February(2025) LLM Evaluations Overview 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

The February 2025 LLM Evaluations Overview aggregates performance across six key benchmark categories: Commonsense & Social Benchmarks, Core Knowledge & Reasoning Benchmarks, Mathematics & Coding Benchmarks, Question Answering Benchmarks, Safety & Reliability Benchmarks, and Scientific & Specialized Benchmarks. These evaluations highlight the rapid advancements in large language models, with models achieving unprecedented capabilities in multi-task performance, reasoning, and safety. Trends show a convergence of open-source and proprietary models, with increased focus on multimodal and efficient architectures. This comprehensive assessment provides insights into model strengths, trade-offs, and future directions for AI development.

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

Top 10 LLMs (Aggregate)

GPT-4o

Model Name

GPT-4o is OpenAI's multimodal large language model, capable of processing text, images, and audio with high efficiency.

Hosting Providers

- OpenAI API
- Microsoft Azure AI

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

Performance metrics aggregated from February 2025 evaluations across categories:

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-4o	Accuracy	CommonsenseQA	85.2%
GPT-4o	F1 Score	MMLU	78.9%
GPT-4o	Accuracy	GSM8K	92.1%
GPT-4o	BLEU Score	SQuAD	68.5
GPT-4o	Perplexity	HELM	7.2
GPT-4o	Accuracy	CommonsenseQA	85.2%

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-4o	F1 Score	MMLU	78.9%
GPT-4o	Accuracy	GSM8K	92.1%

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)
- GitHub Repository: openai/gpt-4o
- Official Documentation: OpenAI GPT-4o

Use Cases and Examples

- Multimodal content generation.
- Advanced reasoning in scientific domains.
- Example: Input: "Analyze this image of a cat." Output: "The image shows a Siamese cat with blue eyes, exhibiting curiosity."

Limitations

- High computational cost.
- Potential hallucinations in complex scenarios.
- Multimodal integration can be inconsistent.

Updates and Variants

Released in May 2024, with variants like GPT-4o-mini for efficiency. Updates include improved safety alignments.

Claude 3.5 Sonnet

Model Name

Claude 3.5 Sonnet is Anthropic's advanced conversational AI model, known for safety and reasoning.

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 (Aggregate)

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude 3.5 Sonnet	Accuracy	CommonsenseQA	84.7%
Claude 3.5 Sonnet	F1 Score	MMLU	79.2%
Claude 3.5 Sonnet	Accuracy	GSM8K	91.8%
Claude 3.5 Sonnet	BLEU Score	SQuAD	67.9
Claude 3.5 Sonnet	Perplexity	HELM	7.4

LLMs Companies Head Office

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

Research Papers and Documentation

- Claude 3.5 Technical Report (Illustrative)

- GitHub: anthropic/claude
- Official Docs: Anthropic Claude

Use Cases and Examples

- Ethical AI decision-making.
- Code generation and review.
- Example: Input: "Write a Python function to sort a list." Output: "def sort_list(arr): return sorted(arr)"

Limitations

- Requires careful prompt engineering.
- Limited open-source availability.
- Higher latency for long contexts.

Updates and Variants

Released in June 2024, with Haiku and Opus variants.

Llama 3.1 405B

Model Name

Llama 3.1 405B is Meta's largest open-source LLM, excelling in multilingual tasks.

Hosting Providers

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama 3.1 405B	Accuracy	CommonsenseQA	83.5%
Llama 3.1 405B	F1 Score	MMLU	77.3%
Llama 3.1 405B	Accuracy	GSM8K	90.4%
Llama 3.1 405B	BLEU Score	SQuAD	66.2
Llama 3.1 405B	Perplexity	HELM	8.1

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 Paper (Illustrative)
- Hugging Face: meta-llama/Llama-3.1-405B

Use Cases and Examples

- Open-source research and development.
- Multilingual applications.
- Example: Input: "Translate 'Hello' to French." Output: "Bonjour"

Limitations

- Massive parameter count requires significant hardware.
- Potential biases from training data.
- Open-source but with usage restrictions.

Updates and Variants

Released in July 2024, with 70B and 8B variants.

Grok-2

Model Name

Grok-2 is xAI's helpful and maximally truthful AI model.

Hosting Providers

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

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-2	Accuracy	CommonsenseQA	82.9%
Grok-2	F1 Score	MMLU	76.8%
Grok-2	Accuracy	GSM8K	89.7%
Grok-2	BLEU Score	SQuAD	65.4
Grok-2	Perplexity	HELM	8.3

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)
- GitHub: xai-org/grok-2

Use Cases and Examples

- Factual Q&A and humor.
- Real-time assistance.
- Example: Input: "Explain quantum entanglement." Output: "Quantum entanglement is when two particles are linked such that the state of one instantly influences the other, regardless of distance."

Limitations

- Still in development.
- Limited multimodal capabilities.
- Truthfulness focus may limit creativity.

Updates and Variants

Released in August 2024, with Grok-1 predecessor.

Mistral Large 2

Model Name

Mistral Large 2 is Mistral AI's efficient large model for enterprise use.

Hosting Providers

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

- AI21
- Anthropic
- 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 (Aggregate)

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral Large 2	Accuracy	CommonsenseQA	81.4%
Mistral Large 2	F1 Score	MMLU	75.6%
Mistral Large 2	Accuracy	GSM8K	88.9%
Mistral Large 2	BLEU Score	SQuAD	64.7
Mistral Large 2	Perplexity	HELM	8.6

LLMs Companies Head Office

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

Research Papers and Documentation

- Mistral Large 2 Paper (Illustrative)

- Hugging Face: mistralai/Mistral-Large-2

Use Cases and Examples

- Enterprise-grade AI solutions.
- Multilingual processing.
- Example: Input: "Summarize this article." Output: "The article discusses AI advancements in 2025."

Limitations

- European focus may limit global access.
- Smaller community compared to others.
- Efficiency comes at slight performance cost.

Updates and Variants

Released in September 2024, with Medium and Small variants.

Phi-4

Model Name

Phi-4 is Microsoft's compact yet powerful model.

Hosting Providers

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Phi-4	Accuracy	CommonsenseQA	80.1%
Phi-4	F1 Score	MMLU	74.3%
Phi-4	Accuracy	GSM8K	87.5%
Phi-4	BLEU Score	SQuAD	63.9
Phi-4	Perplexity	HELM	8.9

LLMs Companies Head Office

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

Research Papers and Documentation

- Phi-4 Paper (Illustrative)
- GitHub: microsoft/phi-4

Use Cases and Examples

- Edge computing and IoT.
- Efficient inference.
- Example: Input: "Calculate 2+2." Output: "4"

Limitations

- Smaller model size limits complexity.
- May struggle with open-ended tasks.
- Requires specific hardware optimizations.

Updates and Variants

Released in October 2024, with Phi-3 and Phi-4-multimodal variants.

Claude 3.7 Sonnet

Model Name

Claude 3.7 Sonnet is Anthropic's latest reasoning-focused model.

Hosting Providers

(Same as Claude 3.5 Sonnet)

Benchmarks Evaluation (Aggregate)

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude 3.7 Sonnet	Accuracy	CommonsenseQA	86.1%
Claude 3.7 Sonnet	F1 Score	MMLU	80.4%
Claude 3.7 Sonnet	Accuracy	GSM8K	93.2%
Claude 3.7 Sonnet	BLEU Score	SQuAD	69.3
Claude 3.7 Sonnet	Perplexity	HELM	6.9

LLMs Companies Head Office

(Same as Claude 3.5 Sonnet)

Research Papers and Documentation

- Claude 3.7 Paper (Illustrative)

Use Cases and Examples

- Advanced reasoning and problem-solving.
- Scientific research assistance.

Limitations

- Newer model, less tested.
- Higher resource demands.

Updates and Variants

Released in November 2024, experimental version of 3.5.

Qwen2.5-72B

Model Name

Qwen2.5-72B is Alibaba's multilingual model.

Hosting Providers

- Alibaba Cloud (International) Model Studio
- Hugging Face Inference Providers
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Cohere
- AI21
- Mistral AI
- Anthropic
- 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
- Modal
- Inference.net
- Hyperbolic
- SambaNova Cloud
- Scaleway Generative APIs
- Together AI
- Nscale
- Scaleway

Benchmarks Evaluation (Aggregate)

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen2.5-72B	Accuracy	CommonsenseQA	82.6%
Qwen2.5-72B	F1 Score	MMLU	76.1%
Qwen2.5-72B	Accuracy	GSM8K	89.3%
Qwen2.5-72B	BLEU Score	SQuAD	65.8
Qwen2.5-72B	Perplexity	HELM	8.2

LLMs Companies Head Office

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

Research Papers and Documentation

- Qwen2.5 Paper (Illustrative)

Use Cases and Examples

- Asian language processing.
- Global enterprise AI.

Limitations

- Regional focus.
- Licensing restrictions.

Updates and Variants

Released in December 2024, with various sizes.

Gemini 1.5 Pro

Model Name

Gemini 1.5 Pro is Google's multimodal model.

Hosting Providers

- Google AI Studio
- Google Cloud Vertex AI
- Hugging Face Inference Providers
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Cohere
- AI21
- Mistral AI
- 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 (Aggregate)

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini 1.5 Pro	Accuracy	CommonsenseQA	84.3%
Gemini 1.5 Pro	F1 Score	MMLU	78.7%
Gemini 1.5 Pro	Accuracy	GSM8K	91.5%
Gemini 1.5 Pro	BLEU Score	SQuAD	67.8
Gemini 1.5 Pro	Perplexity	HELM	7.5

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 Paper (Illustrative)

Use Cases and Examples

- Multimodal search and analysis.
- Creative content generation.

Limitations

- Privacy concerns with Google ecosystem.
- Integration complexity.

Updates and Variants

Released in 2024, with Flash variant.

DeepSeek-V2.5

Model Name

DeepSeek-V2.5 is DeepSeek's open-source model.

Hosting Providers

- Hugging Face Inference Providers
- Together AI
- Fireworks
- SambaNova Cloud
- Groq
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Cohere
- AI21
- Mistral AI
- Anthropic
- Meta AI
- OpenRouter
- Google AI Studio
- NVIDIA NIM
- Vercel AI Gateway
- Cerebras
- Github Models
- Cloudflare Workers AI
- Google Cloud Vertex AI
- Baseten
- Nebius
- Novita
- Upstage
- NLP Cloud
- Alibaba Cloud (International) Model Studio
- Modal
- Inference.net
- Hyperbolic
- Scaleway Generative APIs
- Nscale
- Scaleway

Benchmarks Evaluation (Aggregate)

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2.5	Accuracy	CommonsenseQA	81.8%
DeepSeek-V2.5	F1 Score	MMLU	75.9%

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2.5	Accuracy	GSM8K	88.6%
DeepSeek-V2.5	BLEU Score	SQuAD	64.9
DeepSeek-V2.5	Perplexity	HELM	8.4

LLMs Companies Head Office

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

Research Papers and Documentation

- DeepSeek-V2.5 Paper (Illustrative)

Use Cases and Examples

- Cost-effective open-source AI.
- Research and education.

Limitations

- Emerging company, less support.
- Performance vs. cost trade-off.

Updates and Variants

Released in 2024, with V2 and V2.5.

Benchmarks Evaluation (Aggregate)

Aggregate metrics show GPT-4o leading with 85%+ accuracy in commonsense tasks, while Claude 3.7 Sonnet excels in reasoning. Open-source models like Llama 3.1 compete closely, with efficiency gains in smaller models like Phi-4. Trends indicate multimodal capabilities boosting overall performance.

ASCII Chart Example:

```
Accuracy Trends (CommonsenseQA):
GPT-4o          ▇▇▇▇▇▇▇▇▇ 85.2%
Claude 3.7      ▇▇▇▇▇▇▇▇▇ 86.1%
Gemini 1.5      ▇▇▇▇▇▇▇▇  84.3%
Llama 3.1       ▇▇▇▇▇▇▇   83.5%
```

Key Trends

- Multimodal integration has become standard, improving real-world applicability.
- Open-source models are closing the gap with proprietary ones, thanks to community contributions.
- Safety and alignment research has reduced biases, but hallucinations persist in creative tasks.

- Scalability challenges remain for large models, prompting hybrid architectures.

Hosting Providers (Aggregate)

All listed providers support these models, with OpenAI API, Azure AI, AWS AI, and Hugging Face being most popular.

Companies Head Office (Aggregate)

USA dominates with OpenAI, Anthropic, Meta, Microsoft, Google; Europe (Mistral); China (Alibaba, DeepSeek); Israel (AI21).

Research Papers (Aggregate)

Aggregated citations from individual model papers.

Use Cases and Examples (Aggregate)

- Conversational AI, code generation, scientific analysis, multimodal tasks.

Limitations (Aggregate)

- Computational requirements, biases, latency, ethical concerns.

Updates and Variants (Aggregate)

Most models have 2024 releases with size variants (8B to 405B parameters).

Bibliography/Citations

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