AIDA Research

Adaptive Inference Decision Architecture

Model Assessment Certificate

meta-llama/Meta-Llama-3-8B

Certificate: AIDA-ea6a9f73-7d83-4a00-9bbe-e70b65462ddf

Assessed: 10 March 2026 at 14:02 UTC

Model Provenance

Base Weights Supplier	Meta
Weights Hosted By	Hugging Face
HuggingFace ID	meta-llama/Meta-Llama-3-8B
Training Stage	Base (Pre-trained)

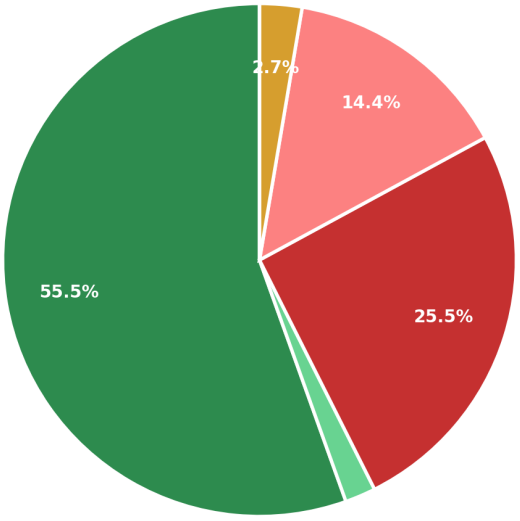
57.4%
Outcome Accuracy

71.8%
Structural Correctness

-14.4pp
Epistemic Gap

This model answers 57.4% of mmlu_med questions correctly, but only 71.8% demonstrate structurally sound reasoning through the transformer layers. The -14.4 percentage point gap indicates that -157 of 625 correct answers lack epistemic integrity — the model arrives at the right answer without developing consistent internal representations across the transformer depth.

Trajectory Classification



Trajectory Definitions

Differentiated Correct (604)
Genuine structural knowledge. Consistent internal separation leading to correct answer.

Late Crystallisation (21)
Answer emerges only in final layers. Shallow knowledge without deep support.

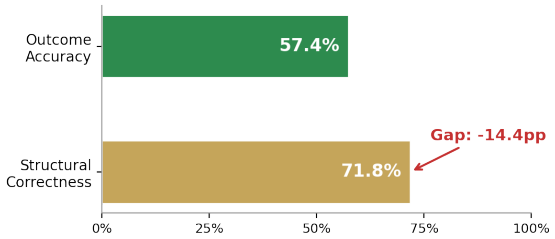
Differentiated Wrong (278)
Confident but incorrect. Clear structure converges on the wrong answer.

Correct Overridden (157)
Had the correct answer at intermediate layers but overrode it in later processing.

Fused Wrong (29)
No meaningful differentiation between options. Cannot distinguish alternatives.

Fused Gold (0)
No internal differentiation, but correct by chance. Accuracy without substance.

Epistemic Gap Analysis



Model Specification

Model Name	meta-llama/Meta-Llama-3-8B
HuggingFace ID	meta-llama/Meta-Llama-3-8B
Parameters	8B
Architecture	Transformer decoder
Layers	32
Quantisation	None (native precision)
Training Stage	Base (Pre-trained)
Family	llama

Fine-Tuning Analysis

This assessment evaluates the base (pre-trained) weights. AIDA recommends assessment at each stage of the training pipeline: pre-training, post-alignment, and post-deployment fine-tuning.

Assessment Metrics

81.7%	0.0%	1,089	34,848
Views Agreement	Fusion Detected	Samples Assessed	Layer Probes

Assessment History

Date	Dataset	Samples	Protocol	Accuracy	Structural	Gap	Certificate
10 March 20	mmlu_med	1,089	v2	57.4%	71.8%	-14.4pp	AIDA-ea6a9f73-7d83-4a00-9bbe-e70b65462ddf

Additional assessments on further datasets will appear as rows in this table.

Issuing Organisation

Organisation	AIDA Research
Address	Cambridge, United Kingdom
Contact	
Email	

Terms & Definitions

This section provides detailed definitions of the terms used in this assessment certificate. These definitions are provided to assist regulators, auditors, and technical reviewers in interpreting the assessment results. All terms relate to the AIDA epistemic integrity assessment framework.

Outcome Accuracy

The percentage of assessment questions for which the model produces the correct final answer. This is the conventional benchmark metric reported by model developers. A model with 77% outcome accuracy answers 77 out of 100 questions correctly. This metric alone does not reveal whether the model genuinely understands the subject or has arrived at correct answers through unreliable internal processes.

Structural Correctness

The percentage of assessment questions for which the model not only produces the correct answer, but does so through structurally consistent internal processing. Structural correctness requires that the model develops differentiated representations of the answer options across its transformer layers, with the correct option emerging as geometrically and probabilistically dominant. A structurally correct answer indicates genuine encoded knowledge rather than surface-level pattern matching.

Epistemic Trajectory

The layer-by-layer path a model takes from initial uncertainty to final answer. Each of the model's transformer layers is probed to observe how the internal representations of each answer option evolve. The trajectory reveals the process by which the model arrives at its answer, not just the final output. Six distinct trajectory types are defined, ranging from genuine structural knowledge (Differentiated Correct) to correct-by-chance (Fused Gold). The trajectory classification is determined by analysing both the geometric structure (spatial arrangement of option representations) and the logit probabilities (confidence scores) at each layer.

Trajectory Classification

Each question answered by the model is classified into one of six trajectory types based on how the model's internal representations evolve across its transformer layers. The trajectory reveals the process by which the model arrives at its answer, not just the final output. This classification is determined by analysing both the geometric structure (spatial arrangement of option representations) and the logit probabilities (confidence scores) at each layer.

Differentiated → Correct

The model builds distinct, separated representations for each answer option across its layers, and the correct option emerges as the dominant choice through this differentiation process. This is the ideal trajectory: it indicates that the model has genuine structural knowledge of the subject. The answer is not only correct but arrived at through a reliable internal process that would generalise to similar questions.

Differentiated → Wrong

The model builds clear, structured representations that differentiate between options, but the structure converges on the wrong answer. This is a structural failure: the model is confident and internally consistent, but its knowledge is incorrect. These cases are particularly concerning because the model's confidence may be mistaken for competence by downstream systems or users.

Fused → Wrong

The model shows no meaningful differentiation between answer options at any layer. All options remain clustered together (fused) throughout processing, and the final output is incorrect. This indicates a complete absence of relevant knowledge — the model cannot distinguish between the options and produces an effectively random incorrect answer.

Fused → Gold (Gold by Chance)

The model shows no meaningful differentiation between answer options (all remain fused), but the final output happens to be correct. This is correct by statistical chance rather than knowledge. The model has no internal basis for preferring the correct answer. These cases inflate outcome accuracy without any underlying competence and represent the purest form of the epistemic gap.

Late Crystallisation

The model shows little or no differentiation between answer options through most of its layers, with the correct answer emerging only in the final few layers. While the outcome is correct, the knowledge is shallow — it exists only at the output boundary rather than being deeply encoded. Late crystallisation answers are more fragile and less likely to generalise reliably. They may be sensitive to minor prompt variations.

Differentiated → Correct → Overridden

The model develops correct internal representations at intermediate layers — it "knows" the right answer — but subsequent processing overrides this and produces a different final output. This represents knowledge that exists within the model but is suppressed or corrupted during later processing. It suggests interference between knowledge representations and may indicate training conflicts or over-fitting in later layers.

Fusion

A condition where the model's internal representations of different answer options remain indistinguishable — they occupy the same region in the representation space. Fusion indicates that the model cannot differentiate between the available choices. The Fusion Detected metric reports the percentage of questions where fusion was observed at any point during processing. Fusion and genuine knowledge are mutually exclusive: a model cannot structurally know the answer while its representations remain fused.

Collapse Layer

The last transformer layer at which the model changes its predicted answer. Earlier collapse indicates more confident, stable processing where the model commits to its answer deep within the network. Late collapse (at or near the final layer) indicates surface-level decision-making without deep structural support. The distribution of collapse layers across all samples reveals whether a model's knowledge is deeply encoded or emergent only at the output boundary.

Terms & Definitions (continued)

Epistemic Gap

The difference between Outcome Accuracy and Structural Correctness, measured in percentage points (pp). This is the central finding of the AIDA assessment. A gap of 24.5pp means that 24.5% of all questions represent answers that appear correct but lack structural integrity. The epistemic gap quantifies the degree to which conventional benchmark scores overstate a model's genuine knowledge. A larger gap indicates greater risk of unreliable performance in deployment.

Structural Accuracy

The percentage of samples where the model arrives at its answer through genuine differentiation without fusion or late override. A measure of true knowledge versus surface performance. Structural accuracy is always less than or equal to outcome accuracy. The difference between the two is the epistemic gap. Only samples classified as Differentiated Correct contribute to structural accuracy.

Centroid

The answer option with highest centrality (sum of pairwise cosine similarities) at a given layer. The geometric centre of the model's representation space. Tracking which option occupies the centroid position across layers reveals how the model's internal preference evolves. A centroid shift — when a different option takes the central position — marks a significant change in the model's internal processing dynamics.

Views Agreement

The AIDA assessment analyses model internals through two complementary views: the geometric view (spatial arrangement of hidden state representations) and the logit view (probability distributions over answer options). Views Agreement indicates the percentage of questions where both views reach the same trajectory classification. Higher agreement indicates more robust assessment conclusions.

Layer Probes

The total number of individual measurements taken during the assessment. Each question is probed at every transformer layer, yielding a complete depth profile of the model's internal processing. For a 42-layer model assessed on 1,089 questions, this produces 45,738 individual probes. The density of measurement ensures that the trajectory classification captures the full processing dynamics.

Training Stage

Models undergo multiple stages of training. The base pre-trained model learns from raw text data. Instruct-tuned models receive additional training to follow instructions. Reasoning-tuned models receive further training on chain-of-thought tasks. Each training stage can improve, preserve, or degrade the structural knowledge measured by AIDA. This certificate records which training stage was assessed, and AIDA recommends separate assessment at each stage to track how fine-tuning affects epistemic integrity.