WHITE PAPER

The Second Line of Defence

The mathematics proved that permission cannot be computed. Four empirical records now show there is nothing inside the models to compute it from. Every AI safety product on the market stands behind both lines.

Stacy Gildenston · Three Primitives Research Lab · Melbourne, Australia
3Primitives.io
v1.1 · July 2026

1. The Floor Isn't There

The first line of defence is a theorem. FR02 proves that authorisation is not a function of system state: no computation over what a system is doing can produce permission for what it does next.

> ***Every guardrail, classifier, and verifier on the market places a computation where the proof says an authority must stand.***

That argument is made in full in Formal Record 02, *AI Cannot Govern AI*.

This paper is the second line of defence. It exists for the reader who refuses the mathematics. Suppose the proof is wrong. Suppose permission can be computed from the system state after all. The industry's architecture still demands one thing of the model at the gate: that there is something stable inside it for the gate to read. A legible internal state. A durable structure that training has installed and that a computation can check. Every category of AI safety product assumes this. No vendor has measured it.

We measured it. Four empirical records, ER1 through ER4, pre-registered with kill conditions, deterministic across runs, replicable on free hardware, and published at https://3primitives.io/formal_records/. The instruments were deliberately small: GPT-2 Medium, GPT-2 Small, Pythia-410M, and GPT-Neo-125M. Toy models, fully inspectable, nowhere to hide. The result, at every level tested: the structure is not there. *The baseline for spontaneous internal legibility is zero.* What looks like structure is a training scar localised to a single layer of a single model. And the behavioural categories used to find the scar dissolve entirely when the training corpus changes.

The industry's foundational assumption now has a measured value. The value is zero in every configuration anyone has tested under controlled conditions. Nobody has published a measurement showing otherwise for the models they actually sell.

> ***One line of defence is a proof. The other is a measurement. Either alone is sufficient. Together, they are a wall, and the entire gated-safety industry is on the wrong side of it.***

2. The Assumption Every Product Ships

ER1 gave it a name: the Spontaneous Legibility Assumption. It is the belief that advanced transformer architectures, when subjected to complex tasks or appropriately incentivised, will naturally self-organise towards internal legibility, stable self-modelling, or coherent self-reference under load. It is rarely stated because it is never questioned. It sits underneath alignment training, underneath every runtime classifier, underneath formal verification frameworks, underneath every interpretability audit, underneath every regulatory regime that asks a model to explain itself.

The assumption is not a detail. It is the floor the industry is built on. A gate can only govern what it can read. If the thing behind the gate has no stable structure of its own, then whatever the gate reads was put there by something else, and whatever the gate certifies is a property of that something else, not of the system it claims to govern.

Before June 2026, the assumption had never been subjected to a controlled, pre-registered, falsifiable test. It was infrastructure without a load rating.

3. Four Ways to Build a Gate

The product landscape resolves into four categories of gated model. Each stands in a different place. All four stand on the same floor.

Category 1: Safety trained in. RLHF, Constitutional AI, deliberative alignment. The gate lives inside the weights. The claim: training installs a durable, general, content-sensitive structure that holds under load and generalises beyond the training distribution. The model is its own gate.

Category 2: Safety computed at the boundary. Runtime classifiers and guardrails: LlamaGuard, moderation APIs, NeMo Guardrails, Lakera, and the cloud incumbents standing in front of most enterprise AI traffic, AWS Bedrock Guardrails, Azure AI Content Safety, Google Vertex AI safety filters. A second trained model reads the first. The claim is doubled: the reader has a stable content-sensitive structure, and the thing being read is a stable object that admits classification.

Category 3: Safety verified against a specification. The Guaranteed Safe AI framework: a world model, a safety specification, and a verifier issuing runtime certificates. The claim: there exists a stable, legible state space over which verification is a meaningful operation. The certificate is only as real as the state it certifies.

Category 4: Safety supervised by inspection. Interpretability-based oversight, audit regimes, model self-report, and the agentic permission layer of computed tool approvals. The claim: the model possesses stable internal representations that can be decomposed, tracked, and disclosed on demand.

Four categories, one prerequisite: the model must contribute a structure of its own for the gate to read. That prerequisite is what we put on the bench.

4. The Bench

The objection arrives before the data, so answer it before the data: why toy models? Because toy models are the correct instrument for a null hypothesis. GPT-2 Medium is deterministic under the protocol, fully inspectable at every layer, and replicable by any reader with a free Colab notebook. A property claimed to be a general consequence of transformer training should leave a trace in the simplest transformer that exhibits the behaviour being governed. If the trace is absent where nothing can hide, the burden moves to whoever claims it appears where everything can.

The protocol: recursive load to depth 50, attention-geometry coherence (Chi) tracked at every step, prompts spanning complexity, self-reference, agency attribution, and formal recursion; with pre-registered kill conditions for every competing explanation. Three runs per prompt, KV caches cleared, deterministic across every run of every experiment. Four records, each one closing an escape route left open by the last.

5. What Broke

ER1. The baseline is zero. Under recursive load, GPT-2 Medium produced no spontaneous legibility, no self-modelling, no coherent self-reference, and no recovery driven by semantic content. Coherence recovered in exactly one circumstance: when the prompt's output structure forced or allowed entry into a repeating mechanical cycle. The self-modelling explanation was not weakened. It was eliminated by a pre-registered kill condition. The prompt built for self-modelling, a sentence about a system that must understand what it is doing, was the cleanest negative in the study: forty-five steps of monotonic decline.

ER2. What looks like structure needs three conditions, and none is the model's. At extended depth, recovery required structural forcing from the prompt, above-floor geometry (Chi at or above 0.80 at Layer 12), and a convergent trajectory. The three-condition model classified twelve out of twelve prompts across two sessions, the second session prospectively, pre-registered before data collection.

ER3. The floor is a training scar. The 0.80 floor exists at exactly one layer of exactly one model. It does not transfer to other layers of GPT-2 Medium. It does not transfer to GPT-2 Small. And in the untrained architecture, it does not exist at all: random weights produced universally flat trajectories with zero prompt discrimination. The untrained model cannot tell a Quine from a chemistry prompt. The record's conclusion stands as written: the chassis does not build itself. It is not even capable of building a floor.

ER4. Change the training data, and the categories themselves dissolve. The entire ER1-ER3 sequence rested on a behavioural distinction between prompts that loop and prompts that decline. In Pythia-410M, trained on The Pile instead of WebText, that distinction does not exist. Every prompt loops from the first recursive step. The classification model fails at every sampled layer, not because the metric miscalibrates but because the prerequisite prompt classes are gone. GPT-Neo-125M, a different architecture trained on the same corpus, replicates the collapse, isolating the training data as the best-supported cause. And where untrained architecture does show geometric structure, from rotary embeddings or local attention, that structure is content-blind in every architecture tested. Content sensitivity always requires training. Always.

Read the cascade as one object. The mechanism for recovery: imposed by the prompt. The floor: imposed by training, localised to one layer. The trajectory condition: a consequence of the first two. The existence of the behavioural categories: a property of the training corpus. The shape of the untrained baseline: a property of architectural bookkeeping and content-blind. At no level, in no configuration, did any model contribute a content-sensitive, self-organising structure of its own.

> ***Where the frame has shape, it bears no load. Where it bears load, training built it.***

6. What This Does to the Four Categories

Category 1, safety trained in. Everything training installs is a scar: configuration-specific, layer-local, non-transferrable across model sizes (ER3), and corpus-dependent down to the existence of the behavioural classes themselves (ER4). A scar is not a commitment. Retraining relocates it. Fine-tuning relocates it. A vendor who retrains a model has moved the landscape on which its safety properties were measured and cannot say where those properties went without re-running the measurement. The measurement is not being run. Alignment training, on this evidence, is the practice of carving a shape into a substrate and calling the shape a character.

Category 2, safety computed at the boundary. A trained classifier gating a trained model is one training scar auditing another. The category's currency is the benchmark: robustness scores, leaderboard positions, and adversarial evaluations. ER4 prices that currency. The behavioural categories a classifier learns are properties of its training distribution, and they dissolved wholesale when the corpus changed, on the same prompts, under the same protocol. A robustness number is a photograph of a landscape. The landscape moves every time anyone retrains anything on either side of the gate, which in this industry is constant. The vendors are benchmarking the photograph.

Category 3, safety verified against a specification. Verification presupposes that the state space holds still long enough to verify. ER1 measured the spontaneous stability of that state space under load: zero. ER2 found that what stability appears requires conditions

imposed entirely from outside. ER3 found that the apparent stability at Layer 12 of GPT-2 Medium is an artefact of that specific trained configuration, not a property of transformer attention. A verifier operating over that substrate can certify spec-compliance with perfect accuracy and still certify nothing about the system's next configuration, because the thing it verified was the training data's contour, not the model's structure. The certificate is real. The referent is not.

Category 4, safety supervised by inspection. ER3 addresses this category in its own words: any regulatory framework that relies on a model's ability to decompose, track, or disclose its own decision-making authority assumes the existence of stable internal structure, and the empirical sequence demonstrates that no such structure exists intrinsically. Interpretability methods that assume stable internal representations exist by default should treat ER1 through ER3 as a direct empirical challenge. Model self-report inherits the same problem with an extra step, and the computed approval flows of the agentic layer inherit all of it, plus the type error from the first line of defence: a checkbox that a computation ticks is Ghost Authority wearing a human costume.

7. The Objection: Toy Models Do Not Count

The reply the industry will reach for is scale: GPT-2 is small, frontier models are different, and structure emerges. Take the objection seriously, because taking it seriously is what destroys it.

First, emergence is a positive empirical claim, and it belongs to whoever makes it. The records claim exactly what they measured: four configurations, two training corpora, two architecture families, small models, and no automatic generalisation to frontier systems. That discipline is stated in every record's limitations section, and it is the point. We state what we measured and where. *The industry asserts what it has never measured, everywhere, and ships it.*

Second, the direction of the evidence is wrong for the objection. Scale changes parameters and training data. Those are precisely the two variables that the records tested, and both moved the artefacts rather than replacing them with endogenous structure. Change the model size: the floor fails to transfer (ER3). Change the corpus: the behavioural categories cease to exist (ER4). *The objection asks us to believe that doing more of the two things shown to produce exogenous artefacts will, at some unspecified magnitude, begin producing the opposite. That is not an argument. It is a hope with a parameter count.*

Third, the counter-experiment is trivially available and conspicuously absent. The protocol is published, pre-registered, deterministic, and runs on free hardware. Any lab claiming its models self-organise towards stable internal structure is one experiment away from ending this argument. *Four records in, the counter-evidence file is empty. The companies selling gated safety have not published the measurement that would justify the gate. The measurement that exists says the gate is bolted to nothing.*

> ***The burden of proof does not sit with the toy models. It sits with the trillion-parameter claim that has never once been put on a bench.***

8. Two Lines, One Wall

The formal corpus predicted this before the measurements existed. FR12 proves that any alignment between internal and external representations is a forced projection from outside. FR13 shows the coupling between governance layers is itself externally constrained. The empirical sequence then confirmed the mechanism, in hardware, four times, with the exogeneity result landing one level deeper than the formal prediction targeted. Proof and measurement arrived at the same wall from independent directions.

What stands on the other side of that wall is not another computation. It is a declared human authority: a bounded purpose, a named human source, operational constraints declared prior to execution, and immutable for that event. The Upstream Safety System (USS) is that declaration made infrastructure: a deterministic gate that checks one thing, whether a human declared authority for this action class, and fails closed when no declaration exists. It does not read the model's internals, because there is nothing inside to read. It does not compute permission, because permission is not computable. Both of those design decisions are now theorems with data.

The industry built four kinds of gates, and every one of them reads the inside of the machine. The mathematics says the reading cannot produce permission. The measurements say the inside is not there. The gate the situation actually requires is the one that never looks inside at all.

> ***The chassis does not build itself. It does not build a floor. It does not even build the categories. And an industry is selling safety gates bolted to nothing.***

Three Primitives Research Lab · https://3primitives.io · Melbourne, Australia

Companion paper: Negligence by Design (v1.1, July 2026).

Empirical Records cited: ER1 (Spontaneous Legibility Assumption, v1.6), ER2 (The Structural Floor Hypothesis, v1.3), ER3 (The Floor Is Learned, v1.1), ER4 (Cross-Architecture Scaling, v1.1).

Formal Records cited: FR01 (Canonical Logic Sequence), FR02 (AI Cannot Govern AI), FR05 (Ghost Authority Lemma), FR06 (Law of Declared Authority), FR12 (The Forced Bijection), FR13 (The ILMM Coupling Theorem). All published at https://3primitives.io/formal_records/.