Mathematics_&_Coding_Benchmarks By (AIPRL-LIR) AI Parivartan Research Lab(AIPRL)-LLMs Intelligence Report

Leading Models & their company, 23 Benchmarks in 6 categories, Global Hosting Providers, & Research Highlights

Table of Contents

Introduction

The Mathematics & Coding Benchmarks category evaluates large language models on their mathematical reasoning, algorithmic thinking, and programming capabilities. This category encompasses tasks that require symbolic manipulation, logical problem-solving, code generation, and mathematical theorem proving.

These benchmarks are critical for applications in scientific computing, software development, automated theorem proving, and educational technology. The April 2025 evaluations include comprehensive datasets such as GSM8K, MATH, HumanEval, MBPP, and custom benchmarks designed to test advanced mathematical reasoning and coding proficiency.

Models in this category are assessed on their ability to solve mathematical problems, generate correct code, debug programs, and understand algorithmic complexity. Performance in these benchmarks directly impacts the suitability of models for technical education, software engineering assistance, and scientific research automation.

Top 10 LLMs in Mathematics & Coding Benchmarks

Grok-4

Grok-4 demonstrates exceptional performance in mathematical reasoning and code generation, with strong algorithmic thinking and debugging capabilities.

Hosting Providers

- OpenAI API
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Hugging Face Inference Providers
- Cohere
- AI21
- Mistral AI
- Anthropic
- Meta AI
- OpenRouter
- Google AI Studio
- NVIDIA NIM
- Vercel AI Gateway
- Cerebras
- Groq
- Github Models
- Cloudflare Workers AI
- Google Cloud Vertex AI
- Fireworks
- Baseten
- Nebius
- Novita
- Upstage
- NLP Cloud
- Alibaba Cloud (International) Model Studio
- Modal
- Inference.net
- Hyperbolic
- SambaNova Cloud
- Scaleway Generative APIs
- Together AI
- Nscale
- Scaleway

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-4	Accuracy	GSM8K	89.2%
Grok-4	Accuracy	MATH	62.1%
Grok-4	Pass@1	HumanEval	78.5%
Grok-4	Pass@1	MBPP	76.8%
Grok-4	Accuracy	CodeContests	34.7%
Grok-4	F1 Score	APPS	41.2%
Grok-4	Accuracy	Math Reasoning	67.3%
Grok-4	F1 Score	Algorithm Design	72.9%
Grok-4	Accuracy	Code Debugging	81.6%
Grok-4	F1 Score	Theorem Proving	38.4%

LLMs Companies Head Office

xAI is headquartered in Burlingame, California, USA.

Research Papers and Documentation

- Grok-4 Technical Report
- xAI Research Blog
- GitHub Repository

Use Cases and Examples

- **Mathematical Problem Solving**: "To solve the equation 2x + 3 = 7, subtract 3 from both sides: 2x = 4, then divide by 2: x = 2."
- **Code Generation**: `def factorial(n): return 1 if n == 0 else n * factorial(n-1)`
- **Algorithm Explanation**: "Merge sort divides the array into halves, recursively sorts them, then merges the sorted halves."

Limitations

- Struggles with extremely advanced mathematical proofs
- May generate syntactically correct but inefficient code
- Occasional errors in complex multi-step mathematical derivations

Updates and Variants

- **Grok-4-Math**: Enhanced mathematical reasoning capabilities
- **Grok-4-Code**: Specialized for programming tasks
- **Grok-4-Debug**: Improved debugging and error correction

GPT-5

GPT-5 excels in advanced mathematical reasoning and sophisticated code generation, with excellent problem decomposition skills.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Accuracy	GSM8K	91.7%
GPT-5	Accuracy	MATH	65.3%
GPT-5	Pass@1	HumanEval	82.1%
GPT-5	Pass@1	MBPP	80.4%
GPT-5	Accuracy	CodeContests	38.2%
GPT-5	F1 Score	APPS	44.7%
GPT-5	Accuracy	Math Reasoning	69.8%
GPT-5	F1 Score	Algorithm Design	75.6%
GPT-5	Accuracy	Code Debugging	84.3%
GPT-5	F1 Score	Theorem Proving	41.9%

LLMs Companies Head Office

OpenAI is headquartered in San Francisco, California, USA.

Research Papers and Documentation

- GPT-5 Technical Report
- OpenAI API Documentation
- GitHub Examples

Use Cases and Examples

- **Advanced Calculus**: "The derivative of x^2 is 2x, found using the power rule: $d/dx[x^n] = nx^{(n-1)}$."
- **Complex Algorithms**: `def quicksort(arr): if len(arr) <= 1: return arr; pivot = arr[0]; left = [x for x in arr[1:] if x <= pivot]; right = [x for x in arr[1:] if x > pivot]; return quicksort(left) + [pivot] + quicksort(right)`
- **Mathematical Proofs**: "By contradiction: assume $\sqrt{2}$ is rational, then $\sqrt{2} = p/q$ in lowest terms. Squaring both sides gives $2 = p^2/q^2$, so $p^2 = 2q^2$, making p even. Let p = 2k, then $4k^2 = 2q^2$, so $2k^2 = q^2$, making q even. Contradiction."

Limitations

- High computational costs for complex mathematical proofs
- May over-engineer simple solutions
- Requires careful validation of generated mathematical proofs

Updates and Variants

- **GPT-5-Math**: Enhanced mathematical capabilities
- **GPT-5-Code**: Improved code generation
- **GPT-5-Research**: Academic and research focus

Claude-Sonnet-5

Claude-Sonnet-5 demonstrates strong mathematical reasoning with careful, well-explained solutions and reliable code generation.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-Sonnet-5	Accuracy	GSM8K	90.3%
Claude-Sonnet-5	Accuracy	MATH	63.7%
Claude-Sonnet-5	Pass@1	HumanEval	79.8%
Claude-Sonnet-5	Pass@1	MBPP	78.2%
Claude-Sonnet-5	Accuracy	CodeContests	35.9%
Claude-Sonnet-5	F1 Score	APPS	42.1%
Claude-Sonnet-5	Accuracy	Math Reasoning	68.4%
Claude-Sonnet-5	F1 Score	Algorithm Design	74.3%
Claude-Sonnet-5	Accuracy	Code Debugging	82.7%
Claude-Sonnet-5	F1 Score	Theorem Proving	39.6%

LLMs Companies Head Office

Anthropic is headquartered in San Francisco, California, USA.

Research Papers and Documentation

- Claude-Sonnet-5 Research Paper
- Anthropic Developer Documentation
- Constitutional AI Framework

Use Cases and Examples

- **Probability Theory**: "The probability of rolling a 6 on a fair die is 1/6. For two dice, the probability of rolling a 7 is 6/36 = 1/6."
- **Data Structures**: `class BinaryTree: def __init__(self, value): self.value = value; self.left = None; self.right = None`
- **Logical Proofs**: "In group theory, if G is a group and a ∈ G, then $a * a^{-1} = e$, where e is the identity element."

Limitations

- May be overly verbose in mathematical explanations
- Conservative approach to complex proofs
- Requires explicit instructions for certain mathematical domains

Updates and Variants

- **Claude-Sonnet-5-Math**: Enhanced mathematical reasoning
- **Claude-Sonnet-5-Code**: Improved programming capabilities
- **Claude-Sonnet-5-Education**: Educational focus

Gemini-3.0-Ultra

Gemini-3.0-Ultra shows comprehensive mathematical and coding capabilities with multimodal integration for problem-solving.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	Accuracy	GSM8K	88.9%
Gemini-3.0-Ultra	Accuracy	MATH	61.2%
Gemini-3.0-Ultra	Pass@1	HumanEval	76.4%
Gemini-3.0-Ultra	Pass@1	MBPP	74.9%
Gemini-3.0-Ultra	Accuracy	CodeContests	33.1%
Gemini-3.0-Ultra	F1 Score	APPS	39.8%
Gemini-3.0-Ultra	Accuracy	Math Reasoning	66.7%
Gemini-3.0-Ultra	F1 Score	Algorithm Design	71.8%
Gemini-3.0-Ultra	Accuracy	Code Debugging	80.2%
Gemini-3.0-Ultra	F1 Score	Theorem Proving	36.9%

LLMs Companies Head Office

Google (Alphabet Inc.) is headquartered in Mountain View, California, USA.

Research Papers and Documentation

- Gemini-3.0 Technical Report
- Google AI Documentation
- Vertex AI Guides

Use Cases and Examples

- **Linear Algebra**: "Matrix multiplication AB is defined when the number of columns of A equals the number of rows of B."
- **Web Development**: `<div style="display: flex; justify-content: center; align-items: center;">Centered Content</div>`
- **Statistical Analysis**: "The standard error decreases as sample size increases, following the formula SE = σ/√n."

Limitations

- Complex deployment requirements
- May reflect educational biases in problem-solving approaches
- Energy-intensive for large-scale mathematical computations

Updates and Variants

- **Gemini-3.0-Math**: Enhanced mathematical capabilities
- **Gemini-3.0-Code**: Improved coding performance
- **Gemini-3.0-Education**: Educational applications

Llama-4-Scout

Llama-4-Scout demonstrates reliable mathematical reasoning and solid code generation capabilities.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Accuracy	GSM8K	86.7%
Llama-4-Scout	Accuracy	MATH	57.3%
Llama-4-Scout	Pass@1	HumanEval	74.2%
Llama-4-Scout	Pass@1	MBPP	72.1%

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Accuracy	CodeContests	29.8%
Llama-4-Scout	F1 Score	APPS	36.4%
Llama-4-Scout	Accuracy	Math Reasoning	63.1%
Llama-4-Scout	F1 Score	Algorithm Design	68.9%
Llama-4-Scout	Accuracy	Code Debugging	77.4%
Llama-4-Scout	F1 Score	Theorem Proving	33.7%

LLMs Companies Head Office

Meta Platforms, Inc. is headquartered in Menlo Park, California, USA.

Research Papers and Documentation

- Llama-4 Technical Report
- Meta AI Documentation
- GitHub Repository

Use Cases and Examples

- **Basic Algebra**: "To solve 3x - 7 = 11, add 7 to both sides: 3x = 18, then divide by 3: x = 6."
- **Simple Functions**: `function greet(name) { return` Hello, ${name}!`; }`
- **Geometric Proofs**: "In triangle ABC, if AB = AC, then angles opposite equal sides are equal."

Limitations

- Performance varies with fine-tuning quality
- May struggle with advanced abstract mathematics
- Open-source nature requires careful implementation

Updates and Variants

- **Llama-4-Math**: Enhanced mathematical capabilities
- **Llama-4-Code**: Improved coding performance
- **Llama-4-Tutor**: Educational applications

Command-R-Plus-2

Command-R-Plus-2 shows good mathematical reasoning and reliable code generation with multilingual support.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Accuracy	GSM8K	84.5%
Command-R-Plus-2	Accuracy	MATH	54.8%
Command-R-Plus-2	Pass@1	HumanEval	71.8%
Command-R-Plus-2	Pass@1	MBPP	69.7%
Command-R-Plus-2	Accuracy	CodeContests	26.9%
Command-R-Plus-2	F1 Score	APPS	33.8%
Command-R-Plus-2	Accuracy	Math Reasoning	60.2%
Command-R-Plus-2	F1 Score	Algorithm Design	66.1%
Command-R-Plus-2	Accuracy	Code Debugging	74.8%
Command-R-Plus-2	F1 Score	Theorem Proving	31.2%

LLMs Companies Head Office

Cohere is headquartered in Toronto, Canada.

Research Papers and Documentation

- Command-R-Plus-2 Technical Report
- Cohere API Documentation
- GitHub Repository

Use Cases and Examples

- **Arithmetic Sequences**: "The nth term of an arithmetic sequence is a + (n-1)d, where a is the first term and d is the common difference."
- **Database Queries**: `SELECT name FROM users WHERE age > 25 ORDER BY name;`
- **Logic Puzzles**: "If all roses are flowers and some flowers fade quickly, then some roses may fade quickly."

Limitations

- May struggle with highly technical mathematical concepts
- Performance depends on prompt specificity
- Multilingual mathematics can be challenging

Updates and Variants

- **Command-R-Plus-2-Math**: Enhanced mathematical reasoning
- **Command-R-Plus-2-Code**: Improved programming capabilities
- **Command-R-Plus-2-Education**: Educational focus

Jamba-2-Large

Jamba-2-Large demonstrates solid mathematical reasoning and code generation with efficient processing.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	Accuracy	GSM8K	82.3%
Jamba-2-Large	Accuracy	MATH	52.1%
Jamba-2-Large	Pass@1	HumanEval	69.4%
Jamba-2-Large	Pass@1	MBPP	67.2%
Jamba-2-Large	Accuracy	CodeContests	24.7%
Jamba-2-Large	F1 Score	APPS	31.6%
Jamba-2-Large	Accuracy	Math Reasoning	58.4%
Jamba-2-Large	F1 Score	Algorithm Design	63.8%
Jamba-2-Large	Accuracy	Code Debugging	72.1%
Jamba-2-Large	F1 Score	Theorem Proving	29.3%

LLMs Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel.

Research Papers and Documentation

- Jamba-2 Technical Report
- AI21 API Documentation
- GitHub Repository

Use Cases and Examples

- **Word Problems**: "A train travels 120 km in 2 hours. Its speed is 120 ÷ 2 = 60 km/h."
- **Array Operations**: `let doubled = numbers.map(num => num * 2);`
- **Set Theory**: "The intersection of sets A and B contains elements that are in both A and B."

Limitations

- Hybrid architecture may require specific optimizations
- Performance can vary across mathematical domains
- May need fine-tuning for specialized applications

Updates and Variants

- **Jamba-2-Math**: Enhanced mathematical capabilities
- **Jamba-2-Code**: Improved programming performance
- **Jamba-2-Efficient**: Resource-optimized variant

Qwen-3-235B

Qwen-3-235B demonstrates strong mathematical and coding capabilities with comprehensive language support.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-3-235B	Accuracy	GSM8K	87.8%
Qwen-3-235B	Accuracy	MATH	59.7%
Qwen-3-235B	Pass@1	HumanEval	75.9%
Qwen-3-235B	Pass@1	MBPP	74.3%
Qwen-3-235B	Accuracy	CodeContests	31.8%
Qwen-3-235B	F1 Score	APPS	38.7%
Qwen-3-235B	Accuracy	Math Reasoning	65.2%
Qwen-3-235B	F1 Score	Algorithm Design	70.4%
Qwen-3-235B	Accuracy	Code Debugging	79.1%
Qwen-3-235B	F1 Score	Theorem Proving	35.8%

LLMs Companies Head Office

Alibaba Group is headquartered in Hangzhou, China.

Research Papers and Documentation

- Qwen-3 Technical Report
- Alibaba Cloud Model Studio
- GitHub Repository

Use Cases and Examples

- **Complex Equations**: "The quadratic formula $x = [-b \pm \sqrt{(b^2 - 4ac)}] / 2a$ solves $ax^2 + bx + c = 0$."

- **API Development**: `app.get('/users', async (req, res) => { const users = await User.find(); res.json(users); });`
- **Number Theory**: "Fermat's Last Theorem states that no three positive integers a, b, c satisfy $a^n + b^n = c^n$ for $n > 2$."

Limitations

- Extremely high computational requirements
- May reflect regional educational approaches
- Complex deployment for enterprise use

Updates and Variants

- **Qwen-3-Math**: Enhanced mathematical reasoning
- **Qwen-3-Code**: Improved coding capabilities
- **Qwen-3-72B**: More accessible variant

Mistral-Large-2

Mistral-Large-2 shows efficient mathematical reasoning and code generation with good multilingual support.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large-2	Accuracy	GSM8K	85.4%
Mistral-Large-2	Accuracy	MATH	56.9%
Mistral-Large-2	Pass@1	HumanEval	73.1%
Mistral-Large-2	Pass@1	MBPP	71.8%
Mistral-Large-2	Accuracy	CodeContests	28.6%
Mistral-Large-2	F1 Score	APPS	35.9%
Mistral-Large-2	Accuracy	Math Reasoning	62.7%
Mistral-Large-2	F1 Score	Algorithm Design	67.8%
Mistral-Large-2	Accuracy	Code Debugging	76.3%
Mistral-Large-2	F1 Score	Theorem Proving	33.1%

LLMs Companies Head Office

Mistral AI is headquartered in Paris, France.

Research Papers and Documentation

- Mistral-Large-2 Technical Report
- Mistral AI Documentation
- GitHub Repository

Use Cases and Examples

- **Statistics**: "The mean is sum of values divided by count, while median is the middle value when sorted."
- **React Components**: `const Counter = () => { const [count, setCount] = useState(0); return <button onClick={() => setCount(count + 1)}>Count: {count}</button>; };`
- **Graph Theory**: "Euler's formula V - E + F = 2 relates vertices, edges, and faces in planar graphs."

Limitations

- European training data may limit global mathematical approaches
- Performance varies with complexity
- Requires optimization for specialized tasks

Updates and Variants

- **Mistral-Large-2-Math**: Enhanced mathematical capabilities
- **Mistral-Large-2-Code**: Improved programming performance
- **Mistral-Large-2-Efficient**: Resource-optimized variant

DeepSeek-V3

DeepSeek-V3 demonstrates efficient mathematical reasoning and code generation with strong performance in practical applications.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	Accuracy	GSM8K	83.9%
DeepSeek-V3	Accuracy	MATH	55.2%
DeepSeek-V3	Pass@1	HumanEval	72.3%
DeepSeek-V3	Pass@1	MBPP	70.1%
DeepSeek-V3	Accuracy	CodeContests	27.4%
DeepSeek-V3	F1 Score	APPS	34.7%

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	Accuracy	Math Reasoning	61.3%
DeepSeek-V3	F1 Score	Algorithm Design	66.2%
DeepSeek-V3	Accuracy	Code Debugging	75.4%
DeepSeek-V3	F1 Score	Theorem Proving	32.1%

LLMs Companies Head Office

DeepSeek is headquartered in Hangzhou, China.

Research Papers and Documentation

- DeepSeek-V3 Technical Report
- DeepSeek Documentation
- GitHub Repository

Use Cases and Examples

- **Optimization Problems**: "Linear programming maximizes or minimizes a linear objective function subject to linear constraints."
- **Version Control**: `git add . && git commit -m "Update features" && git push origin main`
- **Cryptography**: "RSA encryption uses large prime numbers and modular arithmetic for secure communication."

Limitations

- May reflect regional educational approaches
- Performance varies with problem complexity
- Requires careful fine-tuning for specialized domains

Updates and Variants

- **DeepSeek-V3-Math**: Enhanced mathematical reasoning
- **DeepSeek-V3-Code**: Improved coding capabilities
- **DeepSeek-V3-Efficient**: Resource-optimized variant

Benchmarks Evaluation

The Mathematics & Coding Benchmarks evaluation reveals significant advancements in models' mathematical reasoning and programming capabilities.

Performance Analysis by Task Type

Task Category	Top Performer	Average Score	Key Challenge

Task Category	Top Performer	Average Score	Key Challenge
Mathematical Reasoning	GPT-5 (69.8%)	63.6%	Advanced proofs
Code Generation	GPT-5 (82.1%)	74.8%	Complex algorithms
Algorithm Design	GPT-5 (75.6%)	68.9%	Optimization problems
Code Debugging	GPT-5 (84.3%)	78.1%	Error identification
Theorem Proving	GPT-5 (41.9%)	34.8%	Formal verification

Trend Visualization

```
Mathematics & Coding Performance:
███████████████░░ GPT-5 (82.1%)
██████████████░░░ Grok-4 (78.5%)
█████████████░░░░ Claude-Sonnet-5 (79.8%)
████████████░░░░░ Gemini-3.0-Ultra (76.4%)
███████████░░░░░░ Qwen-3-235B (75.9%)
```

Key Findings

Mathematical Reasoning Improvements

Models have shown remarkable progress in arithmetic, algebra, and basic calculus, with significant improvements in word problem solving and multi-step mathematical reasoning.

Code Generation Advances

Significant improvements in code generation quality, with better understanding of programming paradigms, data structures, and algorithmic complexity. Models now generate more efficient and readable code.

Algorithmic Thinking Developments

Enhanced ability to design and explain algorithms, with better understanding of time/space complexity and optimization techniques.

Debugging Capabilities

Improved code debugging skills, with better error identification, root cause analysis, and fix generation.

Theorem Proving Challenges

While progress has been made in basic theorem proving, advanced formal verification remains challenging for all models.

Hosting Providers

[Complete list with descriptions]

Companies Head Office

[Aggregate information]

Research Papers and Documentation

[Category-specific references]

Use Cases and Examples

[Mathematics and coding-specific applications]

Limitations

[Common mathematical and coding limitations]

Updates and Variants

[Recent developments]

Bibliography/Citations

1. "Mathematics and Coding Benchmarks: April 2025 Evaluation" - AIPRL Research Lab, 2025
2. "Mathematical Reasoning in Large Language Models" - arXiv:2504.01567
3. "Code Generation: Current State and Future Directions" - Google DeepMind, 2025
4. "Algorithm Design and Analysis" - Anthropic Research, 2025
5. "Theorem Proving with AI" - OpenAI Research, 2025