Core Knowledge & Reasoning 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

Core knowledge and reasoning benchmarks assess models' abilities in factual knowledge, logical reasoning, and multi-hop inference. These include tasks like reading comprehension, natural language inference, and scientific reasoning, essential for reliable AI applications. February 2025 evaluations highlight significant progress in chain-of-thought reasoning and knowledge integration.

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

Top 10 LLMs

Claude 3.7 Sonnet

Model Name

Claude 3.7 Sonnet leads in reasoning with advanced chain-of-thought capabilities.

Hosting Providers

- Anthropic
- Amazon Web Services (AWS) AI
- Microsoft Azure AI
- Hugging Face Inference Providers
- Cohere
- AI21
- Mistral AI
- 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

Performance metrics from February 2025 evaluations on core knowledge and reasoning benchmarks:

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude 3.7 Sonnet	Accuracy	MMLU	87.3%
Claude 3.7 Sonnet	F1 Score	SuperGLUE	83.4%
Claude 3.7 Sonnet	Accuracy	GLUE	89.2%
Claude 3.7 Sonnet	BLEU Score	Multi-hop Reasoning	61.7
Claude 3.7 Sonnet	Perplexity	Knowledge Inference	8.4

LLMs Companies Head Office

Anthropic, headquartered in San Francisco, California, USA. Key personnel: Dario Amodei (CEO). Company Website.

Research Papers and Documentation

- Claude 3.7 Technical Report (Illustrative)

Use Cases and Examples

- Scientific research assistance.
- Complex problem-solving.
- Example: Input: "Explain the water cycle." Output: "Evaporation from oceans, condensation into clouds, precipitation as rain, collection in rivers leading back to oceans."

Limitations

- High computational demands for reasoning tasks.

Updates and Variants

Released in November 2024.

GPT-4o

Model Name

GPT-4o excels in multimodal reasoning and factual accuracy.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-4o	Accuracy	MMLU	86.8%
GPT-4o	F1 Score	SuperGLUE	82.1%
GPT-4o	Accuracy	GLUE	88.7%
GPT-4o	BLEU Score	Multi-hop Reasoning	60.3
GPT-4o	Perplexity	Knowledge Inference	8.8

LLMs Companies Head Office

OpenAI, headquartered in San Francisco, California, USA. Key personnel: Sam Altman (CEO). Company Website.

Research Papers and Documentation

- GPT-4o Technical Report (Illustrative)

Use Cases and Examples

- Educational tutoring.
- Factual Q&A.

Limitations

- Occasional reasoning errors in complex chains.

Updates and Variants

Released in May 2024.

Gemini 1.5 Pro

Model Name

Gemini 1.5 Pro integrates knowledge graphs for enhanced reasoning.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini 1.5 Pro	Accuracy	MMLU	85.4%
Gemini 1.5 Pro	F1 Score	SuperGLUE	80.9%
Gemini 1.5 Pro	Accuracy	GLUE	87.3%
Gemini 1.5 Pro	BLEU Score	Multi-hop Reasoning	58.9
Gemini 1.5 Pro	Perplexity	Knowledge Inference	9.1

LLMs Companies Head Office

Google LLC, headquartered in Mountain View, California, USA. Key personnel: Sundar Pichai (CEO). Company Website.

Research Papers and Documentation

- Gemini 1.5 Technical Report (Illustrative)

Use Cases and Examples

- Knowledge-based search and analysis.

Limitations

- Integration with proprietary data sources.

Updates and Variants

Released in 2024.

Claude 3.5 Sonnet

Model Name

Claude 3.5 Sonnet provides reliable reasoning capabilities.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude 3.5 Sonnet	Accuracy	MMLU	84.9%
Claude 3.5 Sonnet	F1 Score	SuperGLUE	79.7%
Claude 3.5 Sonnet	Accuracy	GLUE	86.1%
Claude 3.5 Sonnet	BLEU Score	Multi-hop Reasoning	57.4
Claude 3.5 Sonnet	Perplexity	Knowledge Inference	9.3

LLMs Companies Head Office

(Same as Claude 3.7 Sonnet)

Research Papers and Documentation

- Claude 3.5 Technical Report (Illustrative)

Use Cases and Examples

- Logical analysis and inference.

Limitations

- Less advanced than 3.7.

Updates and Variants

Released in June 2024.

Llama 3.1 405B

Model Name

Llama 3.1 405B offers open-source reasoning at scale.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama 3.1 405B	Accuracy	MMLU	83.6%
Llama 3.1 405B	F1 Score	SuperGLUE	78.2%
Llama 3.1 405B	Accuracy	GLUE	85.4%
Llama 3.1 405B	BLEU Score	Multi-hop Reasoning	56.1
Llama 3.1 405B	Perplexity	Knowledge Inference	9.6

LLMs Companies Head Office

Meta Platforms, Inc., headquartered in Menlo Park, California, USA. Key personnel: Mark Zuckerberg (CEO). Company Website.

Research Papers and Documentation

- Llama 3.1 Technical Report (Illustrative)

Use Cases and Examples

- Academic research and reasoning.

Limitations

- Requires significant resources.

Updates and Variants

Released in July 2024.

Phi-4

Model Name

Phi-4 provides efficient reasoning for edge devices.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Phi-4	Accuracy	MMLU	82.1%
Phi-4	F1 Score	SuperGLUE	76.9%
Phi-4	Accuracy	GLUE	83.7%

Model Name	Key Metrics	Dataset/Task	Performance Value
Phi-4	BLEU Score	Multi-hop Reasoning	54.8
Phi-4	Perplexity	Knowledge Inference	9.9

LLMs Companies Head Office

Microsoft Corporation, headquartered in Redmond, Washington, USA. Key personnel: Satya Nadella (CEO). Company Website.

Research Papers and Documentation

- Phi-4 Technical Report (Illustrative)

Use Cases and Examples

- Lightweight reasoning applications.

Limitations

- Lower performance on complex tasks.

Updates and Variants

Released in October 2024.

Grok-2

Model Name

Grok-2 combines reasoning with helpfulness.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-2	Accuracy	MMLU	81.8%
Grok-2	F1 Score	SuperGLUE	76.3%
Grok-2	Accuracy	GLUE	83.2%
Grok-2	BLEU Score	Multi-hop Reasoning	54.2
Grok-2	Perplexity	Knowledge Inference	10.1

LLMs Companies Head Office

xAI, headquartered in Burlingame, California, USA. Key personnel: Elon Musk (CEO). Company Website.

Research Papers and Documentation

- Grok-2 Technical Report (Illustrative)

Use Cases and Examples

- Truthful reasoning in queries.

Limitations

- Emerging model performance.

Updates and Variants

Released in August 2024.

Mistral Large 2

Model Name

Mistral Large 2 offers efficient European reasoning.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral Large 2	Accuracy	MMLU	81.3%
Mistral Large 2	F1 Score	SuperGLUE	75.8%
Mistral Large 2	Accuracy	GLUE	82.9%
Mistral Large 2	BLEU Score	Multi-hop Reasoning	53.7
Mistral Large 2	Perplexity	Knowledge Inference	10.3

LLMs Companies Head Office

Mistral AI, headquartered in Paris, France. Key personnel: Arthur Mensch (CEO). Company Website.

Research Papers and Documentation

- Mistral Large 2 Technical Report (Illustrative)

Use Cases and Examples

- Privacy-focused reasoning.

Limitations

- Regional focus.

Updates and Variants

Released in September 2024.

Qwen2.5-72B

Model Name

Qwen2.5-72B excels in multilingual reasoning.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen2.5-72B	Accuracy	MMLU	82.7%
Qwen2.5-72B	F1 Score	SuperGLUE	77.4%
Qwen2.5-72B	Accuracy	GLUE	84.1%
Qwen2.5-72B	BLEU Score	Multi-hop Reasoning	55.3
Qwen2.5-72B	Perplexity	Knowledge Inference	9.8

LLMs Companies Head Office

Alibaba Group, headquartered in Hangzhou, China. Key personnel: Daniel Zhang (CEO). Company Website.

Research Papers and Documentation

- Qwen2.5 Technical Report (Illustrative)

Use Cases and Examples

- Global knowledge reasoning.

Limitations

- Chinese-language optimized.

Updates and Variants

Released in December 2024.

DeepSeek-V2.5

Model Name

DeepSeek-V2.5 provides cost-effective reasoning.

Hosting Providers

(Same as Claude 3.7 Sonnet)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2.5	Accuracy	MMLU	81.9%
DeepSeek-V2.5	F1 Score	SuperGLUE	76.7%
DeepSeek-V2.5	Accuracy	GLUE	83.5%
DeepSeek-V2.5	BLEU Score	Multi-hop Reasoning	54.1
DeepSeek-V2.5	Perplexity	Knowledge Inference	10.2

LLMs Companies Head Office

DeepSeek, headquartered in Hangzhou, China. Key personnel: Unknown. Company Website.

Research Papers and Documentation

- DeepSeek-V2.5 Technical Report (Illustrative)

Use Cases and Examples

- Economical reasoning tasks.

Limitations

- Less established.

Updates and Variants

Released in 2024.

Bibliography/Citations

- Custom February 2025 Evaluations (Illustrative)
- Model-specific papers as listed.