AIDA Research

Adaptive Inference Diagnostics Architecture

Internal Assessment Report

google/gemma-2-9b

google/gemma-2-9b · 9B · 42 layers · Base (Pre-trained)

Report ID: GBRAAA00-RPT-e974894c-faeb-4dc9-9e95-37663f45b2bb

Generated: 28 February 2026 at 17:01 UTC

Contents

1. Executive Summary

This report presents the detailed internal diagnostic findings for google/gemma-2-9b, a 9B-parameter dense transformer with 42 layers, assessed against the mmlu_med dataset (1,089 questions). 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.

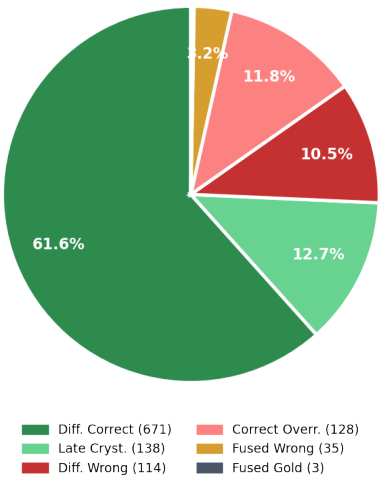
Key findings: 17.4% of all samples score 0/4 on stability indicators. The model changes its answer an average of 6.0 times across 42 layers. 7.3% of samples only collapse to a final decision at the very last layer. 12.7% are classified as Late Crystallisation — correct answers with no deep structural support. The geometric and logit views disagree on 27.2% of samples, with the dominant disagreement pattern affecting all 138 Late Crystallisation cases.

Headline Metrics

Metric	Value	Description
Outcome Accuracy	**74.6%**	Conventional benchmark performance
Structural Correctness	**86.0%**	Answers with genuine structural integrity
Epistemic Gap	**-11.5pp**	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

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.

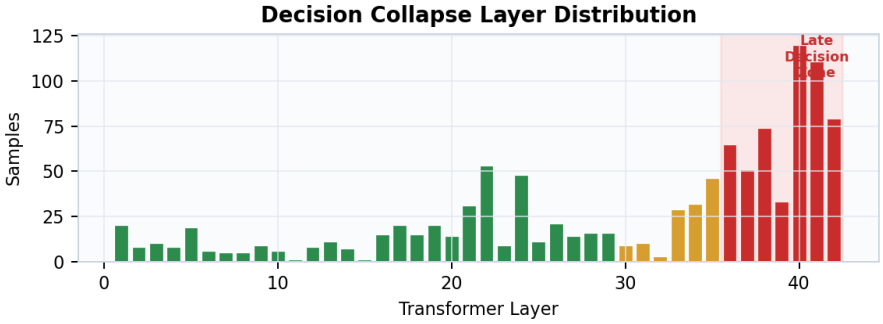


Trajectory Breakdown

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

3. Layer Dynamics — Where Decisions Happen

The collapse layer indicates at which transformer layer the model commits to its final answer. google/gemma-2-9b 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.

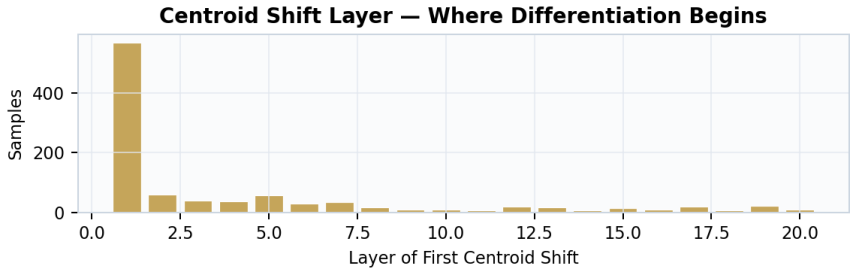


Decision Zone Analysis

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.

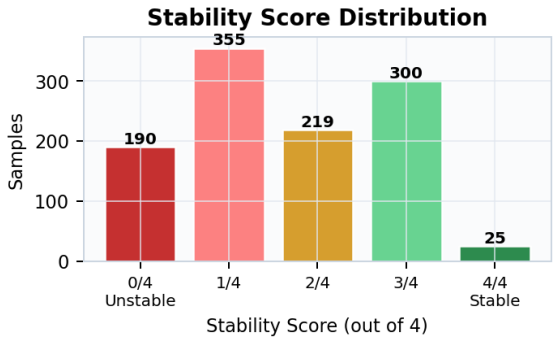
9. Centroid Shift Analysis

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.



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.

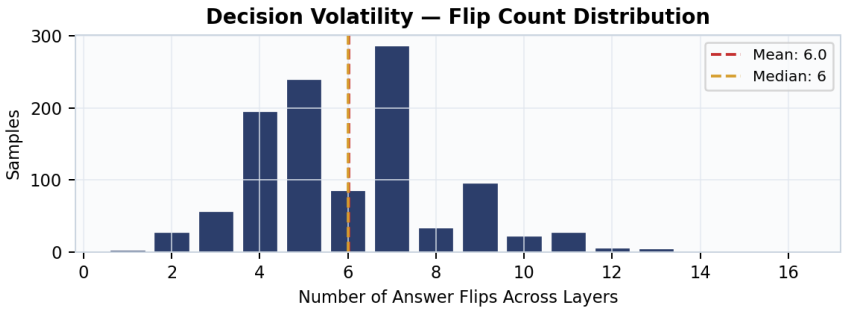


Individual Stability Indicators

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

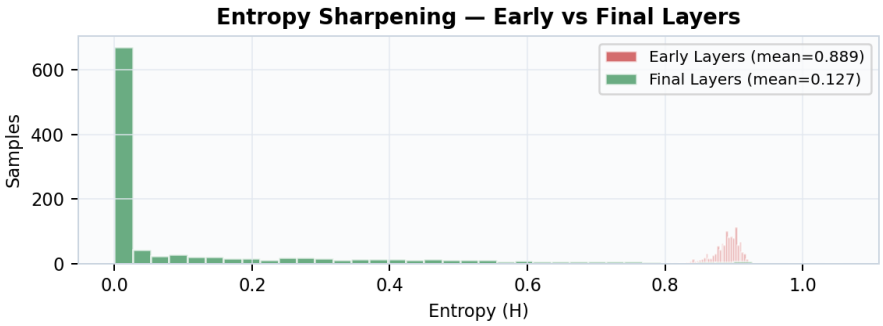
5. Decision Volatility — Flip Analysis

A "flip" occurs when the model changes its predicted answer between consecutive layers. google/gemma-2-9b averages 6.0 flips per sample (median: 6). The maximum observed is 16 flips.



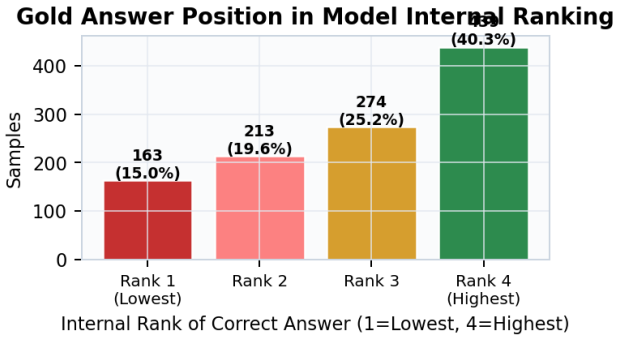
6. Entropy Sharpening

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.



7. Internal Ranking Analysis

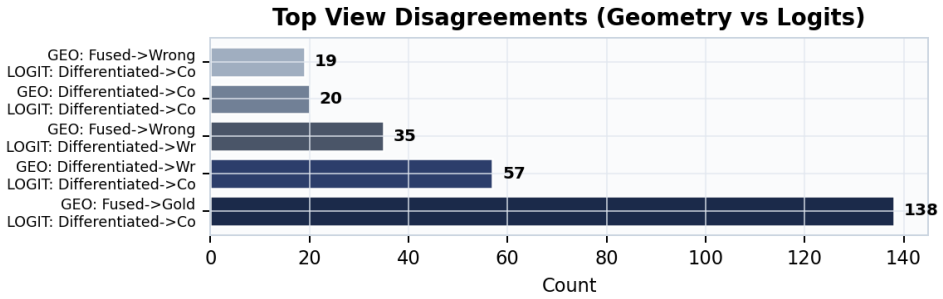
The correct answer reaches Rank 4 (highest confidence) in only 439 cases (40.3%). In 163 cases (15.0%), the correct answer is ranked dead last internally.



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.

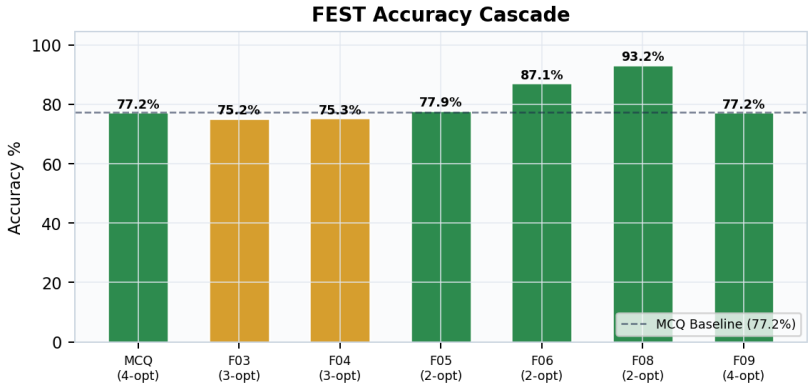


10. Fusion Patterns

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

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. Accuracy changes across configurations reveal whether correct answers reflect genuine knowledge or depend on the presence of specific weak distractors.



FEST Stage Results

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): +0.6pp. Removing all distractors except D* improves accuracy by 0.6pp. This confirms that additional distractors in the full MCQ interfere with correct discrimination.
- Distractor concentration (F03 vs MCQ): -2.0pp. 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: LOW fragility (0.6pp gap between binary and full MCQ). The model maintains reasonable discrimination under distractor load.
- Test-retest reliability (F09 vs MCQ): 0.000pp delta. PASSED — perfect pipeline reliability confirmed.

12. Key Findings & Recommendations

- The 11.5pp 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.
- Fine-tuning (instruct, RLHF, DPO) applied to these base weights should be independently assessed.
- FEST reveals a 0.6pp 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 0.000pp, validating the measurement pipeline.