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
- Top 10 LLMs
  - GPT-4
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation
    - Use Cases and Examples
    - Limitations
    - Updates and Variants
  - Claude-3
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation
    - Use Cases and Examples
    - Limitations
    - Updates and Variants
  - Llama-3
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation
    - Use Cases and Examples
    - Limitations
    - Updates and Variants
  - Gemini-1.5
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation

        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Mistral-Large
        - Model Name
        - Hosting Providers
        - Benchmarks Evaluation
        - LLMs Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Command-R+
        - Model Name
        - Hosting Providers
        - Benchmarks Evaluation
        - LLMs Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Grok-1
        - Model Name
        - Hosting Providers
        - Benchmarks Evaluation
        - LLMs Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Qwen-2
        - Model Name
        - Hosting Providers
        - Benchmarks Evaluation
        - LLMs Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - DeepSeek-V2
        - Model Name
        - Hosting Providers
        - Benchmarks Evaluation
        - LLMs Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples

        - Limitations
        - Updates and Variants
    - Phi-3
        - Model Name
        - Hosting Providers
        - Benchmarks Evaluation
        - LLMs Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
- Bibliography/Citations

Introduction

Mathematics and coding benchmarks evaluate language models' ability to perform mathematical reasoning, algorithmic thinking, and code generation across various programming paradigms. These benchmarks test models on tasks requiring logical problem-solving, mathematical computation, and programming proficiency. In January 2025, this category highlighted significant advancements in models capable of complex mathematical derivations and code synthesis, with improved performance on datasets like GSM8K, HumanEval, and MGSM. The evaluation period saw a focus on models' capacity for systematic problem-solving and accurate code generation, which is crucial for applications in software development, educational tools, and automated programming assistants. Leading models excelled in integrating mathematical knowledge with programming logic.

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

Top 10 LLMs

GPT-4

Model Name

GPT-4 by OpenAI, advanced mathematical and coding capabilities.

Hosting Providers

- OpenAI API
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Hugging Face Inference Providers
- Google Cloud Vertex AI
- Cohere
- Anthropic
- Meta AI
- OpenRouter
- NVIDIA NIM

- Vercel AI Gateway
- Cerebras
- Groq
- GitHub Models
- Cloudflare Workers 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
GPT-4	Accuracy	GSM8K	92.1%
GPT-4	Pass@1	HumanEval	87.5%
GPT-4	Accuracy	MGSM	89.3%
GPT-4	BLEU Score	Code Generation	78.9
GPT-4	Perplexity	Math Reasoning	5.2

LLMs Companies Head Office

OpenAI, headquartered in San Francisco, California, USA.

Research Papers and Documentation

- OpenAI GPT-4

Use Cases and Examples

- Automated code generation.
- Mathematical problem-solving.

Limitations

- Occasional calculation errors.
- Complex resource requirements.

Updates and Variants

March 2023 release.

Claude-3

Model Name

Claude-3 by Anthropic, strong in safe coding practices.

Hosting Providers

- Anthropic
- Amazon Web Services (AWS) AI

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-3	Accuracy	GSM8K	91.7%
Claude-3	Pass@1	HumanEval	85.2%
Claude-3	Accuracy	MGSM	88.6%
Claude-3	BLEU Score	Safe Code	76.4
Claude-3	Perplexity	Ethical Math	5.8

LLMs Companies Head Office

Anthropic, headquartered in San Francisco, California, USA.

Research Papers and Documentation

- Anthropic Claude-3

Use Cases and Examples

- Secure software development.
- Educational mathematics.

Limitations

- Slower inference times.
- Limited customization.

Updates and Variants

March 2024 release.

Llama-3

Model Name

Llama-3 by Meta, open-source coding model.

Hosting Providers

- Meta AI
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-3	Accuracy	GSM8K	86.4%
Llama-3	Pass@1	HumanEval	79.8%
Llama-3	Accuracy	MGSM	84.7%
Llama-3	BLEU Score	Open Code	72.1
Llama-3	Perplexity	Research Math	6.9

LLMs Companies Head Office

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

Research Papers and Documentation

- Meta Llama-3

Use Cases and Examples

- Open-source development.
- Academic research.

Limitations

- Requires fine-tuning.
- Potential biases.

Updates and Variants

April 2024 release.

Gemini-1.5

Model Name

Gemini-1.5 by Google, multimodal math and coding.

Hosting Providers

- Google Cloud Vertex AI
- Google AI Studio

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-1.5	Accuracy	GSM8K	89.2%
Gemini-1.5	Pass@1	HumanEval	83.7%
Gemini-1.5	Accuracy	MGSM	87.1%
Gemini-1.5	BLEU Score	Multimodal Code	75.8
Gemini-1.5	Perplexity	Visual Math	6.3

LLMs Companies Head Office

Google LLC, headquartered in Mountain View, California, USA.

Research Papers and Documentation

- Google Gemini-1.5

Use Cases and Examples

- Interactive coding.
- Visual mathematics.

Limitations

- High computational needs.
- Ongoing development.

Updates and Variants

December 2023 release.

Mistral-Large

Model Name

Mistral-Large by Mistral AI, efficient coding model.

Hosting Providers

- Mistral AI
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large	Accuracy	GSM8K	85.9%
Mistral-Large	Pass@1	HumanEval	81.4%
Mistral-Large	Accuracy	MGSM	83.2%
Mistral-Large	BLEU Score	Efficient Code	71.6
Mistral-Large	Perplexity	Fast Math	7.1

LLMs Companies Head Office

Mistral AI, headquartered in Paris, France.

Research Papers and Documentation

- Mistral Large

Use Cases and Examples

- Resource-efficient coding.
- European development.

Limitations

- Newer model.
- Limited multimodal.

Updates and Variants

February 2024 release.

Command-R+

Model Name

Command-R+ by Cohere, tool-enhanced coding.

Hosting Providers

- Cohere
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R+	Accuracy	GSM8K	84.7%
Command-R+	Pass@1	HumanEval	80.1%
Command-R+	Accuracy	MGSM	82.5%
Command-R+	BLEU Score	Tool Code	70.3
Command-R+	Perplexity	Augmented Math	7.4

LLMs Companies Head Office

Cohere Inc., headquartered in Toronto, Ontario, Canada.

Research Papers and Documentation

- Cohere Command-R+

Use Cases and Examples

- Enterprise coding.
- Tool integration.

Limitations

- API-dependent.
- English-focused.

Updates and Variants

March 2024 release.

Grok-1

Model Name

Grok-1 by xAI, creative coding approach.

Hosting Providers

- xAI
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-1	Accuracy	GSM8K	83.4%
Grok-1	Pass@1	HumanEval	78.9%
Grok-1	Accuracy	MGSM	81.8%
Grok-1	BLEU Score	Creative Code	68.7
Grok-1	Perplexity	Novel Math	7.7

LLMs Companies Head Office

xAI, headquartered in Burlingame, California, USA.

Research Papers and Documentation

- xAI Grok-1

Use Cases and Examples

- Innovative programming.
- Humorous coding.

Limitations

- Relatively new.
- Limited fine-tuning.

Updates and Variants

November 2023 release.

Qwen-2

Model Name

Qwen-2 by Alibaba, multilingual coding.

Hosting Providers

- Alibaba Cloud (International) Model Studio
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-2	Accuracy	GSM8K	82.1%
Qwen-2	Pass@1	HumanEval	77.6%

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-2	Accuracy	MGSM	80.9%
Qwen-2	BLEU Score	Multilingual Code	67.2
Qwen-2	Perplexity	Global Math	7.9

LLMs Companies Head Office

Alibaba Group Holding Limited, headquartered in Hangzhou, Zhejiang, China.

Research Papers and Documentation

- Qwen2

Use Cases and Examples

- International development.
- Multilingual mathematics.

Limitations

- Chinese-centric.
- Less Western adoption.

Updates and Variants

June 2024 release.

DeepSeek-V2

Model Name

DeepSeek-V2 by DeepSeek, efficient coding model.

Hosting Providers

- DeepSeek
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2	Accuracy	GSM8K	80.8%
DeepSeek-V2	Pass@1	HumanEval	76.3%
DeepSeek-V2	Accuracy	MGSM	79.4%
DeepSeek-V2	BLEU Score	Efficient Code	65.9

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2	Perplexity	Resource Math	8.2

LLMs Companies Head Office

DeepSeek, headquartered in Hangzhou, Zhejiang, China.

Research Papers and Documentation

- DeepSeek-V2

Use Cases and Examples

- Cost-effective development.
- Efficient programming.

Limitations

- New model.
- Limited global reach.

Updates and Variants

May 2024 release.

Phi-3

Model Name

Phi-3 by Microsoft, lightweight coding model.

Hosting Providers

- Microsoft Azure AI
- Hugging Face Inference Providers

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Phi-3	Accuracy	GSM8K	79.5%
Phi-3	Pass@1	HumanEval	75.1%
Phi-3	Accuracy	MGSM	78.2%
Phi-3	BLEU Score	Small Model Code	64.3
Phi-3	Perplexity	Efficient Math	8.5

LLMs Companies Head Office

Microsoft Corporation, headquartered in Redmond, Washington, USA.

Research Papers and Documentation

- Microsoft Phi-3

Use Cases and Examples

- Edge development.
- Lightweight applications.

Limitations

- Smaller capacity.
- May need fine-tuning.

Updates and Variants

April 2024 release.

Bibliography/Citations

- OpenAI GPT-4
- Anthropic Claude-3
- Meta Llama-3
- Google Gemini-1.5
- Mistral Large
- Cohere Command-R+
- xAI Grok-1
- Qwen2
- DeepSeek-V2
- Microsoft Phi-3
- Custom January 2025 Evaluations (Illustrative)