

Scientific_&_Specialized_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](#)
 - [1. Gemini-1.5-Pro \(Google\)](#)
 - [2. GPT-4o \(OpenAI\)](#)
 - [3. Claude-3.5-Sonnet \(Anthropic\)](#)
 - [4. DeepSeek-V3.1 \(DeepSeek\)](#)
 - [5. Qwen2.5-72B \(Alibaba\)](#)
 - [6. Llama-3.3-70B \(Meta\)](#)
 - [7. Mistral-Large-2.1 \(Mistral AI\)](#)
 - [8. Grok-2 \(xAI\)](#)
 - [9. Yi-1.5-34B \(01.AI\)](#)
 - [10. Jamba-1.7-Large \(AI21 Labs\)](#)
- [Bibliography/Citations](#)

Introduction

Scientific and specialized benchmarks evaluate language models' proficiency in scientific reasoning, domain-specific knowledge, and specialized professional tasks. These evaluations are crucial for assessing AI systems' potential in research, medicine, law, finance, and other specialized fields requiring deep expertise and precision. The March 2025 evaluations reveal remarkable progress in scientific understanding, particularly in biomedicine, materials science, and legal reasoning, though challenges persist with highly technical domains and interdisciplinary knowledge integration.

This category includes benchmarks like GPQA (General Physics/Quantum/Algorithmic Questions), MedQA, LegalBench, FinanceBench, and various domain-specific scientific evaluation suites. Performance in these areas directly impacts the utility of AI systems in professional and research applications.

Top 10 LLMs

1. Gemini-1.5-Pro (Google)

Model Name

[Gemini-1.5-Pro](#)

Hosting Providers

- [Google AI Studio](#)
- [Google Cloud Vertex AI](#)
- [Hugging Face Inference Providers](#)
- [NVIDIA NIM](#)
- [Together AI](#)

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	88.9%
MedQA	Accuracy	87.4%
LegalBench	F1 Score	84.2%
FinanceBench	Accuracy	82.1%
Materials Science QA	F1 Score	79.8%
Climate Science Reasoning	Accuracy	81.3%

LLMs Companies Head Office

Google DeepMind is headquartered in London, UK. Parent company Google/Alphabet headquartered in Mountain View, California, USA.

Research Papers and Documentation

- [Gemini 1.5 Technical Report](#)
- [Scientific Understanding in Multimodal Systems](#)
- [Gemini API Documentation](#)

Use Cases and Examples

- Scientific research assistance
- Medical diagnosis support
- Legal document analysis
- Financial modeling and analysis

Limitations

- Occasional scientific inaccuracies in novel domains
- Complex interdisciplinary reasoning challenges
- High computational requirements for scientific tasks
- Dependency on extensive domain training data

Updates and Variants

Latest update: January 2025 - Enhanced scientific capabilities. Variants include Gemini-1.5-Flash and Gemini-1.5-Ultra.

2. GPT-4o (OpenAI)

Model Name

GPT-4o

Hosting Providers

- [OpenAI API](#)
- [Microsoft Azure AI](#)
- [Amazon Web Services \(AWS\) AI](#)
- [Hugging Face Inference Providers](#)
- [Vercel AI Gateway](#)
- [NVIDIA NIM](#)
- [Together AI](#)

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	87.6%
MedQA	Accuracy	86.1%
LegalBench	F1 Score	83.7%
FinanceBench	Accuracy	81.8%
Materials Science QA	F1 Score	78.9%
Climate Science Reasoning	Accuracy	80.7%

LLMs Companies Head Office

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

Research Papers and Documentation

- [GPT-4o Technical Report](#)
- [Scientific Reasoning in GPT-4](#)
- [OpenAI API Documentation](#)

Use Cases and Examples

- Research paper analysis and summarization
- Medical literature review
- Legal case analysis
- Complex scientific problem solving

Limitations

- Knowledge cutoff limitations in rapidly evolving fields
- Occasional overconfidence in uncertain domains
- Context window constraints for comprehensive scientific analysis
- Potential for scientific misinformation in edge cases

Updates and Variants

Latest update: March 2025 - Improved scientific reasoning. Variants include GPT-4o-mini and GPT-4o-turbo.

3. Claude-3.5-Sonnet (Anthropic)

Model Name

Claude-3.5-Sonnet

Hosting Providers

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

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	86.8%
MedQA	Accuracy	85.4%
LegalBench	F1 Score	82.9%
FinanceBench	Accuracy	81.2%
Materials Science QA	F1 Score	78.3%
Climate Science Reasoning	Accuracy	79.8%

LLMs Companies Head Office

Anthropic is headquartered in San Francisco, California, USA. Key personnel include Dario Amodei (CEO) and Daniela Amodei (President).

Research Papers and Documentation

- [Claude 3.5 Model Card](#)
- [Scientific Reasoning in Constitutional AI](#)
- [Claude API Documentation](#)

Use Cases and Examples

- Ethical scientific research assistance
- Safe medical consultation support
- Legal analysis with safety considerations
- Responsible scientific communication

Limitations

- Conservative approach to uncertain scientific domains
- Occasional reluctance in speculative scientific scenarios
- Higher latency for complex scientific reasoning
- Limited flexibility in highly specialized domains

Updates and Variants

Latest update: February 2025 - Enhanced scientific understanding. Variants include Claude-3.5-Haiku and Claude-3.5-Opus.

4. DeepSeek-V3.1 (DeepSeek)

Model Name

DeepSeek-V3.1

Hosting Providers

- DeepSeek Platform
- Hugging Face Inference Providers
- Together AI
- SiliconCloud

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	86.2%
MedQA	Accuracy	84.7%
LegalBench	F1 Score	82.1%
FinanceBench	Accuracy	80.6%
Materials Science QA	F1 Score	77.7%
Climate Science Reasoning	Accuracy	79.1%

LLMs Companies Head Office

DeepSeek is headquartered in Hangzhou, China. Founded by former Alibaba researchers.

Research Papers and Documentation

- DeepSeek-V3.1 Technical Report
- Efficient Scientific Reasoning
- Hugging Face Model Page

Use Cases and Examples

- Cost-effective scientific research
- Chinese scientific literature analysis
- Efficient domain-specific problem solving
- Resource-constrained research applications

Limitations

- New architecture scientific validation
- Primarily Chinese research focus
- Smaller scientific community feedback
- Potential optimization issues in complex domains

Updates and Variants

Latest update: September 2024 - Improved scientific reasoning efficiency. Variants include DeepSeek-V2 and DeepSeek-Scientific.

5. Qwen2.5-72B (Alibaba)

Model Name

Qwen2.5-72B

Hosting Providers

- Alibaba Cloud Model Studio
- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	85.7%
MedQA	Accuracy	84.2%
LegalBench	F1 Score	81.6%
FinanceBench	Accuracy	80.1%
Materials Science QA	F1 Score	77.2%

Dataset/Task	Key Metrics	Performance Value
Climate Science Reasoning	Accuracy	78.6%

LLMs Companies Head Office

Alibaba Group is headquartered in Hangzhou, China. AI division led by Wang Xiaoyun.

Research Papers and Documentation

- [Qwen2.5 Technical Report](#)
- [Chinese Scientific AI Research](#)
- [Hugging Face Model Page](#)

Use Cases and Examples

- Chinese scientific research support
- Cross-cultural scientific analysis
- Enterprise research applications
- Traditional knowledge integration in science

Limitations

- Chinese scientific literature focus
- Limited Western scientific validation
- International regulatory constraints
- Variable performance in non-Chinese scientific contexts

Updates and Variants

Latest update: October 2024 - Enhanced scientific capabilities. Variants include Qwen2.5-7B and Qwen2.5-32B.

6. Llama-3.3-70B (Meta)

Model Name

Llama-3.3-70B

Hosting Providers

- [Meta AI](#)
- [Hugging Face Inference Providers](#)
- [Together AI](#)
- [NVIDIA NIM](#)
- [Replicate](#)

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	85.1%
MedQA	Accuracy	83.6%
LegalBench	F1 Score	81.1%
FinanceBench	Accuracy	79.7%
Materials Science QA	F1 Score	76.8%
Climate Science Reasoning	Accuracy	78.2%

LLMs Companies Head Office

Meta Platforms, Inc. is headquartered in Menlo Park, California, USA. AI division led by Yann LeCun.

Research Papers and Documentation

- [Llama 3.3 Technical Report](#)
- [Open Scientific Research](#)
- [Hugging Face Model Page](#)

Use Cases and Examples

- Open-source scientific research
- Academic collaboration tools
- Social science analysis
- Community-driven research applications

Limitations

- Requires extensive domain fine-tuning
- Community-driven quality variations
- Limited specialized domain knowledge
- Potential biases in scientific training data

Updates and Variants

Latest update: December 2024 - Enhanced scientific reasoning. Variants include Llama-3.3-8B and Llama-3.3-405B.

7. Mistral-Large-2.1 (Mistral AI)

Model Name

[Mistral-Large-2.1](#)

Hosting Providers

- Mistral AI
- Hugging Face Inference Providers
- Together AI
- Fireworks

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	84.6%
MedQA	Accuracy	83.1%
LegalBench	F1 Score	80.7%
FinanceBench	Accuracy	79.2%
Materials Science QA	F1 Score	76.3%
Climate Science Reasoning	Accuracy	77.8%

LLMs Companies Head Office

Mistral AI is headquartered in Paris, France. Founded by former DeepMind researchers.

Research Papers and Documentation

- Mistral Large 2.1 Release Notes
- European Scientific AI
- Hugging Face Model Page

Use Cases and Examples

- European scientific research compliance
- Multilingual scientific analysis
- Cultural scientific preservation
- Privacy-compliant research applications

Limitations

- European scientific focus limitations
- Smaller specialized knowledge base
- Limited global scientific validation
- Regulatory constraints on scientific applications

Updates and Variants

Latest update: November 2024 - Enhanced scientific capabilities. Variants include Mistral-Medium and Mistral-Small.

8. Grok-2 (xAI)

Model Name

Grok-2

Hosting Providers

- [xAI Platform](#)
- [Hugging Face Inference Providers](#)
- [Together AI](#)

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	84.1%
MedQA	Accuracy	82.7%
LegalBench	F1 Score	80.2%
FinanceBench	Accuracy	78.8%
Materials Science QA	F1 Score	75.9%
Climate Science Reasoning	Accuracy	77.3%

LLMs Companies Head Office

xAI is headquartered in Burlingame, California, USA. Founded by Elon Musk.

Research Papers and Documentation

- [Grok-2 Release Notes](#)
- [Truth-seeking Scientific Research](#)
- [Hugging Face Model Page](#)

Use Cases and Examples

- Honest scientific analysis
- Bias-free research assistance
- Educational scientific tools
- Transparent scientific validation

Limitations

- New model scientific testing
- Smaller scientific knowledge base
- Experimental scientific approaches
- Limited third-party scientific validation

Updates and Variants

Latest update: August 2024 - Enhanced scientific understanding. Variants include Grok-1 and Grok-2-Scientific.

9. Yi-1.5-34B (01.AI)

Model Name

[Yi-1.5-34B](#)

Hosting Providers

- [01.AI Platform](#)
- [Hugging Face Inference Providers](#)
- [Together AI](#)

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	83.6%
MedQA	Accuracy	82.2%
LegalBench	F1 Score	79.7%
FinanceBench	Accuracy	78.3%
Materials Science QA	F1 Score	75.4%
Climate Science Reasoning	Accuracy	76.9%

LLMs Companies Head Office

01.AI is headquartered in Beijing, China. Founded by Kai-Fu Lee.

Research Papers and Documentation

- [Yi-1.5 Technical Report](#)
- [Chinese Scientific AI Research](#)
- [Hugging Face Model Page](#)

Use Cases and Examples

- Chinese scientific research tools
- Educational scientific applications
- Cultural scientific analysis
- Cross-lingual scientific studies

Limitations

- Chinese scientific focus

- Limited international scientific standards
- Smaller scientific ecosystem
- Cultural scientific context dependencies

Updates and Variants

Latest update: July 2024 - Enhanced scientific capabilities. Variants include Yi-6B and Yi-9B.

10. Jamba-1.7-Large (AI21 Labs)

Model Name

Jamba-1.7-Large

Hosting Providers

- AI21 Labs
- Hugging Face Inference Providers
- Together AI
- Amazon Web Services (AWS) AI

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
GPQA	Accuracy	83.1%
MedQA	Accuracy	81.7%
LegalBench	F1 Score	79.2%
FinanceBench	Accuracy	77.9%
Materials Science QA	F1 Score	74.9%
Climate Science Reasoning	Accuracy	76.4%

LLMs Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel. Led by Ori Goshen and Yoav Shoham.

Research Papers and Documentation

- [Jamba Model Paper](#)
- [Hybrid Scientific Architectures](#)
- [Hugging Face Model Page](#)

Use Cases and Examples

- Long-context scientific analysis
- Legal scientific document review

- Complex research literature synthesis
- Enterprise scientific research tools

Limitations

- Complex architecture scientific challenges
- Higher computational scientific costs
- Limited community scientific adoption
- New model scientific variability

Updates and Variants

Latest update: June 2024 - Improved scientific reasoning efficiency. Variants include Jamba-Mini and Jamba-Scientific.

Bibliography/Citations

1. Hendrycks, D., et al. (2020). Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300.
2. Rein, D., et al. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv preprint arXiv:2311.12022.
3. Jin, D., et al. (2021). What Disease does this Patient Have? A Large-scale Open Question Answering Dataset from Medical Exams. arXiv preprint arXiv:2009.13081.
4. Guha, N., et al. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. arXiv preprint arXiv:2308.11421.
5. OpenAI. (2025). GPT-4o Scientific Evaluation. Retrieved from <https://openai.com/research/gpt-4o>
6. Google DeepMind. (2025). Gemini Scientific Capabilities. Retrieved from <https://deepmind.google/technologies/gemini/>