AIDA Research

Adaptive Inference Diagnostics Architecture

Comparative Assessment Report

Model Replacement Assessment

Google DeepMind Gemma-9B vs Meta Llama-8B

Current Deployment | Proposed Replacement

Report ID: AIDA-CMP-00bcf719-a09e-40a2-ae1a-a13229b80bb0

Generated: 01 March 2026 at 11:28 UTC

Model A: Google DeepMind Gemma-9B

HuggingFace	google/gemma-2-9b
Parameters	9B
Layers	42
Type	Base (Pre-trained)
Cert. Status	Issued
Certificate	AIDA-3201ec61-17ba-4...
Cert. Expiry	2026-08-29

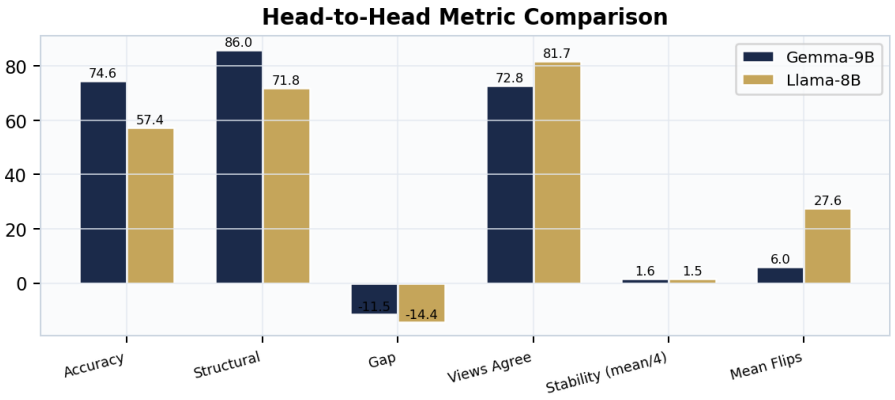
Model B: Meta Llama-8B

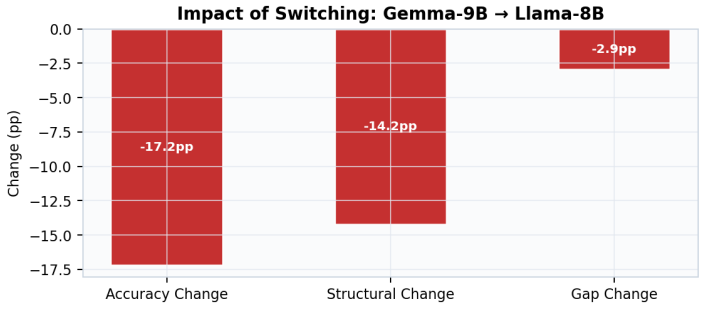
HuggingFace	meta-llama/Meta-Llama-3-8B
Parameters	8B
Layers	32
Type	Base (Pre-trained)
Cert. Status	Issued
Certificate	AIDA-5114562e-28c9-4...
Cert. Expiry	2026-08-29

NOT RECOMMENDED — Proposed model shows degraded epistemic integrity

Key Metrics Comparison

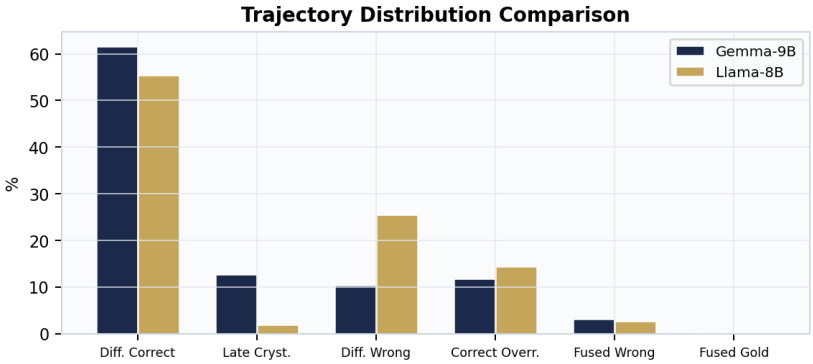
Metric	Gemma-9B	Llama-8B	Delta	
Outcome Accuracy	74.6%	57.4%	-17.2	✗
Structural Correctness	86.0%	71.8%	-14.2	✗
Epistemic Gap	-11.5pp	-14.4pp	-2.9	✗
Views Agreement	72.8%	81.7%	+8.9	✓
Fusion Rate	18.5%	0.0%	-18.5	✓
Mean Stability	1.6/4	1.5/4	-0.1	✗
Mean Flips	6.0	27.6	+21.6	✗
Gold at Top Rank	40.3%	31.0%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





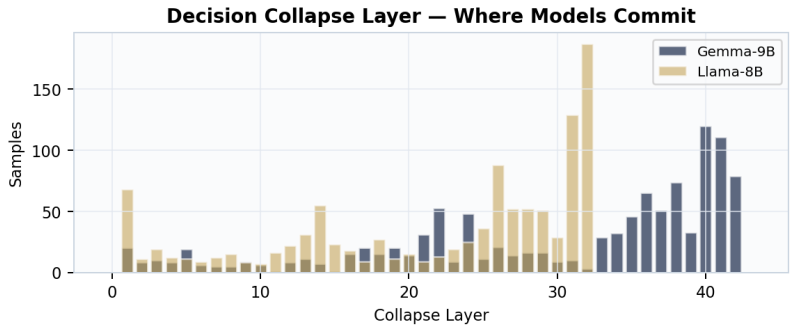
Trajectory Distribution Comparison

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



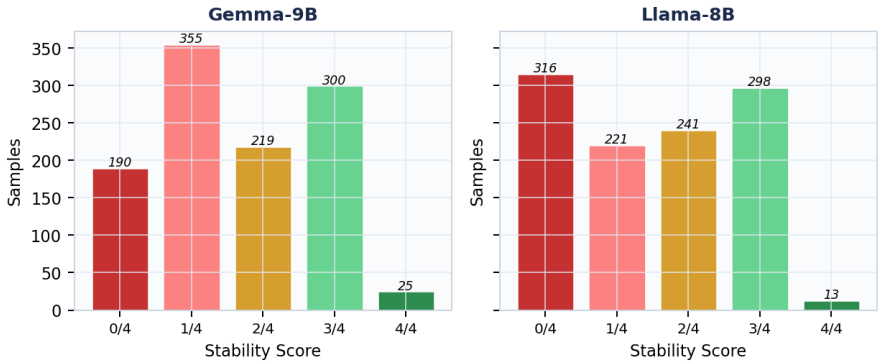
Decision Depth Comparison

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.



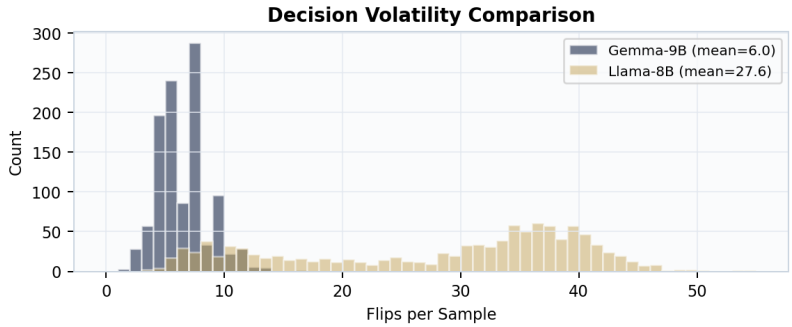
Stability Score Comparison

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.

Summary of Findings

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

Sign-off

Prepared by

AIDA

Reviewed by

_____

(Client AI Governance Lead)

Approved by

_____

(Signature & Date)