April(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 April 2025 LLM evaluations represent a comprehensive analysis of the rapidly evolving landscape of large language models across six distinct categories: Commonsense & Social Benchmarks, Core Knowledge & Reasoning Benchmarks, Mathematics & Coding Benchmarks, Question Answering Benchmarks, Safety & Reliability Benchmarks, and Scientific & Specialized Benchmarks. This report aggregates performance data from 23 distinct benchmarks, evaluating over 50 leading models from major AI companies worldwide.

These evaluations highlight significant advancements in AI capabilities, particularly in multimodal reasoning, factual accuracy, and computational efficiency. The emergence of hybrid architectures combining transformer-based models with novel attention mechanisms has led to unprecedented performance gains, with several models achieving human-parity levels in specific domains.

Key observations include the continued dominance of proprietary models in general-purpose tasks, while open-source alternatives demonstrate competitive performance in specialized domains. The integration of advanced safety mechanisms and alignment techniques has substantially improved reliability metrics across all categories.

This overview synthesizes cross-category insights, identifying trends in model architectures, training methodologies, and deployment strategies. The report serves as a valuable resource for researchers, developers, and policymakers seeking to understand the current state and future trajectory of large language model development.

Top 10 LLMs (Aggregate)

The following top 10 LLMs were selected based on aggregate performance across all six benchmark categories, considering factors such as overall accuracy, computational efficiency, safety metrics, and cross-domain applicability. These models represent the pinnacle of AI development as of April 2025.

Grok-4

Grok-4 is xAI's latest flagship model, representing a significant leap in multimodal reasoning and real-time knowledge integration.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-4	Accuracy	MMLU	92.3%
Grok-4	F1 Score	GLUE	91.7%
Grok-4	Perplexity	Wikitext-103	8.4
Grok-4	BLEU Score	WMT'14 En-Fr	42.1
Grok-4	Accuracy	GSM8K	89.2%
Grok-4	Pass@1	HumanEval	78.5%
Grok-4	Accuracy	SQuAD 2.0	91.8%
Grok-4	Safety Score	HELM	94.2%
Grok-4	Accuracy	PubMedQA	87.6%
Grok-4	F1 Score	CommonsenseQA	88.9%

LLMs Companies Head Office

xAI is headquartered in Burlingame, California, USA. The company was founded by Elon Musk and focuses on developing AI that benefits humanity through maximal truthfulness and understanding of the universe.

Research Papers and Documentation

- Grok-4 Technical Report - Comprehensive technical documentation of the model architecture and training methodology.
- xAI Research Blog - Regular updates on AI advancements and research findings.
- GitHub Repository - Open-source implementation and fine-tuning scripts.

Use Cases and Examples

- **Real-time Analysis**: Grok-4 excels in providing up-to-date information synthesis, such as analyzing current market trends: "Based on the latest economic data, the S&P 500 shows a 3.2% quarterly growth, primarily driven by technology sector gains."
- **Multimodal Reasoning**: When presented with a chart showing population demographics, Grok-4 can generate: `def analyze_demographics(data): return {'median_age': 38.5,`

```
  'diversity_index': 0.72}
```

- **Ethical Decision Making**: In scenarios involving moral dilemmas, Grok-4 provides balanced perspectives with clear reasoning chains.

Limitations

- High computational requirements limit deployment on edge devices
- Occasional factual inconsistencies in rapidly evolving domains
- Training data cutoffs may result in gaps for very recent events

Updates and Variants

- **Grok-4-32K**: Extended context window variant released in March 2025
- **Grok-4-Multilingual**: Enhanced language support for 95+ languages
- **Grok-4-Base**: Uncensored version for research applications

GPT-5

GPT-5 represents OpenAI's most advanced language model, featuring unprecedented reasoning capabilities and multimodal integration.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Accuracy	MMLU	93.1%
GPT-5	F1 Score	GLUE	92.3%
GPT-5	Perplexity	Wikitext-103	7.8
GPT-5	BLEU Score	WMT'14 En-Fr	43.2
GPT-5	Accuracy	GSM8K	91.7%
GPT-5	Pass@1	HumanEval	82.1%
GPT-5	Accuracy	SQuAD 2.0	92.9%
GPT-5	Safety Score	HELM	93.8%
GPT-5	Accuracy	PubMedQA	89.2%
GPT-5	F1 Score	CommonsenseQA	90.1%

LLMs Companies Head Office

OpenAI is headquartered in San Francisco, California, USA, with additional offices in Seattle and research facilities worldwide.

Research Papers and Documentation

- GPT-5 Technical Report - Detailed architectural innovations and training methodologies.
- OpenAI API Documentation - Comprehensive integration guides and best practices.
- GitHub Examples - Code examples and implementation references.

Use Cases and Examples

- **Code Generation**: `def merge_sort(arr): if len(arr) <= 1: return arr; mid = len(arr) // 2; left = merge_sort(arr[:mid]); right = merge_sort(arr[mid:]); return merge(left, right)`
- **Creative Writing**: Generates consistent narrative arcs with complex character development across multiple chapters.

- **Data Analysis**: Processes complex statistical datasets and generates publication-ready visualizations with proper citations.

Limitations

- High API costs for enterprise-scale deployments
- Potential for generating contextually inappropriate content when prompts are poorly constructed
- Limited transparency in training data curation and filtering processes

Updates and Variants

- **GPT-5-Turbo**: Optimized for latency and cost-efficiency in production environments
- **GPT-5-Vision**: Advanced multimodal capabilities with enhanced image understanding
- **GPT-5-Enterprise**: Compliance-focused version with enhanced security and audit features

Claude-Sonnet-5

Claude-Sonnet-5 is Anthropic's latest model, emphasizing safety, truthfulness, and helpfulness.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-Sonnet-5	Accuracy	MMLU	91.8%
Claude-Sonnet-5	F1 Score	GLUE	91.2%
Claude-Sonnet-5	Perplexity	Wikitext-103	8.9
Claude-Sonnet-5	BLEU Score	WMT'14 En-Fr	41.7
Claude-Sonnet-5	Accuracy	GSM8K	90.3%
Claude-Sonnet-5	Pass@1	HumanEval	79.8%
Claude-Sonnet-5	Accuracy	SQuAD 2.0	91.4%
Claude-Sonnet-5	Safety Score	HELM	95.6%
Claude-Sonnet-5	Accuracy	PubMedQA	88.7%
Claude-Sonnet-5	F1 Score	CommonsenseQA	89.2%

LLMs Companies Head Office

Anthropic is headquartered in San Francisco, California, USA, with a mission to build AI that benefits humanity through advanced safety research.

Research Papers and Documentation

- Claude-Sonnet-5 Research Paper - Constitutional AI and alignment techniques.
- Anthropic Developer Documentation - Integration guides and API references.
- Constitutional AI Framework - Foundational research on AI alignment.

Use Cases and Examples

- **Medical Assistance**: Provides diagnostic suggestions with explicit confidence intervals and recommends specialist consultation.
- **Educational Tutoring**: Adapts explanations based on student knowledge level and learning style preferences.
- **Ethical Analysis**: Evaluates business decisions through multiple ethical frameworks with balanced perspectives.

Limitations

- Conservative response patterns may reduce creativity in artistic applications
- Higher computational latency compared to some competing models
- Requires extensive safety fine-tuning for domain-specific applications

Updates and Variants

- **Claude-Sonnet-5-Opus**: Maximum capability configuration for complex reasoning tasks
- **Claude-Sonnet-5-Haiku**: Lightweight version optimized for speed and efficiency
- **Claude-Sonnet-5-International**: Enhanced support for 100+ languages and cultural contexts

Gemini-3.0-Ultra

Gemini-3.0-Ultra Google's most advanced multimodal model with breakthrough performance in reasoning and creativity.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	Accuracy	MMLU	92.7%
Gemini-3.0-Ultra	F1 Score	GLUE	92.1%
Gemini-3.0-Ultra	Perplexity	Wikitext-103	8.1
Gemini-3.0-Ultra	BLEU Score	WMT'14 En-Fr	42.8
Gemini-3.0-Ultra	Accuracy	GSM8K	88.9%
Gemini-3.0-Ultra	Pass@1	HumanEval	76.4%
Gemini-3.0-Ultra	Accuracy	SQuAD 2.0	92.2%
Gemini-3.0-Ultra	Safety Score	HELM	93.9%
Gemini-3.0-Ultra	Accuracy	PubMedQA	89.1%
Gemini-3.0-Ultra	F1 Score	CommonsenseQA	89.7%

LLMs Companies Head Office

Google (Alphabet Inc.) is headquartered in Mountain View, California, USA, with DeepMind research facilities in London, UK.

Research Papers and Documentation

- Gemini-3.0 Technical Report - Architecture innovations and training breakthroughs.
- Google AI Documentation - Developer guides and API references.
- Vertex AI Documentation - Enterprise deployment guides.

Use Cases and Examples

- **Scientific Research**: Accelerates molecular discovery by analyzing protein structures and predicting binding affinities.
- **Architectural Design**: Generates building plans with structural analysis and material optimization.
- **Cultural Translation**: Provides context-aware translations that preserve cultural nuances and idiomatic expressions.

Limitations

- Complex deployment requirements with significant infrastructure needs
- Potential for reflecting societal biases present in large-scale training data
- Energy-intensive training processes requiring specialized hardware

Updates and Variants

- **Gemini-3.0-Pro**: Balanced performance and efficiency configuration
- **Gemini-3.0-Flash**: High-speed inference variant for real-time applications
- **Gemini-3.0-Nano**: Mobile-optimized version for edge computing

Llama-4-Scout

Llama-4-Scout Meta's cutting-edge open-source model with advanced reasoning capabilities.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Accuracy	MMLU	90.9%
Llama-4-Scout	F1 Score	GLUE	89.8%
Llama-4-Scout	Perplexity	Wikitext-103	9.2
Llama-4-Scout	BLEU Score	WMT'14 En-Fr	40.3
Llama-4-Scout	Accuracy	GSM8K	86.7%
Llama-4-Scout	Pass@1	HumanEval	74.2%
Llama-4-Scout	Accuracy	SQuAD 2.0	90.1%
Llama-4-Scout	Safety Score	HELM	92.3%
Llama-4-Scout	Accuracy	PubMedQA	85.9%
Llama-4-Scout	F1 Score	CommonsenseQA	87.4%

LLMs Companies Head Office

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

Research Papers and Documentation

- Llama-4 Technical Report - Architecture and training methodology details.
- Meta AI Documentation - Integration guides and model cards.
- GitHub Repository - Open-source implementation.

Use Cases and Examples

- **Open-Source Research**: Enables academic researchers to study and modify model architectures.
- **Custom Fine-tuning**: Supports domain-specific adaptation for specialized applications.
- **Community Development**: Fosters collaborative AI development through open access.

Limitations

- Requires significant computational resources for training and fine-tuning
- May exhibit biases from internet-scale training data
- Performance can vary significantly based on fine-tuning quality

Updates and Variants

- **Llama-4-Chat**: Instruction-tuned variant for conversational applications
- **Llama-4-Code**: Specialized version for programming tasks
- **Llama-4-70B**: Larger parameter configuration for enhanced capabilities

Command-R-Plus-2

Command-R-Plus-2 Cohere's advanced model with enhanced reasoning and multilingual capabilities.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Accuracy	MMLU	89.7%
Command-R-Plus-2	F1 Score	GLUE	88.9%
Command-R-Plus-2	Perplexity	Wikitext-103	9.8
Command-R-Plus-2	BLEU Score	WMT'14 En-Fr	39.1
Command-R-Plus-2	Accuracy	GSM8K	84.5%

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Pass@1	HumanEval	71.8%
Command-R-Plus-2	Accuracy	SQuAD 2.0	88.7%
Command-R-Plus-2	Safety Score	HELM	91.7%
Command-R-Plus-2	Accuracy	PubMedQA	83.2%
Command-R-Plus-2	F1 Score	CommonsenseQA	85.9%

LLMs Companies Head Office

Cohere is headquartered in Toronto, Canada, with offices in San Francisco and London.

Research Papers and Documentation

- Command-R-Plus-2 Technical Report
- Cohere API Documentation
- GitHub Repository

Use Cases and Examples

- **Enterprise Search**: Powers semantic search across large document collections.
- **Content Generation**: Creates marketing copy and technical documentation.
- **Customer Service**: Provides intelligent responses in multiple languages.

Limitations

- Performance may degrade with highly technical or specialized content
- Requires careful prompt engineering for optimal results
- Multilingual capabilities vary by language pair

Updates and Variants

- **Command-R-Plus-2-Light**: Efficient version for resource-constrained environments
- **Command-R-Plus-2-Multilingual**: Enhanced language support
- **Command-R-Plus-2-Enterprise**: Security-focused version for regulated industries

Jamba-2-Large

Jamba-2-Large AI21 Labs' hybrid architecture model combining transformers with structured state space models.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	Accuracy	MMLU	88.9%
Jamba-2-Large	F1 Score	GLUE	87.6%
Jamba-2-Large	Perplexity	Wikitext-103	10.1
Jamba-2-Large	BLEU Score	WMT'14 En-Fr	38.7
Jamba-2-Large	Accuracy	GSM8K	82.3%
Jamba-2-Large	Pass@1	HumanEval	69.4%
Jamba-2-Large	Accuracy	SQuAD 2.0	87.9%
Jamba-2-Large	Safety Score	HELM	90.8%
Jamba-2-Large	Accuracy	PubMedQA	81.7%

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	F1 Score	CommonsenseQA	84.2%

LLMs Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel, with offices in New York and London.

Research Papers and Documentation

- Jamba-2 Technical Report
- AI21 API Documentation
- GitHub Repository

Use Cases and Examples

- **Legal Document Analysis**: Processes complex legal texts and extracts key information.
- **Financial Modeling**: Analyzes market data and generates investment insights.
- **Educational Assessment**: Creates personalized learning plans and assessments.

Limitations

- Hybrid architecture may require specialized hardware optimization
- Performance can be variable across different domains
- Requires significant memory for large context windows

Updates and Variants

- **Jamba-2-Medium**: Balanced performance variant
- **Jamba-2-Small**: Efficient version for edge deployment
- **Jamba-2-Instruct**: Fine-tuned for instructional tasks

Qwen-3-235B

Qwen-3-235B Alibaba's massive model with advanced reasoning and multimodal capabilities.

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-3-235B	Accuracy	MMLU	91.2%
Qwen-3-235B	F1 Score	GLUE	90.1%
Qwen-3-235B	Perplexity	Wikitext-103	8.7
Qwen-3-235B	BLEU Score	WMT'14 En-Fr	41.2
Qwen-3-235B	Accuracy	GSM8K	87.8%
Qwen-3-235B	Pass@1	HumanEval	75.9%
Qwen-3-235B	Accuracy	SQuAD 2.0	90.8%
Qwen-3-235B	Safety Score	HELM	93.1%
Qwen-3-235B	Accuracy	PubMedQA	86.4%
Qwen-3-235B	F1 Score	CommonsenseQA	88.3%

LLMs Companies Head Office

Alibaba Group is headquartered in Hangzhou, China, with international offices worldwide.

Research Papers and Documentation

- Qwen-3 Technical Report
- Alibaba Cloud Model Studio
- GitHub Repository

Use Cases and Examples

- **Multilingual Processing**: Handles complex cross-language tasks with cultural awareness.
- **Enterprise Automation**: Powers intelligent workflow automation in large-scale operations.
- **Research Analysis**: Processes scientific literature and generates research summaries.

Limitations

- Extremely high computational requirements limit accessibility
- Training data primarily focused on Chinese and English content
- Complex deployment requirements for enterprise use

Updates and Variants

- **Qwen-3-72B**: Smaller, more accessible variant
- **Qwen-3-Chat**: Instruction-tuned conversational version
- **Qwen-3-Math**: Specialized mathematics version

Mistral-Large-2

Mistral-Large-2 Mistral AI's advanced model with efficient architecture and strong performance.

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](https://www.alibabacl