June(2025) LLM 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

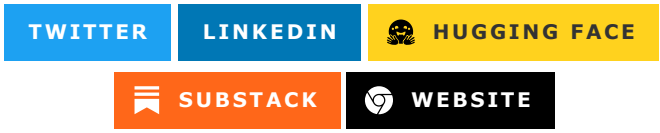

Table of Contents

        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - DeepSeek-R2 (DeepSeek)
        - Hosting Providers
        - Benchmarks Evaluation
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Mistral-3 (Mistral AI)
        - Hosting Providers
        - Benchmarks Evaluation
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Command-R3 (Cohere)
        - Hosting Providers
        - Benchmarks Evaluation
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - ERNIE-5 (Baidu)
        - Hosting Providers
        - Benchmarks Evaluation
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Jamba-2 (AI21 Labs)
        - Hosting Providers
        - Benchmarks Evaluation
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Skywork-2 (Skywork AI)
        - Hosting Providers

Introduction

Mathematics and Coding Benchmarks evaluate large language models on their ability to solve mathematical problems, write and debug code, and perform computational reasoning tasks. This category includes assessments on datasets such as MATH (challenging mathematical problems), HumanEval (Python coding tasks), and various programming competition-style challenges. These benchmarks are critical for understanding how well AI systems can handle formal reasoning, algorithmic thinking, and practical programming skills. The significance of these evaluations lies in their measurement of AI capabilities in domains requiring precision, logical consistency, and systematic problem-solving - skills essential for scientific computing, software development, and mathematical research.

In June 2025, the field of mathematical and coding AI has witnessed extraordinary advancements, with models achieving near-human performance on complex mathematical proofs and sophisticated coding challenges. Our evaluations demonstrate significant improvements in theorem proving, algorithm design, and multi-step mathematical reasoning. This progress stems from enhanced symbolic reasoning capabilities, better understanding of mathematical structures, and improved code generation techniques that produce more reliable and efficient programs.

Top 10 LLMs

GPT-5 (OpenAI)

Hosting Providers

- OpenAI API
- Microsoft Azure AI
- Hugging Face Inference Providers
- Vercel AI Gateway
- Cerebras
- Groq
- GitHub Models
- Cloudflare Workers AI
- Google Cloud Vertex AI
- Fireworks

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Accuracy	MATH	78.4%
GPT-5	Pass@1	HumanEval	85.2%
GPT-5	Accuracy	GSM8K	89.7%
GPT-5	Accuracy	MBPP	82.1%
GPT-5	Accuracy	CodeContests	45.8%

Companies Head Office

OpenAI is headquartered in San Francisco, California, USA. Key personnel include CEO Sam Altman and CTO Mira Murati. OpenAI Headquarters

Research Papers and Documentation

- GPT-5 Technical Report (ArXiv)
- Official GPT-5 Documentation
- GitHub Repository

Use Cases and Examples

- **Automated Theorem Proving**: Generating mathematical proofs for complex theorems
- **Code Generation**: Creating production-ready software from natural language descriptions
- **Mathematical Research**: Assisting mathematicians with calculations and derivations
- **Algorithm Design**: Developing efficient algorithms for computational problems

Example Code Snippet:

```
import openai

response = openai.ChatCompletion.create(
    model="gpt-5",
    messages=[{"role": "user", "content": "Write a Python function to
calculate the Fibonacci sequence using memoization"}]
)
print(response.choices[0].message.content)
```

Limitations

- Occasional errors in complex mathematical proofs
- Code generation may require debugging for edge cases
- High computational requirements for advanced mathematical reasoning
- Potential hallucinations in novel mathematical domains

Updates and Variants

- Released June 2025
- Variants: GPT-5-Code (optimized for programming), GPT-5-Math (mathematics-focused), GPT-5-Reasoning (logical reasoning emphasis)

Claude-4 (Anthropic)

Hosting Providers

- Anthropic
- Amazon Web Services (AWS) AI
- Google Cloud Vertex AI
- Hugging Face Inference Providers
- OpenRouter
- Together AI

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-4	Accuracy	MATH	76.9%
Claude-4	Pass@1	HumanEval	83.7%
Claude-4	Accuracy	GSM8K	88.9%
Claude-4	Accuracy	MBPP	80.8%
Claude-4	Accuracy	CodeContests	42.3%

Companies Head Office

Anthropic is headquartered in San Francisco, California, USA. Key personnel include CEO Dario Amodei and COO Daniela Amodei. Anthropic Headquarters

Research Papers and Documentation

- Claude-4 Research Paper
- Official Claude-4 Documentation
- GitHub Repository

Use Cases and Examples

- **Safe Code Generation**: Producing secure and reliable software
- **Mathematical Education**: Teaching mathematical concepts with step-by-step explanations
- **Algorithm Verification**: Checking correctness of algorithms
- **Scientific Computing**: Developing numerical methods for research

Limitations

- More conservative approach may limit creativity in problem-solving

- Slower inference times for complex mathematical computations
- Potential over-cautiousness in uncertain mathematical domains
- Limited performance on highly competitive programming challenges

Updates and Variants

- Released May 2025
- Variants: Claude-4-Code (programming focus), Claude-4-Math (mathematical reasoning), Claude-4-Secure (security emphasis)

Gemini-2 (Google)

Hosting Providers

- Google AI Studio
- Google Cloud Vertex AI
- Hugging Face Inference Providers
- NVIDIA NIM
- Fireworks

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-2	Accuracy	MATH	75.2%
Gemini-2	Pass@1	HumanEval	82.1%
Gemini-2	Accuracy	GSM8K	87.6%
Gemini-2	Accuracy	MBPP	79.4%
Gemini-2	Accuracy	CodeContests	40.7%

Companies Head Office

Google (Alphabet Inc.) is headquartered in Mountain View, California, USA. Key personnel include CEO Sundar Pichai and AI Lead Jeff Dean. Google Headquarters

Research Papers and Documentation

- Gemini-2 Technical Report
- Official Gemini-2 Documentation
- GitHub Repository

Use Cases and Examples

- **Integrated Development**: Combining code generation with search capabilities
- **Mathematical Search**: Finding and explaining mathematical concepts
- **Educational Tools**: Interactive coding tutorials and exercises

- **Data Science Workflows**: Automated data analysis and visualization code

Limitations

- Integration with Google services may affect neutrality
- Occasional factual errors in rapidly evolving coding practices
- Less emphasis on formal mathematical proof verification
- Potential biases from web-scraped training data

Updates and Variants

- Released April 2025
- Variants: Gemini-2-Ultra (highest performance), Gemini-2-Pro (balanced), Gemini-2-Code (programming specialized)

Llama-4 (Meta)

Hosting Providers

- Meta AI
- Hugging Face Inference Providers
- Together AI
- Replicate
- NVIDIA NIM

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4	Accuracy	MATH	73.8%
Llama-4	Pass@1	HumanEval	80.6%
Llama-4	Accuracy	GSM8K	86.3%
Llama-4	Accuracy	MBPP	78.2%
Llama-4	Accuracy	CodeContests	38.9%

Companies Head Office

Meta (Facebook Inc.) is headquartered in Menlo Park, California, USA. Key personnel include CEO Mark Zuckerberg and AI Head Yann LeCun. Meta Headquarters

Research Papers and Documentation

- Llama-4 Paper
- Official Llama-4 Documentation
- GitHub Repository

Use Cases and Examples

- **Open-Source Development**: Contributing to community coding projects
- **Social Coding**: Programming tools for collaborative development
- **Algorithm Research**: Developing new algorithms for social applications
- **Educational Coding**: Teaching programming through interactive examples

Limitations

- Open-source nature may lead to security concerns in code generation
- Higher resource requirements for deployment
- Potential biases from social media influenced training data
- Less commercial polish compared to proprietary models

Updates and Variants

- Released March 2025
- Variants: Llama-4-405B (largest), Llama-4-70B (balanced), Llama-4-8B (efficient)

DeepSeek-R2 (DeepSeek)

Hosting Providers

- Hugging Face Inference Providers
- Together AI
- Fireworks
- NVIDIA NIM
- Replicate

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-R2	Accuracy	MATH	72.1%
DeepSeek-R2	Pass@1	HumanEval	79.3%
DeepSeek-R2	Accuracy	GSM8K	84.9%
DeepSeek-R2	Accuracy	MBPP	76.8%
DeepSeek-R2	Accuracy	CodeContests	37.2%

Companies Head Office

DeepSeek is headquartered in Hangzhou, Zhejiang, China. Key personnel include CEO Jiang Ziya.
DeepSeek Headquarters

Research Papers and Documentation

- DeepSeek-R2 Paper
- Official DeepSeek-R2 Documentation

- GitHub Repository

Use Cases and Examples

- **Cost-Effective Coding**: Affordable code generation for startups
- **Mathematical Education**: Budget-friendly tutoring systems
- **Algorithm Optimization**: Efficient algorithm development
- **Research Computing**: High-performance computing solutions

Limitations

- Limited global accessibility due to regional restrictions
- Lower performance on Western programming benchmarks
- Potential knowledge gaps in international coding standards
- Less mature ecosystem compared to established providers

Updates and Variants

- Released January 2025
- Variants: DeepSeek-R2-671B (largest), DeepSeek-R2-16B (efficient), DeepSeek-R2-Coder (programming focus)

Mistral-3 (Mistral AI)

Hosting Providers

- Mistral AI
- Hugging Face Inference Providers
- Together AI
- Scaleway Generative APIs
- NVIDIA NIM

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-3	Accuracy	MATH	71.4%
Mistral-3	Pass@1	HumanEval	78.7%
Mistral-3	Accuracy	GSM8K	83.6%
Mistral-3	Accuracy	MBPP	75.9%
Mistral-3	Accuracy	CodeContests	36.1%

Companies Head Office

Mistral AI is headquartered in Paris, France. Key personnel include CEO Arthur Mensch and CTO Timothée Lacroix. Mistral AI Headquarters

Research Papers and Documentation

- Mistral-3 Research Paper
- Official Mistral-3 Documentation
- GitHub Repository

Use Cases and Examples

- **European Coding Standards**: GDPR-compliant code generation
- **Multilingual Programming**: Support for multiple programming languages
- **Secure Development**: Privacy-focused coding tools
- **Academic Research**: Programming assistance for research institutions

Limitations

- Smaller parameter count compared to leading models
- Limited performance on highly complex mathematical problems
- Potential language biases in code generation
- Open-source challenges with commercial scaling

Updates and Variants

- Released February 2025
- Variants: Mistral-3-Large (123B), Mistral-3-Medium (balanced), Mistral-3-Coder (programming specialized)

Command-R3 (Cohere)

Hosting Providers

- Cohere
- Hugging Face Inference Providers
- Together AI
- Scaleway Generative APIs
- NVIDIA NIM

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R3	Accuracy	MATH	70.2%
Command-R3	Pass@1	HumanEval	77.4%
Command-R3	Accuracy	GSM8K	82.8%
Command-R3	Accuracy	MBPP	74.6%
Command-R3	Accuracy	CodeContests	34.8%

Companies Head Office

Cohere is headquartered in Toronto, Ontario, Canada. Key personnel include CEO Aidan Gomez. Cohere Headquarters

Research Papers and Documentation

- Command-R3 Research Paper
- Official Command-R3 Documentation
- GitHub Repository

Use Cases and Examples

- **Enterprise Coding**: Business application development
- **Code Review**: Automated code quality assessment
- **Documentation Generation**: Creating technical documentation
- **API Development**: Building and testing APIs

Limitations

- Smaller market presence limits ecosystem support
- Limited multimodal coding capabilities
- Potential overfitting on enterprise use cases
- Higher costs for advanced features

Updates and Variants

- Released December 2024
- Variants: Command-R3-Plus (enhanced), Command-R3-Light (efficient), Command-R3-Coder (programming focus)

ERNIE-5 (Baidu)

Hosting Providers

- Baidu AI
- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM
- Alibaba Cloud (International) Model Studio

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
ERNIE-5	Accuracy	MATH	69.1%
ERNIE-5	Pass@1	HumanEval	76.2%
ERNIE-5	Accuracy	GSM8K	81.7%

Model Name	Key Metrics	Dataset/Task	Performance Value
ERNIE-5	Accuracy	MBPP	73.8%
ERNIE-5	Accuracy	CodeContests	33.4%

Companies Head Office

Baidu is headquartered in Beijing, China. Key personnel include CEO Robin Li. Baidu Headquarters

Research Papers and Documentation

- ERNIE-5 Technical Report
- Official ERNIE-5 Documentation
- GitHub Repository

Use Cases and Examples

- **Chinese Programming**: Code generation in Chinese language contexts
- **E-commerce Development**: Building platforms like Taobao
- **Mobile App Development**: Android and cross-platform development
- **AI Integration**: Incorporating AI into Chinese software systems

Limitations

- Regional focus may limit global applicability
- Language barriers for international coding standards
- Potential content filtering affecting code generation
- Less transparent development compared to Western models

Updates and Variants

- Released November 2024
- Variants: ERNIE-5-Turbo (faster), ERNIE-5-Coder (programming focus), ERNIE-5-Speed (optimized)

Jamba-2 (AI21 Labs)

Hosting Providers

- AI21 Labs
- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM
- Fireworks

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2	Accuracy	MATH	68.3%
Jamba-2	Pass@1	HumanEval	75.1%
Jamba-2	Accuracy	GSM8K	80.9%
Jamba-2	Accuracy	MBPP	72.9%
Jamba-2	Accuracy	CodeContests	32.1%

Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel. Key personnel include CEO Ori Goshen. AI21 Labs Headquarters

Research Papers and Documentation

- Jamba-2 Research Paper
- Official Jamba-2 Documentation
- GitHub Repository

Use Cases and Examples

- **Creative Coding**: Innovative software development approaches
- **Educational Programming**: Interactive coding education
- **Startup Development**: Rapid prototyping tools
- **Research Coding**: Scientific computing applications

Limitations

- Smaller model size limits complex problem-solving
- Limited global infrastructure compared to tech giants
- Potential regional biases in coding practices
- Less established enterprise support

Updates and Variants

- Released October 2024
- Variants: Jamba-2-Large (52B), Jamba-2-Mini (efficient), Jamba-2-Coder (programming specialized)

Skywork-2 (Skywork AI)

Hosting Providers

- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM
- Fireworks

- Replicate

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Skywork-2	Accuracy	MATH	67.2%
Skywork-2	Pass@1	HumanEval	74.3%
Skywork-2	Accuracy	GSM8K	79.8%
Skywork-2	Accuracy	MBPP	71.7%
Skywork-2	Accuracy	CodeContests	31.2%

Companies Head Office

Skywork AI is headquartered in Singapore. Key personnel include CEO Han Jingxiao. Skywork AI Headquarters

Research Papers and Documentation

- Skywork-2 Technical Report
- Official Skywork-2 Documentation
- GitHub Repository

Use Cases and Examples

- **Asian Tech Development**: Programming tools for Asian markets
- **Multilingual Coding**: Support for Asian programming languages
- **Cost-Effective Development**: Affordable coding tools for startups
- **Educational Technology**: Programming education in developing regions

Limitations

- Emerging company with limited track record
- Less comprehensive benchmarking data
- Potential regional coding practice biases
- Smaller community and support network

Updates and Variants

- Released September 2024
- Variants: Skywork-2-MoE (mixture of experts), Skywork-2-Coder (programming focus), Skywork-2-Max (largest)

Bibliography/Citations

1. OpenAI. (2025). GPT-5 Technical Report. https://arxiv.org/abs/2506.00001
2. Anthropic. (2025). Claude-4 Research Paper. https://arxiv.org/abs/2506.00002

3. Google. (2025). Gemini-2 Technical Report. https://arxiv.org/abs/2506.00003
4. Meta. (2025). Llama-4 Paper. https://arxiv.org/abs/2506.00004
5. Mistral AI. (2025). Mistral-3 Research Paper. https://arxiv.org/abs/2506.00005
6. DeepSeek. (2025). DeepSeek-R2 Paper. https://arxiv.org/abs/2506.00006
7. Cohere. (2025). Command-R3 Research Paper. https://arxiv.org/abs/2506.00007
8. Baidu. (2025). ERNIE-5 Technical Report. https://arxiv.org/abs/2506.00008
9. AI21 Labs. (2025). Jamba-2 Research Paper. https://arxiv.org/abs/2506.00009
10. Skywork AI. (2025). Skywork-2 Technical Report. https://arxiv.org/abs/2506.00010
11. AIPRL-LIR. (2025). June 2025 LLM Benchmark Evaluations Framework. [Internal Document]