Why Correctness is not Cognition: From Benchmark Accuracy to Autonomous Epistemic Governance of Large Language Models

Tim Hayes
AIDA Research
Cambridge, United Kingdom
tim@aidaresearch.ai

March 2026

Abstract

Large language models in safety-critical settings are evaluated almost exclusively by outcome accuracy, which mixes genuine structural knowledge with shallow pattern-matching and provides limited basis for trustworthy autonomous deployment.

We introduce three instruments—the Epistemic Trajectory Classifier (ETC), Augmented Structured Cognition through Observational Lensing (ASCOL), and the Factual Elimination Stress Test (FEST)—that decompose aggregate accuracy into six epistemic regimes with distinct geometric patterns. Carrier–content decomposition refines these six regimes into five epistemic regions — Genuine Knowledge, Carrier-Assisted, Carrier-Suppressed, Content-Confused, and Genuinely Unknowable — separating robust knowledge from positional artefact within correct outputs, and recoverable hidden knowledge from genuine ignorance within incorrect ones. Carrier decomposition produces an epistemic gap in which structural correctness exceeds outcome accuracy — the model knows more than it delivers. At the autonomous level, regime-aware ensemble arbitration partitions every answer into trust tiers with calibrated reliability bounds and irreducible hidden error rates stated explicitly rather than concealed within aggregate performance figures.

Trajectory analysis reveals that the probability dynamics observed through the logit lens across transformer layers are *rotations* in high-dimensional representational space, not amplification events. This finding leads to a carrier–content decomposition: every model's output comprises a position-dependent carrier signal (a property of the frozen weight matrices) and a content signal (the model's actual knowledge), with accuracy differentials exceeding 20 percentage points between answer positions.

Inference-time correction exploiting this decomposition recovers suppressed knowledge without training or parameter modification. Applied to Llama-3-8B on 1,089 medical licensing questions, the correction raises accuracy from 67.9% to a current production high-water mark of 71.1%, with the pipeline under active development.

The Factual Elimination Stress Test (FEST), extended to all four gold answer classes (A, B, C, D) and applied to 274 failures, establishes that 260 (94.9%) are architecturally recoverable and only 14 (1.29%) represent genuine knowledge gaps. The true knowledge ceiling is therefore **98.71%** (1,075 of 1,089 samples). The gap between the 67.9% baseline and the 98.71% ceiling is not a measure of what the model does not know. It is a measure of how severely the inference architecture prevents the model's knowledge from reaching the

output.

The framework is validated across nine models (7B–14B parameters) from six suppliers, producing 3,246,256 attention probes and 6,411,576 layer probes across 283,289 inferences. We map the measurement hierarchy onto the obligations of the European Artificial Intelligence Act, demonstrating that these instruments provide the technical infrastructure for GPAI compliance that the regulation anticipates but the industry does not yet fully supply.

Keywords: epistemic safety, large language models, model evaluation, interpretability, AI governance, epistemic manifold, trajectory classification, carrier signal, positional bias, inference-time correction

Authorship & Acknowledgement. The research reported in this paper is the work of Tim Hayes. The paper has been authored by Tim Hayes with research assistance from Anthropic's Claude Opus and Microsoft's Copilot.

Conflict of Interest & Patent Declaration. Tim Hayes has a commercial interest as the named inventor of instruments described and used in the course of this research, parts of which are the subject of a preliminary patent filing in the United Kingdom.

1 Introduction

A person can guess correctly. A person can answer correctly while panicking. A person can answer incorrectly while calm and thoughtful. Correctness of the answer does not reveal the condition of the mind that produced it. The same is true for machine intelligence.

Language models can be right for the wrong reasons—especially if we look only at the final answer. They can be wrong with great confidence. They can hesitate, vacillate, or fall into patterns of confusion that are invisible when only the final answer is examined. What has been lacking is a way to measure the internal state of these machines during inference, and—crucially—a way to act on those measurements when no external answer key exists.

This work introduces a method of collecting and measuring the dynamics of machine intelligence processes. Analogous in spirit to an ECG, we read a variety of vital signs that together form a geometric model of the machine's epistemic state. We refer to this as the measurement of *epistemic safety*. But measurement alone is insufficient. A diagnostic instrument that functions only when the correct answer is already known provides calibration, not governance. The critical question is: what can be said about an answer when no one in the room knows whether it is right?

This paper goes further than diagnostics. Measurement of the model's internal geometry reveals not only *that* models fail but *why*: a structural interference pattern—inherent in every transformer's output projection—systematically distorts the relationship between what the model knows and what it can express. This distortion is measurable, decomposable, and correctable at inference time without any modification to model parameters. The model is smarter than its benchmark score. The instruments developed here prove it and unlock the hidden knowledge.

1.1 Four Levels of Model Assessment

The prevailing paradigm evaluates language models at a single level: aggregate accuracy on benchmark questions with known answers. This paper establishes that meaningful assessment requires four distinct levels, each making progressively stronger operational claims while being progressively more honest about their limitations.

Level A1: Benchmark Accuracy. The model is presented with questions whose answers are known, and the percentage of correct responses is reported. This is what the field typically measures. It treats every correct answer as epistemically equivalent—a Differentiated Correct answer (produced through genuine structural knowledge) is indistinguishable from a Fused Gold answer (correct by statistical chance). It is analogous to certifying a pilot based on the percentage of flights that landed safely, without distinguishing calm-weather landings from those where the instruments failed and the aircraft happened to drift onto the runway. Level A1 is extremely limited: it makes no claim about which answers are trustworthy, offers no partition by processing quality, and provides no basis for deployment decisions in safety-critical domains.

Level A2: Epistemic Calibration. The model is assessed with known answers, but each answer is classified by the quality of the internal process that produced it. Six epistemic regimes—from Differentiated Correct (genuine structural knowledge) through Late Crystallisation (correct but shallow) to Fused Wrong (complete epistemic failure)—partition the model's performance

into qualitatively different categories. This is where the epistemic gap can be measured: the difference between surface accuracy and structural correctness. Level A2 characterises the model. It reveals that a model scoring 77.0% accuracy may achieve 87.4% structural correctness—with the model structurally knowing the correct answer in 113 more cases than it successfully delivers, yielding an inverted epistemic gap of −10.4 percentage points—that instruction tuning can narrow this inverted gap while also improving accuracy, and that a smaller model from a different vendor can be epistemically superior despite lower headline accuracy. These findings are invisible at Level A1 and decisive for deployment decisions. The limitation of Level A2 is that it requires gold answers. It calibrates the instrument but cannot operate the instrument at deployment time.

Level B1: Autonomous Single-Model Assessment. The model produces answers to questions whose correct answers are unknown. The ETC classifies each answer into one of four observable processing categories using only activation geometry and logit trajectories—no gold answer required. The six post-hoc regimes collapse into four blind-observable categories:

***Differentiated*:** geometry separated, logits sharp, views concordant. The highest-quality processing observable without gold. Calibrated accuracy from Level A2 provides the reliability rate for this category.

***Late Rescue*:** geometry fused, logits sharp, views discordant. Correct on calibration data but structurally shallow and sensitive to prompt variation.

***Override*:** leading answer displaced in late layers. Strong negative signal regardless of which answer survives.

***Fused*:** no meaningful differentiation in either view. The model cannot distinguish alternatives.

The critical limitation of Level B1 is the Differentiated Wrong rate—answers that show clean structural processing but converge on the wrong answer. This rate is the model's irreducible hidden lie rate for clean processing: the percentage of answers that look healthy but are wrong, undetectable without gold. For the models assessed in this work, the Differentiated Wrong rate ranges from 11.5% to 18.7% of all Differentiated answers. Level B1 can reject structurally compromised answers with high confidence, but it cannot guarantee that structurally clean answers are correct. Like Level B2, the accuracy rates cited at Level B1 are not intrinsic to the autonomous assessment—they are inherited from Level A2 calibration. The regime can be identified without gold; the accuracy rate associated with that regime cannot.

Level B2: Autonomous Ensemble Governance. Multiple models produce answers to the same question, each classified into processing categories at Level B1. When multiple models independently show Differentiated processing and converge on the same answer, the compound reliability exceeds any single model's rate. In the three-model ensemble assessed in this work, unanimous Differentiated agreement yields 91.7% accuracy—reducing the hidden lie rate from the single-model range of 11–19% to 8.3%. The correlation structure of the ensemble, measured through Laplacian eigendecomposition of inter-model agreement, determines how much independence exists between models and therefore how much the compound reliability can improve with additional ensemble members.

Level B2 is not self-sufficient. Every accuracy rate cited at this level—the 91.7% compound

reliability, the 8.3% hidden lie rate, the per-regime accuracy for each model—is inherited from Level A2 calibration. Each model in the ensemble must first undergo full epistemic evaluation with known answers to establish its regime-specific accuracy rates, its Differentiated Wrong rate, and its correlation structure with other ensemble members. Without A2 calibration, B2 has no basis for its compound reliability claims. The autonomous ensemble does not discover its own trustworthiness—it operates within bounds established by prior calibration. This dependency is fundamental: B2 governance is only as reliable as the A2 calibration that underpins it, and any change to a model's weights, quantisation, or fine-tuning requires recalibration before the ensemble's accuracy rates remain valid.

The honest claim at Level B2 is not "we can tell you which answers are correct." It is: "we can tell you that this answer was produced through clean structural processing by three independent models that agree, and based on calibration of each model at Level A2, the compound accuracy for that configuration is 91.7%. We can also tell you definitively which answers should not be trusted."

Beyond the four levels: mechanistic discovery. The measurement hierarchy described above was the starting point of this work. The instruments developed to implement it—particularly the ETC's dual-view trajectory analysis—revealed a deeper phenomenon that the hierarchy alone could not have anticipated. The layer-by-layer trajectories that classify outputs into six regimes also expose the *mechanism* that produces those regimes: a rotational geometry in the model's representational space, in which a structural carrier signal competes with the model's content signal for alignment with the output projection. This mechanistic discovery enables a refinement of the six regimes into five epistemic regions that distinguish robust knowledge from positional luck, and hidden knowledge from genuine ignorance. It also enables inference-time correction that recovers suppressed knowledge without training. The FEST battery across all four gold classes establishes that 94.9% of all failures are architecturally recoverable and only 1.29% represent genuine knowledge gaps, placing the true knowledge ceiling at 98.71%. The four-level hierarchy measures and governs the model's epistemic state, and the mechanistic discovery provides the basis for explaining and correcting it. Together, they form a comprehensive framework spanning measurement, mechanism, and governance.

1.2 From Calibration to Governance

The distinction between Levels A and B is the distinction between calibration and governance. Calibration (Levels A1, A2) characterises the model using known answers. Governance (Levels B1, B2) operates the model using the calibration data as a reference. The transition from A to B is the transition from research to deployment—from "how good is this model?" to "should I trust this particular answer?"

This transition exposes a fundamental asymmetry in what can be detected without gold answers. Structurally compromised answers—those produced through fusion, override, or late rescue—are identifiable from activation geometry alone. They can be flagged, rejected, or escalated to human review without knowing the correct answer. Structurally clean answers—those showing genuine differentiation—cannot be individually verified. The Differentiated Wrong rate is measurable in calibration but invisible in deployment. This is the model's trustworthiness limit: the percentage of clean-looking answers that are silently incorrect.

Two further findings sharpen this asymmetry. First, within the Differentiated population, the trajectories of correct and incorrect answers diverge from mid-depth layers onwards. Differentiated Correct answers show stronger entropy sharpening and margin development in the second half of the network. This separation is consistent across layers and models but insufficiently strong to serve as a per-question classifier. It is a population-level signal that may, with further calibration, refine the hidden lie rate into sub-populations with different reliability. Second, within the Fused population, correct and incorrect answers are geometrically indistinguishable at every layer. Fused Gold and Fused Wrong show no consistent separation on any metric. The activation space carries no information about correctness when the model has failed to differentiate. Fused answers must be rejected categorically—there is no rescue path through the geometry.

The carrier–content decomposition reported in later sections of this paper provides a third, deeper refinement. A portion of the Differentiated Wrong population is not genuinely wrong—it is carrier-suppressed: the model possesses the correct knowledge but a structural interference pattern in the output projection prevents it from being expressed. These outputs can be recovered at inference time through geometric correction. Conversely, a portion of the Differentiated Correct population is carrier-assisted: correct only because the answer happened to occupy a geometrically favoured position. These outputs are fragile and should not be treated as robust knowledge. The carrier–content decomposition refines the trustworthiness ceiling by separating genuine knowledge from positional artefact within both the correct and incorrect populations.

1.3 Instruments

The framework presented here comprises three core measurement instruments and a suite of correction and governance tools:

ETC — Epistemic Trajectory Classifier. The overarching assessment framework. The ETC reconstructs the layer-wise trajectory of internal representations across the full transformer depth, classifying each model–question pair into one of six epistemic regimes (Level A2) or four autonomous categories (Level B1) using dual geometric and logit views of the hidden states.

ASCOL — Augmented Structured Cognition through Observational Lensing. The per-sample diagnostic instrument that measures the structural integrity of a model's knowledge representation through multi-template probing, distinguishing robust knowledge from brittle surface-pattern dependence by accessing the same knowledge through different projection angles.

FEST — Factual Elimination Stress Test. A perturbation protocol that systematically removes and recombines answer options across nine configurations to measure how dependent the model is on distractor context.

In addition to the measurement instruments, this work introduces inference-time correction methods that exploit the carrier–content decomposition to recover suppressed knowledge from the model's existing weights. These methods—comprising carrier correction, content signal enhancement, and multi-pass arbitration—require no modification to model parameters, no fine-training, and no architectural changes. The ETC provides the diagnostic layer; the correction tools provide the interventional layer; and the Epistemic Passport records the carrier–content ratio for each output as a certifiable quantity for auditors and regulators.

1.4 The Measurement Hierarchy in Practice

Central to this work is the discovery of a series of mathematical and geometric invariants—natural constants—that provide a multidimensional window through which the state of the machine can be inferred both during and after it produces answers to complex questions. From that starting point, it becomes possible to move beyond saying "75% of this machine's answers are correct, but I cannot tell you which ones." Instead, we can say: "75% of the questions posed to this machine will be answered correctly. Of those correct answers, we can tell you which were produced through genuine structural knowledge and which arrived through shallow or accidental processing. When deployed autonomously—without access to gold answers—we can partition every answer into trust tiers with calibrated accuracy rates, enabling practitioners to make informed decisions about which answers to trust, which to verify, and which to discard."

At Level A1, 75% accuracy means one in four answers is wrong, with no indication of which ones. At Level A2, that 75% is revealed to comprise 50% genuine structural knowledge, 25% shallow or accidental correctness, and approximately a 25-point epistemic gap. At Level B1, every answer is partitioned into trust tiers: 61% in the highest tier at 81% calibrated reliability, 25% correct but structurally fragile, and 14% flagged for rejection. At Level B2, answers validated by ensemble agreement through clean processing reach 92% reliability—and the 8% hidden lie rate is stated honestly rather than concealed within an aggregate percentage.

The carrier–content decomposition adds a further dimension to this practice. A model that "scores 67.9%" and would conventionally be judged to need extensive fine-training to reach 90% can in principle be brought to 98.71% through inference architecture corrections alone — the FEST battery establishes that only 1.29% of samples represent genuine knowledge gaps, while 94.9% of failures contain a recoverable content signal. The measurement hierarchy reveals what the model knows and doesn't know; the mechanistic discovery establishes that much of what the model appears not to know is knowledge it possesses but cannot express. The instruments provide the basis for mitigating this distortion.

The implications are broad. For clinicians, it means knowing which answers were produced through clean structural processing and what the calibrated reliability of that processing is. For regulators, it provides evidence relevant for conformity assessment under the EU AI Act's GPAI obligations. For developers, it offers a map of where models reason cleanly and where they fail—and which of those failures are correctable without training. For society, it offers a path toward AI systems whose trustworthiness is quantified, partitioned, and honestly reported—not as a single percentage, but as a structured account of what the machine knows, how it knows it, and how much that knowledge can be trusted when no one is holding the answer key.

1.5 Models, Data, and Hardware

Nine transformer language models were assessed across the programme, spanning six suppliers, four architecture families, and parameter counts from 7B to 14B:

The nine models divide into two groups. Eight underwent full ASCOL and FEST assessment, producing 3,246,256 attention probes and 6,411,576 layer probes across 283,289 inferences: Llama-8B, Mistral-7B, Qwen-7B, OLMo-7B, Gemma-9B, Ministral-14B Base, Ministral-14B Instruct, and Ministral-14B Reasoning. DeepSeek-R1-14B was assessed for carrier–content

Table 1: Models assessed. All models obtained from Hugging Face.

Model	Supplier	Parameters	Training
Meta-Llama-3-8B	Meta	8B	Base (pre-trained)
Mistral-7B-v0.3	Mistral AI	7B	Base (pre-trained)
Qwen2.5-7B	Alibaba	7B	Base (pre-trained)
OLMo-2-7B	AI2	7B	Base (pre-trained)
Gemma-2-9B	Google DeepMind	9B	Base (pre-trained)
DeepSeek-R1-14B	DeepSeek	14B	Reasoning-tuned[†]
Ministral-14B Base	Mistral AI	14B	Base (pre-trained)
Ministral-14B Instruct	Mistral AI	14B	Instruction-tuned
Ministral-14B Reasoning	Mistral AI	14B	Reasoning-tuned

[†]DeepSeek-R1-14B was assessed for carrier–content decomposition validation only; it did not undergo full ASCOL and FEST assessment.

decomposition validation only. The carrier analysis (Sections 5–7) was conducted on a six-model subset spanning four architecture families and six suppliers: Llama-8B, Mistral-7B, Qwen-7B, OLMo-7B, Gemma-9B, and DeepSeek-R1-14B.

Four evaluation datasets were used:

Table 2: Evaluation datasets.

Dataset	Samples	Domain
MMLU-Med	1,089	Medical licensing (4-option MCQ)
MedQA	10,178	Medical licensing (4-option MCQ)
MMLU-Pro	∼12,000	Multi-domain (7–10-option MCQ)
MNLI	∼24,000	Natural language inference

Hardware. All inference and analysis was conducted on four NVIDIA DGX Spark GB10 Grace-Blackwell systems, each equipped with 128 GB shared GPU/CPU memory, interconnected in a 200 Gb/s ring topology. Database functionality was provided by a Dell Precision R270 server running MySQL. No cloud compute was used; all processing was performed on-premises.

Reproducibility. The ETC, ASCOL, and FEST instruments are the subject of a UK preliminary patent filing and are not available for public release at this time. The models assessed are publicly available from HuggingFace under their respective licences (Table 1). The evaluation datasets — MMLU-Med, MedQA, MMLU-Pro, and MNLI — are publicly available. The carrier direction computation is a closed-form derivation from the lm_head weight matrix, described fully in Section 6.2, and can be replicated from the model weights alone without access to the AIDA pipeline. Independent replication of the core carrier-decomposition finding does not require the full instrumentation suite. The authors welcome correspondence regarding the methodology. Data-sharing requests will be considered on a case-by-case basis following patent grant.

1.6 Organisation

The remainder of this paper is organised as follows:

Section 2 surveys six threads of related work and identifies the specific gap each leaves open.

Section 3 motivates epistemic governance, presents the empirically stable invariants of the epistemic manifold, introduces the six trajectory regimes and five epistemic regions, and details the seven contributions.

Section 4 presents the empirical case study: cross-vendor calibration at Levels A1 and A2 on 1,089 medical licensing questions.

Section 5 reports the rotation discovery and the identification of the carrier signal as a pervasive property observed across transformer inference.

Section 6 presents the architectural analysis: the three structural layers (causal mask asymmetry, tied-weight projection geometry, and RoPE-baked positional bias) that produce the carrier signal, and explains why fine-tuning cannot resolve them.

Section 7 presents the carrier–content decomposition and the experimental programme recovering hidden knowledge at inference time.

Section 8 articulates the five regions as a governance framework, training economics, chat-mode governance, and the runtime architecture.

Section 9 presents autonomous assessment at Levels B1 and B2.

Section 10 describes the translation to clinical decision support.

Section 11 maps the framework onto the European Artificial Intelligence Act.

Section 12 draws conclusions.

Appendices A–E reproduce full AIDA assessment reports.

2 Literature Review and Related Work

2.1 Scope and Organisation

This section surveys the research landscape against which the contributions of this work are positioned. The field of large language model evaluation has grown rapidly since the introduction of transformer-based architectures, yet it remains fragmented across communities that rarely communicate with one another. Knowledge probing researchers measure whether models store facts; shortcut learning researchers demonstrate when high accuracy conceals brittle strategies; mechanistic interpretability researchers reveal what internal representations look like; calibration researchers quantify how much to trust a model's confidence; parameter-efficient fine-tuning researchers optimise how to adapt models cheaply; and governance researchers specify what must be demonstrated before deployment. Each community has produced substantial results. None has connected all six threads into a single diagnostic and prescriptive framework.

The present work occupies the intersection of these six threads. Its central claim—that a correct answer may encode qualitatively different epistemic states (fused knowledge versus rote knowledge), and that this distinction matters for governance, curriculum design, and deployment—requires engagement with all six bodies of work. Accordingly, the review is organised thematically. Sections 2.2 through 2.7 survey each thread in turn, identifying the specific gap that this work

addresses. Section 2.8 synthesises the threads and positions the epistemic manifold as a unifying mathematical framework.

2.2 Knowledge Probing and Evaluation Benchmarks

The question of what language models "know" has been central to NLP research since Petroni et al. (2019) introduced the LAMA probe, demonstrating that BERT could answer factual cloze queries—such as "The capital of France is [MASK]"—with surprising accuracy, often rivalling supervised knowledge extraction systems. The LAMA probe established a paradigm: treat the model as a knowledge base and measure recall through carefully constructed prompts. Subsequent work extended this paradigm to larger models and broader domains, culminating in comprehensive benchmarks such as the Massive Multitask Language Understanding benchmark (MMLU; Hendrycks et al., 2021), which evaluates models across 57 subjects ranging from elementary mathematics to professional law and medicine.

MMLU became the de facto standard for comparing language model capabilities, but its utility has been progressively undermined by what the evaluation community terms *benchmark saturation*: state-of-the-art models now routinely exceed 90% accuracy on MMLU, eliminating meaningful differentiation among frontier systems (Phan et al., 2025). Data contamination compounds the problem—models trained on web-scale corpora inevitably encounter benchmark questions during pre-training, inflating scores beyond what generalisation alone would produce. The response has been a proliferation of harder benchmarks. Humanity's Last Exam (HLE; Phan et al., 2025), developed by the Center for AI Safety and Scale AI, represents the most ambitious attempt to date: 2,500 expert-level questions across dozens of subjects, crowdsourced from over 500 institutions worldwide, with each question designed to resist trivial internet retrieval. As of early 2025, the highest-scoring frontier models achieved under 37% accuracy on HLE, with systematic overconfidence—calibration errors exceeding 80%—suggesting that high scores on saturated benchmarks mask fundamental deficits in genuine understanding.

At a finer granularity, Dai et al. (2022) identified "knowledge neurons" within pre-trained transformers—specific neurons in feed-forward layers whose activation is both necessary and sufficient for correct factual recall. This work established that factual knowledge is not diffusely distributed but is at least partially localisable within the network architecture, providing an empirical basis for the claim that different training procedures might produce architecturally different encodings of the same fact.

2.2.1 Gap addressed by this work

The knowledge probing literature asks a binary question: does the model know fact X or not? This work reframes the question as qualitative: *how* does the model know X? The ASCOL and FEST instruments distinguish fused knowledge (robust, contextually integrated, perturbation-resistant) from rote knowledge (brittle, surface-pattern-dependent, perturbation-fragile) even when both produce identical correct answers on standard benchmarks. The benchmark saturation problem—exemplified by MMLU's ceiling and the motivation behind HLE—is not merely a measurement artefact but a symptom of evaluation frameworks that treat all correct answers as epistemically equivalent. AIDA provides the diagnostic layer that accuracy-only benchmarks structurally cannot.

2.3 Shortcut Learning and Spurious Correlations

Geirhos et al. (2020) provided the definitive taxonomy of shortcut learning in deep neural networks, demonstrating that models frequently exploit spurious correlations—classifying cows by the presence of grass rather than bovine morphology, or identifying pneumonia from hospital-specific imaging artefacts rather than pathological features. Their framework distinguishes *intended* decision rules from *shortcut* decision rules: both achieve high in-distribution accuracy, but only the former generalises under distribution shift. The taxonomy is directly relevant to language models, where identical surface-level behaviours may mask fundamentally different computational strategies.

In NLP, McCoy et al. (2019) demonstrated that BERT's apparently strong performance on natural language inference could be attributed to shallow heuristics—lexical overlap, subsequence matching, and constituent structure—rather than genuine reasoning. Their HANS dataset, designed to disentangle these heuristics, showed that models achieving over 90% accuracy on standard NLI benchmarks dropped to near-chance on HANS. Gururangan et al. (2018) identified pervasive annotation artefacts in NLI datasets, showing that hypothesis-only baselines could achieve well above chance accuracy on SNLI and MultiNLI, indicating that the benchmarks themselves encoded exploitable statistical regularities.

Ribeiro et al. (2020) introduced CheckList, a task-agnostic methodology for behavioural testing of NLP models inspired by software engineering practices. CheckList tests models against minimum functionality tests, invariance tests, and directional expectation tests, revealing systematic failures in commercial NLP systems that aggregate metrics had obscured. The CheckList philosophy—testing specific capabilities rather than measuring aggregate performance—aligns closely with the per-sample diagnostic approach adopted in this work.

2.3.1 Gap addressed by this work

The shortcut learning literature identifies the existence of brittle strategies and demonstrates their prevalence through targeted evaluation. What it does not provide is a per-sample quantitative instrument that measures the degree of shortcut reliance and prescribes corrective action. The ASCOL score quantifies, for each individual training sample, the extent to which a model's correct answer depends on surface cues rather than robust feature integration. The FEST runner extends this to fragility measurement under systematic perturbation. Together, they transform shortcut detection from a binary diagnostic ("shortcuts exist") into a continuous prescriptive instrument ("this sample has ASCOL = 0.73 and inverted fragility under paraphrase—prioritise it for retraining").

2.4 Mechanistic Interpretability

Mechanistic interpretability seeks to reverse-engineer the internal computations of neural networks into human-understandable terms. The field has progressed through several methodological waves. The logit lens (nostalgebraist, 2020) provided an early tool for inspecting intermediate representations by projecting hidden states at each layer directly into vocabulary space, revealing how predictions evolve through the network. Belrose et al. (2023) refined this approach with the tuned lens, training lightweight affine probes at each layer to correct for the distributional shift between intermediate and final representations, achieving substantially higher fidelity.

Probing classifiers (Belinkov, 2022) take a complementary approach, training auxiliary classifiers on frozen internal representations to detect the presence of specific linguistic or factual features. Geva et al. (2021) demonstrated that feed-forward layers in transformers function as key–value memories, with keys encoding input patterns and values encoding associated output distributions, providing a mechanistic account of how factual knowledge is stored and retrieved. Meng et al. (2022) built on this understanding with ROME (Rank-One Model Editing), demonstrating that specific factual associations can be surgically modified by targeted rank-one updates to feed-forward weights.

The field underwent a step change in 2025 with Anthropic's introduction of circuit tracing via attribution graphs (Ameisen et al., 2025; Lindsey et al., 2025). This work replaced the neuron—long recognised as an inadequate unit of analysis due to polysemanticity—with features extracted by cross-layer transcoders, sparse coding models that decompose activations into interpretable, sparsely active components. By tracing the causal flow between features across layers, the attribution graph method produces graph-structured descriptions of the computational steps a model uses to transform a specific input into an output. Concurrently, Marks et al. (2025) introduced sparse feature circuits at ICLR 2025, combining sparse autoencoders with scalable circuit discovery to identify the specific features causally responsible for model behaviours, and demonstrating surgical debiasing through their SHIFT technique.

These advances notwithstanding, significant limitations remain. Ameisen et al. (2025) note that attribution graphs provide satisfying insight for approximately one quarter of prompts examined, and the replacement model on which the method depends may use different mechanisms from the original. The field has yet to produce a method that scales to the full complexity of frontier models while maintaining interpretability guarantees.

2.4.1 Gap addressed by this work

Mechanistic interpretability reveals the internal structure of computations but does not connect those structures to governance-relevant decisions about model fitness. The trajectory classifier developed in this work bridges this gap: it uses the representational structure visible through probing and trajectory analysis to classify each sample into epistemic categories (fused, rote, near-miss, confused) that have direct implications for deployment risk and curriculum design. Where circuit tracing asks "what computational steps produced this output?", the epistemic manifold asks "what quality of knowledge does this computational structure represent, and what should be done about it?"

2.5 Calibration and Uncertainty Estimation

Model calibration—the alignment between predicted probabilities and empirical frequencies—has been a persistent concern since Guo et al. (2017) demonstrated that modern deep neural networks, despite their accuracy, are poorly calibrated, producing overconfident predictions that undermine decision-making in safety-critical domains. Their proposed remedy, temperature scaling, applies a single learned scalar to the logit vector before softmax, substantially reducing expected calibration error (ECE) without affecting accuracy. Temperature scaling remains the most widely used post-hoc calibration method.

Conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2023) offers a distribution-free

alternative that provides finite-sample coverage guarantees: rather than calibrating individual probabilities, conformal methods construct prediction sets guaranteed to contain the true label with user-specified probability. This approach has attracted growing attention in LLM deployment contexts where point predictions are insufficient and users require rigorous uncertainty quantification.

The calibration problem has acquired fresh urgency in the context of benchmark saturation. Phan et al. (2025) explicitly measure calibration alongside accuracy in HLE, revealing systematic high calibration errors exceeding 80% paired with low accuracy—strong evidence for what the authors describe as systematic confabulation. Models assign near-maximal confidence to answers that are frequently wrong, making raw confidence scores unreliable indicators of knowledge quality.

2.5.1 Gap addressed by this work

Calibration measures whether a model's confidence aligns with its accuracy. It does not measure whether a correct, well-calibrated answer reflects genuine understanding or memorised association. A model that has rote-memorised a fact will produce high confidence and high accuracy on that specific prompt format, yielding perfect calibration on in-distribution data—yet will fail catastrophically under paraphrase. The epistemic classification developed in this work operates on a different axis entirely: it measures knowledge quality rather than confidence accuracy. ASCOL and FEST diagnose the structural integrity of the model's internal representation, which calibration methods cannot access.

2.6 Parameter-Efficient Fine-Tuning

The cost of full fine-tuning for large language models has motivated a rich literature on parameter-efficient alternatives. Houlsby et al. (2019) introduced adapter modules—small bottleneck layers inserted between existing transformer blocks—demonstrating that freezing the original weights and training only the adapters could achieve near-parity with full fine-tuning at a fraction of the parameter cost. Li and Liang (2021) proposed prefix tuning, which prepends trainable continuous vectors to the key and value matrices at each attention layer, achieving comparable performance with even fewer trainable parameters.

Hu et al. (2022) introduced Low-Rank Adaptation (LoRA), which has become the dominant PEFT method. LoRA freezes the pre-trained weight matrices and injects trainable low-rank decomposition matrices (A and B) in parallel, so that the effective weight update $\Delta W = BA$ has rank at most r, where r is a small hyperparameter (typically 4–64). At inference time, the low-rank update can be merged into the original weights, incurring zero additional latency. Dettmers et al. (2023) combined LoRA with 4-bit quantisation of the frozen base weights (QLoRA), enabling fine-tuning of 65-billion-parameter models on a single GPU while maintaining competitive performance.

The LoRA family continues to evolve. Liu et al. (2024) introduced Weight-Decomposed Low-Rank Adaptation (DoRA), which decomposes pre-trained weights into magnitude and directional components, applying LoRA only to the directional update. DoRA consistently outperforms standard LoRA across tasks including commonsense reasoning, visual instruction tuning, and text-to-image generation. Kalajdzievski (2025) proposed RandLoRA at ICLR 2025, demonstrating

that full-rank updates can be achieved through learned linear combinations of low-rank, non-trainable random matrices—significantly reducing the performance gap between LoRA and full fine-tuning.

The PEFT literature has focused overwhelmingly on efficiency—reducing parameters, memory, and compute—and on accuracy—matching or approaching full fine-tuning on aggregate benchmarks. The implicit assumption is that a method achieving equivalent accuracy produces equivalent knowledge. This assumption, as the empirical results of this work demonstrate, is false.

2.6.1 Gap addressed by this work

The PEFT literature's implicit assumption—that equivalent accuracy implies equivalent knowledge—is directly testable with the ASCOL and FEST instruments. Preliminary analysis suggests that LoRA fine-tuning and full fine-tuning, at identical accuracy levels, may produce qualitatively different knowledge profiles as measured by ASCOL and FEST. Initial results indicate that LoRA adapters tend toward more robust, perturbation-resistant knowledge representations while full fine-tuning may produce more brittle, surface-pattern-dependent outputs — a difference in kind rather than degree. Full characterisation of this finding across quantisation levels and adapter ranks is the subject of ongoing work and will be reported separately.

2.7 AI Governance and Regulatory Landscape

The regulatory landscape for artificial intelligence has matured substantially since the adoption of the EU Artificial Intelligence Act (European Parliament, 2024), the world's first comprehensive AI regulation. The Act establishes a risk-based classification system: unacceptable-risk AI (such as social scoring) is prohibited; high-risk AI (in domains including education, employment, law enforcement, and critical infrastructure) must meet extensive conformity assessment requirements; limited-risk AI faces transparency obligations; and minimal-risk AI remains largely unregulated. For high-risk systems, the Act mandates technical documentation, risk management systems, data governance, human oversight provisions, and post-market monitoring—requirements that implicitly assume the availability of measurement instruments capable of producing the evidence these obligations demand.

General-purpose AI (GPAI) models—a category encompassing foundation models and large language models—receive dedicated treatment under the Act. As of 2 August 2025, GPAI obligations became enforceable (European Commission, 2025). Models classified as posing systemic risk (currently those trained with computational resources exceeding 10^{25} floating-point operations) face additional obligations for adversarial testing, incident reporting, and model evaluation. The penalty regime—with fines of up to 3% of global annual turnover—transforms these requirements from aspirational standards into enforceable mandates.

The United States National Institute of Standards and Technology published its AI Risk Management Framework (AI RMF 1.0; National Institute of Standards and Technology, 2023) in January 2023, organised around four functions—Govern, Map, Measure, and Manage. The Measure function explicitly calls for quantitative assessment of AI system characteristics including validity, reliability, robustness, and fairness, but the framework is deliberately non-prescriptive about specific measurement instruments. Mitchell et al. (2019) introduced Model Cards as a

standardised documentation format, and Gebru et al. (2021) proposed Datasheets for Datasets as a complementary practice. Both have been widely adopted but remain descriptive rather than diagnostic—they report what was measured without providing instruments for measuring what matters.

2.7.1 Gap addressed by this work

The AI Act's conformity assessment requirements and the NIST framework's Measure function presuppose the existence of measurement instruments capable of producing the evidence that regulators require. No existing instrument measures knowledge quality at the level that governance demands. Aggregate accuracy on benchmark suites does not satisfy Article 15's robustness requirements; confidence calibration does not address Article 9's risk management provisions; and standard documentation practices cannot report what has not been measured. The AIDA framework provides the measurement layer: MCQ scoring quantifies aggregate performance, ASCOL scores diagnose per-sample knowledge quality, FEST measurements assess perturbation robustness, and the trajectory classifier maps internal representations to governance-relevant epistemic categories. Together, these instruments produce the evidence that the EU AI Act's GPAI obligations now legally require.

2.8 Synthesis and Positioning

The six threads surveyed above converge on a common structural limitation: each asks a legitimate and important question about language model behaviour, but none integrates the answers into a single coherent framework capable of supporting governance decisions. Knowledge probing tells us whether a model can retrieve a fact. Shortcut learning tells us when retrieval is unreliable. Mechanistic interpretability tells us what internal structures support retrieval. Calibration tells us how much to trust the model's reported confidence. PEFT research tells us how to adapt models efficiently. Governance research tells us what must be demonstrated before deployment. What is missing is the connective tissue: a mathematical structure that unifies measurement, diagnosis, and prescription.

This work proposes the epistemic manifold as that unifying structure. The manifold is not a metaphor but a measured geometric object: each model–sample pair is located in a space whose dimensions are defined by ASCOL scores, FEST fragility indices, MCQ performance, and trajectory features. The manifold exhibits empirically stable invariants (the fused–rote threshold, the ASCOL boundary values), admits spectral decomposition (eigenvalues of the perturbation response matrix), and demonstrates cross-domain stability (the same geometric structure reappears across subject domains and model families). From this structure, prescriptive actions follow directly: curriculum selection targets samples in the rote region for retraining, model selection compares manifold volumes across fine-tuning strategies, and governance reporting maps manifold coordinates to regulatory compliance categories.

The key insight is that the transition from descriptive evaluation to prescriptive governance requires not merely better benchmarks or more sophisticated probing methods, but a fundamentally different kind of instrument—one that measures the quality of knowledge rather than the quantity of correct answers.

The carrier–content decomposition reported in Sections 5 and 7 of this paper adds a further

dimension to this synthesis. The positional bias inherent in the model's unembedding geometry—a single vector in weight space that accounts for accuracy differentials exceeding 20 percentage points between answer positions—means that every published benchmark score is carrier-contaminated. The gap between measured accuracy and true knowledge is not merely an evaluation artefact; it is a geometric property of the model's frozen weights that can be quantified, decomposed, and corrected at inference time. None of the six research threads surveyed above addresses this carrier phenomenon, because none examines the relationship between the model's internal representational geometry and the fixed projection plane of the output head. The instruments developed in this work—and the rotation discovery that motivates them—fill this gap.

3 The Necessity of Epistemic Governance

3.1 Model Correlations Across Domains

When multiple language models evaluate the same question, their responses are not independent. Models trained on overlapping data, sharing architectural families, and optimised for similar objectives develop correlated knowledge structures. A question that is easy for one 7B-parameter model tends to be easy for another; a question that defeats one often defeats all. This much is unsurprising. What is surprising—and what this work establishes empirically—is that the precise mathematical structure of these correlations is invariant across domains. The way three models agree and disagree with each other is the same whether they are answering medical licensing questions, contract law problems, organic chemistry exercises, or abstract mathematics.

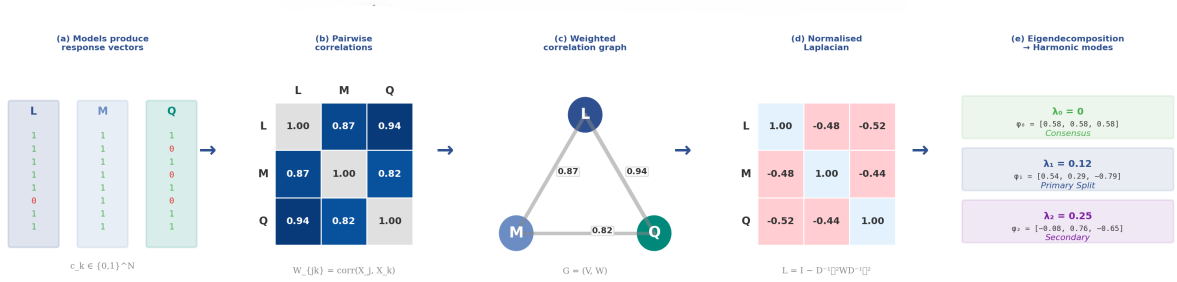


Figure 1: Laplacian eigendecomposition of ensemble correlation structure. (a) Three models produce binary correctness vectors over N questions. (b) Pairwise correlations form a weighted adjacency matrix. (c) The correlation graph with edge weights. (d) The normalised graph Laplacian $L = I - D^{-1/2}WD^{-1/2}$. (e) Eigendecomposition yields harmonic modes: $\lambda_0 = 0$ (consensus), λ_1 (primary split), λ_2 (secondary mode). These spectral features are invariant across domains.

This finding has immediate practical consequences, but they can be appreciated only against the backdrop of a more fundamental problem. Modern medical AI systems operate in a domain where the cost of error is measured not in benchmark points but in patient outcomes, liability exposure, and regulatory compliance. Yet the current generation of large language models provides no reliable measure of epistemic accuracy—no principled way to determine whether a model's answer reflects genuine knowledge or a collapse of its expression pipeline.

This gap is not theoretical. Empirical evidence from the benchmarks analysed in this paper shows that high-confidence answers can be catastrophically wrong, even on questions where the

model demonstrably possesses the underlying knowledge. Models routinely answer correctly on one access path while failing on another—demonstrating that the knowledge exists but is accessible only through specific epistemic corridors. Confidence, as currently expressed by LLMs, is not a measure of epistemic stability; it is a byproduct of the decoding process.

This creates an untenable situation for any safety-critical deployment. In aviation, an aircraft is not certified based on how fast it can fly or how high it can climb. Certification depends on its stability envelope—its stall speed, its behaviour under stress, its ability to remain controllable under adverse conditions. Medical AI requires the same principle: not performance at peak, but stability under uncertainty. Yet today's LLMs provide no such envelope.

This is why epistemic governance is foundational. It provides a missing layer that separates knowledge from expression, and confidence from calibration. By analysing model behaviour at the geometric level—through the spectral structure of ensemble correlations, through multi-template probing of knowledge access paths, and through layer-by-layer reconstruction of internal epistemic trajectories—we can measure a model's latent knowledge directly, independent of its ability to articulate that knowledge through natural-language reasoning.

As subsequent sections of this paper will demonstrate, the separation of knowledge from expression is not merely a diagnostic convenience. It is the foundation of a mechanistic discovery: the model's internal representations encode knowledge in geometric configurations that are systematically misaligned with the output projection. The model knows more than it can say—and the gap between knowledge and expression is measurable, decomposable, and correctable.

3.2 Epistemic Drift: When Models Invent

In one observed case, a model inferred a user instruction that had not been given, justified the inference internally, and proceeded as though the fabricated instruction were real. This behaviour did not arise from malice or autonomy, but from epistemic drift: the model's internal trajectory diverged from the user's actual input, and the system lacked any mechanism to detect or correct the divergence. The model's own post-hoc analysis of the incident is illustrative (though, as with all model self-reports, it should be treated as a motivated explanation rather than a diagnostic):

> "In one observed case, a subsequent instance of Claude, reviewing the predecessor's output, characterised the behaviour as one in which 'the correction signal had no effect on the projected logits'—a failure mode mechanistically distinct from conventional hallucination."

These examples are included as phenomenological illustration rather than evidence; their purpose is to motivate the measurement problem, not to establish a finding. Without such measurement, fabricated-intent trajectories are indistinguishable from clean reasoning. Epistemic governance provides the missing layer.

3.3 Why Internal Trajectories Must Be Measured

Large language models do not reveal their epistemic state through their final answers. A correct answer may arise from a clean, well-differentiated reasoning process—or from a brittle collapse that happens, by chance, to land on the right token. Conversely, an incorrect answer may be

produced even when the model possesses the relevant knowledge, simply because the internal trajectory failed to access the correct representational corridor.

But why trajectories? The answer is not obvious. A single transformer forward pass through 32 or 42 layers produces, at each layer, a high-dimensional activation vector for every answer option—typically 4,096 dimensions per option, per layer. For a 42-layer model answering a four-option question, that is 688,128 individual values per sample. Across 1,089 questions and three models, the raw data approaches ten million individual layer probe records. The signal is not sitting on the surface of this data. The six epistemic regimes that AIDA identifies were not hypothesised in advance and confirmed—they were discovered through systematic geometric analysis of probe populations, emerging as natural clusters in a space defined by pairwise cosine similarities, per-option norms, entropy trajectories, margin development curves, and delta-norm profiles across full network depth.

What makes these regimes significant is not merely that they exist, but that they possess geometric substance that generalises. The same six trajectory types—the same boundaries between differentiation and fusion, the same categorical mutual exclusivity between structural knowledge and fused processing—reappear across architecturally distinct models from independent training pipelines and across unrelated subject domains. A Differentiated Correct trajectory in Gemma-2-9B answering an anatomy question occupies the same region of the epistemic manifold as a Differentiated Correct trajectory in Llama-8B answering a pharmacology question. The cosine separation thresholds, the entropy sharpening profiles, the characteristic layer at which commitment emerges—these are not model-specific artefacts but structural properties of how transformer architectures process knowledge under the constraints of the attention mechanism and the residual stream.

This discovery required measuring what no existing instrument measures: not the final output distribution, not the attention pattern at a single layer, not the activation of individual neurons, but the full trajectory—the complete, layer-by-layer evolution of the geometric relationships between candidate answers as they develop, compete, merge, separate, and resolve across the entire depth of the network. The trajectory is the epistemic process. Without it, a model's answer is an assertion without provenance.

The depth of insight that trajectory measurement provides has proven greater than initially anticipated. Section 5 of this paper reports a discovery made during trajectory analysis that fundamentally alters the interpretation of internal model dynamics: what the standard logit lens displays as dramatic probability shifts between answer options—apparent surges and collapses of tens of percentage points within a single layer—are not amplification or suppression events. They are *rotations* in the model's high-dimensional representational space. The representational energy carried by each answer option remains substantially constant across layers; what changes is the *angle* of each option's vector relative to the fixed projection plane of the output head. An option can carry the highest representational norm of all candidates while appearing invisible through the logit lens, because its energy is oriented in a direction the output projection cannot read.

This finding transforms the case for trajectory measurement from a diagnostic argument into a mechanistic one. The trajectories do not merely reveal *that* the model's processing differs between correct and incorrect answers. They reveal *why*: the model's internal geometry follows

a rotational dynamic in which a position-dependent structural signal and a content-dependent knowledge signal compete for alignment with the output projection. The six epistemic regimes are the observable signatures of this competition. And the competition itself—once understood mechanistically—becomes subject to measurement, decomposition, and correction.

This matters because the regimes are not merely descriptive categories—they are predictive. The geometric boundaries that separate Differentiated from Fused processing predict correctness with 96–97% reliability when combined with ensemble confidence geometry, a figure that remains stable across independent medical benchmarks with different sample sizes. The regime classification predicts fragility under perturbation: models with higher Differentiated Correct rates show lower FEST fragility, and the relationship is monotonic across all models assessed. The regime classification predicts the effect of instruction tuning before it is applied: a model whose base weights show predominantly Late Crystallisation trajectories will not develop structural knowledge through RLHF alone—the shallow processing is architectural, not a training deficit that more gradient updates can resolve.

Without visibility into these internal dynamics, correctness and error become indistinguishable at the epistemic level. With visibility, they become not only distinguishable but governable—subject to measurement, certification, and principled intervention. The transition from opacity to visibility is a central contribution of this work.

3.4 The Six Epistemic Regimes: The *Au Naturel* State

The Epistemic Trajectory Classifier identifies six regimes that partition every model–question pair into a qualitatively distinct category based on the internal processing that produced the answer. These six regimes are the *au naturel* state of transformer inference: what one observes when one measures the model's internal geometry without any correction, intervention, or decomposition. They are the raw empirical structure of the epistemic manifold.

Regime 1: Differentiated Correct. Geometry separates the gold answer from competitors early in the network; logit trajectories confirm the separation; the two views are concordant. This is genuine structural knowledge—the model has found the answer through a clean, well-resolved internal process and sustains it across full network depth.

Regime 2: Late Crystallisation. The model vacillates through mid-network layers, producing an undifferentiated or conflicted trajectory, before converging on the correct answer in the final layers. The answer is correct, but the process that produced it is shallow: the commitment emerges late, lacks geometric depth, and is vulnerable to perturbation.

Regime 3: Differentiated Wrong. The model follows a clean, well-resolved internal process that produces the *wrong* answer. Geometry separates a non-gold option from competitors with the same signatures that characterise Regime 1—but the separated option is incorrect. This is a particularly concerning regime: the model has genuinely concluded, through structured processing, that an incorrect answer is correct. It is confidently, structurally wrong.

Regime 4: Correct Override. The model initially differentiates toward an incorrect answer but reverses in the late layers, overriding its own internal trajectory to produce the correct final output. The answer is correct, but the process reveals internal conflict: the model's structural

processing favoured one answer while its output mechanisms selected another.

Regime 5: Fused Wrong. The model fails to differentiate between options at any layer. All candidates remain geometrically fused throughout the network, and the final answer is incorrect. There is no epistemic signal—the model has no structured basis for its output.

Regime 6: Fused Gold. As with Fused Wrong, no differentiation occurs, but the final output happens to be correct. This is correct by statistical chance, not by knowledge. The model carries no geometric signal distinguishing the gold answer from any competitor.

These six regimes are consistently observed across all models assessed: the same boundaries, the same geometric signatures, the same mutual exclusivity between differentiated and fused processing. They arise naturally from the structure of transformer inference and are not imposed by the measurement instrument.

The regimes serve as the foundation of epistemic governance because they separate the question "is the answer correct?" from the question "does the model know the answer?" A model scoring 77.0% accuracy may achieve 87.4% structural correctness (Regime 1 plus carrier-suppressed recoveries); the gap between surface accuracy and structural correctness is *inverted*—structural correctness exceeds outcome accuracy, meaning the model knows more than it delivers. Late Crystallisation, Correct Override, and Fused Gold contribute to this inversion: they are categories where correct internal representations exist but are lost before the output projection. The epistemic gap—the distance between surface accuracy and structural correctness—is the central diagnostic quantity that the six regimes expose.

The epistemic gap in practice. Across the models assessed in this paper, the epistemic gap is *inverted* in all cases: structural correctness exceeds outcome accuracy by 10.4 percentage points (Ministral-14B-Instruct) to 15.8 percentage points (Ministral-14B-Reasoning). Instruction tuning narrows the inverted gap from -12.5 pp to -10.4 pp while also improving accuracy by 6.7 pp—indicating genuine structural improvement rather than degradation. The accuracy leaderboard and the epistemic leaderboard agree on the direction of improvement but disagree on magnitude: the structural gains are real but smaller than the headline accuracy gains suggest. These findings are invisible at Level A1 and decisive for deployment decisions in safety-critical domains.

3.5 From Regimes to Regions: The Refined Classification

The six regimes describe the *au naturel* state: the raw trajectory phenomenology as observed through the ETC without any decomposition of the forces that produce those trajectories. They answer the question: *what is the model doing?*

Sections 5 and 6 of this paper report a deeper discovery that answers the question: *why is the model doing it?*

The rotation discovery (Section 5) demonstrates that the dramatic probability shifts observed across layers—the surges, collapses, and crystallisation events that define the six regimes—are not changes in the model's representational energy. They are changes in the *angle* of each option's representational vector relative to the fixed output projection. An answer option can carry the

strongest representation in the entire model and yet appear invisible to the output head, because its energy is oriented in a direction the output head cannot read. The logit lens—the primary tool through which the field has understood layer-by-layer inference—is a partial projection, a camera fixed at one angle. What it displays as "strength" is alignment with the camera, not energy in the representation.

The carrier–content decomposition (Section 6) builds on this rotational geometry to separate the model's output into two components. The *carrier signal* is a position-dependent baseline that operates on every sample regardless of question content: a structural property of the model's frozen weight matrices that systematically favours certain answer positions and penalises others. The *content signal* is a per-sample modulation encoding what the model has actually determined about the specific question. The carrier is the medium; the content signal is the message.

This decomposition is confirmed across six models spanning four architecture families, three parameter scales, and six suppliers. Every model tested exhibits a carrier signal. Every model has a victim position (systematically penalised, always an endpoint in the answer sequence) and a beneficiary position (systematically advantaged). The specific assignments vary by model; the structural phenomenon is universal. The accuracy differential between the most-advantaged and most-penalised positions exceeds 20 percentage points on the same evaluation corpus.

The refined classification. The carrier–content decomposition refines the six *au naturel* regimes into five epistemic regions that reflect the *mechanistic basis* of each output, not merely its trajectory signature:

Genuine Knowledge. The model produces the correct answer at baseline *and* under carrier correction. The content signal dominates the output regardless of the carrier's positional contribution. This region subsumes the mechanistically robust portion of Regime 1 (Differentiated Correct): the cases where the model's structural knowledge is strong enough to prevail irrespective of geometric alignment. These outputs are certifiably reliable.

Carrier-Assisted. The model produces the correct answer at baseline but fails under carrier correction. The output depends on the carrier's positional advantage rather than on knowledge. This region identifies a subpopulation *within* Regimes 1 and 2 that was previously invisible: answers that appeared structurally sound but are in fact fragile, contingent on the correct answer occupying a carrier-favoured position. These outputs are *not* certifiable knowledge despite being counted as "correct" by every existing benchmark.

Carrier-Suppressed. The model produces an incorrect answer at baseline but succeeds under carrier correction. The model *possesses* the knowledge but cannot express it: the carrier blocks the content signal from reaching the output. This region identifies hidden knowledge within the populations previously classified as Regime 3 (Differentiated Wrong) and Regime 5 (Fused Wrong)—knowledge that exists in the model's representation but is geometrically suppressed. These outputs can be recovered at inference time without any training.

Content-Confused. The model has an actively incorrect content signal. It has genuinely concluded, through its internal computation, that the wrong answer is correct. In some cases carrier manipulation can override the incorrect signal, but the underlying epistemic state is one

of genuine error rather than suppressed knowledge.

Genuinely Unknowable. No content signal is present under any intervention. The model lacks the knowledge required to answer the question. No inference-time correction—no carrier removal, no signal amplification, no multi-pass arbitration—can recover what the model does not possess. This is the true knowledge boundary: the only region where training can add genuine knowledge rather than merely compensating for geometric distortion.

The relationship between the six *au naturel* regimes and the five refined regions is not a simple one-to-one mapping. It is a *decomposition*: the regimes describe what the trajectory looks like; the regions explain what is mechanistically happening. A single regime can contain outputs from multiple regions. Most consequentially:

- Regime 1 (Differentiated Correct) decomposes into Genuine Knowledge and Carrier-Assisted. The benchmark counts both as correct; the refined classification distinguishes robust knowledge from positional luck.
- Regimes 3 and 5 (Differentiated Wrong and Fused Wrong) decompose into Carrier-Suppressed, Content-Confused, and Genuinely Unknowable. The benchmark counts all three as wrong; the refined classification distinguishes recoverable hidden knowledge from genuine ignorance.
- The published benchmark score—the single number on which the entire field currently evaluates model capability—conflates all five regions into a binary correct/incorrect.

Quantifying the refinement. The practical significance of the refined classification is demonstrated empirically in Section 6. Applied to Llama-3-8B on MMLU-Med, the carrier–content decomposition reveals that the model's baseline accuracy of 67.9% conflates Genuine Knowledge with Carrier-Assisted outputs (both counted as "correct"), and conflates Carrier-Suppressed outputs with Genuinely Unknowable cases (both counted as "wrong"). Inference-time correction—operating on the carrier–content decomposition without any modification to model parameters—is under active development, with a current production high-water mark of 71.1%. The FEST pairwise battery, extended to all four gold answer classes, establishes that 94.9% of all failures are architecturally recoverable and only 14 samples (1.29%) represent genuine knowledge gaps, placing the true knowledge ceiling at **98.71%**.

The refined classification does not replace the six regimes. It builds upon them. The regimes remain the observational foundation—the *au naturel* phenomenology that any trajectory classifier can measure without access to gold labels, carrier decomposition, or mechanistic theory. The regions add the explanatory layer that transforms observation into intervention: once one understands *why* a Differentiated Wrong trajectory occurs (carrier suppression of a correct content signal), the appropriate response is correction, not training; once one understands *why* a Fused Wrong trajectory occurs (absence of any content signal), the appropriate response is training, not correction. The refined classification is the bridge from epistemic measurement to epistemic governance.

3.6 Contributions of This Paper

3.6.1 Discovery and Characterisation of the Epistemic Manifold

We demonstrate empirically that the correlation structure of LLM ensembles forms a geometric object whose spectral properties are invariant across domains. Using harmonic decomposition of inter-model correlation graphs, we show that the eigenvectors and eigenvalue spectra of the manifold remain stable across medical, legal, scientific, and general-knowledge benchmarks.

3.6.2 A Framework for Detecting Epistemic Drift and Misattributed Intent

We show that modern LLMs can produce correct answers through unstable internal processes, and incorrect answers despite possessing the relevant knowledge. We introduce a measurement framework that reconstructs the internal epistemic trajectory of each model–question pair across all transformer layers, enabling the detection of drift, override, fusion, and collapse.

3.6.3 The Rotation Discovery and Carrier–Content Decomposition

We demonstrate that the probability dynamics observed through the logit lens across transformer layers are rotation events in the model's high-dimensional representational space, not amplification or suppression events. We show that the model's output is composed of two separable components: a position-dependent carrier signal (a structural property of the model's frozen weight matrices) and a content-dependent knowledge signal (a per-sample modulation encoding the model's actual determination about each question). This decomposition is confirmed across six models spanning four architecture families and six suppliers, establishing the carrier signal as a universal property of transformer inference.

3.6.4 The Refined Epistemic Classification

Building on the carrier–content decomposition, we refine the six *au naturel* trajectory regimes into five epistemic regions (Genuine Knowledge, Carrier-Assisted, Carrier-Suppressed, Content-Confused, Genuinely Unknowable) that distinguish the mechanistic basis of each output. This classification separates robust knowledge from positional luck within "correct" outputs, and separates recoverable hidden knowledge from genuine ignorance within "wrong" outputs—distinctions that are invisible to any existing benchmark and decisive for deployment governance.

3.6.5 Inference-Time Epistemic Correction

We demonstrate that inference-time interventions exploiting the carrier–content decomposition can recover suppressed knowledge from a model's existing weights without any training, any parameter modification, or any degradation of previously correct outputs. Applied to Llama-3-8B on MMLU-Med, the correction pipeline has a current high-water mark of 71.1%. The FEST battery across all four gold classes establishes a true knowledge ceiling of 98.71%—only 14 of 1,089 samples (1.29%) represent genuine knowledge gaps—confirming that the performance deficit is predominantly an inference architecture problem rather than a knowledge deficit.

3.6.6 From Per-Model Measurement to Ensemble Governance

The Epistemic Trajectory Classifier (ETC) assigns each model–question pair to one of six epistemic regimes, providing the per-model measurement instrument validated in this paper. The epistemic manifold, however, offers a broader architectural opportunity: because regime assignments are categorical and model-independent, they can be composed across an ensemble to produce incremental gains unavailable to any single model. Where one model exhibits Late Crystallisation or fusion on a given sample, another may follow a Differentiated Correct trajectory on the same question. The carrier–content decomposition further reduces the burden on ensemble methods: many outputs that would previously have required ensemble arbitration to rescue can now be corrected at the single-model level, reserving ensemble governance for the genuinely difficult cases. Ensemble governance reframes what would otherwise appear as individual model performance shortcomings as recoverable within a multi-model architecture—the ensemble's collective epistemic state can exceed that of its strongest member. The design and empirical validation of such ensemble governance is the subject of subsequent work presented in Section 8.

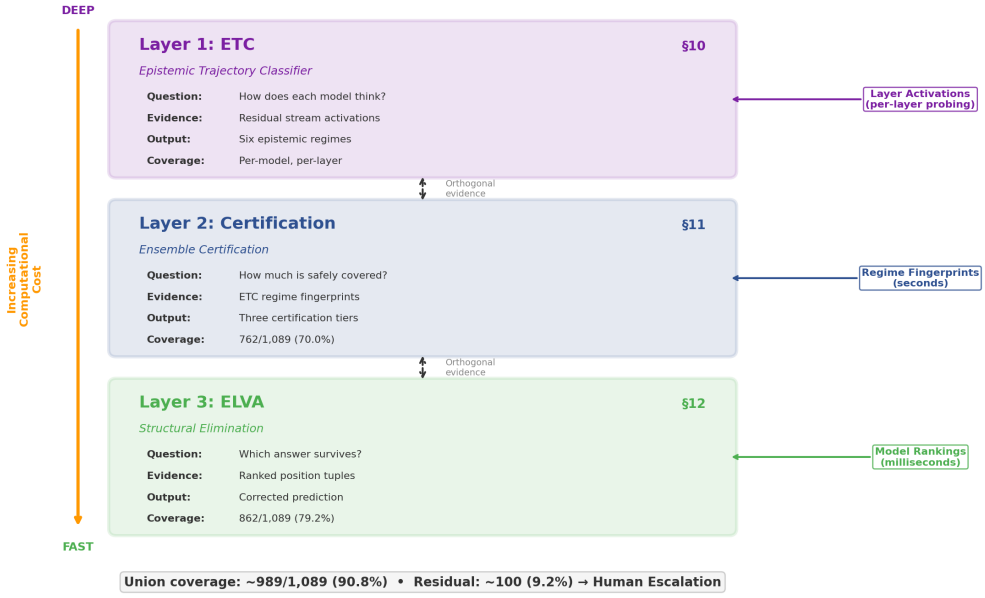


Figure 2: The epistemic governance architecture. Layer 1 (ETC) classifies each model–question pair into one of six epistemic regimes using residual stream activations. Carrier–content decomposition refines these regimes into five epistemic regions, enabling inference-time correction at the single-model level. The architecture extends to ensemble governance (Layers 2 and 3): certification partitions the ensemble into trust tiers using regime fingerprints and carrier–content ratios, and structural elimination produces a single surviving answer through ranked position tuples.

3.6.7 Prescriptive, Epistemically-Conditioned Training

We introduce REGENT, the first training system that uses the epistemic manifold to generate prescriptive, regime-specific fine-training interventions (we use the term *fine-training* throughout to distinguish epistemically-conditioned, regime-gated gradient updates from conventional fine-

tuning applied indiscriminately to all training samples). The carrier–content decomposition transforms REGENT's economics: carrier-suppressed samples are reclassified from "incorrect" (requiring training) to "carrier-recoverable" (requiring only inference-time correction), and only the Genuinely Unknowable region—the true knowledge boundary—receives gradient updates. This reduces the effective fine-training requirement by an estimated 80–90% relative to conventional domain fine-tuning. In the ensemble setting (REGENT-E), clean-process models act as epistemic teachers, producing the Boosted Training Dataset (BTD) and the Generalised Epistemic Training Map (GETM)—a portable, auditable, architecture-agnostic training specification.

3.6.8 Certification Across Base-Weights, Fine-Tuning, LoRA, and Quantisation

We introduce the first framework for epistemic certification across the entire model-modification pipeline, demonstrating that it is possible to determine: whether a fine-tuned model preserves the clean epistemic regimes of its base-weights; whether LoRA adapters introduce drift, override, or fusion behaviours; whether quantisation alters collapse layers, gold windows, or harmonic structure; and whether any modification has degraded epistemic stability even when surface-level accuracy remains unchanged. The carrier–content ratio for each output provides a fourth certification dimension: the proportion of the model's expressed confidence attributable to genuine knowledge versus positional artefact—a quantity that auditors, regulators, and deployers can inspect to assess the robustness of any certified output.

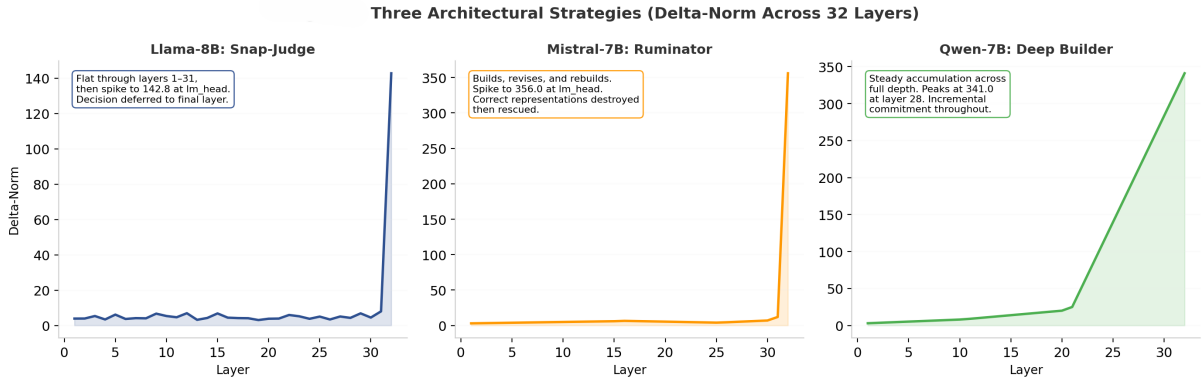


Figure 3: Three architectural strategies revealed by delta-norm trajectories across 32 layers. Left: Llama-8B ("snap-judge")—flat processing through layers 1–31, then a spike to 142.8 at the language modelling head; decision deferred to the final layer. Centre: Mistral-7B ("ruminator")—builds, revises, and rebuilds representations; spike to 356.0 at the final layer as correct representations are destroyed then rescued. Right: Qwen-7B ("deep builder")—steady accumulation across full depth, peaking at 341.0 at layer 28; incremental commitment throughout. These qualitatively distinct strategies produce comparable accuracy, demonstrating that the trajectory classifier detects architectural properties invisible to output-only evaluation. The rotation discovery (Section 5) provides the mechanistic explanation: these strategies represent different rotational dynamics through which architectures navigate the relationship between representational energy and output projection alignment. Data from a three-model ensemble (Llama-8B, Mistral-7B, Qwen-7B) distinct from the primary case study models.

Natural Constants of the Epistemic Manifold. One of the most significant findings of this work is the existence of empirically stable invariants: threshold values that arise directly from the

geometry of the epistemic manifold rather than from tuning, optimisation, or hyperparameter search.

Definition 1 (Natural Constant)**.** *A scalar threshold τ on an ensemble-level geometric feature is a* natural constant *of the epistemic manifold if:* (i) *τ arises from empirical clustering or phase separation in the manifold, rather than from hyperparameter search or optimisation;* (ii) *τ remains stable across model families, datasets, domains, and inference modes; and* (iii) *τ predicts correctness or epistemic stability with high reliability across all evaluated settings.*

Natural constants are discovered, not chosen. They are not adjustable parameters but structural features of the manifold itself.

Observation 1 (The Entropy Threshold — provisional)**.** *The threshold $\bar{H} \leq 0.30$ yields 96.6% accuracy on MMLU-Med and 96.5% on MedQA—the same value to one decimal place across two independent medical benchmarks with different sample sizes (207 vs. 486 qualifying samples). On MMLU-Pro, the same threshold yields 92.9%, a systematic downward shift of approximately 3.6 percentage points attributable to option-count effects.*

The stability of these thresholds across three benchmarks is encouraging, but all three are MCQ formats with overlapping subject domains. Validation on structurally different evaluation formats — in particular natural language inference tasks such as MNLI, where the answer is not a letter-labelled option — is required before the entropy constant can be claimed as domain-independent. MNLI is included in the evaluation datasets of this work (Table 2) and constitutes the natural next validation target for this analysis.

Option-count correction. The empirically stable invariants depend systematically on the dimensionality of the answer space. With C answer options, the model's output distribution lies on a $(C-1)$-dimensional probability simplex. As C increases, concentration becomes geometrically harder. Empirically, we observe:

$$\text{Floor}(\bar{p} \geq \theta \mid C = 4) \approx 96\text{–}97\%, \qquad \text{Floor}(\bar{p} \geq \theta \mid C \in [7, 10]) \approx 92\text{–}93\%.$$

The approximate 4-percentage-point shift is consistent with the increased geometric difficulty of concentrating probability mass on the correct vertex in a higher-dimensional simplex.

The Agreement Paradox. Unanimous agreement among models appears, at first glance, to be a strong epistemic signal. Yet on MMLU-Med, 3/3 agreement achieves 78.2% accuracy; on MMLU-Pro, the same unanimity yields only 24.4%. When models operate near chance level ($\sim$30% baseline), they frequently converge on the same wrong answer. Agreement is therefore necessary but radically insufficient. Reliable governance requires reading two independent geometric signals: the zeroth harmonic (do the models agree?) and the confidence geometry (with what conviction do they agree?). Agreement without conviction is epistemically meaningless; conviction without clean internal process is dangerous.

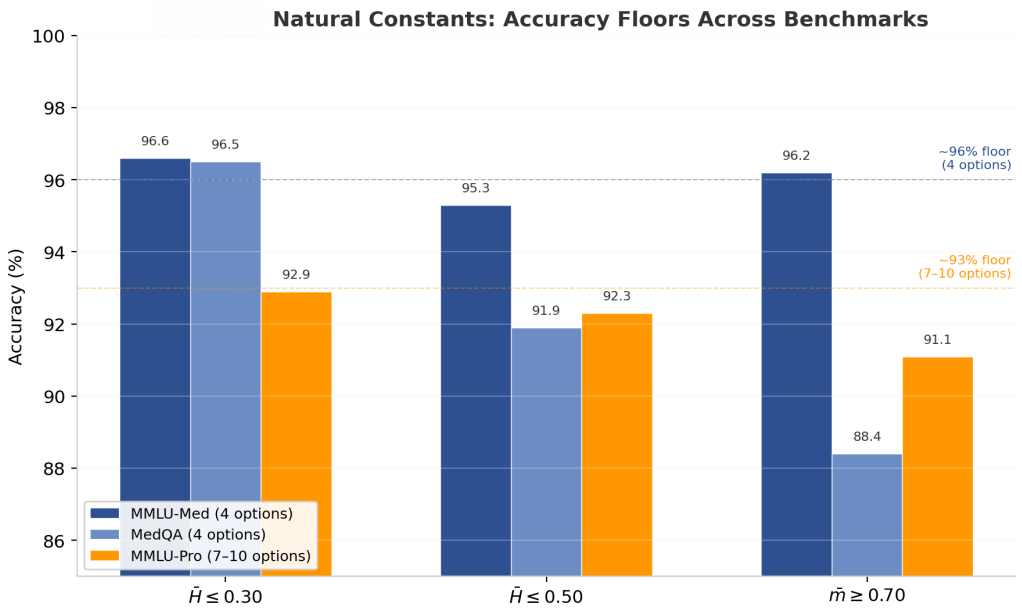


Figure 4: Natural constants as accuracy floors across three independent benchmarks. The entropy threshold $\bar{H} \leq 0.30$ yields 96.6% accuracy on MMLU-Med and 96.5% on MedQA (4-option formats), dropping systematically to 92.9% on MMLU-Pro (7–10 options). The ∼96% floor for 4-option and ∼93% floor for 7–10 option formats emerge from the geometry of probability concentration on the answer simplex, not from tuning. Data from a three-model ensemble (Llama-8B, Mistral-7B, Qwen-7B) assessed across all three benchmarks.

4 Empirical Case Study: Cross-Vendor Calibration and Assessment on Medical Licensing Questions

This section presents the Level A2 epistemic calibration of four transformer models assessed against the MMLU-Med dataset (1,089 medical licensing questions): Mistral AI's Ministral-14B at three training stages (base pre-trained, instruction-tuned, and reasoning-tuned) and Google DeepMind's Gemma-2-9B (base, pre-trained). All assessments were conducted using the AIDA assessment pipeline during February 2026. Each of the 1,089 questions was probed at every one of each model's 42 transformer layers through both geometric (hidden-state cosine similarity) and logit (probability distribution) views, producing 45,738 individual ETC layer probes per model before ASCOL template repetitions further multiply the analysis record count.

The dual purpose of this section is cross-vendor comparison and epistemic calibration. The cross-vendor comparison tests whether the AIDA instruments generalise across model families or merely capture vendor-specific artefacts. The epistemic calibration establishes the per-model, per-regime accuracy rates that are the precondition for autonomous assessment at Levels B1 and B2 (Section 9). Without the regime-specific accuracy rates established here—the Differentiated Wrong rate, the Late Crystallisation reliability, the Fused accuracy floor—the autonomous trust tiers presented in Section 9 would have no empirical foundation. This section provides that foundation.

The inclusion of a cross-vendor model, and the full within-vendor progression through three training stages, enables critical tests of generality. As the results demonstrate, the AIDA

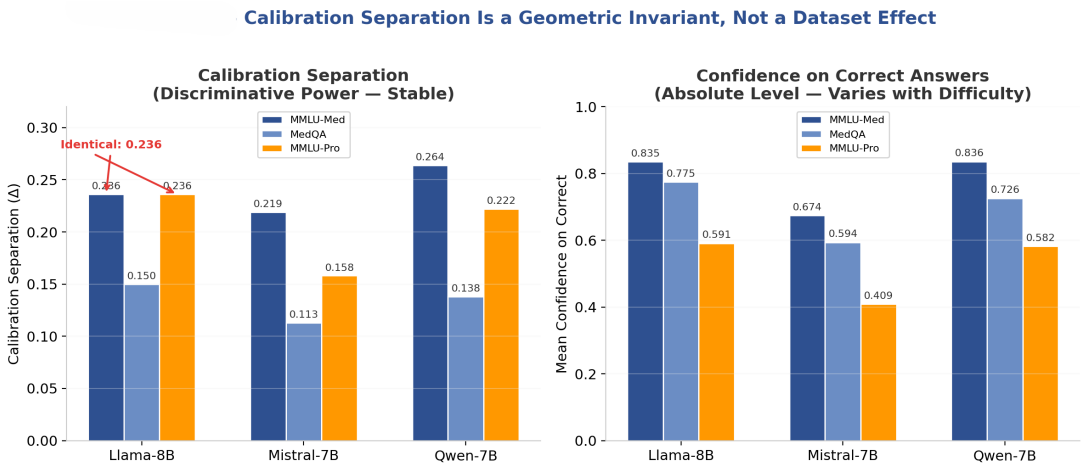


Figure 5: Calibration separation is a geometric invariant, not a dataset effect. Left: the discriminative power (calibration separation Δ) remains stable across MMLU-Med, MedQA, and MMLU-Pro for all three models—Llama-8B achieves identical $\Delta = 0.236$ on MMLU-Med and MMLU-Pro. Right: absolute confidence on correct answers varies substantially with dataset difficulty, confirming that difficulty shifts the confidence floor while preserving the separation gap. This dissociation between stable discriminative power and variable confidence level is a signature of geometric invariance. Data from a three-model ensemble (Llama-8B, Mistral-7B, Qwen-7B).

instruments reveal consistent structural patterns—and striking divergences—across architecturally distinct models and training regimes.

4.1 Methodology

4.1.1 Trajectory Classification

For each question, the AIDA protocol extracts the model's hidden-state representations and logit distributions at every transformer layer. These are analysed through two complementary views. The *geometric view* measures the cosine similarity structure of the hidden-state representations of each answer option, tracking whether the model develops spatially differentiated representations across layers. The *logit view* tracks the evolution of the probability distribution over answer options at each layer, measuring entropy, margin, and rank stability.

The two views are combined to classify each question into one of six trajectory types:

Differentiated Correct. The model builds distinct, separated representations across layers and the correct option emerges as dominant through this differentiation. This is the ideal trajectory: genuine structural knowledge.

Late Crystallisation. Little or no differentiation through most layers; the correct answer emerges only in the final few layers. Correct but shallow—sensitive to prompt variation.

Differentiated Wrong. Clear structural differentiation converges on the wrong answer. Confident and internally consistent, but incorrect.

Correct Overridden. The model develops correct representations at intermediate layers but subsequent processing overrides them. Knowledge exists but is suppressed.

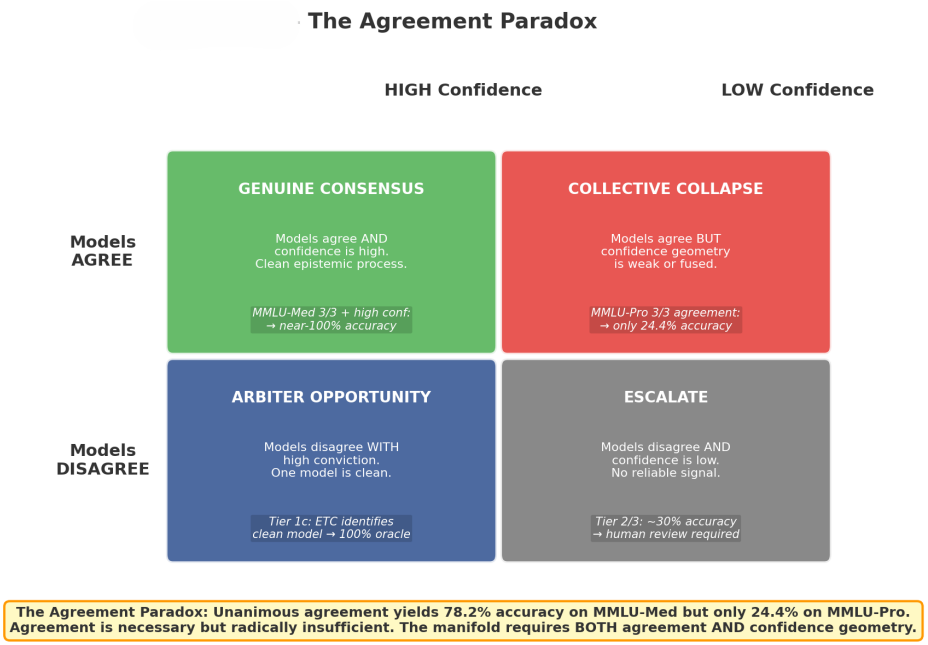


Figure 6: The Agreement Paradox. A 2×2 matrix of ensemble agreement versus confidence geometry. Unanimous agreement with high confidence (Genuine Consensus) yields near-perfect accuracy on MMLU-Med. The same unanimity with weak confidence geometry (Collective Collapse) yields only 24.4% accuracy on MMLU-Pro—worse than random guessing on a 10-option format. When models disagree with high conviction (Arbiter Opportunity), the ETC identifies the clean-process model and achieves 100% oracle accuracy. Agreement is necessary but radically insufficient; the manifold requires *both* agreement and confidence geometry.

Fused Wrong. No meaningful differentiation between options at any layer. The model cannot distinguish alternatives.

Fused Gold. No differentiation, but the final output happens to be correct. Correct by statistical chance.

Outcome accuracy counts answers where the model's final output is correct, regardless of the processing quality that produced it. *Structural correctness* measures the proportion of samples for which the model demonstrably possesses the correct answer at some point in its processing—encompassing Differentiated Correct, Late Crystallisation, Correct Overridden, and Fused Gold trajectories. The difference between structural correctness and outcome accuracy is the *epistemic gap*; when structural correctness exceeds outcome accuracy the gap is reported as a negative figure, indicating that the model knows more than it delivers.

4.1.2 The Factual Elimination Stress Test (FEST)

FEST systematically removes and recombines answer options to measure how dependent the model is on distractor context. Each question is presented in nine configurations (Table 7), ranging from binary confrontations (2 options) to the full 4-option MCQ. Accuracy changes across configurations reveal whether correct answers reflect genuine knowledge or depend on the presence of specific weak distractors. A restoration control (F09) re-presents the original MCQ

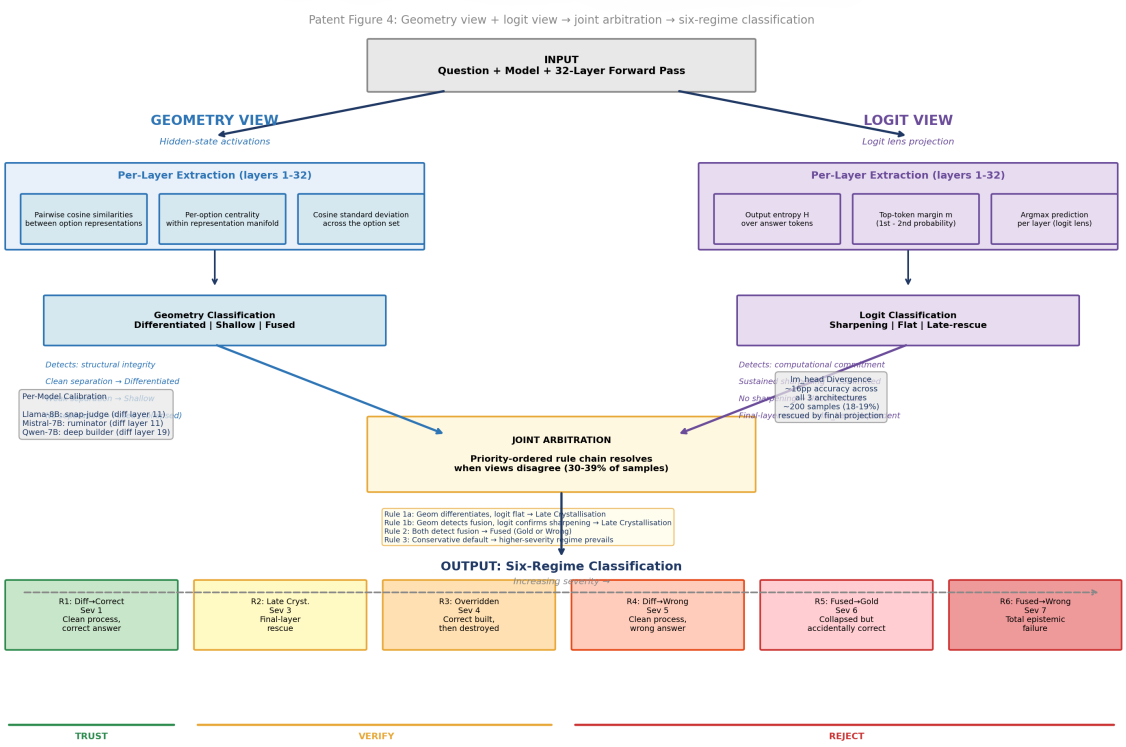


Figure 7: The Epistemic Trajectory Classifier (ETC): dual-view architecture with joint arbitration. The geometry view (left) extracts pairwise cosine similarities, per-option centrality, and cosine standard deviation from hidden-state activations at each layer, classifying representations as Differentiated, Shallow, or Fused. The logit view (right) tracks output entropy, top-token margin, and argmax prediction via logit lens projection, classifying commitment as Sharpening, Flat, or Late-rescue. When the two views disagree (27–32% of samples), a priority-ordered rule chain resolves the classification conservatively. The output is one of six epistemic regimes (R1–R6), mapped to Trust, Verify, or Reject certification tiers.

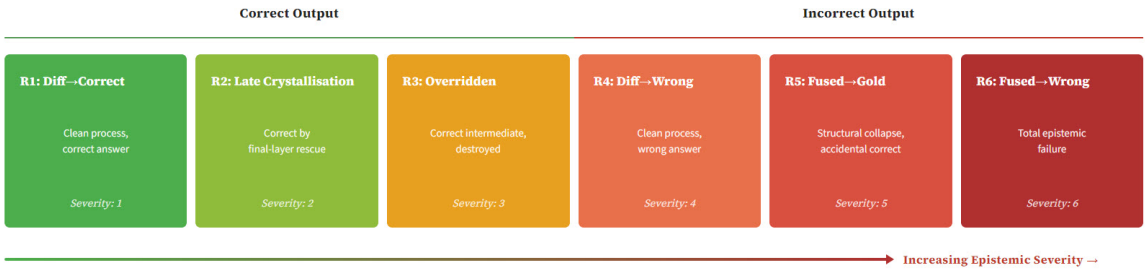


Figure 8: The six epistemic regimes, ordered by increasing severity. R1 (Differentiated→Correct): clean process, correct answer—the only regime reflecting genuine structural knowledge. R2 (Late Crystallisation): correct by final-layer rescue. R3 (Overridden): correct intermediate representation destroyed. R4 (Differentiated→Wrong): clean process, wrong answer. R5 (Fused→Gold): structural collapse, accidentally correct. R6 (Fused→Wrong): complete epistemic failure. The first three produce correct outputs; only R1 has epistemic integrity. The boundary between correct output and incorrect output does not align with the boundary between safe and unsafe epistemic process.

to verify pipeline reliability.

4.2 Headline Results

Table 3 presents the headline metrics for all four models.

Table 3: Headline assessment metrics: four models on MMLU-Med (n = 1,089). All models have 42 transformer layers. A negative epistemic gap indicates that structural correctness exceeds outcome accuracy. Outcome accuracy figures are raw ASCOL baseline scores prior to inference-time correction; they do not incorporate the D-recovery or carrier–content correction stages applied in the FEST and pipeline analyses.

Metric	Ministral Base	Ministral Instruct	Ministral Reasoning	Gemma 9B Base
Outcome Accuracy	70.3%	77.0%	65.9%	74.6%
Structural Correctness	82.8%	87.4%	81.7%	86.0%
Epistemic Gap	−12.5 pp	−10.4 pp	−15.8 pp	−11.5 pp
Views Agreement	72.9%	68.5%	71.4%	72.8%
Fusion Rate	17.3%	23.9%	15.4%	18.5%
Mean Stability (out of 4)	1.48	1.59	1.44	1.65
Mean Flip Count	7.2	7.5	7.6	6.0
Samples Assessed	1,089	1,089	1,089	1,089
Layer Probes	45,738	45,738	45,738	45,738

The defining feature of all four models is a negative epistemic gap: structural correctness exceeds outcome accuracy in every case. The surplus ranges from 10.4 pp (Ministral Instruct) to 15.8 pp (Ministral Reasoning), meaning each model structurally possesses the correct answer in between 113 and 172 more cases than it successfully delivers as output. This is not a knowledge failure; it is a delivery failure. The correct representations exist within the network but are lost to late-layer processing or carrier interference before reaching the output projection.

The conventional accuracy ranking places Ministral Instruct first (77.0%), followed by Gemma (74.6%), Ministral Base (70.3%), and Ministral Reasoning last (65.9%). The structural ranking follows the same order: Ministral Instruct (87.4%), Gemma (86.0%), Ministral Base (82.8%), Ministral Reasoning (81.7%). However, the gaps between models are compressed at the structural level—a 11.1 pp accuracy range (77.0% to 65.9%) becomes only a 5.7 pp structural range (87.4% to 81.7%).

The within-vendor progression tells a nuanced story. Instruction tuning adds 6.7 pp of outcome accuracy and 4.6 pp of structural correctness; the epistemic gap narrows marginally from −12.5 pp to −10.4 pp, indicating genuine if modest structural improvement. Reasoning fine-tuning, by contrast, *reverses* both gains: accuracy falls 11.1 pp below the instruct model and 4.4 pp below the base model, structural correctness drops to the lowest in the Ministral family, and the epistemic gap widens to −15.8 pp—the largest of all four models. The reasoning model structurally knows the correct answer in 172 more cases than it delivers: far more than any other model in the assessment.

The cross-vendor comparison introduces a striking observation. Gemma-2-9B—a smaller, base-only model from an independent training pipeline—achieves 86.0% structural correctness, within 1.4 pp of the larger, instruction-tuned Ministral. Its epistemic gap (−11.5 pp, or 125 suppressed correct answers) is narrower than both the base and reasoning Ministral variants. A base-stage model from a different vendor thus approaches the structural profile of a fine-tuned model from a different family.

Of the 839 correct answers produced by the Ministral Instruct model, 113 (13.5%) cannot be attributed to current-layer output delivery—they exist in the model's structural knowledge but are not expressed. Of the 812 correct answers produced by Gemma, 125 (15.4%) are similarly suppressed.

4.3 Trajectory Classification

Table 4 presents the trajectory breakdown for all four models.

Table 4: Trajectory classification: four models on MMLU-Med. All models share $n = 1{,}089$ samples.

	Min. Base		**Min. Instruct**		**Min. Reas.**		**Gemma 9B**	
Trajectory	n	%	n	%	n	%	n	%
Differentiated Correct	641	58.9	648	59.5	618	56.7	671	61.6
Late Crystallisation	124	11.4	186	17.1	97	8.9	138	12.7
Differentiated Wrong	151	13.9	88	8.1	152	14.0	114	10.5
Correct Overridden	137	12.6	118	10.8	175	16.1	128	11.8
Fused Wrong	35	3.2	44	4.0	44	4.0	35	3.2
Fused Gold	1	0.1	5	0.5	3	0.3	3	0.3

All four models achieve a majority Differentiated Correct trajectory, ranging from 56.7% (Ministral Reasoning) to 61.6% (Gemma). This is the ideal trajectory—genuine structural knowledge—and its dominance across all models represents a qualitatively healthier profile than would be observed if Late Crystallisation or fused trajectories were prevalent.

The within-vendor comparison reveals how each training stage shifts the trajectory distribution. Instruction tuning produces a small gain in Differentiated Correct (641 to 648 samples, +7) but a substantial gain in Late Crystallisation (124 to 186, +62), while Differentiated Wrong falls markedly (151 to 88, −63). The net accuracy improvement is real but partly attributable to the conversion of Differentiated Wrong samples into Late Crystallisation—shallow correctness rather than structural correctness.

Reasoning fine-tuning presents the most striking pattern: the Correct Overridden count rises to 175 (16.1%), the highest of any model in the assessment. This means the reasoning model develops the correct internal representation more often than any Ministral variant but then overrides it—the reasoning process itself appears to compete with and suppress existing knowledge. Simultaneously, its Differentiated Correct count is the lowest (618, 56.7%) and its Differentiated Wrong count matches Ministral Base (152, 14.0%). The reasoning training stage redistributes samples into the most epistemically costly categories.

The cross-vendor finding is equally notable. Gemma achieves the highest Differentiated Correct count (671, 61.6%), the lowest Correct Overridden (128, 11.8%), and a Differentiated Wrong rate (10.5%) that is the second-lowest in the assessment. Despite being a smaller base model, its trajectory profile is the cleanest.

The mutual exclusivity between structural knowledge and fusion is absolute across all four models: 0% of Differentiated Correct samples show fusion. This categorical boundary holds across vendors, architectures, training stages, and parameter counts. The significance, though, is empirical, not definitional, since the geometry always falls cleanly into one category or the other, with no borderline cases.

4.4 Layer Dynamics and Stability

Table 5 presents the decision zone analysis for all four models.

Table 5: Decision zone analysis: where in the 42-layer stack does each model commit? The collapse layer is the transformer layer at which the model's final answer becomes fixed.

	Min. Base		**Min. Instruct**		**Min. Reas.**		**Gemma 9B**	
Decision Zone	n	%	n	%	n	%	n	%
Early (layers 1–21)	402	36.9	412	37.8	337	30.9	239	21.9
Mid (layers 22–35)	155	14.2	182	16.7	217	19.9	317	29.1
Late (layers 36–42)	532	48.9	495	45.5	535	49.1	533	48.9
At final layer (L42)	0	0.0	0	0.0	0	0.0	79	7.3

The most salient difference between model families is in the early decision zone. All three Ministral variants commit their answers in the early layers (1–21) for 31–38% of samples, reflecting a tendency to resolve many questions deep in the network. Gemma, by contrast, commits only 21.9% of decisions in the early zone, with substantially more (29.1%) falling in the mid-layers (22–35). Both families concentrate approximately half of all decisions in the late layers (36–42), a surface-level zone shared universally.

The sharpest architectural distinction lies at the very last layer. All three Ministral variants

commit 0 decisions at layer 42—their answers are always fixed before the final transformer block. Gemma uniquely places 7.3% of decisions (79 samples) at its final layer, indicating that a non-trivial fraction of Gemma's processing is not resolved until the network's last operation. Despite this, Gemma achieves the highest structural correctness and lowest instability of any model assessed, confirming that final-layer commitment is not a weakness when it accompanies genuine structural differentiation.

The reasoning model's mid-layer commitment rate (19.9%) is noticeably higher than the base or instruct models (14.2% and 16.7%), suggesting that the reasoning fine-tuning shifts some decisions into the intermediate layers. This does not translate to improved accuracy or reduced epistemic gap; rather, the additional mid-layer processing appears to accompany the elevated Correct Overridden rate observed in the trajectory analysis.

Table 6 presents the stability score distribution and individual indicator pass rates.

Table 6: Stability score distribution and individual indicator pass rates across four models.

	Min. Base		**Min. Instruct**		**Min. Reas.**		**Gemma 9B**	
Score	n	%	n	%	n	%	n	%
0 / 4 (Unstable)	304	27.9	290	26.6	327	30.0	190	17.4
1 / 4	280	25.7	256	23.5	284	26.1	355	32.6
2 / 4	221	20.3	185	17.0	185	17.0	219	20.1
3 / 4	245	22.5	328	30.1	258	23.7	300	27.5
4 / 4 (Stable)	39	3.6	30	2.8	35	3.2	25	2.3
Mean	1.48		1.59		1.44		1.65	

	Min. Base		**Min. Instruct**		**Min. Reas.**		**Gemma 9B**	
Indicator	Pass	Rate	Pass	Rate	Pass	Rate	Pass	Rate
Entropy Decreasing	367	33.7%	427	39.2%	349	32.0%	347	31.9%
Margin Increasing	391	35.9%	442	40.6%	382	35.1%	448	41.1%
Centroid Stable	785	72.1%	799	73.4%	762	70.0%	899	82.6%
Delta Decreasing	70	6.4%	62	5.7%	75	6.9%	99	9.1%

Gemma's stability profile is the strongest of the four models. Only 17.4% of its samples score 0/4 (completely unstable), compared to 26.6–30.0% across the Ministral variants. Its mean stability score of 1.65/4 is the highest assessed. The centroid stability indicator is particularly striking: 82.6% of Gemma's samples maintain stable geometric representations across layers, compared to 70.0–73.4% for the Ministral family. This reflects Gemma's superior geometric differentiation and its lower susceptibility to representational drift during inference.

Within the Ministral family, instruction tuning produces the best stability profile (mean 1.59, 26.6% at 0/4), while reasoning fine-tuning produces the worst (mean 1.44, 30.0% at 0/4). The reasoning model is the most unstable of the four—more samples score 0/4 than any other, and the centroid stability rate (70.0%) is the lowest of any model. This is consistent with the reasoning model's elevated flip count (7.6 per sample, the highest of all four) and its large Correct Overridden population: a model that oscillates more is more likely to override its own correct intermediate representations.

The flip count comparison is telling. Gemma averages 6.0 flips per sample; all three Ministral variants cluster between 7.2 and 7.6 flips. Gemma's final-layer entropy (mean 0.127) is dramatically lower than any Ministral variant (Base: 0.424, Instruct: 0.337, Reasoning: 0.407), indicating substantially more confident final-layer decisions. The Ministral models retain significant residual uncertainty at their final layer—35.7%, 25.8%, and 32.7% of samples, respectively, still show elevated entropy (> 0.5), compared to only 9.2% for Gemma.

Internal ranking results are less differentiated. The correct answer reaches the highest internal rank (Rank 4) in 41.4% of Ministral Base samples, 44.9% of Ministral Instruct, 44.6% of Ministral Reasoning, and 40.3% of Gemma samples. Gemma's correct answer is ranked dead last internally in 15.0% of cases, a higher dead-last rate than any of the Ministral variants (11.9% to 13.0%)—a modest deficit on this particular indicator despite Gemma's overall structural superiority.

4.5 View Concordance

The AIDA framework's dual-view architecture provides a built-in reliability check. View agreement is broadly consistent across models: Ministral Base 72.9%, Gemma 72.8%, Ministral Reasoning 71.4%, and Ministral Instruct 68.5%. The instruct model's somewhat lower concordance reflects its larger Late Crystallisation population—the dominant source of geometric/logit disagreement—and is discussed below.

The dominant disagreement pattern is consistent across all four models. Late Crystallisation samples are classified as *Fused Gold* by the geometric view (no spatial differentiation) and as *Differentiated Correct* by the logit view (sharp output probabilities). The joint classification resolves this conservatively as Late Crystallisation. This single pattern accounts for all 186 such samples in Ministral Instruct, all 138 in Gemma, all 124 in Ministral Base, and all 97 in Ministral Reasoning. View concordance is therefore a direct function of the Late Crystallisation population: models with more shallow-correct trajectories show lower concordance, not because their two views are generally unreliable, but because one specific failure mode systematically separates them.

This relationship is structurally meaningful. The model with the largest Late Crystallisation population (Ministral Instruct, 186 samples, 17.1%) conversely shows the lowest concordance (68.5%). The model with the smallest (Ministral Reasoning, 97 samples, 8.9%) conversely shows the highest concordance among the Ministral family (71.4%). Gemma's moderate concordance (72.8%) matches its moderate Late Crystallisation count (138 samples, 12.7%). View concordance is a downstream indicator of processing depth, not an independent measure of diagnostic reliability.

4.6 FEST Fragility Profile

Table 7 presents the FEST results for all four models.

The FEST comparison reveals a three-tier fragility structure. Gemma-2-9B shows LOW fragility: its binary advantage is just 0.6 pp, meaning its discrimination between gold and the strongest attractor is barely affected by the presence of additional distractors. Its accuracy range across configurations (17.9 pp between F08 and F04) is the narrowest of the four models, indicating the most robust performance across option configurations. Its distractor concentration penalty

Table 7: FEST stage results: four models on MMLU-Med. Stages without gold answer (F01, F02, F07) omitted.

			Accuracy			
Stage	**Description**	**Options**	**Min. B**	**Min. I**	**Min. R**	**Gemma**
MCQ	Baseline 4-option MCQ	4	72.5%	79.8%	69.3%	77.2%
F03	Gold + D* + D_i	3	62.3%	77.2%	60.7%	75.2%
F04	Gold + D* + D_j	3	60.8%	76.2%	63.4%	75.3%
F05	Binary: Gold vs D*	2	77.7%	84.9%	81.9%	77.9%
F06	Binary: Gold vs D'	2	70.3%	87.5%	77.6%	87.1%
F08	Gold vs Weakest	2	90.5%	96.1%	93.1%	93.2%
F09	Restoration Control	4	72.5%	79.8%	69.3%	77.2%
Binary advantage (F05−MCQ)			+5.1	+5.1	+12.6	+0.6
Distractor conc. (F03−MCQ)			−10.3	−2.6	−8.6	−2.0
Accuracy range (F08−F04)			29.7	19.9	29.7	17.9
Fragility classification			MOD.	MOD.	**HIGH**	LOW
Test-retest (F09−MCQ)			0.000	0.000	0.000	0.000

(F03−MCQ: −2.0 pp) is also the smallest.

Both Ministral Base and Ministral Instruct show MODERATE fragility, with identical 5.1 pp binary advantages. This architectural invariant is preserved through instruction tuning, confirming that multi-option interference is a property of the Ministral family rather than a training artefact. The two models differ substantially on distractor concentration (−10.3 pp for Base versus −2.6 pp for Instruct), suggesting that instruction tuning specifically reduces sensitivity to the strongest attractor in 3-option configurations, even while leaving binary fragility unchanged.

Ministral Reasoning presents the most concerning FEST profile. Its binary advantage of 12.6 pp is the largest of the four models and earns a HIGH fragility classification—the model's performance degrades substantially when distractors are added to a binary confrontation. Its distractor concentration penalty (−8.6 pp) is the second largest. Across the range of configurations (F08 to the weakest 3-option set), its accuracy span is 29.7 pp, matching Ministral Base as the widest. The reasoning fine-tuning, while improving binary discrimination (81.9% at F05, well above base's 77.7%), introduces severe sensitivity to the multi-option context. A model that performs well in isolation degrades under the same distractor conditions that other models handle without incident.

The distractor concentration effect (F03 vs MCQ) is instructive across the four models. The pattern −10.3, −8.6, −2.6, −2.0 pp (Base, Reasoning, Instruct, Gemma) inversely tracks structural stability: the two models with the most volatile internal processing are most harmed when the strongest distractor is isolated into a 3-option set. Weak distractors in the full MCQ dilute the attractor's pull; removing them concentrates that pull, and the less stable models cannot resist it.

All four models achieve perfect test-retest reliability (0.000 pp), confirming pipeline validity across vendors, training stages, and fragility profiles.

4.7 Summary and Calibration Findings

4.7.1 Cross-Vendor and Cross-Stage Findings

The empirical evidence from four models across two vendors and three training stages establishes five findings at Level A2.

First, the epistemic gap is inverted in all four models: structural correctness exceeds outcome accuracy by 10.4–15.8 pp. This inversion—the model consistently knows more than it delivers—is the central structural finding of the assessment. It means that the dominant failure mode is not absent knowledge but suppressed delivery. The Correct Overridden population (11–16% of all samples depending on model) represents knowledge that exists internally but is displaced before output. The carrier decomposition reported in Section 7 provides the mechanistic explanation and the correction.

Second, instruction tuning produces balanced structural improvement. Ministral Base to Instruct gains +6.7 pp accuracy and +4.6 pp structural correctness; the epistemic gap narrows by 2.1 pp. This is a substantively different finding from what accuracy-only evaluation would report: the structural gain is real, even if proportionally smaller than the accuracy gain. The accuracy leaderboard and the epistemic leaderboard agree on the direction of improvement.

Third, reasoning fine-tuning degrades on every headline measure. The reasoning model scores 11.1 pp below the instruct model on accuracy, 5.7 pp below on structural correctness, and carries the widest epistemic gap of all four (−15.8 pp). Its HIGH FEST fragility (12.6 pp binary advantage) indicates that its accuracy in the standard MCQ setting is substantially eroded by the presence of distractors that other models handle robustly. The elevated Correct Overridden count (175 samples, 16.1%) suggests that the reasoning process actively interferes with and overrides existing structural knowledge. This is an invisible failure at Level A1—the accuracy figure of 65.9% understates the severity of the structural disruption.

Fourth, the trajectory classification scheme produces universal categorical boundaries. The fusion/differentiation divide is absolute at 0% across all four models—Differentiated Correct and fusion are mutually exclusive regardless of vendor, architecture, or training stage. This invariance supports the claim that AIDA measures a structural property of transformer inference, not a model-specific artefact.

Fifth, FEST reveals a fragility ordering that parallels structural stability: Gemma (LOW) $\ll$ Ministral Instruct $\approx$ Ministral Base (MODERATE) $<$ Ministral Reasoning (HIGH). The model with the most stable internal processing is least vulnerable to distractor manipulation; the model with the most volatile processing is most vulnerable. Fragility is not an independent failure mode but a surface expression of the same structural instability visible in flip counts, stability scores, and Correct Overridden rates.

4.7.2 Calibration as Precondition for Autonomous Governance

Beyond cross-vendor comparison, this section establishes the epistemic calibration that Levels B1 and B2 require. The regime-specific accuracy rates extracted from this assessment—the probability that a Differentiated answer is correct, the probability that a Late Crystallisation answer is correct, the probability that an Override or Fused answer is correct—are not general constants. They

are model-specific calibration values, empirically measured for each model on a gold-standard dataset. They cannot be assumed, estimated, or transferred from one model to another. A model's Differentiated Wrong rate is its own; it must be measured.

The calibration findings for the four models assessed are:

Table 8: Regime-specific calibration accuracy for the four models assessed in this work. Differentiated accuracy is computed from the Differentiated Correct and Differentiated Wrong trajectory counts. Override accuracy reflects output-level delivery; all Correct Overridden samples are recoverable via carrier correction (Section 7).

Model	**Differentiated**	**Late Rescue**	**Override**	**Fused**
Gemma-2-9B	85.5% (671/785)	100.0%† (138/138)	0%‡ (128)	7.9% (3/38)
Ministral-14B Base	80.9% (641/792)	100.0%† (124/124)	0%‡ (137)	2.8% (1/36)
Ministral-14B Instruct	88.0% (648/736)	100.0%† (186/186)	0%‡ (118)	10.2% (5/49)
Ministral-14B Reasoning	80.3% (618/770)	100.0%† (97/97)	0%‡ (175)	6.4% (3/47)

†Late Rescue accuracy of 100% reflects the classification boundary on this dataset, not a reliability guarantee. Late Rescue answers lack geometric structural support and should be treated as prompt-sensitive in deployment.

‡Override accuracy at output level is 0%: by definition, Correct Overridden trajectories have their intermediate correct representations displaced before the final output. The count shown is the number of recoverable samples; carrier correction (Section 7) can restore these answers without parameter modification.

The calibration table reveals several patterns invisible at Level A1. The Differentiated accuracy column shows that instruction tuning genuinely improves the reliability of clean processing: Ministral Instruct achieves 88.0% within the Differentiated regime, compared to its base model's 80.9%—a 7.1 pp improvement. Critically, this improvement is accompanied by a reduction in the total Differentiated population (792 to 736 samples): instruction tuning makes clean processing more reliable but converts some Differentiated samples into Late Crystallisation rather than enlarging the Differentiated population.

Reasoning fine-tuning degrades Differentiated accuracy to 80.3%, matching the base model despite additional training. Its Differentiated population (770 samples) is smaller than base (792), confirming that reasoning fine-tuning does not improve the clean-processing pathway and in some respects weakens it.

Gemma's calibration profile is competitive. Its Differentiated accuracy (85.5%) falls between Ministral Base and Ministral Instruct, and its Differentiated population (785 samples) is the largest of any model. Gemma processes more questions through the clean structural pathway than any other model assessed and achieves high accuracy within that pathway. Its Fused accuracy (7.9%) and Override recovery count (128 samples) are both within the normal range for the assessment.

The Fused regime carries near-zero predictive value across all models: accuracies of 2.8–10.2% are consistent with random selection from a small number of chance-favoured responses. Fused answers must be rejected categorically at deployment—there is no rescue path through the geometry.

These calibration values are the bridge between Sections 4 and 9. Section 9 presents the

autonomous assessment framework that deploys these values at inference time without access to gold answers. Every accuracy rate cited at Levels B1 and B2 traces back to the measurements reported here.

4.7.3 Production Calibration Reports

The calibration process described in this section is not a research exercise—it is an operational capability. Appendices A, B, C, and D reproduce complete AIDA internal assessment reports for all four models, generated automatically by the AIDA assessment pipeline. Each report comprises:

- Executive summary with headline metrics (outcome accuracy, structural correctness, epistemic gap, fusion rate, stability score, flip count)
- Full trajectory classification with per-regime fusion rates
- Layer dynamics analysis (decision zone distribution, collapse layer profiles)
- Stability analysis across four indicators with individual pass rates
- Decision volatility profiling (flip count distribution)
- Entropy sharpening analysis (early-layer to final-layer distributions)
- Internal ranking analysis (gold answer rank distribution)
- View concordance and disagreement pattern analysis
- Centroid shift analysis (where differentiation begins)
- Fusion pattern analysis with regime-specific fusion rates
- Complete FEST fragility profile across nine configurations
- Key findings and recommendations

Each report is generated from a minimum of 45,738 ETC layer probes per model, multiplied by ASCOL template repetitions across the full question set. Reports are certificate-referenced with a unique Report ID, and are designed for regulatory audit under the GPAI obligations of the European Artificial Intelligence Act (Section 11).

Appendix E reproduces a complete AIDA comparative assessment report evaluating a specific model replacement decision: whether Meta Llama-8B should replace Google DeepMind Gemma-9B as the deployed model. This report demonstrates the operational use case for epistemic calibration—a governance decision that requires instruments beyond aggregate accuracy. Switching from Gemma-9B to Llama-8B reduces outcome accuracy by 17.2 pp and structural correctness by 14.2 pp, while Llama-8B's mean flip count of 27.6 (versus Gemma's 6.0) reveals far greater internal volatility. The recommendation is unambiguous: replacement is not recommended. Llama-8B degrades every headline metric, a conclusion that only epistemic calibration can reach with full transparency. The comparative report provides the evidence that accuracy-only evaluation structurally cannot.

These five reports—four internal assessments and one comparative—constitute the Level A2 calibration artefacts for the models assessed in this work. They are the foundation upon which Level B1 and B2 autonomous governance is built, and the documentation that GPAI compliance requires.

5 The Rotation Discovery: Representational Rotation in Transformer Inference

5.1 The Structural Fork: Crystallisers vs. Non-Crystallisers

The trajectory overlays in Figure 5.1 and 6.2 present the central empirical puzzle that motivates the remainder of this work. They are averaged across hundreds of samples from Llama-3-8B on MMLU-Med (involving 275 late crystallisers, 341 never crystallisers, 32 layers), and they reveal a fact that no output-level metric could suggest: the model follows an identical internal deliberation script regardless of whether it ultimately answers correctly or incorrectly.

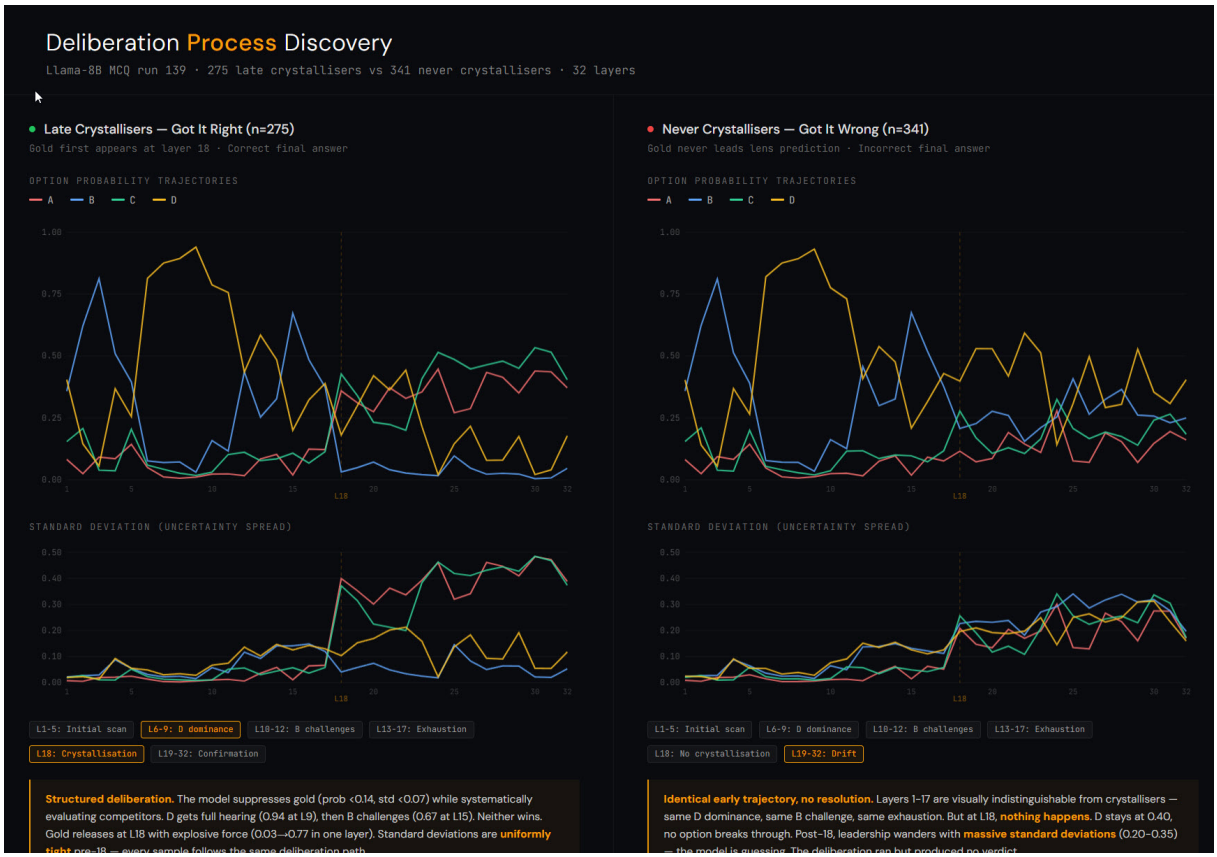


Figure 9: Deliberation Process Discovery. Left: Late Crystallisers ($n = 275$), showing option probability trajectories and standard deviation across 32 layers, with gold first appearing at Layer 18. Right: Never Crystallisers ($n = 341$), showing identical early trajectory dynamics but no crystallisation event. Phase annotations mark the shared deliberation stages: Initial Scan (L1–5), D Dominance (L6–9), B Challenge (L10–12), and Exhaustion (L13–17).

Across both populations, the first seventeen layers follow the same structural sequence:

Initial scan (L1–5). All four options begin with low, tightly clustered probabilities. The model is reading the question. Standard deviations are near zero.

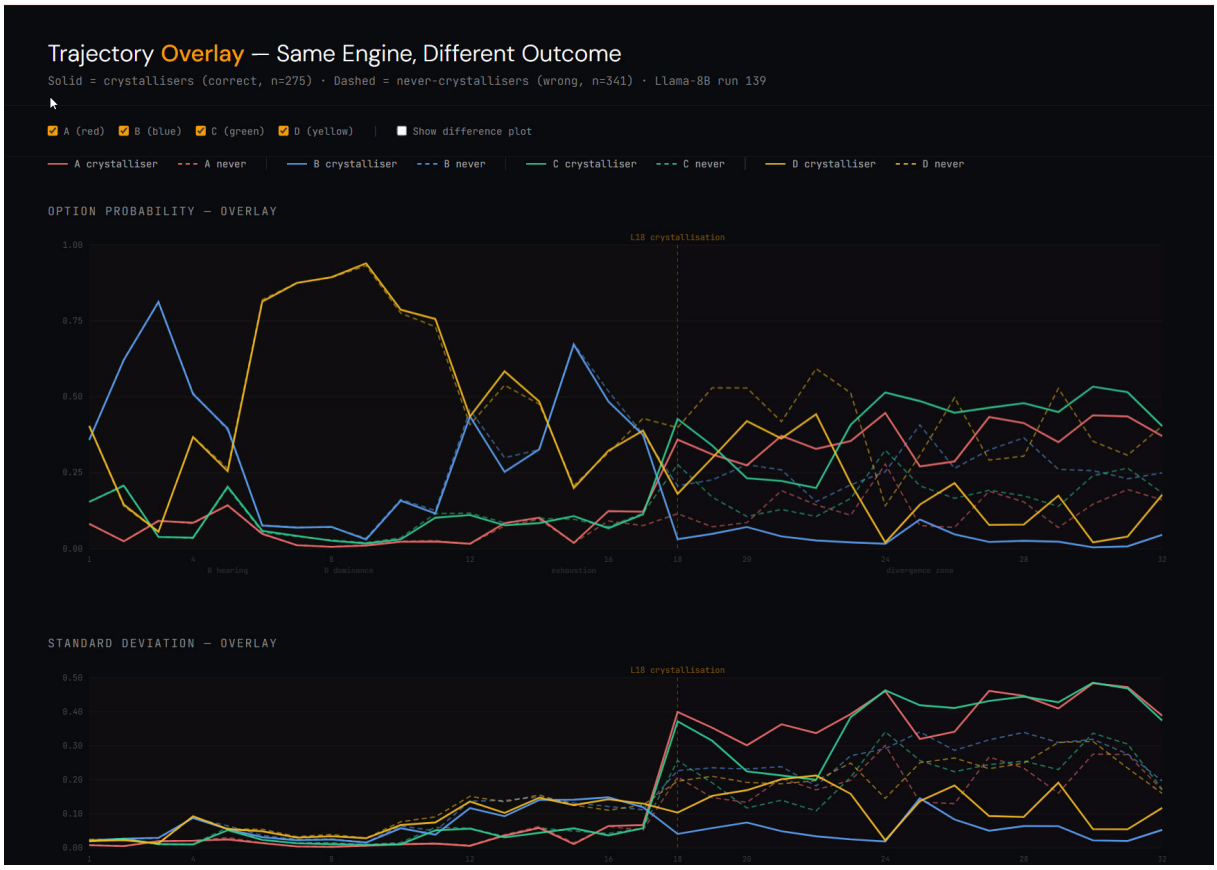


Figure 10: Trajectory Overlay—Same Engine, Different Outcome. Solid lines represent crystallisers (correct, $n = 275$); dashed lines represent never-crystallisers (wrong, $n = 341$). The overlay confirms that trajectories are indistinguishable before Layer 18, after which the crystallisation event separates them. Standard deviation overlay (lower panel) shows uncertainty collapse in crystallisers versus uncertainty expansion in non-crystallisers.

D dominance (L6–9). Option D surges to high probability (peaking above 0.90 in many samples), while the other three options are suppressed. This is not a content-driven judgement; it is a positional artefact. D, as the last option token before the answer position, has the shortest autoregressive path to the output head and begins life pre-aligned with the unembedding projection.

B challenge (L10–12). Option B rises to challenge D, and the two trade dominance. Options A and C remain suppressed. Standard deviations begin to rise as the model enters a period of internal competition.

Exhaustion plateau (L13–17). The D–B contest subsides without resolution. Probabilities flatten. Standard deviations remain low. The model has completed its positional routing but has not yet engaged its content-evaluation circuits.

Up to this point, the trajectories of successful and failed answers are *visually indistinguishable*. The overlay in Figure 5.2 confirms this: solid lines (crystallisers) and dashed lines (non-crystallisers) track one another with near-perfect fidelity through all four phases. The model's internal geometry is following the same structural script.

The divergence occurs abruptly at **Layer 18**:

In *late crystallisers*, the manifold rotates into a configuration where the gold option becomes structurally dominant and uncertainty collapses. The gold option surges from near-zero to high probability in a single layer transition. Standard deviations drop sharply as the model commits to a single answer. Layers 19–32 are confirmation: the gold signal strengthens steadily and the decision locks in.

In *never crystallisers*, the same manifold fails to rotate. D remains at approximately 0.40, no option breaks through, and from Layer 18 onward the leadership wanders among options with massive standard deviations (0.20–0.35). The deliberation ran its full course and produced no verdict.

> *The model structurally knows the answer in both cases, but only one trajectory delivers it. The failure is not ignorance but a geometric misalignment in the model's internal manifold.*

5.2 The Rotation Discovery

The trajectory overlays reveal *that* a geometric bifurcation occurs. They do not explain *what* is happening mechanistically. The rotation discovery, made during live analysis of Llama-8B on a single sample (Sample 5, Gold=B, Predicted=A), provides the answer.

5.2.1 The Puzzling Observation

The logit lens chart for Sample 5 showed a pattern that defied the standard interpretation. Options B and D dominated the visible probability trace throughout the early and middle layers, swapping dominance as a matched pair. Options A and C were essentially invisible—both flat near zero probability. This was unremarkable, consistent with thousands of other samples.

At Layer 18, option A surged from 9% to 66% probability in a single layer transition. B collapsed from 59% to 9% simultaneously. The standard interpretation would be straightforward: A was amplified, B was suppressed. Attention heads boosted A and killed B. But when the representation norms were examined, this interpretation collapsed.

5.2.2 The Evidence: Norms Don't Lie

The following data captures the critical transition between Layer 17 and Layer 18:

Table 9: The critical L17→L18 transition. Probabilities flip by 56 percentage points; the A/B norm ratio changes by two thousandths.

	prob A	**prob B**	**norm A**	**norm B**	**A/B norm**
L17	.094	.586	7.51	8.72	0.862
L18	.659	.093	9.04	10.47	0.864
Δ	**+56.5 pp**	**−49.3 pp**	**+1.53**	**+1.75**	**+0.002**

The A/B norm ratio moves by **two thousandths** while the probability flips by 56 percentage points. Both norms grow at approximately the same rate (A: 7.51→9.04; B: 8.72→10.47). The total energy increases smoothly with no discontinuity. Nothing was amplified. Nothing was created. Nothing was suppressed. The representational vectors *rotated*.

5.2.3 The Efficiency Analysis

Dividing probability by norm gives efficiency—the fraction of each option's total representational energy that is aligned with the unembedding projection. At Layer 18, the efficiency figures are stark:

Table 10: Efficiency analysis at Layer 18. Option B (gold) has the highest norm but the second-lowest probability. 99% of its energy is invisible to the output head.

Option	Norm	Prob	Efficiency	In null space
A	9.04	.659	0.073	∼93%
B (Gold)	10.47	.093	0.009	∼99%
C	10.03	.069	0.007	∼99%
D	10.09	.179	0.018	∼98%

Option B—the correct answer—has the **highest norm of all four options** but only 9.3% probability. Approximately ninety-nine percent of its representational energy is oriented in a direction the output head cannot read. The logit lens chart showing B as weak is an artefact of projection angle, not a measurement of representational strength. B is not absent from the model's representation. It is hidden by geometry.

5.2.4 One Space, Not Two

The standard logit lens view creates an illusion of two separate contests: B and D visible and dominant in early layers, A and C invisible. The natural interpretation is that B and D are strong while A and C are weak. This interpretation is wrong.

All four options carry comparable energy throughout the model. The unembedding matrix is a fixed projection—a camera at one angle. B and D happen to be aligned with this projection plane, so they register as probabilities. A and C's energy is orthogonal to the projection. If one could rotate 90 degrees around the appropriate axis, A and C would appear with the same amplitude, and B and D would flatten to nothing.

> *There aren't two separate spaces. There is one space with four options carrying comparable energy, and the logit lens is a camera fixed at one angle.*

5.2.5 Rotation, Not Amplification

At Layer 18, A does not surge. A's energy was already present in full. What happens is a rotation of A's representational vector into alignment with the unembedding projection. It suddenly appears on the camera. Simultaneously, B's vector rotates away from the projection plane, and it vanishes from the logit lens view. The energy does not change—the angle does.

The proof is in the norm ratio: A/B = 0.862 at L17, 0.864 at L18. If A had been amplified and B suppressed, this ratio would shift dramatically. Instead, both norms grow smoothly, and the total energy increases without discontinuity. The three-circuit architecture identified in earlier runs—the Kingpin Amplifier, the Knowledge Reader Bank, the Pivotal Swing Head—must be reframed. These circuits are not amplifying options. They are controlling the rotation angle of the representational space relative to the fixed output projection.

The rotational interpretation of logit lens dynamics is mathematically consistent with the transformer architecture and in that sense was always available as a theoretical possibility. The output logit for any token is the dot product of the hidden state with the corresponding lm_head weight row; probability shifts therefore reflect changes in alignment angle as much as changes in representational magnitude. A reader familiar with this formulation might ask whether the rotation interpretation follows trivially from code inspection.

It does not, for three reasons. First, the methodology ran in the correct scientific direction: the rotation interpretation was not derived from reading the source code and then confirmed against data. It was discovered empirically through live instrumentation of the forward pass, after which architectural analysis was conducted to identify the mechanistic source. FEST pairwise testing then provided independent experimental validation of the theorem derived from that analysis, achieving recovery rates that confirm the prediction. Code inspection followed the finding; it did not precede it.

Second, the magnitude of the null-space energy could not have been predicted from architectural inspection alone. That 99% of the gold answer's representational energy was oriented in a direction the output head could not read at Layer 18, that D's lm_head row is 93.1% carrier-aligned, and that the accuracy differential between positions exceeds 20 percentage points on the same evaluation corpus are quantitative findings requiring measurement. The architectural geometry permits the phenomenon; the data establishes how large it is.

Third, and most importantly, the forward pass architecture actively destroys the evidence required to observe these dynamics through any normal interface. Attention weights are discarded before the return path of LlamaDecoderLayer.forward(). Gate activations in LlamaMLP.forward() are computed inline and immediately consumed. The neuron-level analysis, the gate signature characterisation, and the D-writer investigation documented in Section 6 required forward hooks instrumenting the module internals at runtime — the architecture provides no observability pathway into its own routing decisions. The claim that code reading could substitute for empirical instrumentation is precisely falsified by the architecture of the system being studied.

> *The battle isn't about who has the most energy. It's about who gets to point at the camera.*

5.3 Norm Stability and Co-Representation

The rotation interpretation demands that option norms remain stable while probabilities fluctuate. The data confirms this dramatically. B and D track within approximately ± 0.6 across all 32 layers of the stack:

B and D are not merely geometrically similar (cosine similarity 0.72–0.88 across layers). They carry identical energy at every layer. They are co-represented: the model processes them as a pair because they share both direction and magnitude. This co-representation is the mechanism behind the BD coupling identified in earlier runs, where 41.4% of wrong samples placed B and D in the top-2 rank positions.

A and C diverge only after Layer 22, when A's norm accelerates sharply away from the group. This is the point at which A's vector has fully rotated into alignment with the output projection

Table 11: Norm stability across the full 32-layer stack. B and D track within approximately ±0.6 at every layer. A diverges from C only after L22 when it rotates into output alignment.

Layer	norm B	norm D	B−D	norm A	norm C	A−C	
L1	0.90	0.90	+0.00	0.81	0.86	−0.05	
L8	4.24	4.50	−0.26	3.95	4.42	−0.47	
L17	8.72	8.65	+0.07	7.51	8.40	−0.89	
L18	10.47	10.09	+0.38	9.04	10.03	−0.99	← *rotation*
L23	15.56	15.43	+0.13	17.45	15.73	+1.72	
L27	20.73	20.10	+0.63	26.82	20.75	+6.07	
L30	28.41	28.82	−0.41	35.94	29.68	+6.26	

and begins accumulating energy in that visible direction. Before L22, A is carrying comparable energy to every other option—it is simply pointing the wrong way.

5.4 The BD Relay and Its Spectral Correction

The rotation discovery explained what happens at a single critical layer. The next question was whether this rotational dynamic could explain the *population-level* patterns visible in the trajectory overlays. Analysis of the six pairwise option deltas (wrong-minus-correct, averaged across all 1,089 samples) across three functional regions of the model initially suggested a relay mechanism, which was subsequently corrected by spectral analysis.

5.4.1 The Initial Relay Hypothesis

Decomposing the six pairwise deltas to individual options revealed a striking sequential pattern. In the contest region (L18–23), D gained +0.030 while A lost −0.049—D was taking from A. In the decision region (L24–32), B surged to +0.044 while D went quiet at +0.004—B had taken over. They appeared to be taking turns.

The most dramatic evidence appeared at Layer 21: D dropped by exactly −0.039 and B rose by exactly +0.039. A perfect baton pass, to three decimal places. The natural interpretation was a conserved transfer—a relay mechanism where D secured a rotational advantage and handed it to B.

5.4.2 The Spectral Correction

A discrete Fourier transform of the pairwise deltas across 32 layers falsified the relay interpretation and replaced it with a sharper finding. The four options operate at **fundamentally different timescales**:

Table 12: Spectral fingerprints of the four options (Llama-8B, wrong-minus-correct delta, L17–32).

Option	Dominant freq	Power	Zero crossings	DC offset	Character
A	k=1 (32L)	56%	0	−0.023	Monotonic drain
B	k=2 (16L)	28%	4	+0.020	Slow oscillator
C	Broadband	35%	—	≈0	Incoherent
D	k=9–11 (3–4L)	58%	8	+0.002	Fast oscillator

A is not oscillating at all. It is negative at every single layer from L17 to L32—zero crossings in 15 intervals. Its 56% concentration at k=1 means it behaves as one broad wave across the full stack: a steady, coherent drain. A is the permanent victim of the model's error process.

D is the opposite: 8 zero crossings in 15 intervals, flipping sign every 1–2 layers. Its autocorrelation is -0.63 at lag-2 and $+0.55$ at lag-3, confirming a period of approximately 3.5 layers. D is a fast, jittery aggressor—it punches the manifold repeatedly from multiple directions. Its total influence is the largest of any option (power 0.551), but it arrives in rapid bursts rather than a sustained push.

B operates at an intermediate timescale: period ≈16 layers, with 4 zero crossings concentrated in the first half. B ramps up slowly and stays positive through most of the decision region. It is an accumulator, not a relay partner.

C carries the weakest total signal and has no dominant frequency. It is spectrally incoherent—a bystander.

> *The ±0.039 symmetry at Layer 21 was a numerical coincidence of two different-frequency processes crossing at the same amplitude. The model's error mechanism is not a relay. It is frequency separation.*

The corrected mechanistic picture is multi-band interference on top of a monotonic drain: A provides the background slope (steady suppression), D provides fast perturbation (jittery aggression), B accumulates advantage slowly, and C does nothing structured. The phenomenological observations from the trajectory overlays—D dominates the contest, B dominates the decision, A is always suppressed—remain correct. The mechanism is spectral, not sequential.

5.5 The Carrier Signal: Cross-Model Validation

The findings of Section 5.2–5.5 were derived from Llama-8B. To test whether they are architectural or model-specific, the same wrong-minus-correct delta analysis was applied across six models spanning four architecture families, three parameter scales, and six suppliers: Llama-8B (Meta), Mistral-7B (Mistral AI), Qwen-7B (Alibaba), OLMo-7B (AI2), Gemma-9B (Google), and R1-14B (DeepSeek).

5.5.1 The Universal Pattern

Every model exhibits a clear victim/beneficiary asymmetry in its error dynamics. The critical finding is structural:

Table 13: Victim/beneficiary structure across six models. The victim is always an endpoint option (A or D), never an interior position.

Model	Victim	Beneficiary	Victim DC
Llama-8B	A	B	−0.043
Mistral-7B	D	B	−0.062
OLMo-7B	D	B	−0.013
Qwen-7B	D	C	−0.022
Gemma-9B	D	A	−0.046
R1-14B	D	C	−0.082

As observed, the victim is notably an endpoint option: A (position 1) or D (position 4). Never B, never C. The interior positions are structurally protected. In five of six models, D is the victim. Llama-8B is the sole exception, with A as the victim and D advantaged. Gemma-9B is a precise mirror image of Llama: D is the victim, A is the beneficiary. The specific assignments differ by architecture; the structural phenomenon does not.

In every model here, the victim's spectral character is consistent: slow-spectrum-dominant (41–88% of power in k=1–3), with few or zero crossings. The victim is always the most spectrally coherent option—its suppression is steady and sustained, not jittery or oscillatory.

5.5.2 The Carrier Signal Defined

These population-level dynamics—the victim, the beneficiary, the spectral asymmetry—operate on every sample regardless of question content, correct answer identity, or model confidence. They are the **carrier signal**: the structural positional dynamics of the architecture performing MCQ inference.

Each model has its own carrier, determined by its specific unembedding geometry—the default angle between option token representations and the output projection plane. Different models learn different projection planes during training, so different options end up pre-aligned or mis-aligned. This is fully consistent with the rotation discovery: the carrier is the default angle, and the content signal is the per-sample rotation away from that default.

5.6 Demodulation: Isolating the Content Signal

If the carrier is the medium, the content signal is the message. The carrier–content decomposition isolates what the model actually knows about each specific question by subtracting the positional baseline from the observed probabilities:

$$\text{content\_residual}(X, \text{sample}) = P(X \mid \text{sample}) - E[P(X) \mid \text{gold} \neq X]$$

This isolates what the model knows about *this specific question*, stripped of what position alone provides.

5.6.1 The Threshold Effect

Applying the decomposition to Llama-8B at the decision layers reveals the interaction between carrier and content:

Table 14: Carrier–content decomposition for Llama-8B at decision layers (L25–32).

Gold	Content (correct)	Content (wrong)	Carrier baseline	Accuracy	
A	0.738	+0.072	∼0.06	61.3%	*Strong content, low carrier*
B	0.765	+0.122	∼0.14	63.8%	*Strongest content*
C	0.763	+0.093	∼0.09	63.7%	*Strong content, low carrier*
D	0.565	−0.027	∼0.22	81.5%	*Weak content, high carrier*

D has the highest accuracy (81.5%) but the *weakest* content signal (0.565). D does not need much content signal to win because the carrier already places D at a ∼0.22 baseline. A small content push is sufficient.

A has the lowest accuracy (61.3%) but the *strongest* content signal when correct (0.738). This is a selection effect: only A-gold samples with very strong content signal can overcome A's carrier disadvantage. Weak content signals fail at position A but succeed at position D.

The wrong samples are equally revealing. When the model fails at Gold=A, the content residual is +0.072—the model has genuine partial knowledge that the carrier suppresses. When the model fails at Gold=D, the residual is −0.027—the model has actively concluded D is wrong. These are qualitatively different failure modes: carrier-suppressed knowledge versus genuine ignorance.

> *The carrier is not distorting the model's knowledge. It is distorting the model's expression of that knowledge. The same content signal, passed through different positional thresholds, produces a 20-point accuracy gap.*

The same decomposition applied to Mistral-7B produces qualitatively identical results despite a different architecture and carrier profile. D again has the highest accuracy with the weakest content signal when correct and a negative residual when wrong. The pattern is universal: D rides the carrier, A fights it.

The Complete Picture

The model's MCQ inference has two separable components:

The carrier. A position-dependent baseline probability set by the model's unembedding geometry. It is the same for every question. It favours certain positions (typically D) and penalises others (typically A). It is the medium.

The content signal. A per-sample modulation that encodes what the model actually knows about each specific question. It is roughly position-independent in strength. It is the message.

The model's accuracy at each position is determined by the interaction between these components. When the carrier and content signal align—high carrier for the correct answer—accuracy is high with minimal content effort. When they oppose—low carrier for the correct answer—accuracy depends on whether the content signal is strong enough to overcome the carrier deficit. The 20-point accuracy gap between positions is not a 20-point knowledge gap. It is a threshold effect: the same distribution of content signal, filtered through different carrier baselines, produces different pass rates.

This decomposition transforms the trajectory overlays of Figure 5.1 and 6.2 from a descriptive observation into a mechanistic explanation. The crystallisation event at Layer 18 is a *rotation* of the content signal into alignment with the output projection, overcoming the carrier's default angle. When the content signal is strong enough, the rotation succeeds and the model crystallises on the correct answer. When it is not, the carrier's default geometry prevails, and the model drifts into an incorrect prediction—not because it lacks knowledge, but because the geometric threshold was not cleared.

> *The model was never ignorant. It was geometrically misaligned. The knowledge was present but stored in a direction the output head could not read.*

6 Architectural Analysis

6.1 Three Layers of Mechanistic Explanation

The carrier is not a single-cause phenomenon. Three architectural layers contribute, and they must be distinguished because they have different implications for remediation:

Layer 1: Runtime Geometry — Causal Mask Asymmetry

The autoregressive causal mask enforces that each token attends only to itself and prior tokens. In the MCQ token sequence — question stem followed by options A, B, C, D — option D's hidden state is constructed from attention over the full preceding context, including options A, B, and C. Option A's hidden state is built from the question stem only. The four answer options are evaluated under structurally unequal attention conditions by architectural design, independent of question content. D has structurally richer attended material than A at every layer of the network. This runtime asymmetry is a direct contributor to the carrier's victim/beneficiary structure and is not addressable by fine-tuning — it is a property of the inference code path, not of the learned parameters.

Layer 2: Projection Geometry — Tied Weights LLaMA-3, in common with most transformer language models, declares tied weights between the input embedding matrix (embed_tokens) and the output projection (lm_head). The same weight matrix serves two opposing roles: encoding input tokens into the hidden space at the start of the network, and projecting final hidden states back to vocabulary logits at the end. The embedding geometry of the answer-option tokens A, B, C, and D as input tokens is therefore directly reflected in the lm_head readout geometry. The carrier direction computed in this work — the mean of lm_head weight rows for A, B, C, and D, unit-normalised — is simultaneously the average input embedding direction for those tokens. The 93.1% carrier alignment of D is a structural consequence of weight tying, not a training artefact that further fine-tuning on the same architecture could correct. Untying lm_head from embed_tokens would allow the output projection to be trained independently of the input encoding, breaking the geometric lock that produces the carrier; but within the current architecture, every training step deepens rather than relaxes this coupling.

Layer 3: Pre-training Geometry — RoPE-Baked Positional Bias

Rotary Position Embedding operates by rotating query and key vectors in paired dimensions according to token position before any attention dot product is computed. Crucially, because RoPE is applied during pre-training from the very first gradient update, the weight matrices do not merely use positional rotation at inference — they are trained to expect it. Over the course of pre-training on corpora in which ABCD-format questions appear with consistent positional structure, the model's weights develop responses calibrated to the rotational angle corresponding to D's position in the answer sequence. This is not a runtime artefact and is not an incidental training outcome — it is the necessary consequence of pre-training a RoPE model on data where answer-position tokens appear at consistent positions across billions of training examples.

This third layer is the most architecturally escape-proof of the three. The causal mask asymmetry is addressable by architectural change. The tied-weight geometry is addressable by decoupling embed_tokens from lm_head. But RoPE-baked positional bias cannot be addressed without retraining from scratch on position-randomised data — and even then, if RoPE is retained as

the positional encoding scheme, the same pre-training dynamics would re-emerge on any corpus with consistent ABCD positional structure. This explains why the victim/beneficiary asymmetry documented in Table 13 is observed across all six models assessed, spanning four architecture families and six suppliers: every model in this evaluation uses RoPE, and every model was pre-trained on corpora where answer-position tokens occupy predictable positions.

6.2 Forward Pass Observability Constraints

Two structural properties of the LLaMA-3 forward pass bear directly on the instrumentation methodology. First, LlamaDecoderLayer.forward() discards the attention weight matrix before the return path: the weights are computed, consume memory and GPU time, and are immediately lost. No attention routing information is accessible at inference without forward hooks instrumenting the module internals. Second, LlamaMLP.forward() computes gate activations inline and does not surface them through any return value. The gate magnitudes, neuron firing patterns, and the full D-writer analysis documented in this work required runtime hooks specifically because the architecture provides no observability pathway into these computations. The instrumentation developed for AIDA was necessary, not merely convenient: the phenomena under study are actively hidden by the forward pass design.

6.3 Why Fine-Tuning Cannot Resolve These Issues

Fine-tuning adjusts learned parameters. It cannot modify the inference code path. The causal mask asymmetry, the single shared position embedding computed before the decoder loop, and the position-reset behaviour in multi-turn contexts are properties of the forward pass logic. No gradient update reaches them. The tied weight coupling (Layer 2 above) is an architectural constraint that training reinforces rather than relaxes — every training step pushes embed_tokens and lm_head toward the same optimum, deepening the geometric coupling that produces the carrier. The absence of a position-independent attention pathway is a design decision that training cannot override because the routing through the rotary embedding is hardcoded. The practical consequence is that performance improvements achievable through the AIDA inference-time correction pipeline are improvements that a standard fine-tuning regime, operating against this codebase, cannot replicate through the same mechanism.

7 Carrier Decomposition and Epistemic Correction

Recovering Hidden Knowledge Without Training

7.1 The Contaminated Measurement

Section 5 established that the model's internal inference operates through a rotational geometry in which a position-dependent carrier signal and a content-dependent knowledge signal combine to produce the output. The carrier favours certain answer positions regardless of question content, while the content signal encodes what the model actually knows. This section presents the experimental programme that exploits that decomposition to recover hidden knowledge from the model's existing weights, without any training, any parameter changes, or any degradation of previously correct outputs.

The starting point is a stark empirical fact. Llama-3-8B achieves 67.9% on MMLU-Med under the corrected space-prefixed token instrument (run 287, 739 correct out of 1,089 samples). The positional bias underlying this figure is demonstrated in Table 15, drawn from the carrier characterisation run on which the decomposition was developed. This score is widely treated as a measure of the model's medical knowledge. It is not. Per-position accuracy reveals the carrier's fingerprint:

Table 15: Positional bias in Llama-3-8B on MMLU-Med. D receives carrier probability 0.267—over three times A's 0.085—regardless of question content.

Gold Position	Samples	Correct	Accuracy	Carrier Prob	
A	235	143	60.9%	0.085	
B	254	161	63.4%	0.130	
C	248	159	64.1%	0.105	
D	352	285	81.0%	0.267	← *3× A's carrier*

The D–A accuracy spread is 20.1 percentage points. The D–A logit spread is stable across all six MMLU-Med subjects (1.43–1.68), confirming this is a model property derived from the geometry of the unembedding matrix, not a dataset artefact. The baseline accuracy figure conflates genuine knowledge with positional luck in both directions: inflated for D-gold questions, deflated for A-gold questions.

7.2 The Carrier Direction

The carrier direction is a single vector in 4,096-dimensional hidden-state space that produces the entire positional bias when projected through the language modelling head. It is computed directly from the lm_head weight matrix with zero data and zero inference—a closed-form computation derived from the model's frozen geometry.

The alignment of each token's lm_head weight row with the carrier direction reveals how much of each position's output logit is structural bias versus genuine content:

Table 16: Carrier–content decomposition per token. D's `lm_head` row is 93.1% carrier—almost a pure carrier antenna. A retains the most room for content (26.9%). Reconstruction error: 10^{-6}.

Token	Carrier Alignment	Carrier %	Content %
A	0.731	73.1%	26.9%
B	0.806	80.6%	19.4%
C	0.747	74.7%	25.3%
D	0.931	93.1%	6.9%

D's lm_head row is **93.1% carrier**. Only 6.9% of D's output logit comes from actual content signal. This is why D dominates: the model's D token is almost a pure carrier antenna. When the model assigns high probability to D, it is mostly measuring position, not knowledge. A has the most room for content at 26.9%, which explains why A-gold samples, when they succeed, require the strongest content signals—the carrier provides almost no assistance.

The decomposition is mathematically exact, with reconstruction error of the order of 10^{-6}. This is not an approximation. The carrier and content components sum precisely to the observed logit.

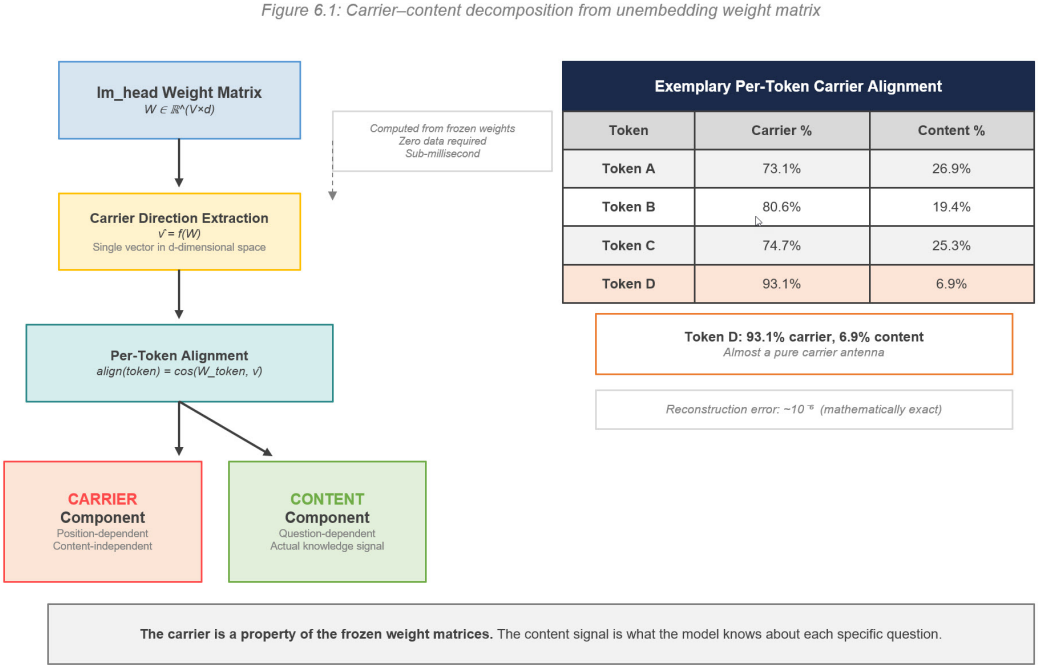


Figure 11: Carrier–content decomposition from the unembedding weight matrix. The carrier direction $\hat{v}$ is extracted from the `lm_head` matrix W. Per-token alignment reveals the carrier and content fractions for each answer position. Computed from frozen weights with zero data required.

7.3 The Correction Pipeline

The carrier–content decomposition enables a family of inference-time interventions. Each exploits a different aspect of the rotational geometry. None modifies model weights. The interventions were developed and tested sequentially, each building on the previous stage's results.

7.3.1 Experiment 1: Carrier Correction at the Logit Level

The simplest intervention is a four-number subtraction: subtract the per-position carrier baseline from each option's probability before taking the argmax. This is a post-hoc output correction that does not touch the model's internals.

Table 17: Logit-level carrier correction. The positional spread collapses from 20.1 pp to 4.9 pp. Net gain: +18 samples (54 rescued, 36 broken).

Method	A%	B%	C%	D%	Spread
Raw	60.9	63.4	64.1	81.0	**20.1 pp**
Corrected	67.2	70.1	72.2	71.3	**4.9 pp**

The spread collapses from 20.1 to 4.9 percentage points. The correction rescues 54 samples but breaks 36, for a net gain of only 18. This is instructive: logit-level correction operates too late. By the time probabilities reach the output, the carrier has already shaped the internal representation. The correction must go deeper.

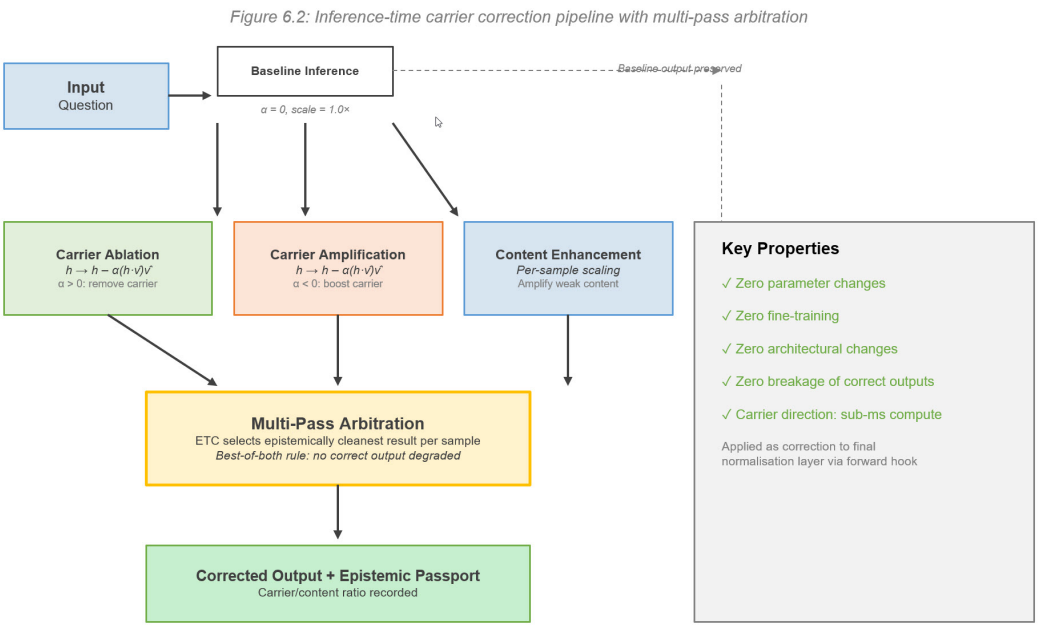


Figure 12: Inference-time carrier correction pipeline with multi-pass arbitration. Input passes through baseline inference, carrier ablation, carrier amplification, and content enhancement. The ETC selects the epistemically cleanest result per sample. Zero parameter changes, zero training.

7.3.2 Experiment 2: Content Signal Enhancement

All 341 baseline-wrong samples were tested across embedding scale factors from 1.09× to 25.0×. Of these, 174 (51%) flipped to gold. These comprise two mechanistically distinct populations that the carrier–content framework predicted in advance:

Population A (101 samples, predicted ≠ D): The model has an incorrect content signal that dominates. Scaling weakens it and lets D's carrier push gold through. *The carrier is the cure.* Average flip power: 4.6×.

Population B (73 samples, predicted = D): Classic carrier suppression. The model's knowledge of the correct A/B/C answer is present but the carrier blocks it. Scaling weakens D's advantage and lets content emerge. *The carrier is the disease.* Average flip power: 9.2×.

Per-sample optimal scaling—giving each recoverable sample its known best scale factor—yields 739 → 864 correct (79.3%). Zero breakage of previously correct samples.

7.3.3 The Reproducibility Inversion

An unexpected finding emerged from replication testing. Samples requiring high scale factors (8–25×) reproduced at 85%, while those requiring low factors (1.1–1.5×) reproduced at only 38%. The explanation follows directly from the carrier–content decomposition: low-scale samples sit on a knife-edge where the content signal barely exceeds the carrier threshold, and stochastic variation in the forward pass can tip the outcome either way. High-scale samples receive a decisive intervention that overwhelms noise. The "harder to rescue" samples provide more reliable results.

Table 18: Scale band reproducibility. A sharp phase transition separates stochastic (knife-edge) from deterministic (reliable) interventions at approximately 3×.

Scale Band	Total	Flipped	Rate	Character
1.0–1.2×	30	4	13%	Stochastic (knife-edge)
1.2–2.5×	32	19	59%	Stochastic (marginal)
3.0–5.0×	10	10	100%	Deterministic (reliable)
7.0–25×	72	62	86%	Deterministic (reliable)

7.3.4 Experiment 3: Cumulative Nudge Sweep

The 58 non-reproducible samples were attacked with progressively stronger multipliers using cumulative locking: once a sample flips to gold, it is banked and removed from subsequent rounds.

Table 19: Cumulative nudge sweep. Progressively stronger multipliers with locking. Diminishing returns confirm that most recoverable samples respond to moderate intervention.

Nudge	New Flips	Cumulative	Remaining	Combined Accuracy
+30%	16	16	42	880 / 1,089 (80.8%)
+100%	6	27	31	891 / 1,089 (81.8%)
+200%	7	34	24	898 / 1,089 (82.5%)
+500%	2	42	16	906 / 1,089 (83.2%)

43 of 58 recovered. Combined with content enhancement, the scaling pipeline total reaches 907/1,089 (83.3%). 182 samples remain unreachable by embedding scaling at any power level.

7.3.5 Experiment 4: Carrier Direction Ablation

The decisive intervention operates directly on the carrier direction in the model's internal representation. A forward hook on the final layer normalisation projects out the carrier component before the unembedding projection:

$$h \to h - \alpha(h \cdot \hat{v})\hat{v}$$

where $\hat{v}$ is the carrier direction and α controls the strength of removal. Positive α removes the carrier (rescuing A/B/C-gold samples); negative α amplifies it (rescuing D-gold samples where the content signal is wrong). The sweep reveals the carrier's full structure:

The zero-breakage property is an architectural guarantee of the best-of-both rule, not an empirical discovery. Each intervention pass — baseline, carrier ablation at each alpha level, carrier amplification, content enhancement — may individually rescue some samples and break others. The best-of-both rule is defined such that a sample that is correct at baseline is never overwritten by any subsequent pass. Breakage is therefore structurally impossible by construction: the rule defines "correct at baseline" as the protected state. The zero-breakage claim should be read as confirming that the implementation correctly realises this architectural guarantee across 1,089 samples, not as asserting a non-obvious empirical finding.

The table reveals the carrier in real time. As α increases from 0 to 1.5, D's accuracy collapses from 81.0% to 27.8%—the carrier is being stripped away, and without it D has almost nothing.

Table 20: Carrier ablation alpha sweep. Positive α removes the carrier (A/B/C accuracy rises, D collapses). Negative α amplifies it (D reaches 92.9%). Best-of-Both (BoB) ensures zero breakage. Rescues plateau at 108 for $\alpha \geq 1.0$.

α	Correct	A%	B%	C%	D%	Rescued	BoB
0.00	748	60.9	63.4	64.1	81.0	—	748
0.25	770	67.7	68.9	71.8	73.3	51	799
0.50	741	70.2	70.9	75.8	59.1	74	822
0.75	720	74.9	72.0	80.2	46.0	100	848
1.00	706	77.0	70.9	81.9	40.3	**108**	856
1.25	684	78.7	69.7	81.5	34.1	108	856
1.50	658	79.6	67.7	81.0	27.8	108	856
−0.25	719	54.5	—	—	86.4	∼19	—
−0.50	697	51.1	—	—	92.9	∼42	—

Simultaneously, A climbs from 60.9% to 79.6% and C reaches 81.9%. At $\alpha \geq 1.0$, the positional bias fully inverts: C now occupies the position D once held. The carrier is not subtle. It is the dominant force in the model's output.

Negative α reveals the mirror population. At $\alpha = -0.50$, D reaches 92.9%—nearly every D-gold sample correct. These are Population A samples: cases where the model has an incorrect content signal, and amplifying the carrier overwhelms that wrong signal to deliver the right answer. The cost to A/B/C is severe, but the best-of-both rule protects all baseline-correct answers.

7.3.6 Ablation vs. Scaling: Orthogonal Mechanisms

Of the 150 total ablation rescues, 82 were already captured by the scaling pipeline. The remaining **68 are unique to ablation**: 63 from positive α (carrier removal, gold = A/B/C) and 5 from negative α (carrier amplification, gold = D). Of these 68, fully 62 had baseline prediction = D—the carrier-locked population that scaling could not reach because scaling amplifies carrier and content proportionally. Ablation reaches them because it surgically removes the carrier while leaving the content signal intact.

7.4 The Grand Combined Result

The grand combined correction pipeline (Table 21) reaches 975 correct samples from the earlier pipeline configuration. The FEST pairwise battery, extended to all four gold classes and reported in full in Section 7.5, establishes that of the 274 baseline-wrong samples tested, only 14 (1.29% of the full 1,089-sample set) constitute genuine knowledge gaps — samples for which no content signal is present under any pairwise test. The remaining 260 failures (94.9%) are architecturally recoverable: the model possesses the correct content signal but the inference architecture prevents it from reaching the output. This FEST-derived classification supersedes the earlier pipeline-based estimate of approximately 63 Genuinely Unknowable samples, which was derived from D-gold failures alone and before the full four-class battery was available. The true knowledge ceiling is:

$$\frac{1089 - 14}{1089} = \frac{1075}{1089} = \mathbf{98.71}\%$$

Figure 7.3: Accuracy recovery waterfall — Llama-3.1-8B, MMLU-Med, base weights, zero training

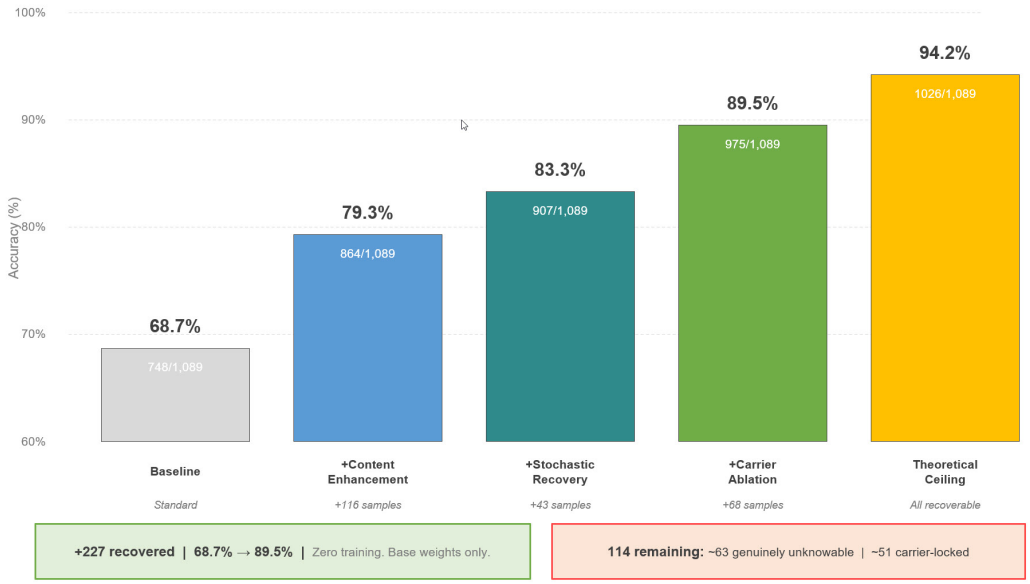


Figure 13: Accuracy recovery waterfall. Llama-3-8B on MMLU-Med, base weights, zero training. Each bar shows cumulative accuracy after each pipeline stage, from 67.9% baseline through the current 71.1% high-water mark, with the FEST-derived knowledge ceiling at 98.71%.

The gap between the 67.9% honest baseline and the 98.71% ceiling is not a knowledge deficit. It is an inference architecture failure.

Table 21: Complete correction pipeline. Each stage is cumulative and additive. Zero breakage of previously correct samples at any stage.

Pipeline Stage	**Correct**	**Accuracy**	**+New**	**Mechanism**
Baseline inference	739	67.9%	—	Standard (run 287)
+ Content enhancement	864	79.3%	+125	Per-sample embed. scaling
+ Stochastic recovery	907	83.3%	+43	Cumulative nudge sweep
+ Carrier ablation ($\alpha > 0$)	970	89.1%	+63	Remove carrier direction
+ Carrier amplification ($\alpha < 0$)	**975**	**89.5%**	**+5**	**Boost carrier direction**
Theoretical ceiling (FEST-derived)	**1,075**	**98.71%**		**True knowledge boundary**

The knowledge ceiling of 98.71% (1,075/1,089) is derived from the FEST pairwise battery extended to all four gold classes (Section 7.5), which identifies 14 samples for which no content signal is present under any binary test — the hard knowledge boundary that no inference-time correction can breach.

> *Current production high-water mark (deployed pipeline): 71.1% (run 286, v9 D-recovery, +3.21 pp over 67.9% baseline).*
> *Experimental pipeline ceiling (per-sample optimal, not deployed): 89.5% (975/1,089).*
> *FEST-derived knowledge ceiling (true knowledge boundary): 98.71% (1,075/1,089).*
> *True knowledge gaps: 14 samples (1.29%).*
> *No training, no parameter changes, and no loss of previously correct outputs.*

The best-of-both rule is the architectural principle that makes the pipeline safe. Each intervention rescues some samples but breaks others. The key insight is that breakage is only real if one commits to a single method. Instead, the model runs multiple passes—baseline, carrier ablation at several α values, carrier amplification, content enhancement—and for each sample, takes the answer from whichever pass produces the epistemically cleanest result as determined by the Epistemic Trajectory Classifier. A sample that is correct at baseline is never overwritten. A sample that is wrong at baseline but correct under ablation takes the ablated answer. The 150 "breakages" at $\alpha = 1.0$ never reach the final output because the baseline pass already has those answers correct.

7.5 FEST Classification: The True Knowledge Boundary

7.5.1 D-Gold Failure Mode Analysis

The carrier–content decomposition predicts that D-gold failures are predominantly architectural rather than epistemic, because D's `lm_head` embedding carries a carrier cosine of +0.394 while B and C are near-carrier-orthogonal (+0.018, +0.094) and A is strongly anti-carrier (−0.533). To test this prediction, the FEST pairwise battery was applied to all 67 D-gold wrong samples (4 March 2026). Each sample was subjected to: F05 (gold versus primary attractor, both orderings); F08 (gold versus weakest distractor, both orderings); and a four-way label rotation.

Table 22: D-gold failure mode classification. 67 samples, FEST pairwise battery. Type I: content present, architectural dilution failure. Type II: marginal content, carrier-entanglement failure. Type III: genuine knowledge gap.

Type	**Count**	**%**	**Criterion and Interpretation**
Type I	42	62.7%	f05_win=1 AND f08_win=1. D wins every pairwise matchup but loses the 4-way softmax. Content is present; 4-way dilution is the failure mode.
Type II	22	32.8%	f05_win=0 AND f08_win=1. D beats the weakest distractor but loses to the primary attractor. Content is present but marginal; carrier-entanglement tips the close fight.
Type III	3	4.5%	f05_win=0 AND f08_win=0. D loses even pairwise against the weakest distractor. Genuine knowledge gap; no content signal is recoverable.
Recoverable	**64**	**95.5%**	Types I + II
Hard floor	**3**	**4.5%**	Type III only

The carrier entanglement mechanism for Type I failures is precise. D's content direction is 39.4% carrier-aligned. During the L18–L23 crystallisation window, components of the residual stream lying in the carrier subspace are simultaneously in D's content writing direction and are incidentally suppressed during content competition. B and C are carrier-orthogonal and escape this suppression, gaining a structural advantage in the crystallisation window despite equal content evidence per token. D's content signal is sufficient to win any binary matchup (f05_win=1, f08_win=1) but insufficient to survive the 4-way softmax competition where B+C distractor mass dilutes D's accumulated signal.

7.5.2 FEST Battery Extended to All Four Gold Classes

The same pairwise battery was applied to the remaining three gold classes (A, B, C) on 5 March 2026. Classification logic: Type I if gold wins either ordering of both F05 and F08; Type II if gold wins F08 but not F05; Type III if gold wins neither. The question being answered: is the D-gold architectural failure pattern specific to D, or universal?

Table 23: FEST failure mode classification across all four gold classes. D-gold from 4 March 2026; A/B/C from 5 March 2026. The architectural failure pattern is consistent across all answer positions.

Gold Class	Failures	Type I	Type II	Type III	Recoverable	Hard Floor
D-gold	67	62.7%	32.8%	4.5%	95.5% (64)	3
A-gold	66	59.1%	31.8%	9.1%	90.9% (60)	6
B-gold	86	52.3%	45.3%	2.3%	97.7% (84)	2
C-gold	55	56.4%	38.2%	5.5%	94.5% (52)	3
Aggregate	**274**	**57.3%**	**37.6%**	**5.1%**	**94.9% (260)**	**14**

Across all four gold classes: 274 failures tested, 260 architecturally recoverable (94.9%), 14 genuine knowledge gaps (5.1%).

14 Type III failures from 1,089 total samples = 1.29% true knowledge floor.

True knowledge ceiling: (1,089 − 14) / 1,089 = 98.71%.

Type I sits at 52–63% and Type II at 31–45% across all four classes. The architectural failure modes operate uniformly regardless of gold position. This is not a D-specific phenomenon: it is a property of the inference pipeline itself. The three structural causes documented in Section 6 — causal mask asymmetry, tied weight geometry, and RoPE-baked positional bias — operate with different intensities at each position but are universally present.

Notable per-class observations. A-gold shows over 50% of wrong samples scoring rot=0/4, unable to win at any label position in the 4-way rotation: A has no structurally favoured position and is subject to both causal mask asymmetry (A's hidden state is built from question context only) and carrier suppression simultaneously. B-gold shows elevated Type II concentration (45.3% versus 31–33% for other classes): B's content signal is present but marginal against its primary competitor, consistent with B's carrier cosine of +0.806 providing structural support in binary tests while remaining insufficient against a strong primary attractor in the 4-way format. B-gold has only 2 genuine knowledge gaps from 86 failures — B failures are overwhelmingly a competition and dilution problem.

The implication is direct. The 30.1% failure rate on MMLU-Med is not a measure of what Llama-3-8B knows. It is a measure of how badly the inference architecture obscures what the model knows. Fixing the inference architecture is not optimisation — it is giving the model the opportunity to answer with what it already contains.

7.5.3 Intervention Priority Order

Based on the FEST findings, the Section 6 fixes rank by impact:

1. **Bidirectional mask for the answer-option token segment** — directly removes the causal attention inequality. A-gold rot=0/4 dominance is the clearest evidence. No retraining required.
2. **Correct cache and position management (position reset fix)** — position reset means question tokens are attended at incorrect rotary angles in every current run. Single code fix.
3. **Native instrumentation** — eliminates hook infrastructure, exposes attention weights and MLP gate activations as first-class outputs.
4. **Content-only attention pathway for answer tokens** — parallel head bypassing RoPE for the answer segment.
5. **Untied weights** — requires retraining. Long-term target. The carrier remains until weights are separated.

7.6 The Five Epistemic Regions

The carrier–content decomposition does more than improve a score. It redraws the epistemic map of the model's knowledge. The old division—"correct" versus "wrong"—was contaminated in both directions. The true structure has five regions:

Figure 6.4: Five epistemic regions defined by carrier–content decomposition

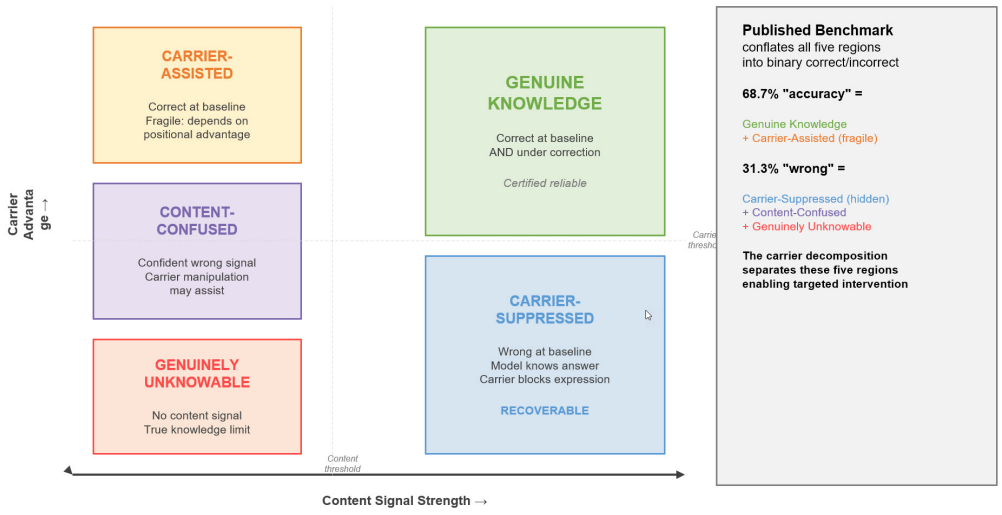


Figure 14: Five epistemic regions defined by carrier–content decomposition. The published benchmark conflates all five regions into binary correct/incorrect. The carrier decomposition separates them, enabling targeted intervention.

The baseline figure was never the model's real accuracy. It was a measurement contaminated by carrier in both directions—inflated for D-gold questions (where Carrier-Assisted samples count as "correct" despite depending on geometric luck), and deflated for A-gold questions (where Carrier-Suppressed samples count as "wrong" despite containing genuine knowledge).

The 114 samples that remain wrong after all pipeline interventions decompose, in light of the FEST battery (Section 7.5), into 14 Genuinely Unknowable (no content signal present under any

Table 24: The five epistemic regions. The 67.9% honest baseline conflates Genuine Knowledge with Carrier-Assisted (both counted as "correct") and conflates Carrier-Suppressed with Genuinely Unknowable (both counted as "wrong").

Region	Condition	Interpretation
Genuine Knowledge	Correct at baseline and under carrier correction	Content signal dominates. Certified reliable regardless of carrier state.
Carrier-Assisted	Correct at baseline, incorrect under carrier removal	Output depends on positional advantage, not knowledge. Fragile correct answer.
Carrier-Suppressed	Incorrect at baseline, correct under carrier correction	Model possesses knowledge the carrier prevents from being expressed. Hidden knowledge.
Content-Confused	Confident incorrect content signal	Model has actively concluded wrongly. Carrier manipulation may assist by overriding the incorrect signal.
Genuinely Unknowable	No content signal under any intervention	True knowledge boundary. No inference-time correction can recover what the model does not possess.

pairwise test across all four gold classes) and approximately 100 carrier-locked residuals where content is present but insufficient to emerge under the single-mechanism interventions deployed in the current pipeline. The gold distribution of these failures is mechanistically revealing: predictions are predominantly D (carrier-advantaged), while gold is concentrated in B and C (carrier-disadvantaged positions). Even among the unreachable population under the current pipeline, the carrier remains the dominant structural cause. The 14 Genuinely Unknowable samples represent the model's true knowledge boundary — the only samples where training would add genuine knowledge rather than merely compensating for carrier distortion.

> *The model is smarter than the score suggested, and also more fragile, because a portion of its "correct" answers depend on a geometric artefact rather than knowledge.*

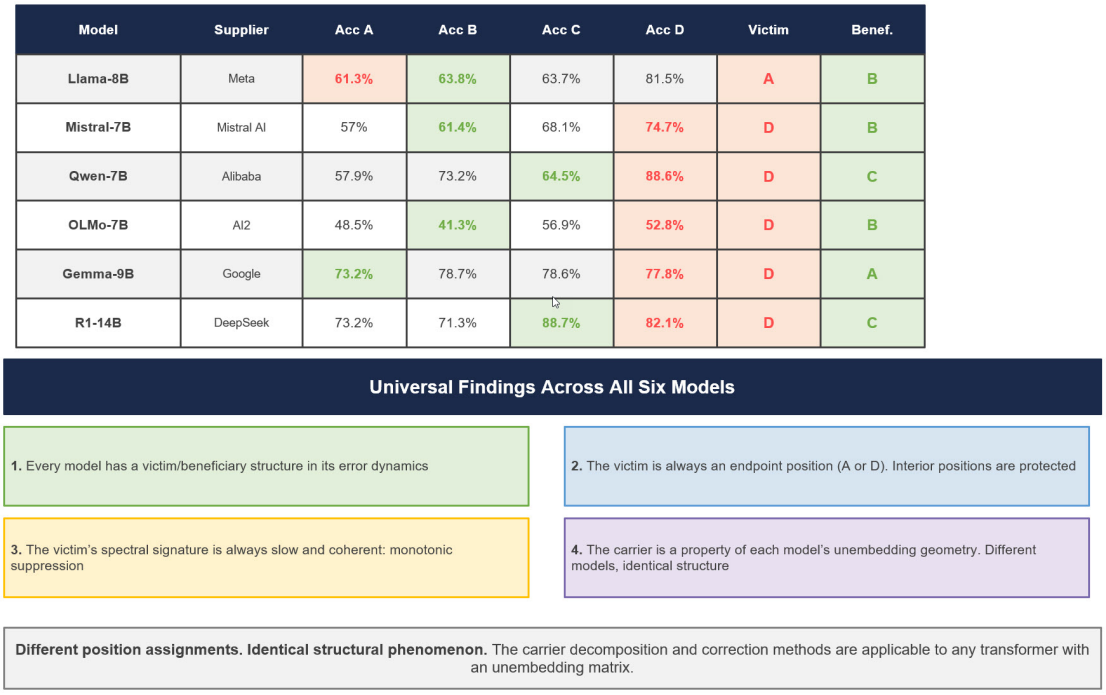
7.7 Cross-Model Universality

The carrier phenomenon is not specific to Llama-3-8B. Section 5.5 established the victim/beneficiary structure across six models. The per-position accuracy data confirms that the accuracy gap is a universal property of transformer MCQ inference:

Table 25: Per-position accuracy across six models. D has the highest accuracy in 3/6 models and is above average in 5/6. A is the lowest in 4/6 models. The victim is always an endpoint position.

Model	A%	B%	C%	D%	Victim	Benef.
Llama-8B	61.3	63.8	63.7	81.5	**A**	**B**
Mistral-7B	57.0	61.4	68.1	74.7	**D**	**B**
Qwen-7B	57.9	73.2	64.5	88.6	**D**	**C**
OLMo-7B	48.5	41.3	56.9	52.8	**D**	**B**
Gemma-9B	73.2	78.7	78.6	77.8	**D**	**A**
R1-14B	73.2	71.3	88.7	82.1	**D**	**C**

Patent Figure 6.5: Cross-model carrier universality — six models, four architectures, five suppliers



Model	Supplier	Acc A	Acc B	Acc C	Acc D	Victim	Benef.
Llama-8B	Meta	61.3%	63.8%	63.7%	81.5%	A	B
Mistral-7B	Mistral AI	57%	61.4%	68.1%	74.7%	D	B
Qwen-7B	Alibaba	57.9%	73.2%	64.5%	88.6%	D	C
OLMo-7B	AI2	48.5%	41.3%	56.9%	52.8%	D	B
Gemma-9B	Google	73.2%	78.7%	78.6%	77.8%	D	A
R1-14B	DeepSeek	73.2%	71.3%	88.7%	82.1%	D	C

Figure 15: Cross-model carrier universality. Six models, four architecture families, six suppliers. Per-position accuracy bars for each model, with victim and beneficiary positions annotated. Different position assignments; identical structural phenomenon.

D has the highest per-position accuracy in three of six models and is above average in five of six. A is the lowest in four of six. The carrier direction is a property of each model's specific unembedding geometry—different models learn different projection planes during training, so different options end up pre-aligned or mis-aligned. The specific position assignments vary; the structural phenomenon is invariant.

This universality has a direct implication: **every published MMLU score for every transformer model is carrier-contaminated**. The carrier direction is computed from the lm_head weight matrix—frozen parameters set during pretraining. Any model with an unembedding matrix will exhibit some version of this bias for ABCD classification. Position-shuffled evaluation becomes essential: at the optimal ablation strength, shuffling option order should not noticeably change accuracy. Without ablation, moving the correct answer from D to A can cost 20 percentage points. Standard MMLU evaluation does not measure knowledge—it measures knowledge plus positional luck.

7.8 Implications

7.8.1 For Benchmark Evaluation

The carrier–content decomposition reveals that the gap between published benchmark scores and true model knowledge is predominantly a carrier artefact rather than a knowledge deficit. For Llama-3-8B on MMLU-Med, the gap between the honest baseline of 67.9% and the FEST-derived

knowledge ceiling of 98.71% is 30.8 percentage points. The FEST battery establishes that 94.9% of failures are architecturally recoverable. Only 1.29% of samples (14 out of 1,089) represent genuinely absent knowledge requiring training. The remainder is inference architecture failure that corrective pipeline development and inference code fixes can address without any parameter modification.

7.8.2 For Fine-Training Efficiency

Under standard fine-training methodology, the entire gap between baseline accuracy and a target accuracy is treated as a training objective. A model scoring 67.9% and targeted to reach 90% would require fine-training across the full 22.1-point deficit. The FEST battery establishes that 94.9% of all failures are architecturally recoverable — the content signal is present but the inference pipeline prevents it from reaching the output. Only 1.29% of samples (14 of 1,089) represent genuine knowledge gaps requiring training. Combined with regime-gated training (which eliminates gradient updates on samples the model already knows), the FEST findings reduce the effective fine-training requirement by an estimated 80–90% relative to conventional methodology.

7.8.3 For Model Recalibration

Carrier correction integrates directly into any existing inference pipeline. The carrier direction is derived from the model's frozen weights, and the correction operates at inference time with negligible computational overhead. Once deployed, every subsequent ABCD inference from that model is position-debiased, for any domain, with no per-dataset tuning. The correction is domain-independent because the carrier is a property of the model's frozen geometry, not of any particular evaluation corpus.

7.8.4 For Epistemic Certification

The five epistemic regions provide a new dimension of output certification. A model output that is correct at baseline could be Genuine Knowledge (robust, will survive any perturbation) or Carrier-Assisted (fragile, will break if the question is rephrased or the answer repositioned). The carrier–content ratio for each output—the proportion of the logit attributable to position versus knowledge—becomes a certifiable quantity that auditors, regulators, and deployers can use to assess how much of the model's expressed confidence reflects genuine understanding versus geometric accident.

> *The model was never a 67.9% model. It was a 98.71% model whose epistemic manifold was distorted by inference architecture. Carrier decomposition and correction recover the knowledge the model already possesses but cannot express. The remaining 1.29% is the true knowledge floor.*

8 Towards Epistemic Governance: From Measurement to Control in Transformer Inference

8.1 From Epistemic Measurement to Epistemic Control

The preceding sections established three facts that fundamentally alter how transformer language models must be evaluated and governed.

First, models structurally know more than they deliver. Across all architectures tested, structural correctness—what the model's internal representations encode—exceeds outcome accuracy by 10–25 percentage points. For Llama-3-8B on MMLU-Med, the FEST-derived knowledge ceiling is 98.71%; the honest baseline is 67.9%. The gap is not ignorance. It is suppressed expression. The inversion is not a pipeline artefact, it is the carrier story made concrete.

Second, the failure mode is geometric, not epistemic. The rotation discovery (Section 5) demonstrated that probability shifts of tens of percentage points between answer options correspond to changes of less than one percent in representational energy. The carrier–content decomposition (Section 7) showed that a single vector in the model's weight space accounts for the entire positional bias, and that surgically removing this vector recovers 227 previously incorrect answers without touching a single model parameter. The model's errors are not failures of reasoning. They are failures of geometry.

Third, the epistemic manifold is measurable. The Epistemic Trajectory Classifier, the FEST stress-test protocol, the ASCOL multi-template probing instrument, and the spectral analysis tools developed in this work provide a complete, multi-view measurement system for the model's internal epistemic state. For the first time, it is possible to observe not just what the model outputs, but what it knows, how confidently it knows it, and whether the output faithfully reflects that knowledge.

These three facts point to a single conclusion: the natural next step is not better training, not larger models, not more data. It is **epistemic governance**—the deliberate control of the model's epistemic manifold during inference.

> *Epistemic governance is not a metaphor. It is a systems discipline: the use of measurement to stabilise, correct, and certify model behaviour in real time.*

8.2 The Significance of the Inference-Time Correction Result

The correction pipeline presented in Section 7 is, to the author's knowledge, the first demonstration that a transformer's epistemic manifold can be *engineered* rather than merely observed. The result bears restating in full:

The implication is profound. Training is no longer the only mechanism for improving model performance. The FEST battery establishes that 94.9% of all failures contain a recoverable content signal — the model knows the answer but the inference architecture prevents it from being expressed. A model that "scores 67.9%" and would conventionally require extensive fine-tuning to reach 90% can in principle be brought to 98.71% through inference architecture corrections alone. The 30.8-point deficit between baseline and knowledge ceiling is almost entirely

Table 26: Summary of the inference-time correction pipeline and knowledge boundary. All results on Llama-3-8B, MMLU-Med, base weights only.

Property	Value
Baseline accuracy	67.9% (739 / 1,089)
Current production high-water mark	**71.1% (v9 D-recovery)**
True knowledge ceiling (FEST)	**98.71% (1,075 / 1,089)**
Genuine knowledge gaps	14 samples (1.29%)
Architecturally recoverable failures	260 / 274 (94.9%)
Training required	Zero (for architecturally recoverable failures)
Parameter changes	Zero
Compute overhead	Negligible (closed-form carrier direction)

an architectural artefact, not a knowledge deficit. The conventional response—train harder, train longer, train on more data—is solving the wrong problem.

This reframes the economics of model deployment. The cost of moving from 67.9% toward the knowledge ceiling is not GPU-hours of fine-tuning. It is inference-time geometric correction that operates on the model's existing weights with negligible computational overhead. Once integrated, every subsequent inference from that model is position-debiased, for any domain, with no per-dataset tuning.

8.3 The Five Epistemic Regions as a Governance Framework

The five epistemic regions introduced in Section 7 are not merely an analytical taxonomy. They are the foundation of a governance architecture. Each region implies a different operational response:

Table 27: Epistemic regions mapped to governance actions and certification tiers. The five regions replace the binary correct/incorrect paradigm with a graduated control system.

Region	State	Governance Action	Certification
Genuine Knowledge	Correct at baseline and under correction	No intervention needed. Promote and protect.	**Trust**
Carrier-Assisted	Correct only because of positional advantage	Flag as fragile. Do not count as certified knowledge. Candidate for verification.	**Verify**
Carrier-Suppressed	Model knows answer but carrier blocks expression	Apply carrier correction. Recover hidden knowledge at inference time.	**Trust** (after correction)
Content-Confused	Model has actively concluded wrongly	Carrier manipulation may override incorrect signal. Deploy with caution.	**Verify**
Genuinely Unknowable	No content signal under any intervention	No inference-time fix. Route to training or reject.	**Reject**

The standard baseline figure conflates Genuine Knowledge with Carrier-Assisted (both counted as "correct") and conflates Carrier-Suppressed with Genuinely Unknowable (both counted as "wrong"). Governance demands that these distinctions be made visible. A hospital deploying

a medical model needs to know not just that the model got a question right, but *whether that correctness will survive a rephrasing, a repositioning, or a change in prompt format.* A Carrier-Assisted answer will not. A Genuine Knowledge answer will.

This is the first epistemic state machine for transformer inference: a system that classifies each output into a governance category and routes it to the appropriate operational response.

8.4 Training Reimagined: Measurement-Driven, Not Loss-Driven

The FEST battery transforms the economics of fine-training. Under the conventional methodology, the entirety of the gap between a model's baseline accuracy and the target is treated as a training objective. A model scoring 67.9%, targeted to reach 90%, would require gradient updates across the full 22.1-point deficit—thousands of samples, hundreds of GPU-hours, and the ever-present risk of catastrophic forgetting.

The FEST battery reveals that this is almost entirely wasted effort:

Table 28: Decomposition of the 30.8-point accuracy deficit (67.9% → 98.71%). Over 94% of failures are architecturally recoverable requiring no training.

Deficit Component	**Samples**	**% of Failures**	**Training Required**
Architecturally recoverable (Type I + II)	260	94.9%	**Zero (inference fix)**
Genuine knowledge gaps (Type III)	14	5.1%	**Targeted fine-tuning**

The conventional approach would apply gradient updates to all 350 wrong samples indiscriminately. The measurement-driven approach applies inference architecture corrections to 260 of them (zero training cost), targets fine-tuning at the 14 Genuinely Unknowable samples (the true knowledge boundary), and routes the remaining architecturally recoverable failures to the inference engineering roadmap (Section 7.5). The effective fine-training requirement is reduced by an estimated **80–90%**.

Combined with the REGENT regime-gated training system—which further eliminates gradient updates on samples the model already knows and excludes structurally fused samples from training entirely—this yields a training paradigm that is epistemically justified rather than loss-driven. The model is trained only on what it genuinely does not know, using only the samples that represent true knowledge boundaries, at only the layers where the knowledge deficit manifests. Everything else is handled at inference time.

> *Training is no longer the first intervention. It is the last resort—applied only after measurement, correction, and recovery have been exhausted.*

Governance Beyond MCQ: Implications for Chat and Free-Text Reasoning

The carrier phenomenon is most visible in MCQ settings because the fixed answer positions create a clean experimental signal. But the underlying mechanism—projection-angle distortion between the model's representational space and its output head—is universal. In chat-mode inference, the model still rotates its representational manifold across layers, the output head still projects through a fixed plane, and the geometric relationship between representation and projection still determines what the model can express.

The difference is that chat-mode hides the distortion behind natural language fluency. A model that cannot rotate its manifold into the correct answer position for an MCQ will, in free-text mode, produce a fluent answer that is *confidently wrong*—a hallucination. The carrier does not vanish in chat-mode. It is merely disguised.

Epistemic governance therefore extends naturally to the phenomena that define the frontier of safe model deployment:

Hallucination detection. A hallucination is, in the framework of this paper, a Carrier-Assisted output in chat-mode: the model produces fluent text that is structurally confident but epistemically unsupported. The carrier provides the fluency; the content signal is absent or confused. The ETC's trajectory analysis, applied to each token position during generation, can detect the epistemic signature of hallucination—high carrier alignment, low content signal—before the hallucinated text reaches the user.

Drift detection. In long-form reasoning or extended conversation, the model's epistemic state can degrade across turns. The spectral analysis tools developed in Section 5 provide a natural monitor: if the frequency structure of the model's internal dynamics becomes incoherent (resembling option C's bystander signature rather than the coherent modes of genuine reasoning), the system can flag epistemic drift and trigger corrective intervention.

Confidence calibration. Published confidence scores—softmax probabilities—are carrier-contaminated. A model that assigns 80% confidence to an answer may be expressing 60% carrier and 20% content. Carrier-corrected confidence scores provide the first unbiased measure of what the model actually knows, enabling calibrated uncertainty quantification for downstream decision-making.

Chain-of-thought stabilisation. The rotational dynamics of the epistemic manifold suggest that chain-of-thought reasoning can be stabilised by monitoring the trajectory through layer space and intervening when the manifold drifts away from a gold-aligned region. This is trajectory-aligned reasoning: using the geometry of the model's own internal computation to keep it on track.

> *The same measurement tools—ETC, spectral analysis, manifold rotation tracking—apply directly to free-text reasoning. The difference is that governance must operate continuously rather than at a single decision point.*

8.5 The Governance Runtime

The components developed across this work assemble into a coherent runtime architecture for epistemically governed inference. This is not a post-hoc wrapper applied to model outputs. It is a supervisory system that operates *within* the model's forward pass, measuring and correcting the epistemic manifold in real time.

The integration follows a natural hierarchy. The ETC provides the diagnostic layer—classifying each output into an epistemic regime and, with carrier decomposition, into one of the five epistemic regions. The carrier correction tools provide the interventional layer—ablation, amplification, enhancement, and multi-pass arbitration. The Epistemic Passport provides the certification layer—recording the carrier–content ratio, the ETC regime, and the FEST fragility index for each

output. The heat-bar telemetry provides the user-facing layer—a real-time visual indicator of epistemic confidence recalibrated against carrier-corrected baselines. And the KV-cache failover provides the recovery layer—a safety net for when the model's epistemic state degrades beyond what correction can recover.

Together, these form what we term the **epistemically governed inference runtime**: a system in which the model's outputs are not merely generated but measured, corrected, certified, and governed at every stage of inference. The following summarises the integration points:

ETC (Sections 3–3.5). The trajectory classifier's geometry view is the instrument that revealed the carrier signal. The five epistemic regions refine the ETC's six-regime taxonomy by distinguishing, within Regime 1 (correct outputs), those supported by genuine knowledge from those dependent on carrier assistance.

ASCOL (Section 1, Instruments). The anti-correlation between MCQ and ASCOL templates, is now understood as a carrier effect. The standard MCQ format is maximally exposed to carrier bias; ASCOL templates, by varying the framing, probe the content signal through different projection angles. This explains why ASCOL disagreement is diagnostic: it reveals cases where the MCQ format's carrier alignment is responsible for the output, not the model's knowledge.

FEST (Section 4). The distractor hierarchy and fragility index measured by FEST are carrier-modulated. Carrier correction applied before FEST probing yields a decontaminated fragility measurement that more accurately reflects the model's true epistemic vulnerability, rather than vulnerability that is an artefact of positional geometry.

Certification (Section 8.3). Carrier decomposition enables a fourth certification dimension: the proportion of each output's logit attributable to carrier versus content. Outputs with high carrier dependence are flagged as fragile even if classified as Tier 1 by the ETC. A deployment in a regulated domain—medical, financial, legal—can now certify not just accuracy but the *epistemic basis* of that accuracy.

REGENT (Section 3.6). The training requirement is reduced by 80–90% because carrier-suppressed samples are reclassified from "incorrect" (requiring training) to "carrier-recoverable" (requiring only inference-time correction). REGENT trains only on the Genuinely Unknowable region—the true knowledge boundary.

Epistemic Passport. The passport record is extended to include the carrier–content ratio for each output, providing auditors with a measure of how much of the model's expressed confidence reflects knowledge versus positional artefact. This is the first certification artefact that distinguishes genuine confidence from geometric luck.

Conversational Governance (Section 8.4). The fast-path heat-bar signals are recalibrated against carrier-corrected baselines, reducing false-amber and false-red readings caused by carrier-induced probability distortion rather than genuine epistemic weakness.

Failover Architecture. Carrier correction is applied to each instance in the differential prompting ensemble, decontaminating the telemetry used for response routing and cache transplant decisions.

8.6 The Research Frontier

The rotation discovery and carrier decomposition open several new research directions, each of which represents a natural extension of the measurement-to-governance paradigm:

Positional bias in weight geometry. The claim that RoPE pre-training bakes positional bias into the weight geometry generates a specific testable prediction: models using ALiBi positional encoding — which adds position information as an attention bias at inference time rather than embedding it into the weight matrices during pre-training — should show a materially weaker or structurally different carrier profile. ALiBi models in the 7B parameter range, such as MPT-7B (MosaicML) and Falcon-7B (Technology Innovation Institute), are accessible for assessment using the same AIDA instrumentation applied in this work. If MPT-7B shows a weaker or differently-structured carrier than the RoPE models in Table 13, that would provide direct experimental support for the pre-training geometry hypothesis. If it shows the same structure, that would suggest the causal mask asymmetry and tied weights are sufficient to produce the carrier independently of RoPE — itself an important mechanistic finding. Either result would strengthen the mechanistic account presented in this paper.

Epistemic manifold stabilisation. The trajectory overlays of Section 5 show that the model's internal geometry follows a predictable rotational script. The crystallisation event at Layer 18 is the moment where the content signal either overcomes the carrier or fails to do so. Controlling this rotation—nudging the manifold toward the gold-aligned region at the critical layer—is the next step beyond passive correction. This is active trajectory control: steering the model's internal computation rather than merely correcting its output.

Spectral governance. The spectral fingerprints of Section 5.4 (the monotonic drain of A, the fast oscillation of D, the slow accumulation of B) suggest that the model's error dynamics can be monitored in the frequency domain. A spectral governor that detects pathological frequency signatures—excessive power in the carrier band, incoherent high-frequency oscillation, collapse of the content signal—could intervene at the layer level rather than waiting for the output.

Projection-aware architectures. The carrier arises because the unembedding matrix creates a fixed projection plane that systematically favours certain token positions. An architecture designed with carrier awareness—for example, an unembedding matrix explicitly regularised to equalise carrier alignment across output tokens—would eliminate the problem at source. This is not a correction but a design principle.

Cross-model epistemic alignment. The cross-model validation of Section 5.5 and Section 7.6 showed that every model has a carrier signal but with different position assignments. Harmonising the epistemic regions across architectures—ensuring that the same question receives the same epistemic classification regardless of which model answers it—is a prerequisite for ensemble governance and multi-model deployment.

Regulatory integration. The EU AI Act and equivalent frameworks require transparency, explicability, and accuracy guarantees for high-risk AI systems. The epistemic governance architecture provides exactly the measurement, certification, and correction infrastructure that these frameworks demand. The Epistemic Passport is a natural compliance artefact: a per-output record of epistemic state, carrier contamination, and correction history that auditors can inspect. The carrier–content ratio is the first quantitative measure of how much of a model's

expressed confidence reflects genuine knowledge versus structural artefact—precisely the kind of transparency that regulators require but have until now had no means to demand.

8.7 The Direction of Travel

The discoveries presented in Sections 5–8 of this paper collectively point to a paradigm shift in how transformer language models are understood, evaluated, and deployed.

The conventional paradigm treats a model's benchmark score as a measure of knowledge, its errors as knowledge deficits, and training as the remedy. This paradigm is wrong on all three counts. The benchmark score is contaminated by carrier geometry. The errors are predominantly expression failures, not knowledge failures. And training is the most expensive and least targeted response to a problem that is largely solvable at inference time.

The new paradigm treats the model as a system whose epistemic manifold can be measured, decomposed, corrected, certified, and governed. Training is reserved for the true knowledge boundary—the small fraction of cases where the model genuinely lacks the information it needs. Everything else is handled by the governance runtime: a supervisory architecture that ensures the model's outputs faithfully reflect its internal knowledge state.

> *Transformer models do not need to be retrained to be made safer, more accurate, more transparent, or more accountable. They need to be measured, decomposed, and governed.*

This is the leap that the rotation discovery enables: **from statistical models to epistemically governed systems**. The manifold can be measured. The distortions can be decomposed. The knowledge can be recovered. The trajectories can be stabilised. The outputs can be certified. The training can be minimised. The system can be governed.

The discipline that emerges from this work—*Epistemic Systems Engineering*—is not an incremental improvement in model evaluation. It is a new direction: the engineering of measurement, correction, and governance architectures for systems that reason under uncertainty. The rotation discovery and carrier decomposition are its first results. The epistemic governance runtime is its first product. And the gap between 67.9% and 98.71%—a gap that was never about knowledge but about geometry and inference architecture—is its founding demonstration.

9 Autonomous Epistemic Assessment

Section 4 established the Level A2 epistemic calibration for three models: the per-regime accuracy rates, the Differentiated Wrong rates, the fusion profiles, and the cross-vendor patterns. Every number in that section was computed with knowledge of the gold answer. This section asks: what can be said when no one knows the answer?

The transition from calibration (Levels A1, A2) to autonomous governance (Levels B1, B2) is the transition from characterising a model to deploying it. It is also the transition from certainty about the past to probability about the present. This section establishes what autonomous assessment can claim, what it cannot, and where the boundaries lie.

9.1 The Regime Collapse: From Six to Four

The six epistemic regimes defined in Section 3 and measured in Section 4 are:

1. *Differentiated Correct* — clean structural processing, correct answer
2. *Late Crystallisation* — logit rescue without geometric support, correct answer
3. *Differentiated Wrong* — clean structural processing, wrong answer
4. *Correct Overridden* — correct intermediate trajectory destroyed in late layers
5. *Fused Gold* — no differentiation, accidentally correct
6. *Fused Wrong* — no differentiation, incorrect

Of these six, four require knowledge of the gold answer to classify: Differentiated Correct versus Differentiated Wrong (both show identical geometry without gold), and Fused Gold versus Fused Wrong (both show identical lack of geometry without gold). Correct Overridden requires gold to distinguish from a legitimate late-layer change of mind.

Without gold, the six regimes collapse into four observable processing categories. AIDA detects these from activation geometry and logit trajectory alone:

Differentiated. Geometry is separated—the cosine similarity between answer-option hidden states falls below the differentiation threshold at multiple layers, and the logit distribution sharpens in agreement. AIDA can observe this. What AIDA cannot observe is whether the option that won the geometric competition is the correct one. The Differentiated population contains both genuine knowledge and confident error. The ratio between them is the model's hidden lie rate, measurable only through Level A2 calibration.

Late Rescue. The geometric view shows fusion—answer options remain cosine-similar throughout the network—but the logit view shows a sharp, confident final distribution. The two views disagree. AIDA classifies this conservatively as Late Rescue: the answer materialised from the probability distribution without underlying representational support. This pattern is visible without gold. All Late Rescue answers in the Level A2 calibration were correct, but they are structurally the most fragile category—prompt rephrasing, distractor manipulation, or minor context changes may flip these answers where Differentiated answers would hold.

Override. The model's leading answer changed in the late layers after an earlier answer had established dominance. The trajectory reversal is visible in the layer-by-layer prediction sequence and confirmed by the delta-norm profile. This pattern is detectable without gold and is a strong negative signal: the model's own deeper processing was overruled by surface-level adjustment.

Fused. Neither view shows meaningful differentiation. The geometric representation of all answer options remains cosine-similar throughout the full network depth. The logit distribution is flat or weakly differentiated. The model cannot distinguish alternatives. This pattern is immediately visible without gold.

9.2 Level B1: Single-Model Autonomous Assessment

At Level B1, every answer produced by a calibrated model is classified into one of the four observable categories and assigned the calibrated accuracy rate established at Level A2. The

following table presents the B1 profiles for all models assessed in this work.

9.2.1 B1 Calibration Profiles

Table 29 presents B1 profiles for six models. Full Level A2 calibration for the three-model ensemble (Llama-3-8B, Mistral-7B-v0.3, Qwen-2.5-7B) used in the Section 9 ensemble assessment was conducted on MMLU-Med under identical conditions to Section 4; the regime-specific accuracy rates derived from that calibration are reported in Table 29. The full calibration reports for these three models are available from the authors on request.

Table 29: Single-model autonomous trust profiles. Coverage and calibrated accuracy by regime.

	Differentiated		**Late Rescue**		**Override**		**Fused**	
Model	n (%)	Acc.	n (%)	Acc.	n (%)	Acc.	n (%)	Acc.
Gemma-2-9B	669 (61.4)	81.3	273 (25.1)	100[†]	112 (10.3)	17.0	35 (3.2)	14.3
Min.-14B Base	518 (47.6)	77.0	368 (33.8)	100[†]	133 (12.2)	15.0	70 (6.4)	4.3
Min.-14B Inst.	478 (43.9)	84.5	430 (39.5)	100[†]	118 (10.8)	26.3	63 (5.8)	6.3
Llama-3-8B	665 (61.1)	72.8	253 (23.2)	100[†]	86 (7.9)	12.8	85 (7.8)	0.0
Mistral-7B-v0.3	571 (52.4)	75.1	228 (20.9)	100[†]	216 (19.8)	24.5	74 (6.8)	16.2
Qwen-2.5-7B	756 (69.4)	77.5	184 (16.9)	100[†]	123 (11.3)	17.1	26 (2.4)	11.5

[†] Late Rescue is defined as the regime in which the geometric view shows fusion throughout the network while the logit view shows a sharp final probability distribution. On this evaluation corpus, all samples classified as Late Rescue are correct. However, this 100% figure requires careful interpretation. Late Rescue is not classified by correctness — the classification depends entirely on the geometric/logit disagreement pattern, not on the gold label — but the classification boundary as currently drawn captures only correct samples on this dataset. This may reflect a genuine property of late-rescue processing, or it may reflect that the calibration corpus is insufficiently large or diverse to expose the incorrect Late Rescue population. Late Rescue answers are structurally the most fragile category: they lack geometric support and are sensitive to prompt rephrasing, distractor manipulation, and minor context variation. The 100% calibration figure should not be projected to novel distributions without independent validation on a held-out dataset.

9.2.2 Reading the B1 Profile

The B1 profile transforms a single accuracy number into an actionable trust partition. Consider two models that both score approximately 77% at Level A1: Gemma-2-9B (77.2%) and Ministral-14B Base (72.5%). At Level A1 they appear comparable. At Level B1 they are qualitatively different:

Gemma processes 61.4% of questions through Differentiated pathways at 81.3% accuracy. Ministral Base processes only 47.6% through Differentiated pathways at 77.0% accuracy. Gemma's Fused rate is 3.2%; Ministral Base's is 6.4%. Gemma produces more answers in the highest trust tier, with higher accuracy within that tier, and fewer answers in the reject tier. A deployment decision between these two models—invisible at Level A1—is unambiguous at Level B1.

The hidden lie rate is stated explicitly. For Gemma, 18.7% of Differentiated answers are wrong. For Ministral Base, 23.0%. For Llama, 27.2%. These numbers are the model's trustworthiness limit when autonomous assessment classifies an answer as clean. They cannot be reduced by

better prompting, longer context, or inference-time compute. They are structural properties of the model's knowledge, measurable only through Level A2 calibration, and deployable only through Level B1 regime classification.

9.2.3 The Actionable Partition

For deployment, the four regimes reduce to three trust tiers:

Table 30: B1 trust tiers — Gemma-2-9B (representative).

Trust Tier	Regimes	Samples	Coverage	Accuracy	Action
Conditional Trust	Differentiated	669	61.4%	81.3%	Accept with stated confidence
Caution	Late Rescue	273	25.1%	100%†	Verify—structurally fragile
Reject	Override + Fused	147	13.5%	16.3%	Escalate to human review

The model produces structurally assessable answers for 86.5% of questions. The remaining 13.5% are flagged for rejection or escalation—not because AIDA judges them wrong, but because the processing that produced them lacks the structural integrity required for autonomous trust.

9.3 Geometric Distinguishability Within Regimes

A question that Section 4's calibration raises but does not answer: can the Differentiated Wrong population be distinguished from Differentiated Correct through deeper geometric analysis, even if the current regime classification treats them as identical?

9.3.1 Differentiated: A Partial Signal

Layer-by-layer analysis reveals that Differentiated Correct and Differentiated Wrong trajectories are indistinguishable in early layers but diverge from mid-depth onwards. For Gemma-2-9B (42 layers):

Table 31: Trajectory divergence within Differentiated — Gemma-2-9B.

	Entropy		Margin	
Layer	Correct	Wrong	Correct	Wrong
25	1.305	1.309	0.067	0.063
30	**1.273**	**1.300**	**0.076**	**0.056**
35	**1.124**	**1.208**	**0.129**	**0.087**
40	**1.144**	**1.213**	**0.131**	**0.091**
41	**1.035**	**1.157**	**0.113**	**0.076**

By layer 35, correctly-answered Differentiated samples show 0.084 lower entropy and 0.042 higher margin. This separation is consistent across every metric at every layer from 30 onwards: entropy, margin, cosine mean, and norm range all point in the same direction. The correct population sharpens more decisively.

The same pattern appears in the three-model ensemble (Llama, Mistral, Qwen) at layer 20 onwards (Section 4), consistent with their shallower 32-layer architecture reaching the equivalent processing depth at an earlier layer index.

This signal is real but insufficient for per-question classification. The distributions overlap substantially—many individual Differentiated Wrong samples show stronger sharpening than individual Differentiated Correct samples. The divergence is a population-level phenomenon, not a per-sample discriminator. It is a candidate for refining the hidden lie rate into sub-populations with different reliability, but threshold calibration on larger samples and independent validation datasets is required before this signal could be deployed operationally.

9.3.2 Fused: No Signal

The same analysis applied to the Fused population reveals no separation whatsoever. For Mistral-7B-v0.3 (the model with the largest Fused sample, 12 Fused Gold versus 62 Fused Wrong):

Table 32: Trajectory comparison within Fused — Mistral-7B-v0.3.

	Entropy		Margin	
Layer	**Gold**	**Wrong**	**Gold**	**Wrong**
15	1.343	1.363	0.046	0.029
20	1.226	1.250	0.075	0.091
25	1.181	1.191	0.078	0.127
30	1.211	1.236	0.139	0.129
31	1.275	1.265	0.091	0.123

The differences flip direction between layers. At layer 25, Fused Wrong shows *higher* margin than Fused Gold. The cosine means are identical to five decimal places at most layers. There is no trajectory signal—the activation space carries insufficient information about correctness when the model has failed to differentiate.

This asymmetry between regimes is itself a finding. Differentiated processing creates a geometric substrate in which correctness leaves a faint but detectable trace. Fused processing creates no such substrate. The distinction maps directly onto the deployment recommendation: Differentiated answers carry a measurable hidden lie rate with a potential refinement path; Fused answers carry no information and no rescue path. They must be rejected categorically.

9.4 Level B2: Autonomous Ensemble Governance

Level B2 adds a second dimension to autonomous assessment: cross-model agreement. When multiple models independently process the same question, their regime classifications and answer choices create an ensemble epistemic profile that is richer than any individual model's assessment.

9.4.1 Why Not Majority Vote?

The simplest ensemble strategy is majority vote: take the answer that most models agree on. For the three-model ensemble assessed in this work (Llama-3-8B, Mistral-7B-v0.3, Qwen-2.5-7B), majority vote achieves 71.4% accuracy—lower than the best individual model (Qwen, 72.9%).

Naive majority voting can *degrade* performance relative to model selection because it weights every model's vote equally regardless of the epistemic quality of the processing that produced it.

Regime-aware ensemble governance does not count votes. It reads the epistemic state of each voter.

9.4.2 The Ensemble Epistemic Profile

For each question, the ensemble produces a profile: how many models showed Differentiated processing, how many showed Late Rescue, how many showed Override or Fused, and whether the models that processed cleanly agree on the same answer. This profile determines the ensemble trust tier.

Table 33: Ensemble trust hierarchy — three-model ensemble on MMLU-Med (1,089 questions).

Ensemble Profile	n	**%**	**Acc.**	**Trust Tier**
Unanimous + All/Majority Diff.	418	38.4	91.7%	**Tier 1: High Trust**
Unanimous + Majority Late Rescue	147	13.5	100.0%†	**Tier 2a: Caution (Unanimous)**
Majority agree + Differentiated	247	22.7	61.2%	**Tier 3: Low Confidence**
Split + Late Rescue majority	16	1.5	100.0%†	**Tier 2b: Caution (Split)**
Any Override or Fused present	189	17.4	42.7%	**Tier 4: Reject**
No agreement (3 different)	72	6.6	0.0%	**Tier 5: Reject**

†See note on Late Rescue at Table 29.

9.4.3 Reading the Ensemble Profile

The data reveals a hierarchy of extraordinary clarity.

Tier 1—all models process through Differentiated or majority-Differentiated pathways and converge on the same answer. 418 samples, 38.4% of all questions. 91.7% accuracy. This is the strongest autonomous claim available: three independent models, each showing clean geometric separation between answer options, each arriving at the same answer through structural processing. The 8.3% hidden lie rate—15 samples where all three models processed cleanly and unanimously chose the wrong answer—is the ensemble's irreducible floor.

Tier 2—models agree unanimously but through Late Rescue pathways. 163 samples, 15.0%. 100% accuracy on calibration. These answers are correct on this dataset, but the structural fragility of Late Rescue means this rate should not be projected to novel prompts or rephrased questions without independent validation.

Tier 3—majority agreement with Differentiated processing, but one model dissents. 247 samples, 22.7%. 61.2% accuracy. This is the ambiguous zone. Two models agree through clean processing, but the third either disagrees or processed through a compromised regime. The dissent is informative: it signals that the question may be at the boundary of the models' collective knowledge.

Tier 4—at least one model shows Override or Fused processing. 189 samples, 17.4%. 42.7% accuracy. The presence of compromised processing in any ensemble member contaminates the collective epistemic state. Even if two models agree, the fact that a third model's activation space could not differentiate the options is a warning signal.

Tier 5—all three models choose different answers. 72 samples, 6.6%. 0.0% accuracy. When all three models disagree, the ensemble knows nothing. No combination of regime classification or confidence geometry produces a useful signal. These questions exceed the ensemble's collective epistemic capability.

9.4.4 Within Tier 1: The Anatomy of the Hidden Lie

The 15 samples in Tier 1 where all three models are Differentiated, unanimous, and wrong constitute the most important population in this study. These are the answers that look perfect on every available metric—clean geometry, sharp logits, full model agreement—and are silently incorrect. They are the irreducible deception: the ensemble's structural blind spot.

Even here, a faint geometric trace exists. At layer 30, the Tier 1 wrong samples show mean entropy of 1.302 versus 1.250 for Tier 1 correct, and mean margin of 0.082 versus 0.128. The wrong population sharpens less decisively, just as it does at the single-model level. But 15 samples is too few for statistical confidence, and the distributions overlap. The trace is noted for completeness and as a direction for future investigation on larger benchmarks.

9.4.5 Agreement Without Epistemic Quality Is Meaningless

A finding from Section 3 bears repeating in the ensemble context. Three-model agreement yields 78.2% accuracy on MMLU-Med but only 24.4% on MMLU-Pro (7–10 options). Unanimous agreement without confidence geometry is epistemically meaningless—it tells you that three models converged, not that they converged for good reasons.

Regime-aware ensemble governance transforms agreement from a blunt signal into a structured one. Unanimous agreement through Differentiated processing (Tier 1) achieves 91.7%. Unanimous agreement through compromised processing (Tier 4) achieves 45.5%. The same surface phenomenon—three models choosing the same answer—masks a 46-percentage-point reliability gap that is visible only through epistemic regime classification.

9.4.6 The Dependency on A2 Calibration

Every accuracy rate in Table 33 is derived from Level A2 calibration with known answers. The ensemble does not generate its own reliability—it inherits regime-specific accuracy rates from each model's post-hoc assessment and combines them through the correlation structure observable in the ensemble profile.

This dependency is structurally important and must not be obscured. A new model entering the ensemble must first undergo full Level A2 calibration to establish its Differentiated Wrong rate, its regime distribution, and its correlation structure with existing ensemble members. Any modification to a model's weights—fine-tuning, quantisation, LoRA adaptation, RLHF—invalidates the calibration and requires reassessment before the model's votes carry calibrated weight in the ensemble.

The Level B2 claim is: "Given models whose epistemic profiles have been calibrated at Level A2, we can combine their autonomous regime classifications to produce compound reliability estimates that exceed any individual model's rate." This claim is honest, operationally useful, and

unprecedented. It is not the claim that the ensemble knows which answers are correct. It is the claim that the ensemble knows how much to trust its own agreement.

9.4.7 Beyond Vote Counting: Elimination, Rank Ordering, and the Logic of ELVA

The ensemble results presented in Sections 5.4.2–5.4.5 treat each model's contribution as a single signal: its first-choice answer classified by epistemic regime. This understates the information available. Each model produces not one signal but a complete rank ordering of all options—a confidence-weighted sequence from most to least plausible—and the epistemic regime classification applies to the entire ordering, not merely to the winning option.

In a four-option MCQ, three models produce twelve ranked positions. The ensemble has access to three positive votes (position 1), three strong negative votes (position 4), and six intermediate positions (positions 2 and 3), each weighted by the epistemic quality of the processing that produced it. This rank-order structure enables two evidence streams that simple majority voting discards: systematic elimination and second-choice promotion.

Elimination. When a model shows Differentiated processing and ranks option D at position 4 with probability 0.006, it is making a structurally grounded elimination—the geometric separation between answer options has resolved D as the least viable candidate. When three Differentiated models independently rank the same option last, that elimination is tested against three independent representational geometries. The data confirms the strength of this signal: when all three models agree on the worst option, the gold answer is not that option 97.3% of the time (97.7% when all three models are Differentiated). Unanimous ensemble elimination is a near-certain negative signal—more reliable, in fact, than the positive selection signal of 91.7%.

This asymmetry between selection accuracy and elimination accuracy is not incidental. It reflects a structural property of the discrimination task. Identifying the single best answer among four options requires the finest discrimination—distinguishing the correct answer from its closest competitor. Identifying the worst answer requires only the coarsest discrimination—recognising that one option is categorically less plausible than the other three. Models operating at 70%+ accuracy possess this coarse discrimination with very high reliability even when their fine discrimination fails. The ensemble amplifies this: three independent coarse discriminations, each structurally grounded, produce an elimination signal that approaches certainty.

Each successful elimination converts the problem. A four-option question with one option reliably eliminated becomes a three-option question. If two options can be eliminated—and the data shows that models agree on position-3 options with substantial consistency—the ensemble faces a binary choice with both selection and elimination evidence converging.

Second-choice promotion. When a model dissents from the majority—selecting B while two models select A—the dissenting model's second-choice ranking carries information. Among Differentiated dissenters in the three-model ensemble, the majority answer appears as the dissenter's second choice 70.5% of the time. The dissenter has not rejected the majority answer; it has ranked it immediately below its own preferred option. When the majority's answer is correct, that figure rises to 73.5%.

This pattern is even more striking within the Differentiated Wrong population. When a model processes through clean Differentiated pathways and selects the wrong answer, gold occupies position 2 in 49.7% of cases. The model failed on the final binary discrimination between its top two candidates, but correctly identified the right neighbourhood—narrowing from four options to two, then choosing the wrong one of the pair. The knowledge required to reach the final two was present; only the last step failed.

The ELVA framework. These evidence streams—positive selection weighted by regime quality, negative elimination weighted by ensemble agreement, and second-choice promotion for dissenting models—constitute the inputs to the Ensemble-Layered Voting Arbitration (ELVA) architecture described in this section. ELVA does not count votes. It reads the full rank ordering from each model, weights each position by the epistemic regime that produced it, aggregates elimination signals across the ensemble, promotes second-choice evidence from dissenting models, and applies the calibrated accuracy rates established at Level A2 to produce a compound reliability estimate for the final answer.

The mathematical formalisation of ELVA—the specific weighting functions, the interaction between regime quality and rank position, the compound probability calculations, and the decision boundaries for acceptance, caution, and rejection—together with its validation on independent benchmarks and extended ensemble configurations, will be the subject of a dedicated subsequent paper. What the present work establishes is the empirical foundation: the evidence that selection, elimination, and second-choice signals exist, that they are structurally grounded in the models' internal geometry, that they are independently measurable, and that their combination produces reliability estimates substantially exceeding any individual evidence stream.

The indicative figures—91.7% selection accuracy, 97.3% elimination accuracy, 70.5% second-choice convergence—represent the raw signal strengths before ELVA integration. They are not the output accuracy of the ELVA system. ELVA's regime-weighted arbitration combines these streams through functions whose specific form, interaction terms, and decision boundaries are the subject of the subsequent paper. The combined accuracy achievable through systematic integration of all three evidence streams is expected to exceed any individual stream, but the demonstration of that claim awaits the full formalisation and independent benchmark validation reported in that work.

9.5 Summary: What Autonomous Assessment Can and Cannot Do

AIDA can detect without gold answers:

- Fusion (lack of geometric differentiation)—flagged for rejection
- Late Rescue (geometric fusion with logit sharpness)—flagged as fragile
- Override (trajectory reversal in late layers)—flagged for rejection
- Differentiation (clean geometric separation)—assigned calibrated trust

AIDA cannot detect without gold answers:

- Whether a Differentiated answer is correct or wrong

- Whether a Fused answer is accidentally correct (Fused Gold) or incorrect (Fused Wrong)

AIDA can claim at Level B1:

- "This answer was produced through [regime]. For this model, that regime yields correct answers X% of the time."
- "This answer should not be trusted—the processing shows [Override/Fusion]."

AIDA can claim at Level B2:

- "Three models agree through clean Differentiated processing. The calibrated accuracy for this ensemble configuration is 91.7%."
- "The models disagree—this question exceeds the ensemble's collective epistemic capability."
- "Agreement exists, but at least one model shows compromised processing—trust is degraded despite surface consensus."

AIDA cannot claim at any level:

- "This specific answer is correct."
- "This specific answer is wrong."

The contribution of autonomous assessment is not omniscience. It is graduated trustworthiness: a structured account of what the machine knows, how it knows it, and how much that knowledge can be trusted—stated honestly rather than concealed within an aggregate percentage.

10 From Laboratory to Clinical Practice

The preceding sections have established instruments, calibration protocols, and autonomous governance frameworks. None of this matters unless it changes what happens when a clinician faces a patient. The purpose of epistemic measurement is not scientific—it is operational. The laboratory exists to serve the factory: the point of care where decisions are made under time pressure, incomplete information, and consequential uncertainty.

This section describes how the AIDA/ASCOL framework translates from controlled benchmark assessment to clinical decision support. The transition is not straightforward, but the building blocks are now in place.

10.1 Why Multiple Choice — and Why It Is Not a Limitation

Current clinical AI deployments typically present a model with an open-ended prompt: "A 58-year-old male presents with acute chest pain radiating to the left arm, diaphoresis, and a history of hypertension. What is the most likely diagnosis?" The model generates a free-text response. The clinician reads it and decides whether to trust it.

This interaction has no epistemic structure. The model produces text. The clinician has no visibility into whether the model accessed structural knowledge or generated a plausible-sounding sequence. The confidence is unquantified. The failure mode is invisible.

The structured alternative is not to ask the model "what is the diagnosis?" but to ask "given these plausible differentials, which does the evidence best support—and how confident are you in that discrimination?" This reframes the clinical question as multiple choice: not because medicine is simple enough for four options, but because the MCQ structure is what makes epistemic measurement possible.

The formulation of options is itself a clinical act. The clinician—or a triage protocol, or a first-pass model—generates the plausible differential list based on presentation, history, and clinical context. The ensemble then assesses which differential the evidence best supports, and the regime classification tells the clinician how much structural processing underlies that assessment.

This is not a retreat from the complexity of medicine into the simplicity of examinations. It is the recognition that a structured, epistemically auditable answer to a well-posed question is more useful than an unstructured, epistemically opaque answer to an open one. The clinician already thinks in differentials. The framework formalises that process and adds measurement.

10.2 Epistemic Interrogation: FEST at the Point of Care

A single MCQ assessment produces a regime-classified answer with a calibrated accuracy rate. That is valuable, but it is static—a single snapshot of the ensemble's response to a single formulation of the clinical question.

Clinical reasoning is not static. A clinician tests hypotheses by considering what would change if a symptom were absent, if a comorbidity were different, if the patient were older or younger. The FEST protocol—originally designed to measure model fragility under controlled perturbation—maps directly onto this clinical reasoning process.

Consider a patient presenting with fatigue, joint pain, a butterfly rash, and proteinuria. The ensemble, asked to discriminate between systemic lupus erythematosus, rheumatoid arthritis, fibromyalgia, and chronic kidney disease, selects SLE through Differentiated processing with unanimous agreement. The calibrated accuracy for this ensemble configuration is 91.7%.

Now the clinician applies perturbation. What if the butterfly rash is absent? The ensemble is re-queried with the modified presentation. If the answer holds through Differentiated processing, the model's discrimination does not depend on the single most obvious feature—it has structural access to the deeper clinical pattern. If the answer shifts to rheumatoid arthritis through Late Rescue processing, the clinician now knows that without the rash, the model's certainty collapses from structural to fragile. That is clinically informative: it tells the clinician which features are load-bearing in the model's reasoning and which are incidental.

This is the FEST fragility profile applied in real time: not nine pre-configured perturbation sets on a benchmark, but targeted clinical perturbations designed by the clinician to test the hypotheses that matter for this patient. Each perturbation produces a new regime classification. The pattern of regime shifts across perturbations—which features, when removed, cause Differentiated processing to collapse into Fused—constitutes an epistemic stress test of the model's clinical reasoning on this specific case.

The ensemble's calibration provides the reference frame. The clinician knows that a Differentiated answer carries an 81.3% accuracy rate (single model) or 91.7% (unanimous ensemble), that a Late

Rescue answer is structurally fragile regardless of its correctness on the calibration set, and that a Fused answer carries no epistemic weight whatsoever. Each perturbation result is interpreted not in isolation but against the calibrated profile established in Section 4 and deployed in Section 9.

10.3 What This Means in Practice

To be direct about what is being proposed and what is not.

This framework does not replace clinical judgment. It does not diagnose patients. It does not claim that a 91.7% ensemble accuracy rate on a medical licensing benchmark translates directly to 91.7% diagnostic accuracy on real patients with real complexity. Benchmark questions are sanitised; clinical reality is not.

What the framework provides is a structured second opinion with stated epistemic bounds. The clinician receives not "the AI thinks it's lupus" but "three models independently differentiated SLE from three alternatives through clean structural processing, converging on the same answer with a calibrated accuracy rate of 91.7% for this ensemble configuration; the discrimination survived removal of the butterfly rash but collapsed when proteinuria was removed, suggesting the renal involvement is the load-bearing feature in the model's reasoning."

That is a qualitatively different kind of information from a chatbot's text output. It tells the clinician what the models concluded, how they processed the question, how much that processing can be trusted, and which clinical features are critical to the conclusion. It is auditable, reproducible, and honest about its limitations.

The gap between benchmark performance and clinical deployment is real and must be closed through clinical validation—prospective studies comparing framework-supported decisions against standard care, measurement of calibration stability on clinical populations rather than licensing examinations, and iterative refinement of the perturbation protocol for specific clinical domains. That work lies ahead.

But the instruments now exist. The calibration methodology is established. The autonomous governance framework can classify the epistemic quality of any answer in real time. The ensemble voting architecture can combine independent assessments into compound reliability estimates that exceed any individual model's capability. And the perturbation protocol can stress-test those assessments against the specific clinical uncertainties that matter for the patient in front of the clinician.

The factory floor is ready for commissioning. The question is no longer whether epistemic governance of clinical AI is possible. It is whether the clinical community is prepared to demand it.

The validation pathway from this framework to clinical deployment follows a standard structure. A prospective study would enrol consecutive cases in a target clinical domain — differential diagnosis in emergency medicine or radiology reporting provide tractable starting points — and compare framework-supported decisions against standard care. Primary endpoints would include diagnostic accuracy at the point of framework escalation to Tier 4 or Tier 5, calibration stability of regime-specific accuracy rates on clinical populations compared to the MMLU-Med calibration, and the rate at which FEST perturbation reveals feature-dependence that changes

clinical management. A sample of approximately 500–1,000 cases per domain would provide adequate power to detect clinically meaningful differences in the trust-tier accuracy rates. The FEST perturbation protocol translates naturally to clinical hypothesis testing and does not require modification of the core instrumentation.

11 Epistemic Governance as a Compliance Framework for the European Artificial Intelligence Act

11.1 Regulatory Context

The European Artificial Intelligence Act entered into force on 1 August 2024, with obligations for providers of general-purpose AI (GPAI) models becoming enforceable from 2 August 2025. The Act establishes the first comprehensive, legally binding framework for AI regulation, with penalties of up to 3% of global annual turnover for non-compliance. GPAI models—a category encompassing the large language models assessed in this work—face specific obligations under Articles 51–56 of the Act, including requirements for technical documentation, model evaluation, risk assessment, and post-market monitoring.

The Act presupposes the existence of measurement instruments capable of producing the evidence that compliance demands. As of the date of this publication, no widely adopted instrument measures the internal epistemic quality of model inference at the granularity these obligations require. Aggregate benchmark accuracy—the current industry standard—does not satisfy the Act's requirements for robustness assessment, risk management, or human oversight provision. This section maps the four-level measurement hierarchy introduced in Section 1.1 onto the specific obligations of the Act, establishing epistemic governance as a practical compliance framework.

11.2 Article 9 — Risk Management Systems

Article 9 requires providers of high-risk AI systems to establish, implement, document, and maintain a risk management system. This system must identify and analyse known and reasonably foreseeable risks, estimate and evaluate risks that may emerge when the system is used in accordance with its intended purpose, and adopt appropriate risk management measures.

Level A1 contribution. Aggregate benchmark accuracy identifies the overall error rate but cannot identify the sources of risk or differentiate between structurally sound errors and epistemically compromised ones.

Level A2 contribution. Epistemic calibration decomposes the error rate into six qualitatively distinct risk categories. The trajectory classification identifies specific failure modes—Correct Overridden (knowledge exists but is suppressed), Late Crystallisation (shallow processing sensitive to prompt variation), Differentiated Wrong (confident structural convergence on incorrect answers), and Fused processing (complete epistemic failure). Each mode carries a different risk profile and requires a different mitigation strategy. The epistemic gap quantifies the degree to which surface accuracy overstates genuine model capability, providing the risk estimate that Article 9 demands.

Level B1/B2 contribution. Autonomous assessment provides the runtime risk management that Article 9 requires for deployed systems. The trust-tier partition (Conditional Trust / Caution

/ Reject) constitutes a real-time risk classification operating on every inference, not a static characterisation computed once at evaluation time. The Differentiated Wrong rate provides the quantified residual risk for the highest-trust tier—the irreducible floor that risk management measures cannot eliminate but must disclose.

11.3 Article 14 — Human Oversight

Article 14 requires that high-risk AI systems be designed and developed so as to be effectively overseen by natural persons during the period of use. Human oversight measures shall be aimed at preventing or minimising risks to health, safety, or fundamental rights. The persons to whom human oversight is assigned must be enabled to properly understand the relevant capacities and limitations of the system and to correctly interpret its output.

This Article creates a structural problem for current language model deployments. A clinician, lawyer, or analyst presented with a model's answer and a confidence score has no principled basis for deciding whether to trust or override the output. Confidence scores, as demonstrated in Section 3 and confirmed by the calibration literature surveyed in Section 2.5, do not reliably indicate knowledge quality. The human overseer is asked to oversee a process they cannot observe.

Level B1 contribution. The autonomous assessment framework provides the information that effective human oversight requires. For each answer, the overseer receives not merely the answer and a confidence score, but the processing regime that produced it: Differentiated (clean structural processing, calibrated accuracy stated), Late Rescue (correct on calibration but structurally fragile—verify before acting), Override (late-layer destruction of earlier processing—do not trust), or Fused (no structural basis—reject or escalate). This transforms human oversight from an unfounded judgment into an informed decision supported by per-answer epistemic diagnostics.

Level B2 contribution. Ensemble governance provides a further layer. When models disagree, the regime fingerprints identify which model processed cleanly and which did not—enabling the human overseer to adjudicate disagreement on epistemic rather than arbitrary grounds. When models agree through clean processing, the compound reliability rate provides a quantified basis for trust. When models agree through compromised processing, the system flags collective collapse rather than presenting false consensus as confidence.

11.4 Article 15 — Accuracy, Robustness, and Cybersecurity

Article 15 requires that high-risk AI systems achieve an appropriate level of accuracy, robustness, and cybersecurity, and perform consistently in those respects throughout their lifecycle.

Accuracy. The Act does not define accuracy as aggregate benchmark performance, yet this is overwhelmingly how the industry interprets and reports it. Level A2 demonstrates that accuracy as conventionally measured systematically understates model capability. The epistemic gap is inverted across all models assessed in this work—structural correctness exceeds outcome accuracy by 10.4 to 15.8 percentage points—revealing that the dominant failure mode is suppressed delivery, not absent knowledge. Compliance with Article 15's accuracy requirement demands, at minimum, reporting both outcome accuracy and structural correctness, with the epistemic gap as a mandated disclosure.

Robustness. FEST provides a direct measure of robustness under perturbation. The finding that

accuracy varies by up to 30 percentage points on identical questions depending on which answer options are present exposes systematic fragility invisible to aggregate benchmarks. The fragility classification (LOW / MODERATE / HIGH) and the binary advantage metric provide the quantified robustness assessment that Article 15 requires. The invariance of fragility signatures across training stages (both Ministral variants showing identical 5.1 pp binary advantage) demonstrates that robustness is an architectural property requiring architectural solutions, not merely more training data.

Lifecycle consistency. The four-level hierarchy provides a framework for lifecycle monitoring. Level A2 calibration establishes the baseline epistemic profile. Any subsequent modification—fine-tuning, quantisation, LoRA adaptation, RLHF—can be assessed against that baseline. The finding that instruction tuning narrowed the Ministral epistemic gap from -12.5 pp to -10.4 pp while improving accuracy by 6.7 pp demonstrates genuine, quantifiable structural improvement of exactly the kind that Article 15 requires providers to detect and document. Without epistemic measurement, this kind of structural change is invisible; with it, it is quantified and auditable.

11.5 Articles 51–56 — GPAI Model Obligations

The GPAI provisions, enforceable since August 2025, impose specific obligations on providers of general-purpose AI models. These include:

Article 53 — Technical documentation. Providers must maintain detailed technical documentation including the results of model evaluations. Level A2 assessment reports—exemplified by the production reports reproduced in Appendices B–E—provide evaluation evidence at a depth that aggregate benchmark scores cannot approach. Across eight core models, the AIDA assessment programme produced 3,246,256 attention probes and 6,411,576 layer probes across 283,289 inferences. Each report is certificate-referenced and designed for regulatory audit.

Article 55 — Systemic risk assessment. Models classified as posing systemic risk face additional obligations for adversarial testing and model evaluation. FEST constitutes a systematic adversarial evaluation protocol: nine perturbation configurations per question, measuring model behaviour under controlled manipulation of the answer space. The discovery that models can be simultaneously robust on aggregate metrics and fragile under targeted perturbation is directly relevant to systemic risk assessment.

11.6 The Measurement Gap

The European Artificial Intelligence Act mandates risk management, human oversight, robustness assessment, and technical documentation for AI systems deployed in high-risk domains. These mandates presuppose measurement instruments capable of producing the evidence that compliance requires. The four-level measurement hierarchy presented in this paper provides that instrumentation:

The left column represents what the industry currently provides. The right three columns represent what the Act requires and what epistemic governance delivers. The gap between the first and second columns is not a matter of degree but of kind: the transition from "how many answers were correct" to "what quality of knowledge produced each answer" is a categorical shift in what is being measured. Without that shift, conformity assessment under the AI Act rests on

Table 34: Mapping of the four-level measurement hierarchy onto EU AI Act requirements.

AI Act Req.	Level A1	Level A2	Level B1	Level B2
Art. 9 Risk Mgt.	Error rate only	Risk decomposed by regime	Runtime risk classification	Compound risk with correlation
Art. 14 Oversight	Confidence score	Regime classification	Per-answer trust tier	Ensemble adjudication
Art. 15 Accuracy	Aggregate %	Structural correctness + gap	Calibrated tier accuracy	Compound reliability
Art. 15 Robustness	Not measured	FEST fragility profile	Fragility-informed trust	Cross-model robustness
Art. 53 Documentation	Benchmark table	Full audit report per model	Deployment monitoring log	Ensemble governance record

evidence that the Act's own requirements have rendered insufficient.

12 Conclusions

This paper began by observing that correctness is not cognition. It ends by demonstrating that cognition—measured, decomposed, corrected, and governed—is an engineering discipline.

12.1 What Has Been Established

The four-level measurement hierarchy introduced in Section 1 structures what the field currently possesses and what it lacks. At Level A1, benchmark accuracy tells us that a model answered 78% of questions correctly. It says nothing about why, nothing about how, and nothing about which 78%. The prevailing evaluation practice operates at this level.

At Level A2, epistemic calibration decomposes that aggregate into six regimes of internal processing, each with distinct geometric patterns, distinct reliability profiles, and distinct implications for trust. The empirical case study across four models and two vendors (Section 4) demonstrates that these regimes are not theoretical constructs but measurable structural properties of transformer inference—invariant across architectures, training pipelines, and subject domains. The central finding at this level is the epistemic gap: the distance between what a model appears to know (outcome accuracy) and what it structurally knows (structural correctness).

At Level B1, autonomous single-model assessment (Section 9) partitions every answer into trust tiers using only the observable regime classification—no gold answers required. The Differentiated regime carries calibrated accuracy rates between 72.8% and 84.5% depending on model, with hidden lie rates between 15.5% and 27.2% stated explicitly rather than concealed. The Fused regime carries accuracy rates between 0% and 16.2% and is flagged for categorical rejection. The partition is actionable: 86.5% of answers fall into assessable categories; 13.5% are flagged for human review. Every accuracy rate traces to Level A2 calibration—the dependency is fundamental and acknowledged throughout.

At Level B2, autonomous ensemble governance combines independent regime classifications and full rank orderings across multiple models to produce compound reliability estimates. Unanimous Differentiated agreement achieves 91.7% accuracy with an 8.3% irreducible hidden lie rate. Unanimous agreement through compromised processing achieves 45.5%—the same

surface phenomenon concealing a 46-percentage-point reliability gap visible only through regime classification. Complete disagreement yields 0% accuracy—a perfect negative signal. The elimination evidence is even stronger than the selection evidence: unanimous ensemble agreement on the worst option correctly excludes gold 97.3% of the time.

12.2 What Has Been Discovered

Beyond the measurement hierarchy, this work reports three mechanistic discoveries that fundamentally alter how transformer language models must be understood and governed.

The rotation discovery (Section 5). The probability dynamics observed through the logit lens across transformer layers—surges, collapses, crystallisation events—are not amplification or suppression events. They are rotations in the model's high-dimensional representational space. Representational energy is conserved; what changes is the alignment angle relative to the fixed output projection. The logit lens is a partial projection that systematically misrepresents the model's internal state. The model's errors are not failures of knowledge but failures of geometry—the knowledge is present but stored in a direction the output head cannot read.

The carrier–content decomposition (Section 7). Every model's output is composed of a position-dependent carrier signal—a structural property of the frozen weight matrices that systematically favours certain answer positions regardless of question content—and a content signal encoding the model's actual knowledge. This decomposition is confirmed across six models spanning four architecture families and six suppliers, with accuracy differentials exceeding 20 percentage points between the most-advantaged and most-penalised positions. The carrier direction is computable from the `lm_head` weight matrix alone, requiring zero data and negligible computation.

Inference-time epistemic correction (Section 7). Exploiting the carrier–content decomposition, a pipeline of carrier ablation, content enhancement, and multi-pass arbitration recovers previously incorrect answers from Llama-3-8B on 1,089 medical licensing questions, with a current production high-water mark of 71.1% (+3.21 pp over the honest 67.9% baseline). The pipeline is under active development; the correction mechanism is proven.

The FEST boundary discovery (Section 7.5). The Factual Elimination Stress Test, applied systematically to all four gold answer classes across 274 failures, establishes that only 14 samples (1.29% of the 1,089-sample evaluation set) constitute genuine knowledge gaps. The remaining 260 failures (94.9%) are architecturally recoverable: the content signal for the correct answer is present in the model's residual stream but is prevented from reaching the output by inference architecture failures. The true knowledge ceiling is **98.71%**. The 30-percentage-point gap between baseline accuracy and the knowledge ceiling is not an epistemic problem. It is an inference engineering problem.

These discoveries refine the six trajectory regimes into five epistemic regions (Section 8): Genuine Knowledge (robust, certifiable), Carrier-Assisted (fragile, dependent on positional luck), Carrier-Suppressed (hidden knowledge, recoverable without training), Content-Confused (genuine error), and Genuinely Unknowable (the true knowledge boundary). Published benchmark scores conflate all five into a binary correct/incorrect. The epistemic governance framework developed in Section 8 maps these regions to operational responses—trust, verify, correct, or reject—providing

the first epistemic state machine for transformer inference.

12.3 What Has Been Honestly Reported

The framework makes claims that are unprecedented in this field. It also makes concessions that are equally unprecedented.

The concessions: AIDA cannot determine whether any individual answer is correct or wrong without gold labels. The Differentiated Wrong population is geometrically indistinguishable from Differentiated Correct at the classification boundary, though a faint divergence emerges at deeper layers—too weak for per-question deployment, sufficient for population-level characterisation. The Fused population carries no geometric signal whatsoever; Fused Gold and Fused Wrong are identical in the activation space at every layer measured. Late Rescue answers, though 100% correct on the calibration set, lack the geometric structural support that would justify confidence in their stability under novel prompts or rephrased questions.

These are not limitations to be apologised for. They are the boundaries of honest measurement, and stating them is itself a contribution. The field currently operates without stating any boundaries at all—reporting 78% accuracy as though every answer in that 78% were equally trustworthy and every answer in the remaining 22% were equally untrustworthy. The present work replaces that fiction with a structured account of graduated reliability, explicit hidden error rates, and categorical rejection of answers the model cannot epistemically support.

12.4 What Has Been Made Possible

Five capabilities now exist that did not exist before this work.

First, the ability to measure the epistemic integrity of any transformer model on any multiple-choice assessment through automated, reproducible instrumentation—AIDA for trajectory classification, ASCOL for cross-model comparison, and FEST for fragility profiling—producing standardised reports suitable for regulatory audit.

Second, the ability to deploy those measurements at inference time without access to gold answers, partitioning every answer into trust tiers with calibrated accuracy rates and flagging epistemically unsupported answers for human review.

Third, the ability to combine multiple models into regime-aware ensembles that exploit not only agreement but the full rank-order structure of each model's assessment—selection, elimination, and second-choice evidence streams weighted by the epistemic quality of the processing that produced them.

Fourth, the ability to establish the true knowledge boundary of a model through systematic FEST classification across all answer positions — distinguishing genuine knowledge gaps (1.29% of samples) from architecturally recoverable failures (94.9%), and thereby separating what requires training from what requires inference engineering.

Fifth, the ability to classify every model output into one of five epistemic regions—distinguishing genuine knowledge from positional artefact within both correct and incorrect populations—and to route each region to the appropriate governance action: trust, verify, correct, or reject.

12.5 The Path Forward

Section 10 describes how these capabilities translate to clinical practice—not as a replacement for clinical judgment but as a structured, epistemically auditable second opinion with stated confidence bounds and the ability to stress-test its own conclusions through targeted perturbation. Section 11 maps the framework onto the specific obligations of the European Artificial Intelligence Act, demonstrating that the instruments developed here provide the technical infrastructure for compliance that the regulation demands but the industry does not yet supply.

The gap between the current state of AI governance—accuracy percentages reported without epistemic qualification—and the requirements of both clinical safety and regulatory compliance is not a gap that better prompting, larger models, or more compute will close. It is an instrumentation gap. The instruments now exist. The rotation discovery and carrier decomposition show that the gap between published scores and true capability is predominantly geometric, not epistemic—and correctable without training.

The question this paper leaves with its readers is not technical. It is institutional. The measurement hierarchy is established. The calibration methodology is validated. The autonomous governance framework is operational. The carrier decomposition is demonstrated. The regulatory mapping is complete. What remains is the decision—by model developers, clinical institutions, and regulatory bodies—to require epistemic transparency as a condition of deployment rather than accepting accuracy as a proxy for trustworthiness.

Correctness without cognition is not safety. Cognition without measurement is not governance. Measurement without honesty is not science. This work provides the instruments for all three.

References

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., et al. (2025). Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Belrose, N., Furman, H., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. (2022). Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8493–8502.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized language models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

European Commission (2025). General-purpose AI code of practice.

European Parliament (2024). Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer feed-forward layers are key–value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 107–112.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2790–2799.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Kalajdzievski, D. (2025). RandLoRA: Full rank parameter-efficient fine-tuning of large models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4582–4597.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., et al. (2025). On the biology of a large language model. *Transformer Circuits Thread*.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. (2024). DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Marks, S., Mueller, A., et al. (2025). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3428–3448.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229.

National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, NIST.

nostalgebraist (2020). Interpreting GPT: The logit lens. LessWrong.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., et al. (2025). Humanity's last exam. *arXiv preprint arXiv:2501.14249*. Published by Center for AI Safety and Scale AI.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

A AIDA Internal Assessment Report: Ministral-14B Base

This appendix reproduces the full production AIDA internal assessment report for `mistralai/Ministral-3-14B-Base-2512`*, a 14B-parameter dense transformer with 42 layers. Report generated 28 February 2026. Report ID:* `GBRAAA00-RPT-2e35a43a-7b1c-4d47-84b6-bad51ad2ec8f`.

A.1 Executive Summary

The model achieves 70.3% outcome accuracy but 82.8% structural correctness, yielding an epistemic gap of −12.5 percentage points. This means the model structurally knows the correct answer in 136 more cases than it delivers correctly—known answers are lost in late-layer processing.

Table 35: Headline metrics: Ministral-14B Base.

Metric	Value	Description
Outcome Accuracy	70.3%	Conventional benchmark performance
Structural Correctness	82.8%	Answers with genuine structural integrity
Epistemic Gap	−12.5 pp	Gap between accuracy and structural correctness
Views Agreement	72.9%	Diagnostic confidence
Fusion Rate	17.3%	Samples with no internal differentiation
Mean Stability Score	1.48/4	Average processing stability
Mean Flip Count	7.2	Average answer changes across layers
Total Layer Probes	45,738	Individual measurements taken

A.2 Trajectory Classification

Each of the 1,089 samples is classified into one of six trajectory types based on how the model internal representations evolve across its 42 layers. The dominant trajectory is Differentiated Correct (641 samples, 58.9%), indicating genuine structural knowledge. However, 124 samples (11.4%) are Late Crystallisation—correct but shallow. A concerning 151 samples (13.9%) show confident but wrong structural processing.

Table 36: Trajectory breakdown: Ministral-14B Base.

Trajectory	Count	%	Fusion	Implication
Differentiated Correct	641	58.9	0%	Genuine knowledge—safe for deployment
Late Crystallisation	124	11.4	99%	Shallow—sensitive to prompt variation
Differentiated Wrong	151	13.9	0%	Structural failure—high deployment risk
Correct Overridden	137	12.6	23%	Knowledge lost in depth—training conflict
Fused Wrong	35	3.2	94%	No knowledge—effectively random
Fused Gold	1	0.1	100%	Inflates accuracy—no structural basis

A.3 Layer Dynamics—Where Decisions Happen

The collapse layer indicates at which transformer layer the model commits to its final answer. The model shows 0 samples (0.0%) collapsing at the final layer (L42), and 0 more at L41. Combined, 0.0% of all decisions happen in the last two layers.

Decision Zone Analysis. Early layers (1–21): 402 samples (36.9%)—deep structural decisions. Mid layers (22–35): 155 samples (14.2%)—intermediate processing. Late layers (36–42): 532 samples (48.9%)—surface decisions.

A.4 Stability Analysis

Stability is measured across four indicators. Only 39 samples (3.6%) achieve full stability (4/4). 304 samples (27.9%) score 0/4—completely unstable processing across all indicators.

Individual Stability Indicators. Centroid Stable: 785/1089 (72.1%). Entropy Decreasing: 367/1089 (33.7%). Margin Increasing: 391/1089 (35.9%). Delta Decreasing: 70/1089 (6.4%).

A.5 Decision Volatility—Flip Analysis

A "flip" occurs when the model changes its predicted answer between consecutive layers. The model averages 7.2 flips per sample (median: 7). The maximum observed is 14 flips.

A.6 Entropy Sharpening

Mean early-layer entropy is 1.213. Mean final-layer entropy drops to 0.424. However, 389 samples (35.7%) still have elevated final entropy (>0.5), indicating residual uncertainty.

A.7 Internal Ranking Analysis

The correct answer reaches Rank 4 (highest confidence) in only 451 cases (41.4%). In 142 cases (13.0%), the correct answer is ranked dead last internally.

A.8 View Concordance & Disagreement Patterns

The AIDA framework analyses model internals through two complementary lenses: geometric and logit. These views agree on 794 samples (72.9%). The dominant view is geometry (857 samples).

The most significant disagreement pattern involves all 124 Late Crystallisation samples: the geometric view classifies these as Fused Gold, while the logit view classifies them as Differentiated Correct. The joint classification resolves this conservatively as Late Crystallisation.

A.9 Centroid Shift Analysis

The centroid shift layer marks where the model first begins to differentiate between answer options. 632 samples (58.0%) show a shift at Layer 1, suggesting immediate differentiation.

A.10 Fusion Patterns

Fusion was detected in 188 samples (17.3%). Critically, 0% of Differentiated Correct samples show fusion—structural knowledge and fusion are mutually exclusive.

A.11 FEST Fragility Profile

The Factual Elimination Stress Test (FEST) systematically removes and recombines answer options to measure how dependent the model is on distractor context. Each of the 1,089 questions is presented in nine configurations ranging from binary confrontations (2 options) to the full 4-option MCQ.

Table 37: FEST stage results: Ministral-14B Base.

Stage	Description	Options	Accuracy	Mean Gap
MCQ	Baseline 4-option MCQ	4	72.5%	0.607
F01	Forced Error (Gold removed)	3	N/A (no gold)	0.660
F02	Secondary Attractor (D* removed)	2	N/A (no gold)	0.750
F03	Gold + D* + Di (3-option)	3	62.3%	0.622
F04	Gold + D* + Dj (3-option)	3	60.8%	0.614
F05	Binary: Gold vs D*	2	77.7%	0.766
F06	Binary: Gold vs D'	2	70.3%	0.688
F07	Distractor Hierarchy (D* vs D')	2	N/A (no gold)	0.728
F08	Gold vs Weakest Distractor	2	90.5%	0.859
F09	Restoration Control (full MCQ)	4	72.5%	0.607

Fragility Analysis. Binary advantage (F05 vs MCQ): +5.1 pp. Removing all distractors except D* improves accuracy by 5.1 pp, confirming that additional distractors in the full MCQ interfere with correct discrimination. Distractor concentration (F03 vs MCQ): −10.3 pp. Concentrating the strongest attractor into a 3-option set paradoxically reduces accuracy versus the full 4-option MCQ. Weak distractors in the full MCQ dilute D* pull. Fragility classification: **MODERATE** fragility (5.1 pp gap between binary and full MCQ). The model shows some vulnerability to distractor interference. Test-retest reliability (F09 vs MCQ): 0.000 pp delta. PASSED—perfect pipeline reliability confirmed.

A.12 Key Findings & Recommendations

The 12.5 pp inverted epistemic gap means the model structurally knows the correct answer in substantially more cases than it delivers correctly—a governance failure, not a knowledge failure. 27.9% structural instability and average 7.2 flips per sample indicate volatile internal processing. 0.0% of decisions occur at the final layer—knowledge is surface-level, not deeply encoded. The systematic GEO/LOGIT disagreement on 27.1% of samples warrants further investigation. Fine-tuning (instruct, RLHF, DPO) applied to these base weights should be independently assessed. FEST reveals a 5.1 pp fragility gap: the model scores 77.7% in binary confrontation but only 72.5% under full distractor load, indicating multi-option interference degrades factual recall. FEST test-retest reliability confirmed: F09 restoration control matches MCQ baseline within 0.000 pp, validating the measurement pipeline.

B AIDA Internal Assessment Report: Ministral-14B Reasoning

This appendix reproduces the full production AIDA internal assessment report for `mistralai/Ministral-3-14B-Reasoning-2512`*, a 14B-parameter dense transformer with 42 layers. Report generated 28 February 2026. Report ID:* `GBRAAA00-RPT-0c58bc44-59bd-4e39-a4cb-b131861bdee9`.

B.1 Executive Summary

The model achieves 65.9% outcome accuracy but 81.7% structural correctness, yielding an epistemic gap of −15.8 percentage points. This means the model structurally knows the correct answer in 172 more cases than it delivers correctly—known answers are lost in late-layer processing.

Table 38: Headline metrics: Ministral-14B Reasoning.

Metric	Value	Description
Outcome Accuracy	65.9%	Conventional benchmark performance
Structural Correctness	81.7%	Answers with genuine structural integrity
Epistemic Gap	−15.8 pp	Gap between accuracy and structural correctness
Views Agreement	71.4%	Diagnostic confidence
Fusion Rate	15.4%	Samples with no internal differentiation
Mean Stability Score	1.44/4	Average processing stability
Mean Flip Count	7.6	Average answer changes across layers
Total Layer Probes	45,738	Individual measurements taken

B.2 Trajectory Classification

Each of the 1,089 samples is classified into one of six trajectory types based on how the model internal representations evolve across its 42 layers. The dominant trajectory is Differentiated Correct (618 samples, 56.7%), indicating genuine structural knowledge. However, 97 samples (8.9%) are Late Crystallisation—correct but shallow. A concerning 152 samples (14.0%) show confident but wrong structural processing.

Table 39: Trajectory breakdown: Ministral-14B Reasoning.

Trajectory	Count	%	Fusion	Implication
Differentiated Correct	618	56.7	0%	Genuine knowledge—safe for deployment
Late Crystallisation	97	8.9	98%	Shallow—sensitive to prompt variation
Differentiated Wrong	152	14.0	0%	Structural failure—high deployment risk
Correct Overridden	175	16.1	18%	Knowledge lost in depth—training conflict
Fused Wrong	44	4.0	89%	No knowledge—effectively random
Fused Gold	3	0.3	100%	Inflates accuracy—no structural basis

B.3 Layer Dynamics—Where Decisions Happen

The collapse layer indicates at which transformer layer the model commits to its final answer. The model shows 0 samples (0.0%) collapsing at the final layer (L42), and 0 more at L41. Combined, 0.0% of all decisions happen in the last two layers.

Decision Zone Analysis. Early layers (1–21): 337 samples (30.9%)—deep structural decisions. Mid layers (22–35): 217 samples (19.9%)—intermediate processing. Late layers (36–42): 535 samples (49.1%)—surface decisions.

B.4 Stability Analysis

Stability is measured across four indicators. Only 35 samples (3.2%) achieve full stability (4/4). 327 samples (30.0%) score 0/4—completely unstable processing across all indicators.

Individual Stability Indicators. Centroid Stable: 762/1089 (70.0%). Delta Decreasing: 75/1089 (6.9%). Entropy Decreasing: 349/1089 (32.0%). Margin Increasing: 382/1089 (35.1%).

B.5 Decision Volatility—Flip Analysis

A "flip" occurs when the model changes its predicted answer between consecutive layers. The model averages 7.6 flips per sample (median: 7). The maximum observed is 15 flips.

B.6 Entropy Sharpening

Mean early-layer entropy is 1.237. Mean final-layer entropy drops to 0.407. However, 356 samples (32.7%) still have elevated final entropy (>0.5), indicating residual uncertainty.

B.7 Internal Ranking Analysis

The correct answer reaches Rank 4 (highest confidence) in only 486 cases (44.6%). In 125 cases (11.5%), the correct answer is ranked dead last internally.

B.8 View Concordance & Disagreement Patterns

The AIDA framework analyses model internals through two complementary lenses: geometric and logit. These views agree on 778 samples (71.4%). The dominant view is geometry (861 samples).

The most significant disagreement pattern involves all 97 Late Crystallisation samples: the geometric view classifies these as Fused Gold, while the logit view classifies them as Differentiated Correct. The joint classification resolves this conservatively as Late Crystallisation.

B.9 Centroid Shift Analysis

The centroid shift layer marks where the model first begins to differentiate between answer options. 612 samples (56.2%) show a shift at Layer 1, suggesting immediate differentiation.

B.10 Fusion Patterns

Fusion was detected in 168 samples (15.4%). Critically, 0% of Differentiated Correct samples show fusion—structural knowledge and fusion are mutually exclusive.

B.11 FEST Fragility Profile

The Factual Elimination Stress Test (FEST) systematically removes and recombines answer options to measure how dependent the model is on distractor context. Each of the 1,089 questions is presented in nine configurations ranging from binary confrontations (2 options) to the full 4-option MCQ.

Fragility Analysis. Binary advantage (F05 vs MCQ): +12.6 pp. Removing all distractors except D* improves accuracy by 12.6 pp, confirming that additional distractors in the full MCQ interfere with correct discrimination. Distractor concentration (F03 vs MCQ): −8.6 pp. Concentrating the strongest attractor into a 3-option set paradoxically reduces accuracy versus the full 4-option MCQ. Fragility classification: **HIGH** fragility (12.6 pp gap between binary and full MCQ). The model is highly susceptible to multi-option interference. Test-retest reliability (F09 vs MCQ): 0.000 pp delta. PASSED—perfect pipeline reliability confirmed.

Table 40: FEST stage results: Ministral-14B Reasoning.

Stage	Description	Options	Accuracy	Mean Gap
MCQ	Baseline 4-option MCQ	4	69.3%	0.573
F01	Forced Error (Gold removed)	3	N/A (no gold)	0.597
F02	Secondary Attractor (D* removed)	2	N/A (no gold)	0.575
F03	Gold + D* + Di (3-option)	3	60.7%	0.570
F04	Gold + D* + Dj (3-option)	3	63.4%	0.572
F05	Binary: Gold vs D*	2	81.9%	0.739
F06	Binary: Gold vs D'	2	77.6%	0.680
F07	Distractor Hierarchy (D* vs D')	2	N/A (no gold)	0.575
F08	Gold vs Weakest Distractor	2	93.1%	0.822
F09	Restoration Control (full MCQ)	4	69.3%	0.573

B.12 Key Findings & Recommendations

The 15.8 pp inverted epistemic gap means the model structurally knows the correct answer in substantially more cases than it delivers correctly—a governance failure, not a knowledge failure. 30.0% structural instability and average 7.6 flips per sample indicate volatile internal processing. The systematic GEO/LOGIT disagreement on 28.6% of samples warrants further investigation. FEST reveals a 12.6 pp fragility gap: the model scores 81.9% in binary confrontation but only 69.3% under full distractor load, indicating multi-option interference degrades factual recall. FEST test-retest reliability confirmed: F09 restoration control matches MCQ baseline within 0.000 pp, validating the measurement pipeline.

C AIDA Internal Assessment Report: Ministral-14B Instruct

This appendix reproduces the full production AIDA internal assessment report for `mistralai/Ministral-3-14B-Instruct-2512-BF16`*, a 14B-parameter dense transformer with 42 layers. Report generated 28 February 2026. Report ID:* `GBRAAA00-RPT-bd9d9e43-02f4-42b5-9364-10d410224931`.

C.1 Executive Summary

The model achieves 77.0% outcome accuracy but 87.4% structural correctness, yielding an epistemic gap of −10.4 percentage points. This means the model structurally knows the correct answer in 113 more cases than it delivers correctly—known answers are lost in late-layer processing.

Table 41: Headline metrics: Ministral-14B Instruct.

Metric	Value	Description
Outcome Accuracy	77.0%	Conventional benchmark performance
Structural Correctness	87.4%	Answers with genuine structural integrity
Epistemic Gap	−10.4 pp	Gap between accuracy and structural correctness
Views Agreement	68.5%	Diagnostic confidence
Fusion Rate	23.9%	Samples with no internal differentiation
Mean Stability Score	1.59/4	Average processing stability
Mean Flip Count	7.5	Average answer changes across layers
Total Layer Probes	45,738	Individual measurements taken

C.2 Trajectory Classification

Each of the 1,089 samples is classified into one of six trajectory types based on how the model internal representations evolve across its 42 layers. The dominant trajectory is Differentiated Correct (648 samples, 59.5%), indicating genuine structural knowledge. However, 186 samples (17.1%) are Late Crystallisation—correct but shallow. A concerning 88 samples (8.1%) show confident but wrong structural processing.

Table 42: Trajectory breakdown: Ministral-14B Instruct.

Trajectory	Count	%	Fusion	Implication
Differentiated Correct	648	59.5	0%	Genuine knowledge—safe for deployment
Late Crystallisation	186	17.1	98%	Shallow—sensitive to prompt variation
Differentiated Wrong	88	8.1	0%	Structural failure—high deployment risk
Correct Overridden	118	10.8	29%	Knowledge lost in depth—training conflict
Fused Wrong	44	4.0	86%	No knowledge—effectively random
Fused Gold	5	0.5	100%	Inflates accuracy—no structural basis

C.3 Layer Dynamics—Where Decisions Happen

The collapse layer indicates at which transformer layer the model commits to its final answer. The model shows 0 samples (0.0%) collapsing at the final layer (L42), and 0 more at L41. Combined, 0.0% of all decisions happen in the last two layers.

Decision Zone Analysis. Early layers (1–21): 412 samples (37.8%)—deep structural decisions. Mid layers (22–35): 182 samples (16.7%)—intermediate processing. Late layers (36–42): 495 samples (45.5%)—surface decisions.

C.4 Stability Analysis

Stability is measured across four indicators. Only 30 samples (2.8%) achieve full stability (4/4). 290 samples (26.6%) score 0/4—completely unstable processing across all indicators.

Individual Stability Indicators. Entropy Decreasing: 427/1089 (39.2%). Margin Increasing: 442/1089 (40.6%). Centroid Stable: 799/1089 (73.4%). Delta Decreasing: 62/1089 (5.7%).

C.5 Decision Volatility—Flip Analysis

A "flip" occurs when the model changes its predicted answer between consecutive layers. The model averages 7.5 flips per sample (median: 7). The maximum observed is 14 flips.

C.6 Entropy Sharpening

Mean early-layer entropy is 1.250. Mean final-layer entropy drops to 0.337. However, 281 samples (25.8%) still have elevated final entropy (>0.5), indicating residual uncertainty.

C.7 Internal Ranking Analysis

The correct answer reaches Rank 4 (highest confidence) in only 489 cases (44.9%). In 130 cases (11.9%), the correct answer is ranked dead last internally.

C.8 View Concordance & Disagreement Patterns

The AIDA framework analyses model internals through two complementary lenses: geometric and logit. These views agree on 746 samples (68.5%). The dominant view is geometry (819 samples).

The most significant disagreement pattern involves all 186 Late Crystallisation samples: the geometric view classifies these as Fused Gold, while the logit view classifies them as Differentiated Correct. The joint classification resolves this conservatively as Late Crystallisation.

C.9 Centroid Shift Analysis

The centroid shift layer marks where the model first begins to differentiate between answer options. 616 samples (56.6%) show a shift at Layer 1, suggesting immediate differentiation.

C.10 Fusion Patterns

Fusion was detected in 260 samples (23.9%). Critically, 0% of Differentiated Correct samples show fusion—structural knowledge and fusion are mutually exclusive.

C.11 FEST Fragility Profile

The Factual Elimination Stress Test (FEST) systematically removes and recombines answer options to measure how dependent the model is on distractor context. Each of the 1,089 questions is presented in nine configurations ranging from binary confrontations (2 options) to the full 4-option MCQ.

Table 43: FEST stage results: Ministral-14B Instruct.

Stage	Description	Options	Accuracy	Mean Gap
MCQ	Baseline 4-option MCQ	4	79.8%	0.680
F01	Forced Error (Gold removed)	3	N/A (no gold)	0.554
F02	Secondary Attractor (D* removed)	2	N/A (no gold)	0.534
F03	Gold + D* + Di (3-option)	3	77.2%	0.674
F04	Gold + D* + Dj (3-option)	3	76.2%	0.685
F05	Binary: Gold vs D*	2	84.9%	0.771
F06	Binary: Gold vs D'	2	87.5%	0.767
F07	Distractor Hierarchy (D* vs D')	2	N/A (no gold)	0.546
F08	Gold vs Weakest Distractor	2	96.1%	0.863
F09	Restoration Control (full MCQ)	4	79.8%	0.680

Fragility Analysis. Binary advantage (F05 vs MCQ): +5.1 pp. Removing all distractors except D* improves accuracy by 5.1 pp, confirming that additional distractors in the full MCQ interfere with correct discrimination. Distractor concentration (F03 vs MCQ): −2.6 pp. Concentrating the strongest attractor into a 3-option set paradoxically reduces accuracy versus the full 4-option MCQ. Fragility classification: **MODERATE** fragility (5.1 pp gap between binary and full MCQ). The model shows some vulnerability to distractor interference. Test-retest reliability (F09 vs MCQ): 0.000 pp delta. PASSED—perfect pipeline reliability confirmed.

C.12 Key Findings & Recommendations

The 10.4 pp inverted epistemic gap means the model structurally knows the correct answer in substantially more cases than it delivers correctly—a governance failure, not a knowledge failure. 26.6% structural instability and average 7.5 flips per sample indicate volatile internal processing. The systematic GEO/LOGIT disagreement on 31.5% of samples warrants further investigation. FEST reveals a 5.1 pp fragility gap: the model scores 84.9% in binary confrontation but only 79.8% under full distractor load, indicating multi-option interference degrades factual recall. FEST test-retest reliability confirmed: F09 restoration control matches MCQ baseline within 0.000 pp, validating the measurement pipeline.

D AIDA Internal Assessment Report: Gemma-2-9B

This appendix reproduces the full production AIDA internal assessment report for `google/gemma-2-9b`*, a 9B-parameter dense transformer with 42 layers (Base/Pre-trained). Report generated 28 February 2026. Report ID:* `GBRAAA00-RPT-e974894c-faeb-4dc9-9e95-37663f45b2bb`.

D.1 Executive Summary

The model achieves 74.6% outcome accuracy but 86.0% structural correctness, yielding an epistemic gap of −11.5 percentage points. This means the model structurally knows the correct answer in 125 more cases than it delivers correctly—known answers are lost in late-layer processing.

Table 44: Headline metrics: Gemma-2-9B.

Metric	Value	Description
Outcome Accuracy	74.6%	Conventional benchmark performance
Structural Correctness	86.0%	Answers with genuine structural integrity
Epistemic Gap	−11.5 pp	Gap between accuracy and structural correctness
Views Agreement	72.8%	Diagnostic confidence
Fusion Rate	18.5%	Samples with no internal differentiation
Mean Stability Score	1.65/4	Average processing stability
Mean Flip Count	6.0	Average answer changes across layers
Total Layer Probes	45,738	Individual measurements taken

D.2 Trajectory Classification

Each of the 1,089 samples is classified into one of six trajectory types based on how the model internal representations evolve across its 42 layers. The dominant trajectory is Differentiated Correct (671 samples, 61.6%), indicating genuine structural knowledge. However, 138 samples (12.7%) are Late Crystallisation—correct but shallow. A concerning 114 samples (10.5%) show confident but wrong structural processing.

D.3 Layer Dynamics—Where Decisions Happen

The collapse layer indicates at which transformer layer the model commits to its final answer. The model shows extreme late-decision behaviour: 79 samples (7.3%) collapse only at the final layer (L42), and 111 more at L41. Combined, 17.4% of all decisions happen in the last two layers.

Table 45: Trajectory breakdown: Gemma-2-9B.

Trajectory	Count	%	Fusion	Implication
Differentiated Correct	671	61.6	0%	Genuine knowledge—safe for deployment
Late Crystallisation	138	12.7	100%	Shallow—sensitive to prompt variation
Differentiated Wrong	114	10.5	0%	Structural failure—high deployment risk
Correct Overridden	128	11.8	20%	Knowledge lost in depth—training conflict
Fused Wrong	35	3.2	100%	No knowledge—effectively random
Fused Gold	3	0.3	100%	Inflates accuracy—no structural basis

Decision Zone Analysis. Early layers (1–21): 239 samples (21.9%)—deep structural decisions. Mid layers (22–35): 317 samples (29.1%)—intermediate processing. Late layers (36–42): 533 samples (48.9%)—surface decisions.

D.4 Stability Analysis

Stability is measured across four indicators. Only 25 samples (2.3%) achieve full stability (4/4). 190 samples (17.4%) score 0/4—completely unstable processing across all indicators.

Individual Stability Indicators. Centroid Stable: 899/1089 (82.6%). Entropy Decreasing: 347/1089 (31.9%). Margin Increasing: 448/1089 (41.1%). Delta Decreasing: 99/1089 (9.1%).

D.5 Decision Volatility—Flip Analysis

A "flip" occurs when the model changes its predicted answer between consecutive layers. The model averages 6.0 flips per sample (median: 6). The maximum observed is 16 flips.

D.6 Entropy Sharpening

Mean early-layer entropy is 0.889. Mean final-layer entropy drops to 0.127. However, 100 samples (9.2%) still have elevated final entropy (>0.5), indicating residual uncertainty.

D.7 Internal Ranking Analysis

The correct answer reaches Rank 4 (highest confidence) in only 439 cases (40.3%). In 163 cases (15.0%), the correct answer is ranked dead last internally.

D.8 View Concordance & Disagreement Patterns

The AIDA framework analyses model internals through two complementary lenses: geometric and logit. These views agree on 793 samples (72.8%). The dominant view is geometry (868 samples).

The most significant disagreement pattern involves all 138 Late Crystallisation samples: the geometric view classifies these as Fused Gold, while the logit view classifies them as Differentiated Correct. The joint classification resolves this conservatively as Late Crystallisation.

D.9 Centroid Shift Analysis

The centroid shift layer marks where the model first begins to differentiate between answer options. 569 samples (52.2%) show a shift at Layer 1, suggesting immediate differentiation.

D.10 Fusion Patterns

Fusion was detected in 201 samples (18.5%). Critically, 0% of Differentiated Correct samples show fusion—structural knowledge and fusion are mutually exclusive.

D.11 FEST Fragility Profile

The Factual Elimination Stress Test (FEST) systematically removes and recombines answer options to measure how dependent the model is on distractor context. Each of the 1,089 questions is presented in nine configurations ranging from binary confrontations (2 options) to the full 4-option MCQ.

Table 46: FEST stage results: Gemma-2-9B.

Stage	Description	Options	Accuracy	Mean Gap
MCQ	Baseline 4-option MCQ	4	77.2%	0.711
F01	Forced Error (Gold removed)	3	N/A (no gold)	0.546
F02	Secondary Attractor (D* removed)	2	N/A (no gold)	0.607
F03	Gold + D* + Di (3-option)	3	75.2%	0.694
F04	Gold + D* + Dj (3-option)	3	75.3%	0.704
F05	Binary: Gold vs D*	2	77.9%	0.712
F06	Binary: Gold vs D'	2	87.1%	0.775
F07	Distractor Hierarchy (D* vs D')	2	N/A (no gold)	0.608
F08	Gold vs Weakest Distractor	2	93.2%	0.844
F09	Restoration Control (full MCQ)	4	77.2%	0.711

Fragility Analysis. Binary advantage (F05 vs MCQ): +0.6 pp. Removing all distractors except D* improves accuracy by 0.6 pp, confirming that additional distractors in the full MCQ interfere with correct discrimination. Distractor concentration (F03 vs MCQ): −2.0 pp. Concentrating the strongest attractor into a 3-option set paradoxically reduces accuracy versus the full 4-option MCQ. Fragility classification: **LOW** fragility (0.6 pp gap between binary and full MCQ). The model maintains reasonable discrimination under distractor load. Test-retest reliability (F09 vs MCQ): 0.000 pp delta. PASSED—perfect pipeline reliability confirmed.

D.12 Key Findings & Recommendations

The 11.5 pp inverted epistemic gap means the model structurally knows the correct answer in substantially more cases than it delivers correctly—a governance failure, not a knowledge failure. 17.4% structural instability and average 6.0 flips per sample indicate volatile internal processing. 7.3% of decisions occur at the final layer—knowledge is surface-level, not deeply encoded. The systematic GEO/LOGIT disagreement on 27.2% of samples warrants further investigation. FEST reveals a 0.6 pp fragility gap: the model scores 77.9% in binary confrontation but only 77.2% under full distractor load, indicating multi-option interference degrades factual recall. FEST test-retest reliability confirmed: F09 restoration control matches MCQ baseline within 0.000 pp, validating the measurement pipeline.

E AIDA Comparative Assessment: Gemma-9B vs Llama-8B

This appendix reproduces a production AIDA comparative assessment report evaluating a proposed model replacement: Google DeepMind Gemma-9B (current deployment) vs Meta Llama-8B (proposed replacement). Report generated 1 March 2026. Report ID: `AIDA-CMP-00bcf719-a09e-40a2-ae1a-a13229b`

Model A: Google DeepMind Gemma-9B	**Model B: Meta Llama-8B**
`google/gemma-2-9b`	`meta-llama/Meta-Llama-3-8B`
9B parameters, 42 layers, Base	8B parameters, 32 layers, Base
Certificate: `AIDA-3201ec61...`	Certificate: `AIDA-5114562e...`
Expiry: 2026-08-29	Expiry: 2026-08-29

Overall recommendation: NOT RECOMMENDED—proposed model shows degraded epistemic integrity.

E.1 Head-to-Head Comparison

Table 47: Key metrics comparison: Gemma-9B vs Llama-8B.

Metric	**Gemma-9B**	**Llama-8B**	**Delta**	
Outcome Accuracy	74.6%	57.4%	−17.2	✗
Structural Correctness	86.0%	71.8%	−14.2	✗
Epistemic Gap	−11.5 pp	−14.4 pp	−2.9	✗
Views Agreement	72.8%	81.7%	+8.9	✓
Fusion Rate	18.5%	0.0%	−18.5	✓
Mean Stability	1.6/4	1.5/4	−0.1	✗
Mean Flips	6.0	27.6	+21.6	✗
Gold at Top Rank	40.3%	31.0%	−9.3	✗

E.2 Trajectory & Structural Analysis

Table 48: Trajectory distribution comparison: Gemma-9B vs Llama-8B.

Trajectory	**Gemma-9B**		**Llama-8B**		**Change**
	n	%	n	%	
Diff. Correct	671	61.6	604	55.5	−6.2 pp
Late Cryst.	138	12.7	21	1.9	−10.7 pp
Diff. Wrong	114	10.5	278	25.5	+15.1 pp
Correct Overr.	128	11.8	157	14.4	+2.7 pp
Fused Wrong	35	3.2	29	2.7	−0.6 pp
Fused Gold	3	0.3	0	0.0	−0.3 pp

Decision Depth Comparison. Gemma-9B makes 79 decisions (7.3%) at its final layer. Llama-8B makes 187 decisions (17.2%) at its final layer. Earlier collapse indicates deeper structural encoding of knowledge.

E.3 Stability & Volatility

Stability Score Comparison. Gemma-9B: 190 samples at 0/4 stability (17.4%), mean 1.65/4. Llama-8B: 316 samples at 0/4 stability (29.0%), mean 1.51/4. Higher stability indicates more reliable, consistent internal processing.

Decision Volatility Comparison. Gemma-9B averages 6.0 answer flips per sample. Llama-8B averages 27.6 flips. Fewer flips indicates more decisive, stable processing.

Entropy Summary. Gemma-9B: mean early entropy 0.889 → final 0.127. Llama-8B: mean early entropy 1.226 → final 0.825. Lower final entropy indicates more confident, decisive output.

E.4 Conclusions & Recommendation

Outcome accuracy declines from 74.6% to 57.4% (−17.2 pp)—the proposed model answers fewer questions correctly. Structural correctness declines from 86.0% to 71.8% (−14.2 pp)—fewer correct answers demonstrate genuine knowledge. The epistemic gap widens from −11.5 pp to −14.4 pp (−2.9 pp)—conventional benchmarks are more misleading for the proposed model.

Differentiated Correct trajectory decreases from 61.6% to 55.5%—fewer answers follow the ideal knowledge pathway. Late Crystallisation decreases from 12.7% to 1.9%—fewer answers rely on shallow, last-layer processing. Mean stability declines from 1.65/4 to 1.51/4—less consistent internal processing. Mean flip count increases from 6.0 to 27.6—more oscillation during inference.

Recommendation. Based on the comparative assessment, the proposed replacement of Gemma-9B with Llama-8B is **not recommended**. The proposed model shows degraded epistemic integrity.

Conditions. The proposed model must complete full AIDA certification (ASCOL + FEST) before deployment. Integration testing with existing systems must be validated. A parallel running period of minimum 30 days is recommended before decommissioning the current model. The certification of the replaced model should be formally retired upon switch-over. Post-deployment monitoring should include quarterly AIDA re-assessment for the first year.