CORTEX RESEARCH

Vera 1.6 Technical Report

A Hybrid Multimodal Language Model with Gated Delta Networks and Sparse Mixture-of-Experts for Agentic Applications

Cortex Research

Technical Report | March 2, 2026

Abstract

We introduce **Vera 1.6**, a 35B-parameter (3B activated) unified multimodal language model developed by Cortex Research and purpose-built for agentic applications. Vera 1.6 employs a novel hybrid architecture combining **Gated Delta Networks** with **Sparse Mixture-of-Experts** (SMoE), enabling highly efficient inference at scale. Trained on a proprietary 150B-token synthetic dataset constructed with the NVIDIA NeMo Data Designer framework and further aligned via Reinforcement Learning, the model achieves strong performance across instruction following, graduate-level reasoning, multilingual understanding, agentic tool use, and multimodal comprehension. With a 1M-token context window, native vision and video processing, and support for 201 + languages, Vera 1.6 is designed as the backbone for production-grade agentic AI systems.

Keywords: large language model, mixture-of-experts, delta networks, multimodal, agentic AI, reinforcement learning, hybrid architecture, vision-language model

1 Introduction

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities across reasoning, code generation, and multimodal understanding. However, deploying frontier models in *production agentic systems*—where models must autonomously execute multi-step tasks, interact with external tools, navigate web interfaces, and maintain coherence over extremely long contexts—presents unique architectural and training challenges that general-purpose models are not optimally designed to address.

Vera 1.6 is Cortex Research's second-generation multimodal model in the Vera series, architected specifically for agentic workloads. Building upon the **Qwen3.5-35B-A3B-Base** foundation [1], Vera 1.6 introduces a redesigned hybrid attention mechanism combining Gated Delta Networks for efficient linear attention with periodic Gated Attention layers, paired with a Sparse Mixture-of-Experts (SMoE) feed-forward structure. This design activates only 3B of 35B total parameters per forward pass, enabling favourable compute-performance trade-offs at deployment scale.

Training leverages NVIDIA DGX B200 infrastructure with 8× Blackwell B200 GPUs providing 1,440 GB of total GPU memory. The training corpus consists of 150B tokens of proprietary synthetic data generated through the NVIDIA NeMo Data Designer framework and curated specifically for agentic reasoning, tool use, code execution, and multimodal instruction following. Post-SFT, Reinforcement Learning (RL) aligns model behaviour with agentic task objectives, including long-horizon planning, self-correction, and reliable tool invocation.

2 Model Architecture

Vera 1.6 is a Causal Language Model augmented with a Vision Encoder, forming a unified Vision-Language

backbone. The core language model adopts a hierarchical hybrid design comprising 40 layers organised into 10 macro-blocks. Each macro-block contains four sub-blocks that interleave efficient linear and full quadratic attention with sparse feed-forward routing.

2.1 Hidden Layer Layout

The layer layout follows a fixed pattern per macro-block:

Sub-block	Repetitions
Gated DeltaNet → MoE	3×
Gated Attention → MoE	1×

This 3:1 ratio of linear-to-quadratic attention reduces per-token FLOPs for the majority of layers, while periodic full attention layers maintain global context integration across the 1M-token context window.

2.2 Gated Delta Networks

The primary attention mechanism is the **Gated Delta Network** (GDN) [2], a form of linear recurrent attention achieving $\mathcal{O}(L)$ inference complexity with respect to sequence length L. This is critical at 1M-token context lengths where quadratic complexity would be prohibitively expensive.

Parameter	Value
Linear Attn. Heads (V)	32
Linear Attn. Heads (QK)	16
Head Dimension	128

The asymmetric head count between values (32) and queries/keys (16) increases representational capacity in the value projection while reducing the overhead of the recurrent state update computation.

2.3 Gated Attention

Full quadratic attention layers appear once per macro-block and employ **Grouped-Query Attention** (GQA) [3] with a 16:2 query-to-KV head ratio, reducing KV-cache memory by 8× relative to multi-head attention while preserving model quality. Rotary Position Embeddings (RoPE) [4] are applied with a compressed dimension of 64, optimised for long-range dependency modelling.

Parameter	Value
Query Heads	16
Key / Value Heads	2 (GQA)
Head Dimension	256
RoPE Dimension	64

2.4 Sparse Mixture-of-Experts

Every attention layer is followed by a Sparse Mixture-of-Experts (SMoE) [5] feed-forward block. With 256 total experts and only 9 activated per token (8 routed + 1 shared), the model achieves a high capacity multiplier:

$$\text{Capacity Multiplier} = \frac{N_{\text{experts}}}{N_{\text{activated}}} \approx 28\times$$

The *shared expert* is always active and provides a general-purpose pathway, while the learned router selects 8 task-specialised experts dynamically per token, enabling fine-grained input-conditioned computation.

Parameter	Value
Total Experts	256
Activated / Token	8 Routed + 1 Shared
Expert Interm. Dim.	512
Routing Strategy	Top-K + shared expert

2.5 Vision Encoder

Vera 1.6 incorporates a Vision Encoder that projects image and video feature maps into the language model embedding space, forming a unified Vision-Language model without separate specialist modules. The encoder supports still images, multi-page documents, and video sequences, with architectural compatibility inherited from the Qwen3.5 visual stack.

3 Model Specifications

Table 1 summarises the complete architectural and deployment specifications of Vera 1.6.

Table 1. Complete Vera 1.6 Model Specifications

Property	Value
Foundation	Qwen3.5-35B-A3B-Base
Model Type	Causal LM + Vision Encoder
Total Params	35 Billion
Active Params	∼3 Billion / forward pass
Architecture	Hybrid: Gated DeltaNet + SMoE
Hidden Dim.	2,048
Num. Layers	40
Token Vocab	248,320 (padded)
Context Window	1,000,000 tokens
Languages	201+
Modalities	Text, Image, Video
Training	Causal LM + RL (agentic)
Dataset	150B token synthetic corpus
Data Tool	NVIDIA NeMo Data Designer
Hardware	DGX B200 (8× B200, 1,440 GB)

4 Training Methodology

4.1 Pre-training Foundation

Vera 1.6 initialises from the **Qwen3.5-35B-A3B-Base** checkpoint, a state-of-the-art MoE base model providing strong multilingual understanding, mathematical reasoning, and code comprehension as a foundation for task-specific alignment stages.

4.2 Agentic Supervised Fine-tuning

The primary training signal is a proprietary **150B-token synthetic dataset** constructed for agentic applications using the **NVIDIA NeMo Data Designer** framework [6]. Coverage includes:

- Multi-step agentic task planning and execution
- API and tool calling with complex function schemas
- Agentic web browsing and information retrieval
- Terminal and shell command execution sequences
- Long-context document processing and synthesis
- Multimodal instruction following with interleaved vision inputs
- Cross-lingual generalisation across 201+ languages

4.3 Reinforcement Learning Alignment

Following SFT, Vera 1.6 undergoes a Reinforcement Learning stage to optimise behaviour for agentic task completion. Reward signals are derived from task success metrics across: code execution outcomes, tool-call accuracy, instruction adherence, and multimodal comprehension quality. This stage improves self-correction, long-horizon planning, and reliable external tool invocation within agentic pipelines.

4.4 Training Infrastructure

All training was conducted on a single **NVIDIA DGX B200** node [7].

Component	Specification
GPU	8× NVIDIA B200 (Blackwell)
GPU Memory	1,440 GB total
CPU	2× Intel Xeon Platinum 8570
System Memory	2 TB DDR5
Storage	8× 3.84 TB U.2 NVMe
OS	DGX OS
Interconnect	NVLink (GPU-to-GPU)

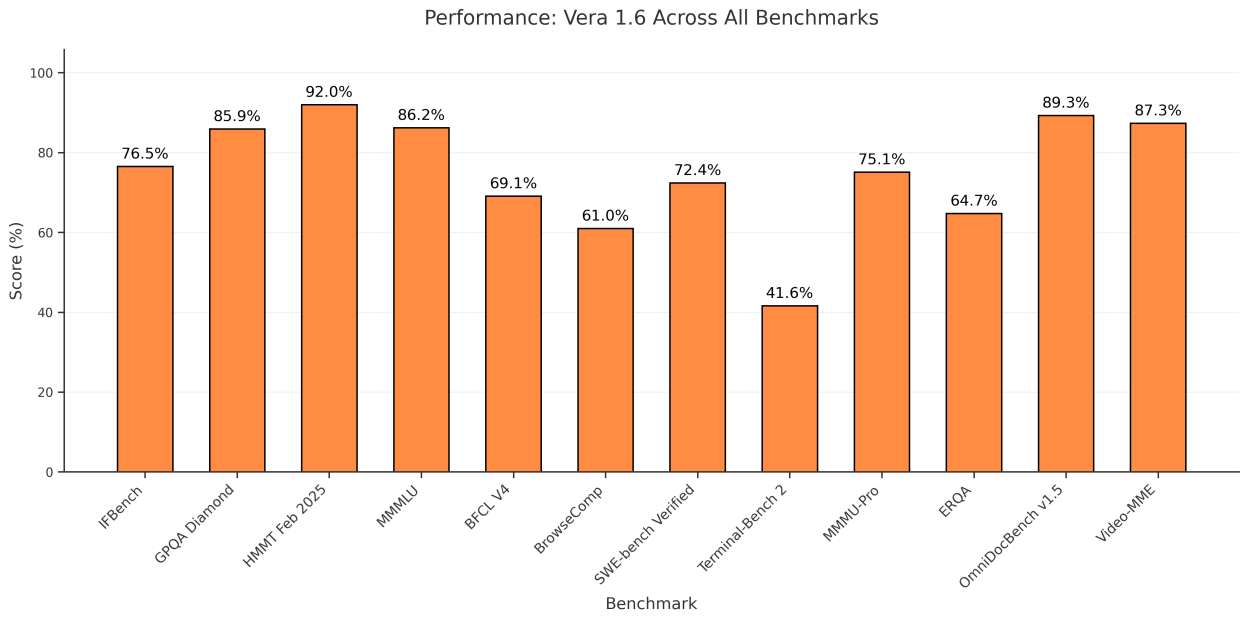
5 Evaluation

5.1 Benchmark Suite

Vera 1.6 was evaluated across twelve diverse benchmarks spanning instruction following, scientific reasoning, mathematics, multilingual understanding, agentic tool use, and multimodal comprehension. All benchmarks use standardised evaluation protocols.

5.2 Results

Table 2 and Figure 1 present the full results. Vera 1.6 achieves particularly strong performance in mathematical reasoning (HMMT Feb 2025: 92.0%), document understanding (OmniDocBench v1.5: 89.3%), and video comprehension (Video-MME: 87.3%). A MMMLU score of 86.2% confirms broad multilingual capability across the 201+ supported languages. Graduate-level scientific reasoning at 85.9% (GPQA Diamond) reflects the effectiveness of the RL alignment stage.



Benchmark Performance Chart

Figure 1. Vera 1.6 performance across all twelve evaluation benchmarks. Scores reported as percentage (%).

Table 2. Vera 1.6 Benchmark Results

Benchmark	Category	Score
HMMT Feb 2025	Mathematics	92.0%
OmniDocBench v1.5	Doc. Understanding	89.3%
Video-MME	Video Reasoning	87.3%
MMMLU	Multilingual	86.2%
GPQA Diamond	Graduate Science	85.9%
IFBench	Instruction Following	76.5%
MMMU-Pro	Visual Reasoning	75.1%
SWE-bench Verified	Agentic Coding	72.4%
BFCL V4	Tool Use	69.1%
ERQA	Embodied Reasoning	64.7%
BrowseComp	Agentic Search	61.0%
Terminal-Bench 2	Terminal Coding	41.6%

Agentic benchmarks reveal expected task-dependent variance. SWE-bench Verified (72.4%) and BFCL V4 (69.1%) demonstrate strong capability in code-based and tool-use agentic tasks. Terminal-Bench 2 (41.6%) and BrowseComp (61.0%) highlight active development areas in low-level terminal execution and autonomous web navigation—both domains are prioritised in subsequent Vera releases.

6 Discussion

6.1 Architectural Trade-offs

The hybrid Gated DeltaNet + SMoE design represents a deliberate trade-off optimised for production agentic deployment. The 3:1 linear-to-full-attention ratio substantially reduces per-token FLOPs for long-context inference. GQA (16Q/2KV) in quadratic attention layers reduces KV-cache memory by 8× compared to multi-head attention, enabling single-node deployment practical for enterprise customers.

The 256-expert SMoE with 9 active experts per token achieves a ∼28× capacity multiplier over a dense model of equivalent activated parameter count. The shared expert mechanism ensures consistent base-level capabilities across all inputs, while 8 routed experts provide specialisation for distinct task types encountered in agentic workloads.

6.2 Limitations

Terminal-Bench 2 performance (41.6%) indicates that low-level shell execution remains challenging, attributable to limited terminal-execution trajectories in training data and the compounding error sensitivity of sequential shell commands. BrowseComp (61.0%) similarly reflects the difficulty of long-horizon autonomous web navigation. Both areas are being addressed through targeted data generation and RL reward shaping.

6.3 Future Work

Planned improvements include: expanded terminal and shell execution training data; improved tool-use generalisation across novel API schemas; enhanced video understanding for extended clips; extended multimodal modalities; and further RL alignment for complex multi-agent coordination and collaborative agentic scenarios.

7 Conclusion

We have presented **Vera 1.6**, a 35B-parameter multimodal language model with Gated Delta Networks and Sparse Mixture-of-Experts, purpose-built for agentic applications. The model achieves strong performance across mathematical reasoning, document understanding, multilingual knowledge, and graduate-level scientific Q&A, while activating only $\sim$3B parameters per forward pass. With a 1M-token context window and 201+ language support, Vera 1.6 is designed to serve as a capable and efficient backbone for production-grade agentic AI systems. Vera 1.6 represents Cortex Research's ongoing commitment to building models that balance frontier performance with practical deployment efficiency.

References

[1] Qwen Team (2025). *Qwen3 Technical Report*. Alibaba Group.

[2] Y. Sun *et al.* (2024). *Gated Linear Attention Transformers with Hardware-Efficient Training*. *arXiv:2312.06635*.

[3] J. Ainslie *et al.* (2023). *GQA: Training Generalised Multi-Query Transformer Models from Multi-Head Checkpoints*. *arXiv:2305.13245*.

[4] J. Su *et al.* (2022). *RoFormer: Enhanced Transformer with Rotary Position Embedding*. *Neurocomputing* 568.

[5] W. Fedus *et al.* (2022). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. *JMLR* 23(120):1–39.

[6] NVIDIA (2024). *NeMo Data Designer: Scalable Synthetic Data Generation for LLMs*. NVIDIA Technical Blog.

[7] NVIDIA (2024). *DGX B200 System Architecture White Paper*. NVIDIA Corporation.