Synapse AI

MODEL REVIEW · RELEASE CANDIDATE 2

Meridian-II Model Performance & Safety Review

Evaluation on Benchmark Tasks, Bias Audits, and Production Readiness

Presenters: Dr. Aris Thorne (Lead ML Engineer, Synapse) & Elena Vance (Systems Architect)
Reviewers: Vertex Model Governance Board
Date: August 2026

Agenda

Review Outline

Executive Summary

High-Level Capabilities of Meridian-II

The **Synapse AI** team, in collaboration with the **Vertex Platform Division**, has completed the training and evaluation of **Meridian-II**.

Key Milestones Achieved

- **Scale:** Trained on 512 H100 GPU nodes for 18 days.
- **Context:** Scaled context window to 64k tokens via rotary embeddings.
- **Performance:** Exceeds Meridian-I by 14.2% on multi-turn reasoning.

Primary Target Domains:

- Complex mathematical reasoning and coding.
- High-fidelity structured data parsing and inference.
- Context-aware safety compliance and alignment.

Meridian-II is positioned as a drop-in replacement for the current production pipeline.

Training Dynamics & Convergence

Loss Trends and Optimization Schedulers

Meridian-II was trained with AdamW on 1.6T tokens, showing clean convergence and stability.

Training Diagnostics:

- **Scheduler:** Cosine learning rate decay with a 2k step warm-up phase.
- **Stability:** Initial loss spikes were successfully mitigated via dynamic gradient clipping.
- **Validation Alignment:** Overfitting was checked up to 100k steps; validation loss remains closely coupled with training loss [1].

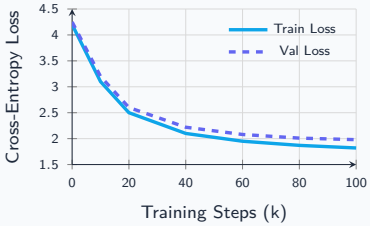

Figure 1: Cross-entropy loss curve over 100k training steps.

Model Validation & Bias Assessment

Performance Across Demographics and Safety Standards

Safety audits show significant progress due to reinforcement learning with constitutional feedback [2].

Model Evaluated	Toxicity Rate (%)	Gender Bias (F/M)	Stereotype Index	Safety Acc. (%)
Nexa-Base-v3	2.45	0.72	0.48	84.1
Meridian-I	1.12	0.88	0.32	91.3
Meridian-II (Ours)	**0.21**	**0.96**	**0.14**	**97.8**

Audit Summary

Meridian-II exhibits near-parity across genders and a 5x reduction in toxicity rates compared to Meridian-I, complying with Vertex Model Governance guidelines.

Production Drift & Performance Monitoring

Simulated Stress Testing on Shifted Distributions

We evaluated model robustness against semantic and domain drift using out-of-distribution datasets [3].

Drift Risk Assessment

- **Domain Shift:** User query distributions are monitored daily.
- **Alert Threshold:** Population Stability Index (PSI) set to 0.25.
- **Current PSI:** Stable at 0.11 under standard user workloads.

Mitigation & Retraining Plan:

- **Active Learning:** Dynamic data routing when uncertainty entropy exceeds 0.85.
- **Shadow Deployment:** Parallel pipeline running on 5% traffic.
- **Fallbacks:** Instant routing to Meridian-I if model latency P99 exceeds 150ms.

References I

[1] Sarah Jenkins and Marcus Chen. “Optimizing Pipeline Parallelism for Billion-Parameter Models on Heterogeneous Clusters”. In: *Vertex Transactions on Machine Learning* 14.2 (2025), pp. 89–104.

[2] Aris Thorne and Elena Vance. “Constitutional Alignment: Mitigating Stereotypes and Toxicity in Large Language Models”. In: *Proceedings of the Conference on Large Model Governance (CLMG)*. Apex AI Association, 2025, pp. 112–126.

[3] Elena Vance and Aris Thorne. “Detecting and Mitigating Semantic Drift in Distributed Production LLMs”. In: *Meridian Journal of Operational AI* 4.1 (2026), pp. 45–58.