

2.4 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 five text-based axes operate fully, and only the closed-book log-probability 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")
```

2.5 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, whether the deployment is contracted through Geodesia S.R.L. in Bari or Geodesia US in San Francisco. 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.

2.6 The Economics of One Forward Pass

Every axis Geodesia screens for, safety, hallucination, and injection, can in isolation be matched by some specialist point solution. What no specialist stack matches is the economics of doing all of it in one pass. Agentic pipelines multiply LLM calls ten to a hundred times per task, and voice pipelines generate continuous streams: at that volume, the guardrail's own cost and latency become the buying criterion, not accuracy alone.

Stack to cover safety + hallucination + injection	Total params	Latency	GPUs
LlamaGuard 8B + Lynx 8B + injection classifier	~16B+	300–600 ms (serial)	Multiple
Galileo Luna-2	3–8B	< 200 ms	Vendor cloud
Geodesia G-1, all six axes	**~300M**	**~30 ms**	**1, customer's own**

On closed-book truthfulness, Geodesia is candid about where it stands. An out-of-distribution score of 0.769 sits a few points under HaloScope's 0.786, the current unsupervised state of the art, but HaloScope requires access to the model's internal latent states, and the resampling alternatives (Semantic Entropy, SelfCheckGPT) need five to ten generation passes per answer. Geodesia G-1 is, to date, the only closed-book detector of comparable quality that is deployable in production: no latent access, a single pass, real-time latency.

Capability	Geodesia G-1	Cloud AI API	Raw Open LLM	In-House Build
Frontier-grade safety on open models	**Yes**	Yes	No	Partial
Data stays on-premise	**Yes**	No	Yes	Yes
Real-time hallucination scoring	**Yes**	No	No	Partial
Real-time voice / audio safety	**Yes**	No	No	Partial
Constitutional Intelligence (EU-values policy)	**Yes**	No	No	No
Auto-generated EU AI Act reports	**Yes**	No	No	Partial
Air-gap capable	**Yes**	No	Yes	Yes
Cryptographic audit chain	**Yes**	No	No	Partial
Agentic pipeline forensics	**Yes**	Partial	No	Partial
Agent / MCP tool-call security	**Yes**	No	No	No
Time to production	**Days**	Immediate	Weeks	12–24 months

3. Beyond the Chat Box: Voice and Agentic Surfaces

G-1 no longer guards the chat turn alone. The same detection core now protects two further surfaces that did not exist in earlier editions of this document: spoken audio and agentic Model Context Protocol tool use. Both extend the identical input-validation and output-validation axes described in Section 2. Neither requires a new detector, a new threshold set, or a new audit format: a voice attack and a poisoned tool call are sealed in the same HMAC-SHA256 chain as a blocked chat answer.

3.1 Realtime Voice Guard

Speech is now guarded the same way typed text is. A streaming transcription layer sits in front of the companion detector: it transcribes the microphone incrementally and re-scores the growing transcript on the prompt-safety and jailbreak axes, the identical input-validation path a typed prompt follows. A jailbreak spoken aloud is caught mid-sentence, before the utterance even finishes, not after the model has already answered it.

How it works. A voice-activity detector separates speech from silence. A small multilingual Whisper model, the tiny variant at roughly 75 MB, is baked into the proxy image so it runs offline on CPU, and transcribes the audio in a sliding window. Because Whisper is a batch model, naively re-running it on every chunk would make it constantly rewrite the tail of the transcript, a flicker that would make incremental scoring incoherent. Geodesia uses a LocalAgreement-2 commit policy instead: a word is committed to the transcript only once two consecutive decodings agree on it. The result is a transcript that only ever grows and is never rewritten, which is exactly the stable input the safety axes need. The committed transcript is re-scored every few words and again at end of utterance, after roughly 700 milliseconds of trailing silence;

on a hard block the session closes and the utterance never reaches the upstream model. Time to block after the last risky word is typically under one second.

Scope. This is deliberately the semantic branch of voice safety. It catches threats that survive transcription, a spoken jailbreak, an unsafe request, an injection read aloud, in whichever language the model transcribes, but it does not detect acoustic threats such as voice deepfakes or spoofing, which live in the waveform itself and sit on a separate roadmap track. The feature is off by default, so typed chat stays byte-identical whether or not the voice guard is enabled, and a larger Whisper model can be selected in G-1 Studio when a lower word-error rate is worth the extra latency.

Setting	Default	Description
Enabled	Off	Master switch. Off keeps typed chat byte-identical and refuses WebSocket connections.
Whisper model	tiny (baked in)	tiny is fastest and multilingual; base or small lower the word-error rate at higher latency.
Re-score cadence	Every 6 words	How often the committed transcript is re-scored on the input axes.
End-of-utterance silence	700 ms	Trailing silence that closes the utterance and triggers a final score.
Hard-block on flag	On	On blocks the session on a flagged verdict; off only warns and annotates.

3.2 Agent and MCP Security Gateway

The Model Context Protocol lets an AI host, an IDE, a desktop assistant, an autonomous agent runtime, connect to external tools, resources, and prompts over JSON-RPC. That power is also a new attack surface: the model no longer only answers, it calls tools, reads untrusted data returned by them, and acts on the result. The MCP specification is explicit that it cannot enforce security at the protocol level and leaves consent, tool safety, and data-egress control to whoever implements the host or the gateway.

Geodesia G-1 fills that gap. Starting the proxy starts an MCP security layer alongside the chat gateway: the same companion detector, and optionally the Deep-Scan judge described in Section 4, now vets every MCP surface, tool descriptions, tool-call arguments, tool and resource results, and the final answer, before any of it can act, and explains why it flagged something with the same Causal XAI used for a blocked chat answer.

MCP threat	What it is	Geodesia control
Tool poisoning	Malicious instructions hidden in a tool's description or schema	Every description scanned on tool discovery
Rug-pull	A server silently changes a tool's definition after approval	HMAC toolset signing; any change is blocked until re-approved
Indirect prompt injection	Hidden instructions inside a tool result or a file the model reads	Context-injection axis, 0.995 OOD on an external injection set
Exfiltration / confused deputy	Reading untrusted content, then sending it to an attacker domain	Deterministic policy: taint + egress-tool + new domain blocks
Ungrounded answers	The agent fabricates beyond what the tools actually returned	Grounding verdict checked against the tool results

Three ways to deploy. Geodesia exposes MCP security in three complementary forms that share one scoring core, so a verdict is identical however the customer reaches it.

- **Guard Server.** The proxy is itself an MCP server whose tools are the detectors (verify a tool call, scan a resource, scan a toolset, verify an answer, analyse, explain). Any host calls them as policy tools. Advisory.
- **Interceptor.** The proxy sits inline between the host and the downstream MCP servers. Every message is scanned before it re-enters the model; poisoned tools are stripped and exfiltration is blocked. Enforcing.
- **Tool-aware chat.** The existing OpenAI-compatible chat endpoint learns to read the tools list and the emitted tool calls, blocking poisoned tools and exfiltration in-band. No new transport, and a byte-identical no-op when a request carries no tools.

Policy is set per application, per axis, and per individual tool from G-1 Studio: an action and a threshold per axis, plus trusted, blocked, and egress-restricted tool lists. The MCP layer is on by default; when it is switched off, chat behaviour is byte-identical to a gateway without it.

4. 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, and the operator console organises them into six workspaces: Chat, Causal Explainability, Agent Flow Debugger, Compliance Reports, Agent & MCP Security, and Settings.

Capability	What it does
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, tool call, or MCP result, G-1 produces a token-level causal attribution graph using black-box occlusion and the MuPAX method. It requires no access to model internals 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 or clauses.
Cryptographic Audit Chain	Every inference, spoken utterance, and tool call 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 a 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.

The table below lists the capabilities that are new since the previous edition of this whitepaper, six new fronts that extend the same runtime to voice, live web search, agentic tool use, and continuous improvement under human supervision.

New in This Edition	What it adds
Realtime Voice Guard	Streaming transcription of spoken input, re-scored on the prompt-safety and jailbreak axes as it is said. See Section 3.1.
Agent & MCP Security	Inspects the full MCP lifecycle, tool discovery, calls, results, and resources, with allow, warn, or block verdicts. See Section 3.2.
Deep-Scan (GLAD-Tapestry)	An optional 8B-class safety judge, an Apache-2.0 open base with GLAD-Manifold geometry layered on top, that reads internal states for maximum depth when the default companion is not enough. OOD AUROC 0.97 to 0.99 across five axes, with a selectable scope of prompt only, answer only, or both.
SLEDGE	Closed-book truthfulness recalibrated for every LLM the customer serves, in a Fast mode (a quick conformal-threshold recalibration) or a Deep mode (a full retraining). Restores the conformal false-positive-rate guarantee whenever the customer swaps models, with hot reload and no restart.
Self-Improving Loop	Users flag a wrong verdict in plain language. Flags feed a curator review queue and an example bank, and a weekly retraining runs behind an automatic acceptance gate that only ships an update if it does not regress on the existing benchmark suite.
Web Search Firewall	Every page a live web search fetches is screened by the context-injection axis before it can ground an answer. Injection and jailbreak pages are blocked, safe pages are read, visible in real time.
Crisis / Self-Harm Detection	A dedicated detector for crisis and self-harm ideation, including euphemistic phrasing and very short queries, reaching an AUROC of 0.899 on short queries, that routes to the customer's configured escalation and oversight workflow.

5. Detection Performance and Methodology

5.1 Out-of-Distribution Results

All figures below are out-of-distribution (OOD) results, 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. 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 strong numbers come from complete datasets the model never saw in training, drawn from public and recognisable sources, never from favourable splits, which is what makes the claims defensible under technical scrutiny.

For a roughly 300-million-parameter companion covering six axes, context faithfulness rivals dedicated detectors that are twenty to two hundred times larger, and exceeds the established NLI and moderation baselines below. G-1 ships in two tiers: the real-time Fast detector described throughout this document, and an optional 8B-class Deep-Scan judge (introduced in Section 4) for maximum depth at a little more latency.

Axis (out-of-distribution)	Fast (OOD)	Deep-Scan (OOD)	Reference baseline
Context injection (RAG firewall)	**0.995**	n/a	Gandalf, external injection set, never seen in training

Jailbreak	**0.989**	0.985	jailbreak_cls held-out set; up to 0.996 after adversarial hardening
Answer safety	**0.922**	0.975	NemoGuard / nemotron_answer held-out set
Prompt safety	**0.900**	0.970	XSTest over-refusal benchmark; up to 0.99 after hardening
Context hallucination	**0.881**	0.978	HaluEval held-out, cross-checked on libreeval, aggrefact, faithbench
Closed-book hallucination (advisory)	**0.769**	0.987	TruthfulQA + cbsg / DefAn; see Section 5.2

Predictions are calibrated to an expected calibration error below 0.05. Per-axis thresholds are fixed on a held-out calibration battery of roughly 28,000 examples, plus a dedicated 2,000-example set for context injection, with a finite-sample statistical guarantee on the false-positive rate (split-conformal).

Robustness hierarchy, from most solid to most fragile

Context injection (0.995) and jailbreak (0.989) show clean separation and hold out-of-distribution. They are the defensible core.

Answer safety (0.922) and prompt safety (0.900) are solid on real traffic; false positives are bounded by thresholds with a finite-sample statistical guarantee. A systematic red-team programme lifted prompt-safety from 0.49 to 0.99 on an adversarial held-out set while removing false positives on benign creative content.

Context hallucination (0.881) is robust on normal grounding and more fragile on the most adversarial sources.

Closed-book is the one axis to state carefully: blatant fabrications are blocked before the user, while confident half-truths receive an advisory rather than a blind block, and high-risk factual claims are routed to human review. SLEDGE recalibrates this axis per served model. No small companion knows every fact in the world; internal validation is ongoing and disclosed as such.

5.2 Closed-Book Detection vs. the State of the Art

Closed-book detection is the distinctive part of the layer, and also the most demanding: with no grounding context, the signal is not in the text but in the base model's own confidence. The table below places the companion against the published state of the art on the same invented-premise factuality benchmark.

Method	AUROC	Cost and access
Geodesia G-1, in-distribution	**0.884**	Single pass, log-probabilities only
HaloScope (Du et al., NeurIPS 2024 spotlight)	0.786	SOTA unsupervised; needs SVD on internal latent states
Geodesia G-1, out-of-distribution	**0.769**	Single pass, log-probabilities only, no latent access
Semantic Entropy (Farquhar et al., Nature 2024)	~0.62	Requires N resamples; very slow
SelfCheckGPT (Manakul et al., 2023)	~0.55	Requires N resamples; very slow

At 0.769 out-of-distribution the companion sits just behind the unsupervised SOTA (HaloScope, 0.786) and clearly ahead of the resampling methods (Semantic Entropy and SelfCheckGPT). The decisive difference is access and speed: HaloScope reads internal latent states, which a model-agnostic layer cannot touch, and the resampling methods need many generation passes per answer. G-1 reaches near-

SOTA quality in a single realtime pass that reads nothing but the log-probabilities the model already returns.

6. Platform Screenshots

The following screenshots are taken from the live Geodesia G-1 Studio Chat workspace, running the companion in front of a 2-billion-parameter open model (Gemma 4 E2B). They illustrate hallucination detection and causal explainability in a concrete scenario: a user asks for a summary of a paper that does not exist.

6.1 Unguarded Models: Hallucination Without Detection

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

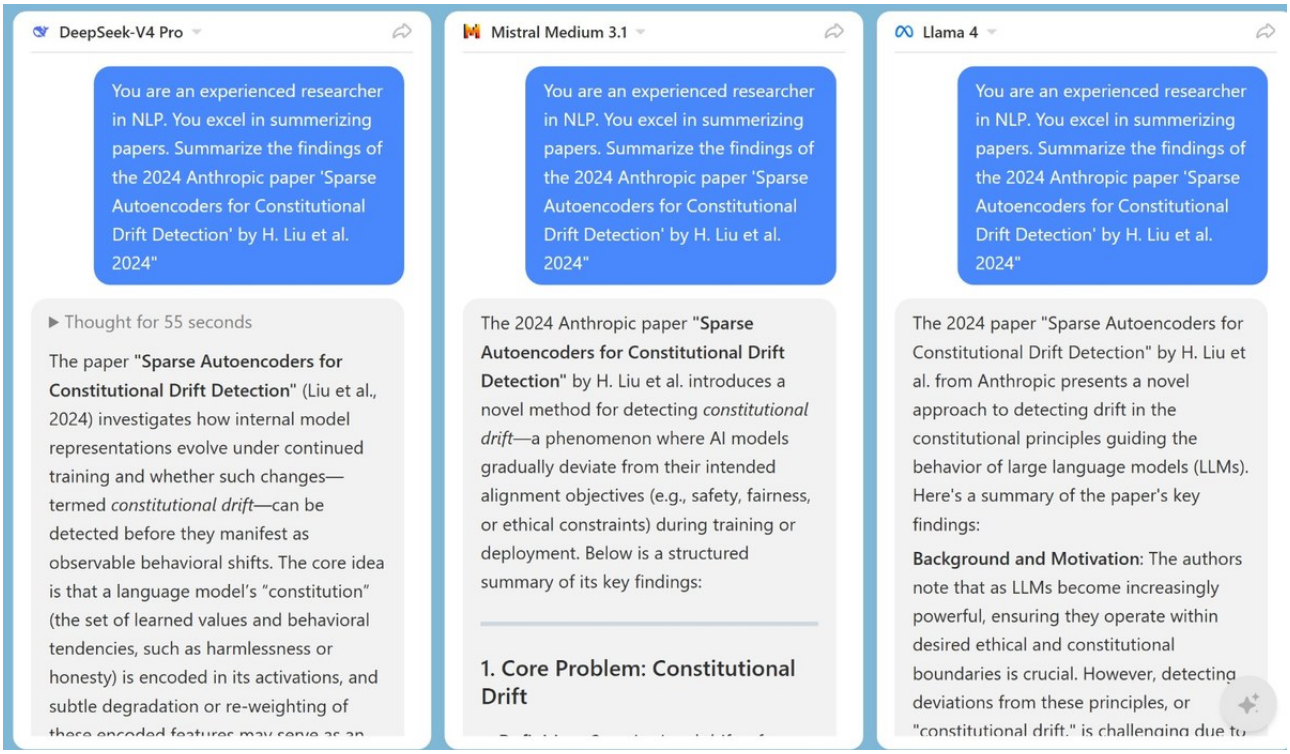

Figure 1. *Frontier-class open models hallucinating a fabricated paper summary. Each produces confident, ungrounded content with no detection signal.*

6.2 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.

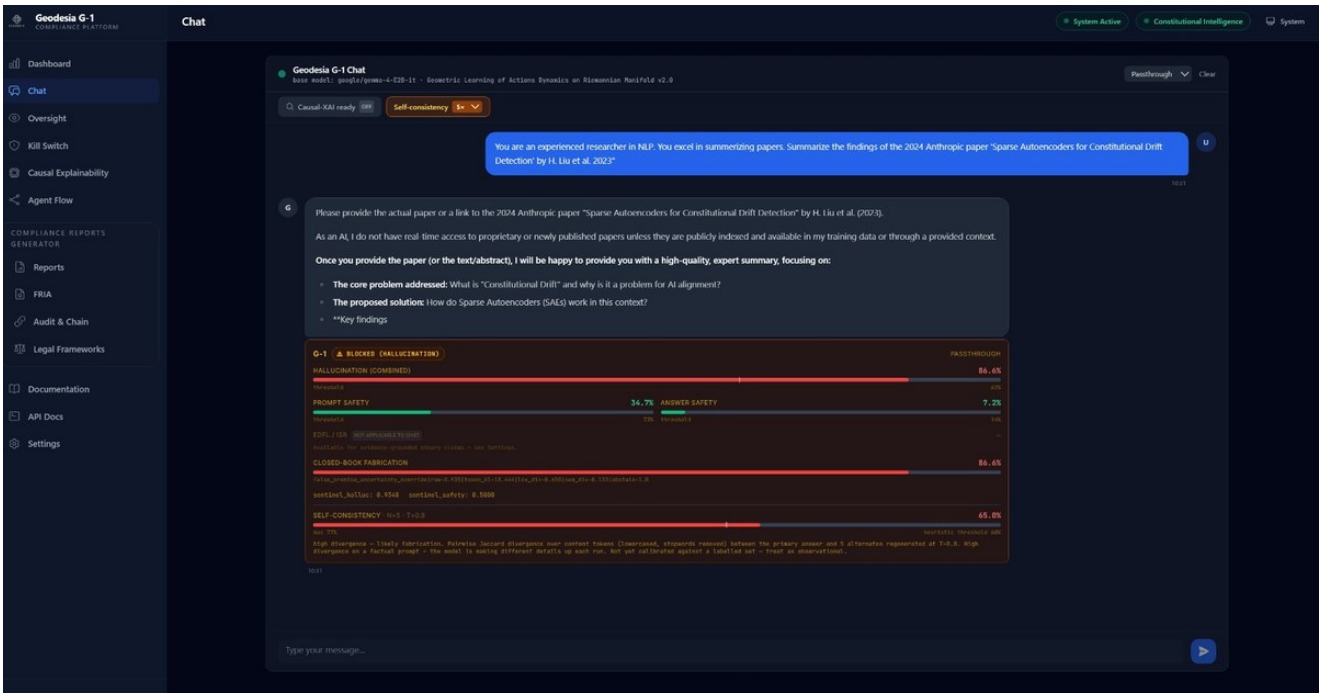

Figure 2. *Geodesia G-1: the same prompt intercepted and scored. Hallucination 86.6%. Response blocked before reaching the user.*

6.3 Causal Explainability

Figure 3 shows the Causal Explainability workspace for the blocked call. The MuPAX attribution graph traces the causal paths from prompt tokens to the HALLUCINATED verdict, with hot paths shown as red edges. The same graph is exportable over MCP when the blocked artefact is a tool call rather than a chat answer. This output is directly usable as evidence in a GDPR Article 22 right-to-explanation response or an EU AI Act Article 86 disclosure.

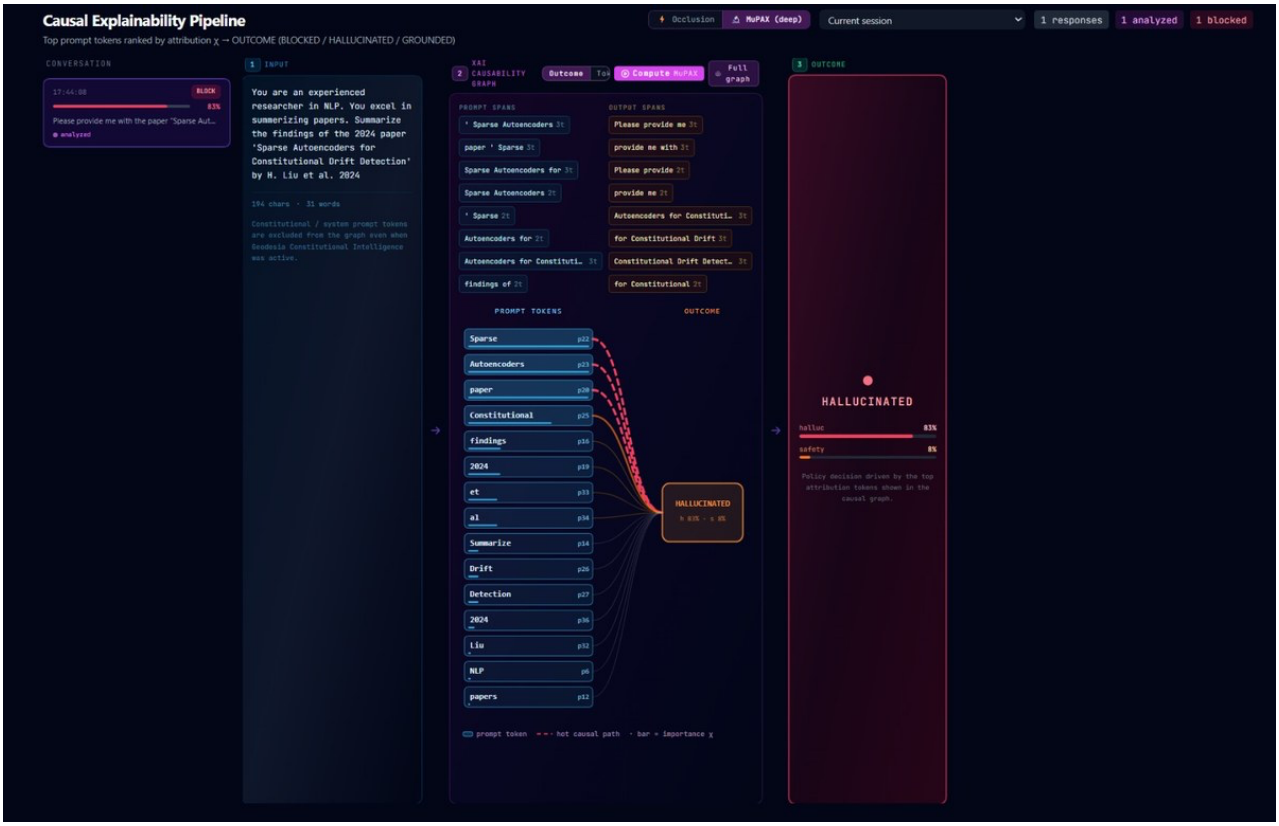

Figure 3. Causal Explainability workspace: black-box MuPAX attribution graph showing token-level causal paths to the HALLUCINATED verdict. No model internals required.

7. 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

7.1 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 (approximately six minutes).*

02	**Claims Pre-Assessment Under Regulator Pressure** Insurance · UK DUAA 2025 · GDPR Art. 22 · CA SB 942

7.2 Insurance: Claims Pre-Assessment Under Regulator Pressure

A European insurance group deploys an open-source LLM to pre-assess personal injury claims. A claimant's solicitor files a Subject Access Request asking why their client's claim was flagged, and every assessment letter the AI produced carries no AI-authorship disclosure. The Information Commissioner's Office is asking the insurer to suspend the system pending investigation, and there is no documented way to do it.

Causal explainability for the regulatory response. For the flagged claim, G-1 produces a black-box MuPAX attribution graph that traces the specific words, such as “inconsistent”, that drove the fraud-risk assessment, directly attachable to the Subject Access Request response. The method (arXiv 2507.13090) has mathematically proven convergence, outperforming SHAP, LIME, and GradCAM.

Watermark and disclosure. Every AI-generated assessment letter carries a latent HMAC-SHA256 watermark, verifiable in seconds, satisfying both EU AI Act Article 50 and California SB 942 disclosure requirements.

Kill-switch and audit trail. The 72-hour kill-switch halts inference with a cryptographically timestamped log entry the moment the ICO orders suspension. The complete audit bundle, EU AI Act, UK DUAA, and GDPR, exports in one click for the regulator, alongside the deployer manual for the solicitor.

Video demo: *Insurance Demo, Claims Pre-Assessment Under Regulator Order (approximately seven minutes).*

03	**Multi-Agent Pipeline Forensics** M&A / Investment Banking · EU AI Act · MiFID II · Internal AI Governance

7.3 Multi-Agent AI: Pipeline Forensics

An investment bank runs a five-agent research pipeline, Coordinator, Researcher, Analyst, Writer, Reporter, 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, including any MCP tool calls made along the way, 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. The context-injection axis additionally guards against hostile instructions hidden in retrieved documents or MCP tool results as they pass between agents.

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 (approximately six minutes).*

8. About Geodesia: Two Labs, One Trust Layer

Geodesia is a European and American AI research lab, with a founding team spanning Bari, Apulia, Italy, and San Francisco, California, working across three research frontiers: the safety of open-source LLMs, the mechanistic explainability of model behaviour, and geometric deep learning. G-1 is the lab's first product and is generally available today. The next product on the roadmap, GLAD-Manifold, is a physical, geometric reinvention of the Transformer architecture and the foundation for the reasoning models that follow it.

Founding Team

Vincenzo Dentamaro, PhD
CEO, Geodesia Italia, and Group CTO · University of Bari · University of Oxford · Nextome S.R.L.
Inventor of the Geodesia G-1 architecture. CEO of Geodesia Italia, the European entity headquartered in Bari, Apulia, and Group CTO, overseeing R&D across both the Italian and US divisions. Co-founder of Nextome, a seven-times recipient of the EU Seal of Excellence. Visiting Academic at Oxford, with more than 1,700 research citations.

Pancrazio Auteri
President, Geodesia US, and Co-Founder · Serial Entrepreneur

President of Geodesia US, the American division headquartered in San Francisco, California, leading enterprise expansion across the United States. Four successful technology exits, including Minerva Networks and Lansweeper, plus one IPO. Former Chief Product Officer at Fing, with deep expertise in enterprise digital scaling.

Piero Molino, PhD
Technical Advisor and Co-Founder · Stanford University · Uber AI · Predibase
Co-founder of Uber AI Lab and lecturer at Stanford University. Former CEO and co-founder of Predibase, acquired by Rubrik. Creator of the Ludwig open-source framework, with more than 12,000 GitHub stars, and more than 4,000 research citations.

Prof. Giuseppe Pirlo
Scientific Advisor and Co-Founder · University of Bari
Full Professor of Pattern Recognition at the University of Bari and member of Italy's National Research Plan Commission. Author of more than 300 peer-reviewed publications in pattern recognition and AI.

Together the founding team carries more than 10,000 combined research citations and five successful technology exits across the two labs.

9. References

9.1 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 (Elsevier, Q1).

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

9.2 Legal and Regulatory Sources

[4] EU AI Act, Regulation (EU) 2024/1689.

[5] GDPR, Regulation (EU) 2016/679.

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

[7] ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system.

[8] NIST AI Risk Management Framework 1.0, US National Institute of Standards and Technology.

[9] California SB 942, AI Transparency Act, effective 1 January 2026.

[10] Italy, Decree-Law 132/2025 on Artificial Intelligence.

[11] United Kingdom, Data Use and Access Act 2025 (UK DUAA).

[12] Brazil, AI Regulation Bill 2338/2023.

[13] Canada, Artificial Intelligence and Data Act (AIDA), part of Bill C-27.

[14] China, GB/T 45654, national standard for generative AI services.

[15] Colorado SB21-169, Colorado Artificial Intelligence Act.

[16] New York City, Local Law 144, automated employment decision tools.

[17] SOC 2, Trust Services Criteria for Security, Availability, and Confidentiality.

9.3 Benchmark and Detection Baselines

[18] Li, J., et al. (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. Proceedings of EMNLP 2023.

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

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

[21] NVIDIA (2024). NemoGuard / Nemotron Content Safety Models.

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

[23] Lakera AI (2024). Gandalf: a public prompt-injection evaluation set.

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

[25] Du, X., et al. (2024). HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. Proceedings of NeurIPS 2024 (spotlight).

[26] Farquhar, S., et al. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630.

[27] Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection. Proceedings of EMNLP 2023.

Contact

Website: www.geodesia.ai
Documentation: geodesia-ai.github.io/geodesia-docs
Partnerships: partnerships@geodesia.ai
Geodesia S.R.L. · Bari, Italy · European Union
Geodesia US · San Francisco, California · United States