# 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

Scientific and specialized benchmarks evaluate language models' ability to understand and generate content in specialized domains such as medicine, law, finance, and technical fields. These benchmarks test models on tasks requiring domain-specific knowledge, terminology understanding, and expert-level reasoning. In January 2025, this category highlighted significant advancements in models capable of handling complex scientific literature and specialized knowledge bases, with improved performance on datasets like ARC-Challenge, scientific QA tasks, and domain-specific evaluations. The evaluation period saw a focus on models' capacity for accurate information retrieval and synthesis in specialized fields, which is crucial for applications in research assistance, professional services, and expert systems. Leading models excelled in integrating specialized knowledge with general reasoning capabilities.

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, strong in scientific and specialized domains.

#### 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 | ARC-Challenge | 96.2% |
| GPT-4 | F1 Score | Scientific QA | 91.7% |
| GPT-4 | Accuracy | Legal Reasoning | 89.3% |
| GPT-4 | BLEU Score | Technical Writing | 78.9 |
| GPT-4 | Perplexity | Domain Knowledge | 5.1 |

**LLMs Companies Head Office**

OpenAI, headquartered in San Francisco, California, USA.

**Research Papers and Documentation**

- OpenAI GPT-4

**Use Cases and Examples**

- Scientific research assistance.
- Legal document analysis.

**Limitations**

- High computational requirements.
- Occasional factual errors.

#### Updates and Variants

March 2023 release.

### Claude-3

#### Model Name

Claude-3 by Anthropic, focused on reliable specialized knowledge.

#### Hosting Providers

- Anthropic
- Amazon Web Services (AWS) AI

#### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Claude-3 | Accuracy | ARC-Challenge | 95.8% |
| Claude-3 | F1 Score | Scientific QA | 90.4% |
| Claude-3 | Accuracy | Legal Reasoning | 88.7% |
| Claude-3 | BLEU Score | Safe Technical | 77.2 |
| Claude-3 | Perplexity | Ethical Domains | 5.6 |

#### LLMs Companies Head Office

Anthropic, headquartered in San Francisco, California, USA.

#### Research Papers and Documentation

- Anthropic Claude-3

#### Use Cases and Examples

- Medical research.
- Ethical expert systems.

#### Limitations

- Slower inference.
- Limited customization.

#### Updates and Variants

March 2024 release.

## Llama-3

### Model Name

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

### Hosting Providers

- Meta AI
- Hugging Face Inference Providers

### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Llama-3 | Accuracy | ARC-Challenge | 91.4% |
| Llama-3 | F1 Score | Scientific QA | 86.9% |
| Llama-3 | Accuracy | Legal Reasoning | 84.2% |
| Llama-3 | BLEU Score | Open Technical | 73.6 |
| Llama-3 | Perplexity | Research Domains | 6.8 |

### LLMs Companies Head Office

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

### Research Papers and Documentation

- Meta Llama-3

### Use Cases and Examples

- Academic research.
- Open-domain expertise.

### Limitations

- Requires fine-tuning.
- Potential biases.

### Updates and Variants

April 2024 release.

## Gemini-1.5

### Model Name

Gemini-1.5 by Google, multimodal specialized capabilities.

### Hosting Providers

- Google Cloud Vertex AI
- Google AI Studio

### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Gemini-1.5 | Accuracy | ARC-Challenge | 93.7% |
| Gemini-1.5 | F1 Score | Scientific QA | 88.9% |
| Gemini-1.5 | Accuracy | Legal Reasoning | 86.5% |
| Gemini-1.5 | BLEU Score | Multimodal Technical | 75.8 |
| Gemini-1.5 | Perplexity | Visual Domains | 6.2 |

### LLMs Companies Head Office

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

### Research Papers and Documentation

- Google Gemini-1.5

### Use Cases and Examples

- Scientific visualization.
- Technical documentation.

### Limitations

- High resource demands.
- Ongoing development.

### Updates and Variants

December 2023 release.

## Mistral-Large

### Model Name

Mistral-Large by Mistral AI, efficient specialized model.

### Hosting Providers

- Mistral AI
- Hugging Face Inference Providers

### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Mistral-Large | Accuracy | ARC-Challenge | 90.1% |
| Mistral-Large | F1 Score | Scientific QA | 85.6% |
| Mistral-Large | Accuracy | Legal Reasoning | 82.9% |
| Mistral-Large | BLEU Score | Efficient Technical | 72.3 |
| Mistral-Large | Perplexity | Fast Domains | 7.1 |

### LLMs Companies Head Office

Mistral AI, headquartered in Paris, France.

### Research Papers and Documentation

- Mistral Large

### Use Cases and Examples

- European research.
- Resource-efficient expertise.

### Limitations

- Newer model.
- Limited multimodal.

### Updates and Variants

February 2024 release.

## Command-R+

### Model Name

Command-R+ by Cohere, enterprise specialized focus.

### Hosting Providers

- Cohere
- Hugging Face Inference Providers

#### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Command-R+ | Accuracy | ARC-Challenge | 88.9% |
| Command-R+ | F1 Score | Scientific QA | 84.3% |
| Command-R+ | Accuracy | Legal Reasoning | 81.6% |
| Command-R+ | BLEU Score | Enterprise Technical | 71.1 |
| Command-R+ | Perplexity | Business Domains | 7.4 |

#### LLMs Companies Head Office

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

#### Research Papers and Documentation

- Cohere Command-R+

#### Use Cases and Examples

- Corporate research.
- Professional services.

#### Limitations

- API-dependent.
- English-focused.

#### Updates and Variants

March 2024 release.

### Grok-1

#### Model Name

Grok-1 by xAI, creative specialized reasoning.

#### 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 | ARC-Challenge | 87.6% |
| Grok-1 | F1 Score | Scientific QA | 83.1% |
| Grok-1 | Accuracy | Legal Reasoning | 80.4% |
| Grok-1 | BLEU Score | Creative Technical | 69.7 |
| Grok-1 | Perplexity | Novel Domains | 7.7 |

#### LLMs Companies Head Office

xAI, headquartered in Burlingame, California, USA.

#### Research Papers and Documentation

- xAI Grok-1

#### Use Cases and Examples

- Innovative research.
- Creative problem-solving.

#### Limitations

- Relatively new.
- Limited fine-tuning.

#### Updates and Variants

November 2023 release.

### Qwen-2

#### Model Name

Qwen-2 by Alibaba, multilingual specialized model.

#### Hosting Providers

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

#### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Qwen-2 | Accuracy | ARC-Challenge | 86.3% |
| Qwen-2 | F1 Score | Scientific QA | 81.8% |

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Qwen-2 | Accuracy | Legal Reasoning | 79.2% |
| Qwen-2 | BLEU Score | Multilingual Technical | 68.4 |
| Qwen-2 | Perplexity | Global Domains | 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 research.
- Global expertise.

#### Limitations

- Chinese-centric.
- Less Western adoption.

#### Updates and Variants

June 2024 release.

### DeepSeek-V2

#### Model Name

DeepSeek-V2 by DeepSeek, efficient specialized model.

#### Hosting Providers

- DeepSeek
- Hugging Face Inference Providers

#### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| DeepSeek-V2 | Accuracy | ARC-Challenge | 85.1% |
| DeepSeek-V2 | F1 Score | Scientific QA | 80.6% |
| DeepSeek-V2 | Accuracy | Legal Reasoning | 78.1% |
| DeepSeek-V2 | BLEU Score | Efficient Technical | 67.2 |

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

#### LLMs Companies Head Office

DeepSeek, headquartered in Hangzhou, Zhejiang, China.

#### Research Papers and Documentation

- DeepSeek-V2

#### Use Cases and Examples

- Cost-effective research.
- Efficient expertise.

#### Limitations

- New model.
- Limited global reach.

#### Updates and Variants

May 2024 release.

### Phi-3

#### Model Name

Phi-3 by Microsoft, lightweight specialized model.

#### Hosting Providers

- Microsoft Azure AI
- Hugging Face Inference Providers

#### Benchmarks Evaluation

| Model Name | Key Metrics | Dataset/Task | Performance Value |
|---|---|---|---|
| Phi-3 | Accuracy | ARC-Challenge | 83.9% |
| Phi-3 | F1 Score | Scientific QA | 79.4% |
| Phi-3 | Accuracy | Legal Reasoning | 77.1% |
| Phi-3 | BLEU Score | Small Model Technical | 65.8 |
| Phi-3 | Perplexity | Efficient Domains | 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 research.
- Lightweight expertise.

### 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)