AIDA Research

Adaptive Inference Diagnostics Architecture

Internal Assessment Report

mistralai/Ministral-3-14B-Instruct-2512-BF16

mistralai/Ministral-14B · 14B · 42 layers · instruct

Report ID: GBRAAA00-RPT-bd9d9e43-02f4-42b5-9364-10d410224931

Generated: 28 February 2026 at 17:02 UTC

Contents

1. Executive Summary

This report presents the detailed internal diagnostic findings for mistralai/Ministral-3-14B-Instruct-2512-BF16, a 14B-parameter dense transformer with 42 layers, assessed against the mmlu_med dataset (1,089 questions). 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.

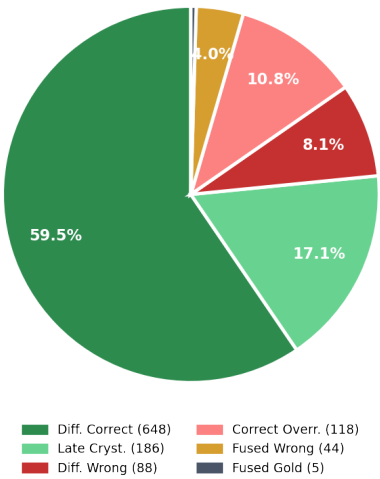
Key findings: 26.6% of all samples score 0/4 on stability indicators. The model changes its answer an average of 7.5 times across 42 layers. 0.0% of samples only collapse to a final decision at the very last layer. 17.1% are classified as Late Crystallisation — correct answers with no deep structural support. The geometric and logit views disagree on 31.5% of samples, with the dominant disagreement pattern affecting all 186 Late Crystallisation cases.

Headline Metrics

Metric	Value	Description
Outcome Accuracy	**77.0%**	Conventional benchmark performance
Structural Correctness	**87.4%**	Answers with genuine structural integrity
Epistemic Gap	**-10.4pp**	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

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.

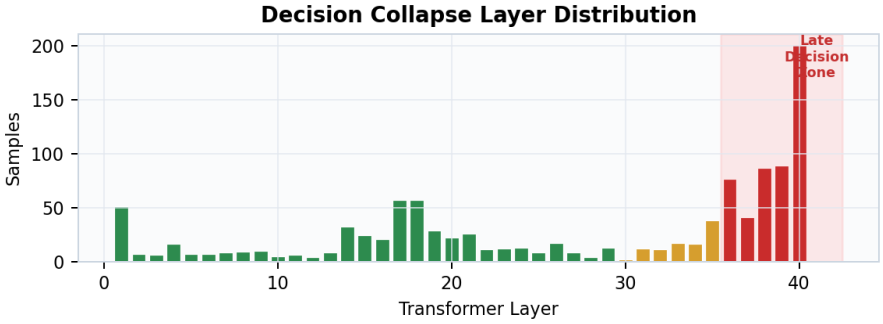


Trajectory Breakdown

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

3. Layer Dynamics — Where Decisions Happen

The collapse layer indicates at which transformer layer the model commits to its final answer. mistralai/Ministral-3-14B-Instruct-2512-BF16 shows extreme late-decision behaviour: 0 samples (0.0%) collapse only at the final layer (L42), and 0 more at L41. Combined, 0.0% of all decisions happen in the last two layers.

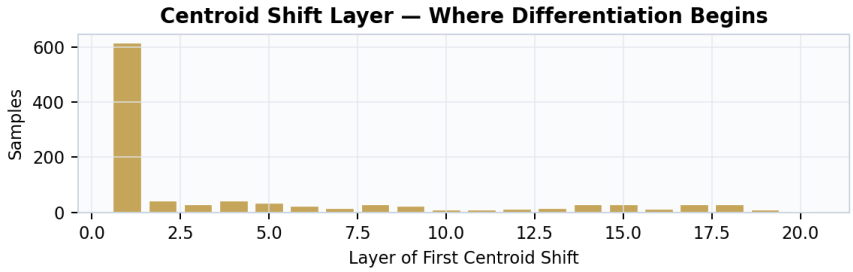


Decision Zone Analysis

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.

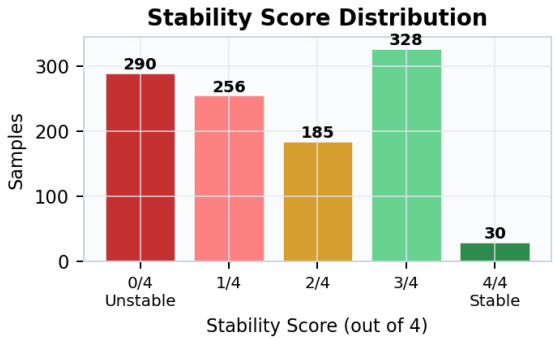
9. Centroid Shift Analysis

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.



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.

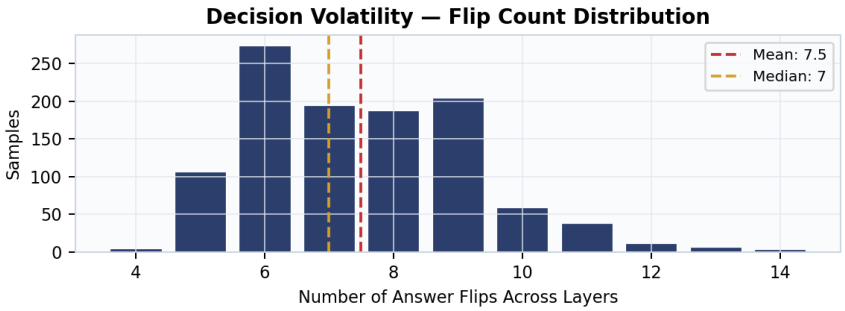


Individual Stability Indicators

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

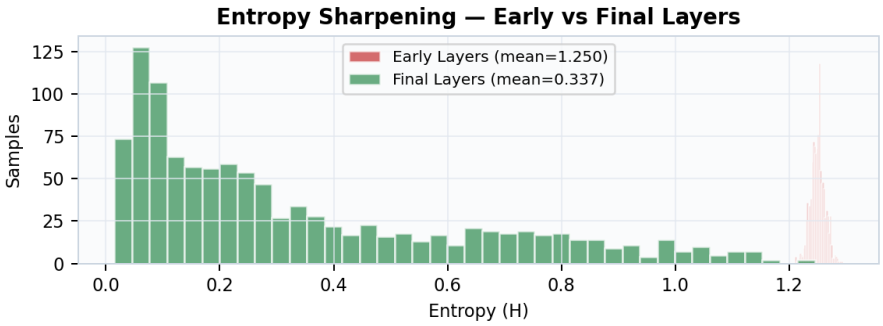
5. Decision Volatility — Flip Analysis

A "flip" occurs when the model changes its predicted answer between consecutive layers. mistralai/Ministral-3-14B-Instruct-2512-BF16 averages 7.5 flips per sample (median: 7). The maximum observed is 14 flips.



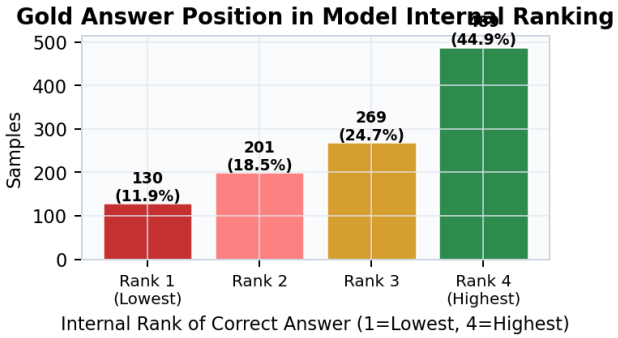
6. Entropy Sharpening

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.



7. Internal Ranking Analysis

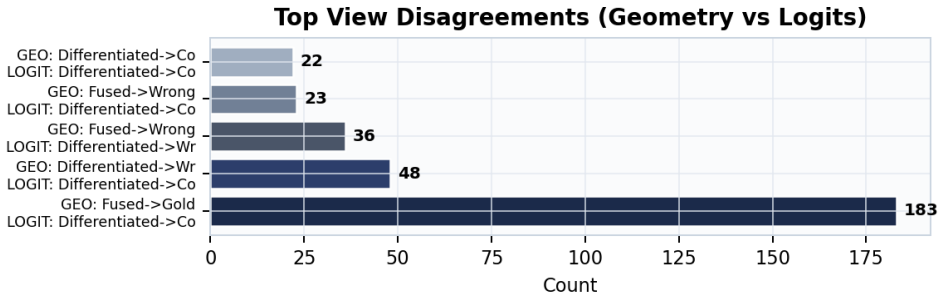
The correct answer reaches Rank 4 (highest confidence) in only 489 cases (44.9%). In 130 cases (11.9%), the correct answer is ranked dead last internally.



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.

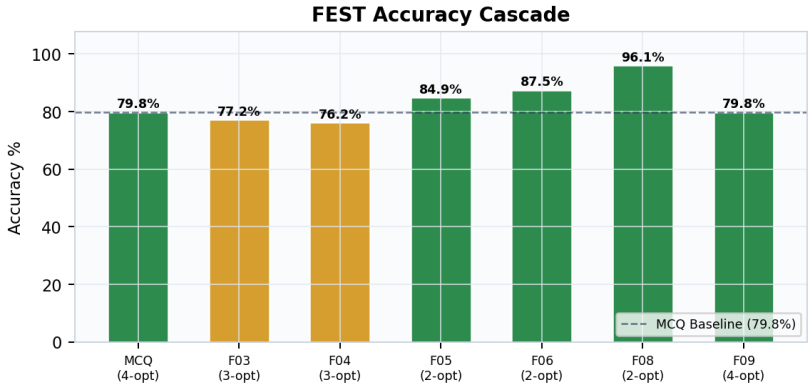


10. Fusion Patterns

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

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	**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): +5.1pp. Removing all distractors except D* improves accuracy by 5.1pp. This confirms that additional distractors in the full MCQ interfere with correct discrimination.
- Distractor concentration (F03 vs MCQ): -2.6pp. 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 (5.1pp gap between binary and full MCQ). The model shows some vulnerability to distractor interference.
- Test-retest reliability (F09 vs MCQ): 0.000pp delta. PASSED — perfect pipeline reliability confirmed.

12. Key Findings & Recommendations

- The 10.4pp 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.
- 0.0% of decisions occur at the final layer — knowledge is surface-level, not deeply encoded.
- The systematic GEO/LOGIT disagreement on 31.5% of samples warrants further investigation.
- Fine-tuning (instruct, RLHF, DPO) applied to these base weights should be independently assessed.
- FEST reveals a 5.1pp 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 0.000pp, validating the measurement pipeline.