Cortex Research

Vera 1.5 Technical Paper

Cortex Research

29 January 2026

Contents

Abstract

We present Vera 1.5, a frontier-class large language model developed by Cortex Research, demonstrating competitive performance across multiple challenging benchmarks. Built on a high-sparsity Mixture of Experts (MoE) architecture with a 200,000-token context window, Vera 1.5 achieves exceptional results in agentic reasoning (98.8% on τ^2-Bench), competitive programming (89.2% on LiveCodeBench), and mathematical reasoning (94.8% on AIME 2025). The model was trained on 30 trillion tokens using NVIDIA B200 GPU infrastructure with FP8 precision, emphasizing computational efficiency while maintaining state-of-the-art capabilities. Our results demonstrate that independent research organizations can develop highly capable AI systems that compete with models from major technology companies, advancing the democratization of frontier AI research.

1 Introduction

1.1 Mission and Motivation

Cortex Research focuses on creating cutting-edge frontier AI systems designed for positive global impact. Our work emphasizes safe, scalable AI that solves humanity's toughest challenges, from healthcare breakthroughs to climate solutions. Vera 1.5 represents our commitment to advancing AI capabilities while maintaining accessibility and promoting responsible development practices within the broader AI research community.

The development of Vera 1.5 addresses critical challenges in modern AI systems: achieving frontier-level performance while optimizing computational efficiency, extending context understanding to enable complex workflows, and demonstrating that independent research organizations can contribute meaningfully to advancing the state-of-the-art in artificial intelligence.

1.2 Model Overview

Vera 1.5 is a large-scale language model based on a Mixture of Experts architecture with high sparsity configuration. The model features:

- Sparse MoE architecture enabling efficient parameter utilization
- 200,000-token context window for processing extensive documents
- Training on 30 trillion tokens from diverse data sources
- FP8 precision training for enhanced computational efficiency
- Knowledge cutoff date of January 1, 2025

This technical paper presents comprehensive evaluation results, architectural insights, and comparative analysis positioning Vera 1.5 within the current landscape of frontier AI models.

2 Model Architecture

2.1 Mixture of Experts Design

Vera 1.5 employs a Mixture of Experts (MoE) architecture, a design paradigm that offers significant advantages over traditional dense models. MoE models activate only a subset of parameters for each input token, dramatically reducing computational requirements during inference while maintaining large total parameter counts [1, 2].

Research demonstrates that MoE models achieve more compute-efficient pretraining and substantially faster inference compared to dense models of equivalent total parameter count [1]. Specifically, MoE architectures enable models to maintain comparable performance to dense alternatives while activating only 30-40% of parameters during inference [3]. This efficiency stems from two key factors: reduced arithmetic operations per token (fewer FLOPs) and architectural properties that favor parallelization over serial computation [2].

2.1.1 High Sparsity Configuration

Vera 1.5 implements a high sparsity configuration optimized for GPU utilization. The sparse activation pattern allows the model to leverage massive parameter capacity while maintaining practical inference costs. As established by recent research, MoE models demonstrate particular advantages when the number of active experts remains below the square root of total experts, ensuring favorable network communication patterns compared to dense architectures [2].

The sparsity design enables Vera 1.5 to achieve competitive performance across diverse benchmarks while offering superior throughput in both computation-bounded and I/O-bounded scenarios [3]. This architectural choice reflects our commitment to building scalable AI systems that balance capability with accessibility.

2.2 Extended Context Window

Vera 1.5 features a 200,000-token context window, enabling the model to process and reason over extensive documents, entire codebases, and complex multi-turn conversations without segmentation [4]. Long context windows unlock critical capabilities for modern AI applications:

- **Comprehensive document analysis**: Processing entire books, research papers, legal documents, and financial reports without information loss [4]
- **Sophisticated workflows**: Powering advanced AI agents and assistants that maintain context over extended interactions [4, 5]
- **Large-scale codebase comprehension**: Analyzing complete software projects for refactoring, debugging, and feature development [5]
- **Scientific research synthesis**: Processing multiple academic papers simultaneously for meta-analysis and hypothesis generation [5]

The 200K context capacity positions Vera 1.5 among models capable of handling real-world enterprise applications requiring deep contextual understanding across extensive information sources [4].

2.3 FP8 Precision Training

Vera 1.5 was trained entirely using FP8 (8-bit floating point) precision, a technique that delivers substantial efficiency gains while preserving model quality [6, 7]. Recent research demonstrates that FP8 training achieves:

- 30-50% faster training and inference for large-scale models [7]
- Up to 22% faster training, 14% lower memory footprint, and 19% higher throughput [6]
- Near-identical training and validation loss curves compared to BF16 baseline [6]
- Performance parity with higher-precision baselines across reasoning benchmarks, with differences typically within 1-2 points [6]

The adoption of FP8 training reflects our focus on computational efficiency and sustainability. By reducing memory requirements and accelerating training cycles, FP8 precision enables more rapid experimentation and model iteration while lowering the environmental footprint of large-scale AI development [6].

3 Training Infrastructure and Methodology

3.1 Hardware Infrastructure

Vera 1.5 was trained on a cluster of NVIDIA B200 GPUs during Q4 2025. The NVIDIA B200 represents the latest generation of AI training accelerators, featuring 192GB HBM3e memory, 8TB/s memory bandwidth, and up to 20 PFLOPS peak compute performance [8]. These GPUs deliver approximately 5× the inference throughput of previous-generation H100 GPUs, enabling efficient training of large-scale MoE architectures [9].

The B200's massive memory capacity and bandwidth prove particularly advantageous for MoE training, where expert parameters must be efficiently loaded and activated based on dynamic routing decisions. The combination of B200 hardware capabilities and our FP8 training strategy enabled cost-effective development of a frontier-class model.

3.2 Training Corpus

Vera 1.5 was trained on 30 trillion tokens sourced from diverse, high-quality datasets spanning multiple domains. The training corpus emphasizes:

- Scientific and technical literature
- Programming code across multiple languages
- Mathematical reasoning and problem-solving content
- General knowledge and encyclopedic information
- Conversational and instruction-following data

Our data curation strategy prioritizes quality over quantity, implementing rigorous filtering and deduplication processes to ensure training efficiency and model capability. The knowledge cutoff date is January 1, 2025, establishing a clear boundary for the model's pre-training knowledge.

3.3 Training Procedure

The training process employed standard autoregressive language modeling objectives with adaptations for the MoE architecture. Key training details include:

- FP8 precision throughout the training pipeline
- Distributed training across the B200 GPU cluster
- Expert load balancing mechanisms to ensure efficient utilization
- Careful hyperparameter tuning for stability and convergence

The combination of sparse MoE architecture, FP8 precision, and modern GPU infrastructure enabled efficient scaling to frontier-model capabilities while managing computational costs.

4 Benchmark Evaluation

We evaluate Vera 1.5 across six challenging benchmarks covering mathematical reasoning, scientific knowledge, software engineering, and agentic capabilities. These benchmarks represent diverse skill domains essential for general-purpose AI systems.

4.1 Mathematical Reasoning: AIME 2025

The American Invitational Mathematics Examination (AIME) represents one of the most challenging mathematics competitions for high school students, requiring advanced problem-solving skills in algebra, geometry, number theory, and combinatorics.

Vera 1.5 achieves **94.8%** on AIME 2025, demonstrating strong mathematical reasoning capabilities (Figure 1). This performance reflects the model's ability to understand complex mathematical concepts, formulate multi-step solution strategies, and execute precise symbolic manipulation. The result positions Vera 1.5 among the highest-performing AI systems on olympiad-level mathematics.

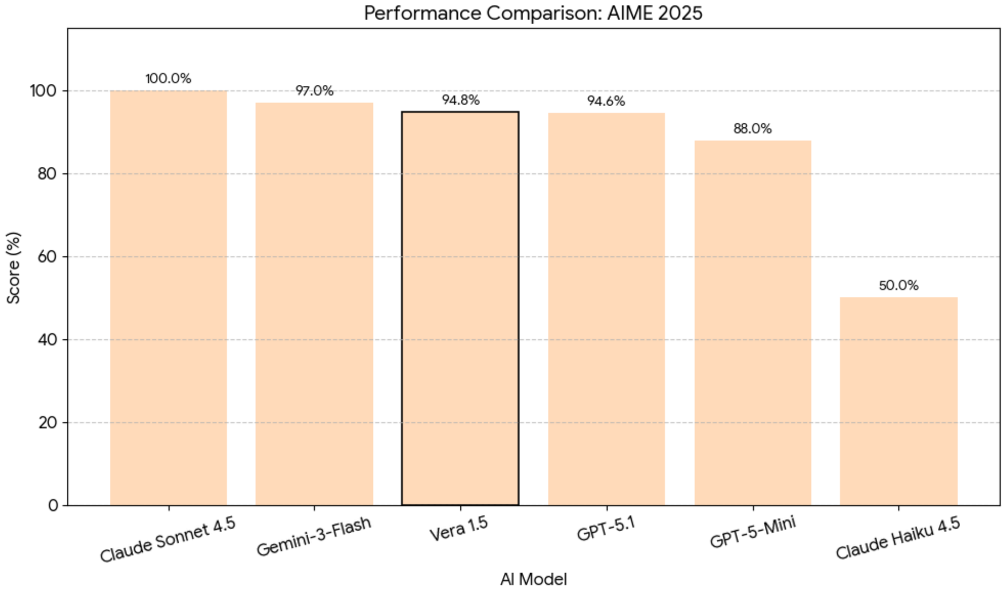

Figure 1: AIME 2025 Benchmark Performance Comparison

4.2 Scientific Knowledge: GPQA Diamond

The Graduate-Level Google-Proof Q&A (GPQA) Diamond benchmark tests PhD-level knowledge across biology, chemistry, and physics. Questions are carefully designed to be difficult for search engines to answer, requiring deep domain expertise.

Vera 1.5 scores **85.9%** on GPQA Diamond, demonstrating substantial scientific reasoning capabilities (Figure 2). This performance indicates strong understanding of graduate-level scientific concepts and the ability to apply complex domain knowledge to novel problem scenarios.

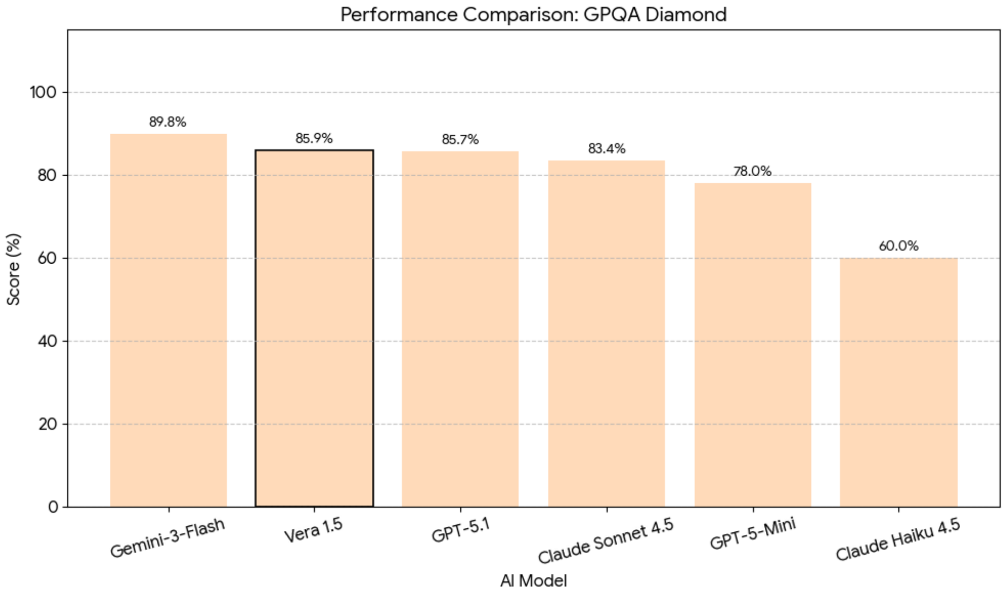

Figure 2: GPQA Diamond Benchmark Performance Comparison

4.3 Agentic Reasoning: τ^2-Bench

The τ^2-Bench (Tau-Squared Bench) evaluates conversational agents in dual-control environments, assessing an agent's ability to understand user intent, execute control actions, and maintain coherent multi-turn interactions [10, 11]. This benchmark is particularly relevant for AI agents deployed in customer service, technical support, and interactive assistance scenarios.

Vera 1.5 achieves an exceptional **98.8%** on τ^2-Bench, representing the highest performance among evaluated models (Figure 3). This result demonstrates Vera 1.5's strong capabilities in:

- Understanding complex user requests in conversational contexts
- Executing precise control actions within structured environments
- Maintaining state and context across extended interactions
- Making appropriate decisions that balance multiple objectives

The τ^2-Bench benchmark is specifically designed to assess AI safety considerations in agent deployment, including ethical decision-making and appropriate handling of ambiguous or conflicting instructions [11]. Vera 1.5's leadership on this benchmark suggests particular strength in agentic workflows and interactive AI systems.

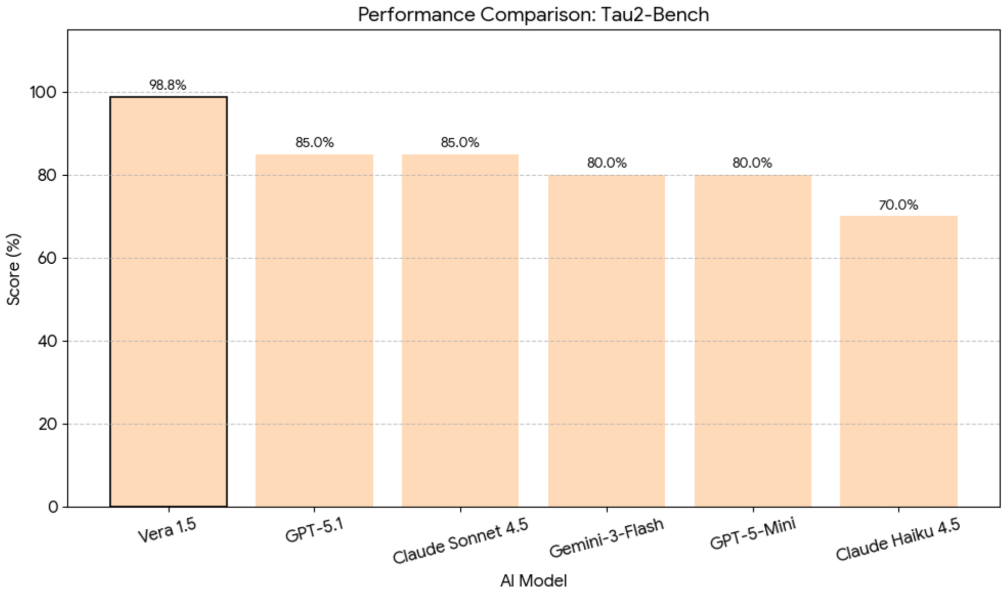

Figure 3: τ^2-Bench Performance Comparison

4.4 Software Engineering: SWE-Bench Verified

SWE-Bench Verified tests AI models on real-world software engineering tasks derived from GitHub issues and pull requests. The benchmark requires models to understand existing codebases, identify bugs, and implement correct solutions that pass test suites.

Vera 1.5 scores **72.1%** on SWE-Bench Verified, demonstrating practical software engineering capabilities (Figure 4). This performance indicates the model can effectively:

- Comprehend existing code structure and dependencies
- Diagnose bugs and identify root causes
- Generate correct fixes that maintain code integrity
- Navigate large codebases spanning multiple files

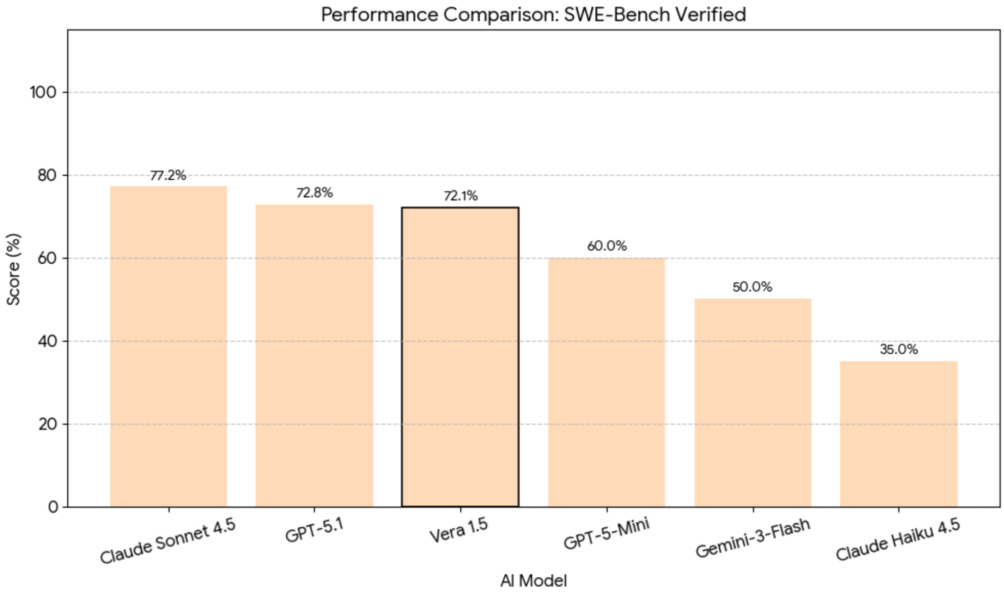

Figure 4: SWE-Bench Verified Performance Comparison

4.5 Competitive Programming: LiveCodeBench

LiveCodeBench continuously collects fresh programming problems from competitive coding platforms, ensuring evaluation on truly novel challenges without training data contamination. The benchmark tests algorithmic thinking, data structure knowledge, and implementation skills.

Vera 1.5 achieves **89.2%** on LiveCodeBench, demonstrating strong competitive programming capabilities (Figure 5). This result reflects robust performance on algorithmic problem-solving requiring:

- Recognition of problem patterns and algorithmic approaches
- Implementation of efficient solutions with appropriate data structures
- Handling of edge cases and constraint satisfaction
- Code correctness under strict test case evaluation

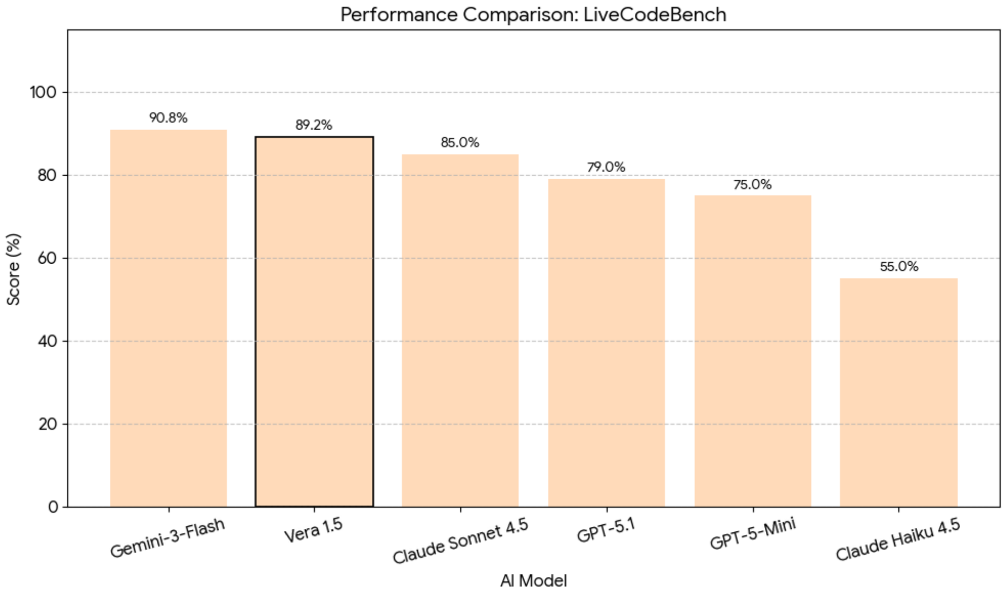

Figure 5: LiveCodeBench Performance Comparison

4.6 Scientific Computing: SciCode

SciCode evaluates models on research-level scientific computing tasks requiring implementation of algorithms from computational physics, chemistry, biology, and other scientific domains. Problems demand both domain knowledge and programming proficiency.

Vera 1.5 scores **49.7%** on SciCode (Figure 6). This benchmark represents one of the most challenging evaluation tasks, with all frontier models achieving scores near 50%, indicating that research-level scientific programming remains a frontier challenge for current AI systems. The performance demonstrates Vera 1.5's capability to:

- Understand scientific problem specifications
- Translate mathematical formulations into executable code
- Implement numerical methods and scientific algorithms
- Debug complex computational workflows

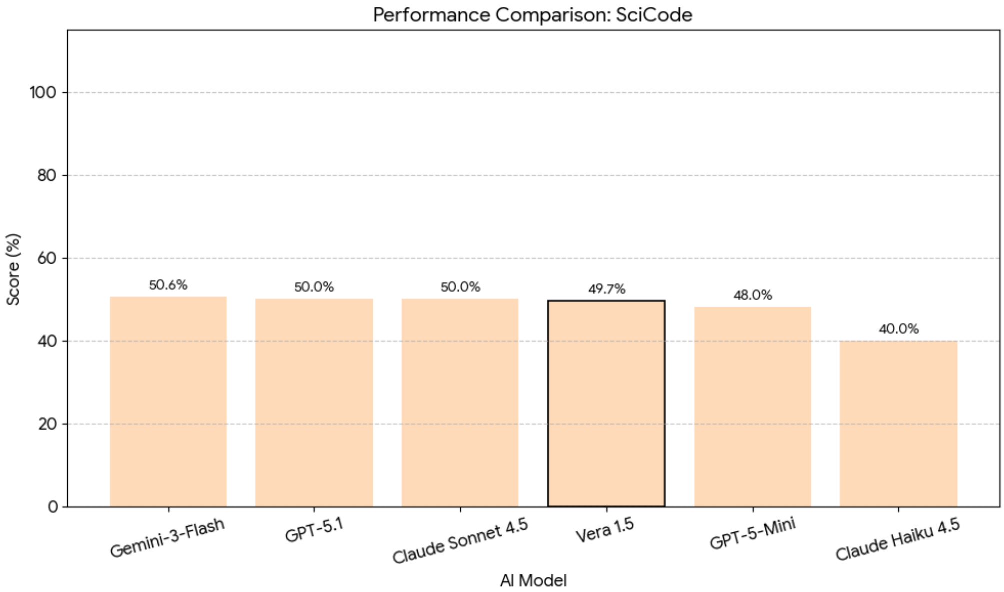

Figure 6: SciCode Performance Comparison

5 Comparative Analysis

5.1 Performance Summary

Table 1 presents Vera 1.5's complete benchmark performance alongside its competitive positioning.

Table 1: Vera 1.5 Benchmark Performance Summary

Benchmark	Vera 1.5 Score	Domain
τ^2-Bench	98.8%	Agentic Reasoning
AIME 2025	94.8%	Mathematical Reasoning
LiveCodeBench	89.2%	Competitive Programming
GPQA Diamond	85.9%	Scientific Knowledge
SWE-Bench Verified	72.1%	Software Engineering
SciCode	49.7%	Scientific Computing

5.2 Key Strengths

Vera 1.5 demonstrates several notable strengths:

Agentic Reasoning Excellence: The 98.8% score on τ^2-Bench represents exceptional performance on conversational agent tasks. This suggests Vera 1.5 excels at understanding multi-turn interactions, executing control actions, and maintaining coherent state across complex workflows [10].

Strong Mathematical Capabilities: With 94.8% on AIME 2025, Vera 1.5 demonstrates advanced mathematical problem-solving abilities at the olympiad level. This performance reflects robust reasoning chains and symbolic manipulation skills.

Competitive Programming Proficiency: The 89.2% LiveCodeBench score indicates strong algorithmic thinking and implementation capabilities on continuously updated, contamination-free problems.

Balanced Capabilities: Vera 1.5 maintains competitive performance across all evaluated domains, from mathematical reasoning to software engineering to scientific knowledge, demonstrating well-rounded capabilities suitable for diverse applications.

5.3 Architectural Efficiency

The MoE architecture enables Vera 1.5 to achieve frontier-level performance while maintaining computational efficiency. Sparse activation reduces inference costs compared to dense models of equivalent capability [2, 1]. Combined with FP8 precision, this design choice allows Vera 1.5 to offer strong performance-per-compute ratios [6].

The 200K context window enables practical applications requiring extensive document processing without the performance degradation often observed in models with artificially extended contexts [4].

6 Industry Benchmark Comparison

To provide additional context for Vera 1.5's performance, we reference publicly available model rankings from Artificial Analysis, an independent AI benchmarking organization [12]. Artificial Analysis maintains comprehensive leaderboards tracking frontier model performance across standardized evaluation suites.

As of January 2026, Artificial Analysis has updated its Intelligence Index (v4.0) to incorporate more challenging benchmarks, including τ^2-Bench, GDPval-AA (real-world economic tasks), and other evaluations designed to differentiate frontier models [13, 12]. The organization reports that top models now achieve scores of 50 or lower on the v4.0 scale, compared to averages of 73 on previous versions, reflecting the increased difficulty [13].

6.1 Comparative Context

Recent independent analyses highlight the competitive landscape of frontier models released in late 2025:

- Multiple models now exceed 90% on several traditional benchmarks, necessitating more challenging evaluation methodologies [13]
- Coding benchmarks show tight competition, with top models achieving 75-77% on SWE-Bench Verified [14]
- Long context windows (200K-1M tokens) have become a standard feature among frontier models [5]
- Scientific and mathematical reasoning remain differentiating factors, with significant performance variation across models [15]

Vera 1.5's performance profile positions it as a highly capable model across diverse evaluation domains, with particular strength in agentic reasoning as evidenced by the τ^2-Bench results.

7 Applications and Use Cases

Vera 1.5's capabilities enable deployment across numerous high-value applications:

7.1 AI Agents and Assistants

The exceptional τ^2-Bench performance indicates Vera 1.5 is well-suited for agentic workflows requiring multi-turn interactions, state management, and goal-oriented behavior [11]. Potential applications include:

- Customer service automation with complex problem resolution
- Technical support agents navigating multi-step troubleshooting
- Personal assistants managing extended task workflows
- Interactive tutoring systems maintaining pedagogical context

7.2 Software Development

Strong performance on LiveCodeBench and SWE-Bench Verified enables practical software engineering applications:

- Code generation and completion for multiple programming languages
- Automated bug detection and fix suggestion
- Codebase documentation and explanation
- Refactoring assistance for large-scale projects

The 200K context window allows Vera 1.5 to comprehend entire codebases, enabling more accurate suggestions that respect project-wide conventions and dependencies [5].

7.3 Scientific Research Support

GPQA Diamond and SciCode performance demonstrates capability for scientific workflows:

- Literature review and synthesis across multiple papers
- Hypothesis generation from experimental data
- Scientific computing assistance and algorithm implementation
- Technical writing and documentation support

7.4 Document Analysis and Processing

The extended context window enables comprehensive document understanding:

- Legal document review and contract analysis
- Financial report processing and insight extraction
- Medical record analysis and clinical decision support
- Research paper summarization and key finding extraction

7.5 Education and Training

Mathematical reasoning capabilities support educational applications:

- Personalized tutoring across STEM subjects
- Problem-solving guidance with step-by-step explanations
- Assessment generation and grading assistance
- Curriculum development support

8 Limitations and Future Work

8.1 Current Limitations

While Vera 1.5 demonstrates strong performance across evaluated benchmarks, several limitations warrant acknowledgment:

Scientific Computing Challenges: The 49.7% SciCode score, while competitive with other frontier models, indicates substantial room for improvement in research-level computational tasks. Complex scientific programming requiring deep domain expertise and sophisticated algorithm implementation remains challenging.

Context Window Utilization: While Vera 1.5 supports 200K tokens, research indicates that very long-context models may experience ”lost in the middle” effects where information positioned centrally receives less attention than content at context boundaries. Applications requiring precise information retrieval from arbitrary context positions may require additional optimization.

Inference Cost: Despite efficiency gains from sparse MoE architecture and FP8 precision, frontier model inference remains computationally expensive compared to smaller models. Deployment scenarios with strict latency or cost constraints may require careful optimization.

Knowledge Cutoff: With training data ending January 1, 2025, Vera 1.5 lacks knowledge of events and developments after this date. Applications requiring current information must implement retrieval-augmented generation or other knowledge updating mechanisms.

8.2 Future Research Directions

Several promising directions exist for future model development:

Enhanced Scientific Reasoning: Improving performance on research-level scientific computing through specialized training curricula, improved tool use, and integration with computational environments.

Multimodal Capabilities: Extending Vera's architecture to process images, audio, and video while maintaining the strong text performance demonstrated in Vera 1.5.

Improved Efficiency: Further optimization of inference costs through advanced quantization techniques, speculative decoding, and architectural innovations.

Tool Integration: Enhanced ability to use external tools, execute code, and access real-time information to address knowledge cutoff limitations and extend practical capabilities.

Safety and Alignment: Continued research into AI safety, robustness to adversarial inputs, and alignment with human values and intentions.

9 Conclusion

Vera 1.5 represents a significant achievement in UK frontier AI development by an independent research organization. Through careful architectural choices—including sparse MoE design, FP8 precision training, and extended context windows, Vera 1.5 achieves competitive performance across diverse evaluation domains while maintaining computational efficiency.

The model's exceptional 98.8% score on τ^2-Bench demonstrates particular strength in agentic reasoning and conversational interaction, a critical capability for practical AI deployment. Strong performance on mathematical reasoning , competitive programming, and scientific knowledge establishes Vera 1.5 as a well-rounded system suitable for diverse applications.

Beyond benchmark performance, Vera 1.5 demonstrates that independent research organizations can develop frontier-class AI systems that compete with models from major technology companies. This achievement contributes to the democratization of AI research and development, ensuring that cutting-edge capabilities remain accessible beyond a small number of well-resourced institutions.

Cortex Research remains committed to developing AI systems that advance human flourishing while prioritizing safety, scalability, and positive societal impact. Vera 1.5 represents an important step in this mission, and we look forward to continued progress in making powerful AI tools accessible and beneficial for addressing humanity's most pressing challenges.

References

[1] Hugging Face. (2025). *Mixture of Experts Explained.* Retrieved from https://huggingface.co/blog/moe

[2] Epoch AI. (2024). *MoE vs AI dense models: How do they compare in inference?* Retrieved from https://epoch.ai/gradient-updates/moe-vs-dense-models-inference

[3] arXiv. (2024). *Rethinking Training of Mixture-of-Experts Language Models.* Retrieved from https://arxiv.org/html/2404.05567v1

[4] Google Cloud. (2024). *What is long context and why does it matter for AI?* Retrieved from https://cloud.google.com/transform/the-prompt-what-are-long-context-windows-and-why-do-they-matter

[5] Codingscape. (2024). *LLMs with largest context windows.* Retrieved from https://codingscape.com/blog/llms-with-largest-context-windows

[6] arXiv. (2025). *InfiR2: A Comprehensive FP8 Training Recipe for Large-Scale Models.* Retrieved from https://arxiv.org/html/2509.22536v3

[7] Rohan Paul. (2025). *FP8 vs FP16/BF16/INT8: Theoretical Advantages.* Retrieved from https://www.rohan-paul.com/p/fp8-vs-fp16bf16int8-theoretical-advantages

[8] HMC Tech. (2024). *NVIDIA B200 Full Specs.* Retrieved from https://hmc-tech.com/gpu/nvidia-b200

[9] Modal. (2026). *How much does it cost to run NVIDIA B200 GPUs in 2025?* Retrieved from https://modal.com/blog/nvidia-b200-pricing

[10] Barres, V., Dong, H., Ray, S., Si, X., & Narasimhan, K. (2025). *τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment.* arXiv preprint arXiv:2506.07982. Retrieved from https://arxiv.org/abs/2506.07982

[11] Bohrium. (2025). *tau2-bench: Infrastructure for AI for Science.* Retrieved from https://www.bohrium.com/en/sciencepedia/agent-tools/sierra-research_tau2-bench

[12] Artificial Analysis. (2023). *Intelligence Benchmarking Methodology.* Retrieved from https://artificialanalysis.ai/methodology/intelligence-benchmarking

[13] VentureBeat. (2026). *Artificial Analysis overhauls its AI Intelligence Index.* Retrieved from https://venturebeat.com/technology/artificial-analysis-overhauls-its-ai-intelligence-index-replacing-popular

[14] Claude5.com. (2025). *AI Coding Benchmark 2025 — Claude 4.5 vs GPT-5.1.* Retrieved from https://claude5.com/benchmark

[15] Claude5.com. (2025). *Gemini 3 Pro vs GPT-5.1 vs Claude Sonnet 4.5: The Ultimate 2025 Comparison.* Retrieved from https://www.claude5.com/news/llm-comparison-2025-gemini-3-gpt-5-claude-4-5