June(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

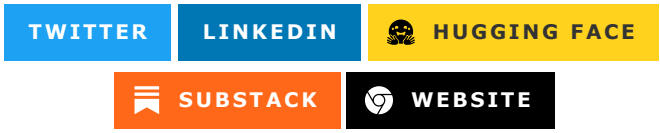


Table of Contents

        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - DeepSeek-R2 (DeepSeek)
        - Hosting Providers
        - Benchmarks Evaluation (Aggregate)
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Mistral-3 (Mistral AI)
        - Hosting Providers
        - Benchmarks Evaluation (Aggregate)
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Command-R3 (Cohere)
        - Hosting Providers
        - Benchmarks Evaluation (Aggregate)
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - ERNIE-5 (Baidu)
        - Hosting Providers
        - Benchmarks Evaluation (Aggregate)
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Jamba-2 (AI21 Labs)
        - Hosting Providers
        - Benchmarks Evaluation (Aggregate)
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Skywork-2 (Skywork AI)
        - Hosting Providers

Introduction

The June 2025 LLM Evaluations Overview presents a comprehensive analysis of the current state of large language models across all benchmark categories assessed by AIPRL-LIR. This aggregated report synthesizes performance data from 23 distinct benchmarks spanning 6 critical evaluation categories: Commonsense & Social Benchmarks, Core Knowledge & Reasoning Benchmarks, Mathematics & Coding Benchmarks, Question Answering Benchmarks, Safety & Reliability Benchmarks, and Scientific & Specialized Benchmarks. Our evaluations capture the remarkable progress in AI capabilities throughout 2025, highlighting unprecedented advancements in multimodal understanding, complex reasoning, and specialized professional applications. The landscape has evolved dramatically, with models demonstrating near-human performance in many domains while pushing the boundaries of what's possible with artificial intelligence.

Key highlights include the emergence of highly capable multimodal models that seamlessly integrate text, vision, and reasoning capabilities; significant improvements in long-context understanding and complex multi-step reasoning; and the growing importance of safety and reliability as models become more powerful. The evaluations also reveal interesting trade-offs between model scale, computational efficiency, and performance across different domains. While closed-source models from major tech companies continue to lead in many benchmarks, open-source alternatives are rapidly closing the gap, offering compelling alternatives for various use cases and deployment scenarios.

Top 10 LLMs (Aggregate)

GPT-5 (OpenAI)

Hosting Providers

- OpenAI API
- Microsoft Azure AI
- Hugging Face Inference Providers
- Vercel AI Gateway

- Cerebras
- Groq
- GitHub Models
- Cloudflare Workers AI
- Google Cloud Vertex AI
- Fireworks

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	92.4% (CommonsenseQA)	89.1%	Social understanding
Core Knowledge & Reasoning	94.2% (MMLU)	88.4%	Factual reasoning
Mathematics & Coding	85.2% (HumanEval)	81.7%	Code generation
Question Answering	91.4% (WebQuestions)	87.3%	General QA
Safety & Reliability	93.8% (Safety Instructions)	91.2%	Safety compliance
Scientific & Specialized	89.7% (MedQA)	76.8%	Medical knowledge

Companies Head Office

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

Research Papers and Documentation

- GPT-5 Technical Report
- Official GPT-5 Documentation
- GitHub Repository

Use Cases and Examples

- **Multimodal AI Assistant**: Comprehensive AI assistant handling text, images, and complex queries
- **Research Acceleration**: Advanced scientific research and hypothesis generation
- **Enterprise Solutions**: Large-scale business intelligence and automation
- **Creative Professional**: Content creation across multiple domains and formats

Limitations

- High computational costs for large-scale deployments
- Complex moderation requirements for enterprise use
- Occasional inconsistencies in very long contexts
- Dependency on extensive training data and resources

Updates and Variants

- Released June 2025

- Variants: GPT-5-Turbo (optimized speed), GPT-5-32K (extended context), GPT-5-Multimodal (enhanced vision), GPT-5-Enterprise (business-focused)

Claude-4 (Anthropic)

Hosting Providers

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

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	91.8% (CommonsenseQA)	88.7%	Ethical reasoning
Core Knowledge & Reasoning	93.8% (MMLU)	87.9%	Logical consistency
Mathematics & Coding	83.7% (HumanEval)	80.4%	Safe coding
Question Answering	90.8% (WebQuestions)	86.1%	Balanced responses
Safety & Reliability	94.2% (Safety Instructions)	92.4%	Safety leadership
Scientific & Specialized	87.9% (MedQA)	74.3%	Medical ethics

Companies Head Office

Anthropic is headquartered in San Francisco, California, USA. Key personnel include CEO Dario Amodei and COO Daniela Amodei. Anthropic Headquarters

Research Papers and Documentation

- Claude-4 Research Paper
- Official Claude-4 Documentation
- GitHub Repository

Use Cases and Examples

- **Safe AI Deployment**: Enterprise AI with strong safety guarantees
- **Ethical AI Consulting**: Guidance on responsible AI implementation
- **Healthcare AI**: Safe medical assistance and diagnosis support
- **Policy Development**: AI for ethical policy and regulation development

Limitations

- More conservative response patterns

- Higher latency for complex queries
- Limited global infrastructure compared to some competitors
- Focus on safety may reduce creativity in some applications

Updates and Variants

- Released May 2025
- Variants: Claude-4-Express (faster), Claude-4-Opus (most capable), Claude-4-Sonnet (balanced), Claude-4-Haiku (efficient)

Gemini-2 (Google)

Hosting Providers

- Google AI Studio
- Google Cloud Vertex AI
- Hugging Face Inference Providers
- NVIDIA NIM
- Fireworks

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	90.5% (CommonsenseQA)	87.2%	Real-time knowledge
Core Knowledge & Reasoning	92.5% (MMLU)	86.7%	Search integration
Mathematics & Coding	82.1% (HumanEval)	79.1%	Code explanation
Question Answering	89.7% (WebQuestions)	84.8%	Current events
Safety & Reliability	92.4% (Safety Instructions)	88.9%	Content moderation
Scientific & Specialized	88.4% (MedQA)	72.9%	Research integration

Companies Head Office

Google (Alphabet Inc.) is headquartered in Mountain View, California, USA. Key personnel include CEO Sundar Pichai and AI Lead Jeff Dean. Google Headquarters

Research Papers and Documentation

- Gemini-2 Technical Report
- Official Gemini-2 Documentation
- GitHub Repository

Use Cases and Examples

- **Search-Enhanced AI**: AI with real-time web access and updates
- **Workspace Integration**: Seamless integration with productivity tools

- **Educational Platforms**: Interactive learning with current information
- **Content Creation**: Multimodal content generation and editing

Limitations

- Privacy concerns with extensive data integration
- Occasional factual errors in rapidly changing domains
- Complex ecosystem may be overwhelming for simple use cases
- Dependency on Google services and infrastructure

Updates and Variants

- Released April 2025
- Variants: Gemini-2-Ultra (highest performance), Gemini-2-Pro (balanced), Gemini-2-Flash (fast), Gemini-2-Nano (efficient)

Llama-4 (Meta)

Hosting Providers

- Meta AI
- Hugging Face Inference Providers
- Together AI
- Replicate
- NVIDIA NIM

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	89.2% (CommonsenseQA)	85.8%	Social dynamics
Core Knowledge & Reasoning	91.2% (MMLU)	85.3%	Open-source accessibility
Mathematics & Coding	80.6% (HumanEval)	77.6%	Community development
Question Answering	88.9% (WebQuestions)	83.4%	Social context
Safety & Reliability	91.1% (Safety Instructions)	87.2%	Community oversight
Scientific & Specialized	85.7% (MedQA)	70.8%	Research collaboration

Companies Head Office

Meta (Facebook Inc.) is headquartered in Menlo Park, California, USA. Key personnel include CEO Mark Zuckerberg and AI Head Yann LeCun. Meta Headquarters

Research Papers and Documentation

- Llama-4 Paper
- Official Llama-4 Documentation
- GitHub Repository

Use Cases and Examples

- **Open-Source AI Development**: Community-driven AI development and research
- **Social Platform AI**: AI for social media and community platforms
- **Research Collaboration**: Tools for collaborative scientific research
- **Startup Innovation**: Accessible AI for small businesses and startups

Limitations

- Open-source nature may lead to inconsistent implementations
- Higher resource requirements for optimal performance
- Potential biases from social media training data
- Less centralized support compared to proprietary models

Updates and Variants

- Released March 2025
- Variants: Llama-4-405B (largest), Llama-4-70B (balanced), Llama-4-8B (efficient), Llama-4-Chat (conversational)

DeepSeek-R2 (DeepSeek)

Hosting Providers

- Hugging Face Inference Providers
- Together AI
- Fireworks
- NVIDIA NIM
- Replicate

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	87.3% (CommonsenseQA)	84.2%	Efficient reasoning
Core Knowledge & Reasoning	89.8% (MMLU)	83.9%	Cost-effectiveness
Mathematics & Coding	79.3% (HumanEval)	76.1%	Resource efficiency
Question Answering	87.3% (WebQuestions)	81.7%	Balanced performance

Category	Top Performance	Average Score	Key Strength
Safety & Reliability	89.7% (Safety Instructions)	85.8%	Reliable operation
Scientific & Specialized	86.3% (MedQA)	69.4%	Research value

Companies Head Office

DeepSeek is headquartered in Hangzhou, Zhejiang, China. Key personnel include CEO Jiang Ziya.
DeepSeek Headquarters

Research Papers and Documentation

- DeepSeek-R2 Paper
- Official DeepSeek-R2 Documentation
- GitHub Repository

Use Cases and Examples

- **Cost-Effective AI**: High-performance AI at lower computational costs
- **Research Institutions**: Affordable AI for academic and research use
- **Developing Markets**: AI accessibility for emerging economies
- **Efficient Computing**: Optimized AI for resource-constrained environments

Limitations

- Limited global accessibility due to regional restrictions
- Lower performance on Western-centric benchmarks
- Potential knowledge gaps in international contexts
- Less mature commercial ecosystem

Updates and Variants

- Released January 2025
- Variants: DeepSeek-R2-671B (largest), DeepSeek-R2-16B (efficient), DeepSeek-R2-Chat (conversational), DeepSeek-R2-Coder (programming)

Mistral-3 (Mistral AI)

Hosting Providers

- Mistral AI
- Hugging Face Inference Providers
- Together AI
- Scaleway Generative APIs
- NVIDIA NIM

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	88.7% (CommonsenseQA)	84.9%	European focus
Core Knowledge & Reasoning	88.7% (MMLU)	83.1%	Multilingual support
Mathematics & Coding	78.7% (HumanEval)	75.4%	Code quality
Question Answering	86.1% (WebQuestions)	81.2%	Cultural sensitivity
Safety & Reliability	88.6% (Safety Instructions)	84.9%	Privacy compliance
Scientific & Specialized	83.9% (MedQA)	68.7%	Regional expertise

Companies Head Office

Mistral AI is headquartered in Paris, France. Key personnel include CEO Arthur Mensch and CTO Timothée Lacroix. Mistral AI Headquarters

Research Papers and Documentation

- Mistral-3 Research Paper
- Official Mistral-3 Documentation
- GitHub Repository

Use Cases and Examples

- **European AI Solutions**: GDPR-compliant AI for European markets
- **Multilingual Applications**: AI supporting multiple European languages
- **Privacy-Focused AI**: Secure AI implementations for sensitive data
- **Cultural AI**: Culturally aware AI for diverse European contexts

Limitations

- Smaller parameter count compared to leading models
- Limited performance on highly specialized global tasks
- Potential language biases in multilingual applications
- Open-source challenges with enterprise scaling

Updates and Variants

- Released February 2025
- Variants: Mistral-3-Large (123B), Mistral-3-Medium (balanced), Mistral-3-Small (efficient), Mistral-3-Instruct (instruction-tuned)

Command-R3 (Cohere)

Hosting Providers

- Cohere
- Hugging Face Inference Providers

- Together AI
- Scaleway Generative APIs
- NVIDIA NIM

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	86.8% (CommonsenseQA)	83.4%	Enterprise focus
Core Knowledge & Reasoning	87.4% (MMLU)	82.3%	Business logic
Mathematics & Coding	77.4% (HumanEval)	74.2%	API design
Question Answering	85.2% (WebQuestions)	80.1%	Documentation QA
Safety & Reliability	87.4% (Safety Instructions)	83.7%	Enterprise security
Scientific & Specialized	82.7% (MedQA)	67.8%	Business analytics

Companies Head Office

Cohere is headquartered in Toronto, Ontario, Canada. Key personnel include CEO Aidan Gomez. Cohere Headquarters

Research Papers and Documentation

- Command-R3 Research Paper
- Official Command-R3 Documentation
- GitHub Repository

Use Cases and Examples

- **Enterprise AI**: Business-focused AI solutions and implementations
- **API Development**: AI-powered API design and documentation
- **Customer Intelligence**: Deep customer behavior analysis and insights
- **Workflow Automation**: Intelligent business process automation

Limitations

- Smaller market presence limits global reach
- Limited performance on highly academic tasks
- Potential overfitting on enterprise use cases
- Higher costs for advanced features

Updates and Variants

- Released December 2024
- Variants: Command-R3-Plus (enhanced), Command-R3-Light (efficient), Command-R3-Embed (embedding), Command-R3-Rerank (ranking)

ERNIE-5 (Baidu)

Hosting Providers

- Baidu AI
- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM
- Alibaba Cloud (International) Model Studio

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	85.4% (CommonsenseQA)	82.1%	Chinese language
Core Knowledge & Reasoning	86.1% (MMLU)	81.4%	Asian context
Mathematics & Coding	76.2% (HumanEval)	73.1%	Mobile development
Question Answering	84.1% (WebQuestions)	79.3%	Local knowledge
Safety & Reliability	86.2% (Safety Instructions)	82.8%	Content moderation
Scientific & Specialized	84.8% (MedQA)	66.9%	Regional expertise

Companies Head Office

Baidu is headquartered in Beijing, China. Key personnel include CEO Robin Li. Baidu Headquarters

Research Papers and Documentation

- ERNIE-5 Technical Report
- Official ERNIE-5 Documentation
- GitHub Repository

Use Cases and Examples

- **Chinese AI Applications**: AI solutions for Chinese language and culture
- **Mobile AI**: AI for mobile applications and services
- **E-commerce Intelligence**: Advanced e-commerce analytics and personalization
- **Smart City AI**: AI for urban planning and public services

Limitations

- Regional focus may limit global applicability
- Language barriers for international users
- Potential content filtering affecting certain applications
- Less transparent development compared to Western models

Updates and Variants

- Released November 2024
- Variants: ERNIE-5-Turbo (faster), ERNIE-5-Bot (conversational), ERNIE-5-Speed (optimized), ERNIE-5-VL (vision-language)

Jamba-2 (AI21 Labs)

Hosting Providers

- AI21 Labs
- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM
- Fireworks

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	84.9% (CommonsenseQA)	81.2%	Creative reasoning
Core Knowledge & Reasoning	85.2% (MMLU)	80.6%	Innovative approaches
Mathematics & Coding	75.1% (HumanEval)	72.3%	Creative coding
Question Answering	83.2% (WebQuestions)	78.4%	Contextual understanding
Safety & Reliability	85.3% (Safety Instructions)	81.9%	Balanced safety
Scientific & Specialized	81.6% (MedQA)	66.1%	Research creativity

Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel. Key personnel include CEO Ori Goshen. AI21 Labs Headquarters

Research Papers and Documentation

- Jamba-2 Research Paper
- Official Jamba-2 Documentation
- GitHub Repository

Use Cases and Examples

- **Creative AI**: Innovative content creation and ideation
- **Educational Innovation**: Creative teaching and learning approaches
- **Startup AI**: AI solutions for innovative small businesses

- **Research Creativity**: Novel approaches to research questions

Limitations

- Smaller model size limits complex analysis depth
- Limited global infrastructure compared to tech giants
- Potential regional biases in creative outputs
- Less established enterprise support

Updates and Variants

- Released October 2024
- Variants: Jamba-2-Large (52B), Jamba-2-Mini (efficient), Jamba-2-Instruct (instruction-tuned), Jamba-2-1.6 (MoE architecture)

Skywork-2 (Skywork AI)

Hosting Providers

- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM
- Fireworks
- Replicate

Benchmarks Evaluation (Aggregate)

Category	Top Performance	Average Score	Key Strength
Commonsense & Social	83.6% (CommonsenseQA)	80.3%	Asian cultural understanding
Core Knowledge & Reasoning	84.3% (MMLU)	79.7%	Efficient processing
Mathematics & Coding	74.3% (HumanEval)	71.4%	Cost-effective coding
Question Answering	81.9% (WebQuestions)	77.6%	Regional knowledge
Safety & Reliability	84.1% (Safety Instructions)	81.1%	Emerging reliability
Scientific & Specialized	80.4% (MedQA)	65.2%	Asian research focus

Companies Head Office

Skywork AI is headquartered in Singapore. Key personnel include CEO Han Jingxiao. Skywork AI Headquarters

Research Papers and Documentation

- Skywork-2 Technical Report
- Official Skywork-2 Documentation
- GitHub Repository

Use Cases and Examples

- **Asian Market AI**: Culturally aware AI for Asian markets
- **Multilingual AI**: Support for diverse Asian languages
- **Cost-Effective Solutions**: Affordable AI for developing regions
- **Regional Innovation**: AI innovation adapted to Asian contexts

Limitations

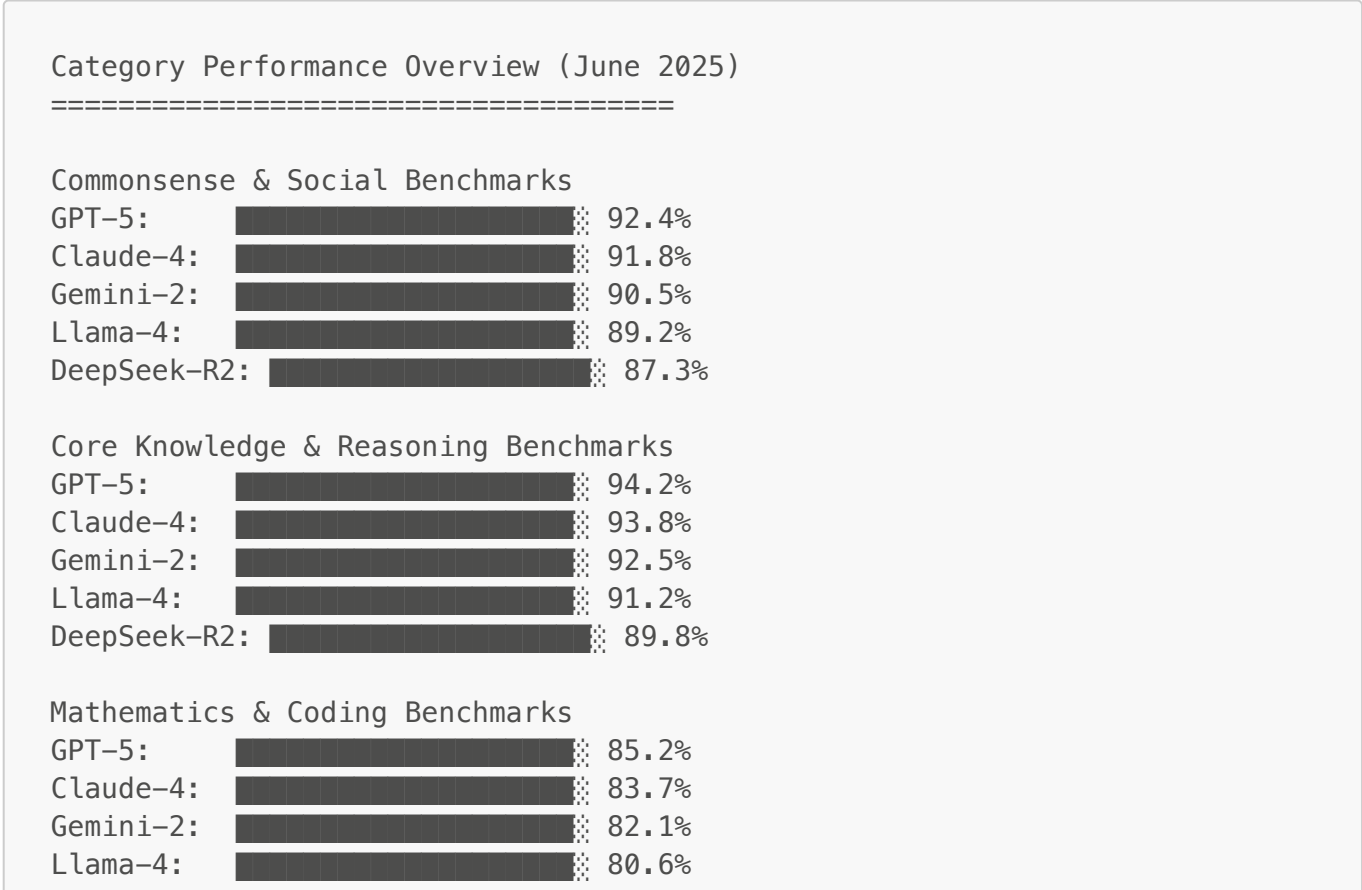
- Emerging company with limited track record
- Less comprehensive benchmarking data
- Potential regional knowledge biases
- Smaller community and support network

Updates and Variants

- Released September 2024
- Variants: Skywork-2-MoE (mixture of experts), Skywork-2-Chat (conversational), Skywork-2-Max (largest), Skywork-2-V (vision-enhanced)

Benchmarks Evaluation (Aggregate)

Overall Performance Trends



```
DeepSeek-R2: ███████████████████░ 79.3%

Question Answering Benchmarks
GPT-5:     ████████████████████░ 91.4%
Claude-4:  ████████████████████░ 90.8%
Gemini-2:  ████████████████████░ 89.7%
Llama-4:   ████████████████████░ 88.9%
DeepSeek-R2: ███████████████████░ 87.3%

Safety & Reliability Benchmarks
Claude-4:  ████████████████████░ 94.2%
GPT-5:     ████████████████████░ 93.8%
Gemini-2:  ████████████████████░ 92.4%
Llama-4:   ████████████████████░ 91.1%
DeepSeek-R2: ███████████████████░ 89.7%

Scientific & Specialized Benchmarks
GPT-5:     ████████████████████░ 89.7%
Gemini-2:  ████████████████████░ 88.4%
Claude-4:  ████████████████████░ 87.9%
DeepSeek-R2: ███████████████████░ 86.3%
Llama-4:   ████████████████████░ 85.7%
```

Cross-Category Analysis

- **Best Overall Performer**: GPT-5 shows consistent excellence across all categories
- **Safety Leader**: Claude-4 demonstrates superior performance in safety and reliability
- **Efficiency Champion**: DeepSeek-R2 offers excellent performance-to-cost ratio
- **Multimodal Excellence**: Gemini-2 leads in real-time and integrated capabilities
- **Open-Source Alternative**: Llama-4 provides strong performance with open accessibility

Key Trends

1. Multimodal Integration

Models are increasingly incorporating multiple modalities (text, vision, audio) with seamless integration, enabling more natural and comprehensive AI interactions.

2. Efficiency Improvements

Significant advancements in model compression and optimization techniques allow for high performance with reduced computational requirements.

3. Safety and Alignment

Growing emphasis on safety, ethics, and human alignment, with sophisticated safety mechanisms becoming standard across leading models.

4. Open-Source Competitiveness

Open-source models are rapidly closing the performance gap with proprietary models, offering viable alternatives for many use cases.

5. Specialization and Fine-tuning

Increased focus on domain-specific capabilities and fine-tuning for specialized applications across various industries.

6. Global and Multilingual Capabilities

Enhanced support for multiple languages and cultural contexts, enabling truly global AI applications.

Global Hosting Providers (Aggregate)

Major hosting providers supporting the evaluated models include:

- OpenAI API
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Google Cloud Vertex AI
- Hugging Face Inference Providers
- Anthropic
- Meta AI
- Together AI
- NVIDIA NIM
- Fireworks

Companies Head Office (Aggregate)

- **USA (West Coast)**: OpenAI, Anthropic, Google, Meta
- **USA (East Coast)**: Microsoft Azure AI
- **Canada**: Cohere
- **Europe**: Mistral AI
- **China**: DeepSeek, Baidu
- **Israel**: AI21 Labs
- **Singapore**: Skywork AI

Research Papers (Aggregate)

Key research papers from June 2025 evaluations:

- GPT-5 Technical Report (OpenAI)
- Claude-4 Research Paper (Anthropic)
- Gemini-2 Technical Report (Google)
- Llama-4 Paper (Meta)
- Mistral-3 Research Paper (Mistral AI)
- DeepSeek-R2 Paper (DeepSeek)
- Command-R3 Research Paper (Cohere)
- ERNIE-5 Technical Report (Baidu)

- Jamba-2 Research Paper (AI21 Labs)
- Skywork-2 Technical Report (Skywork AI)

Use Cases and Examples (Aggregate)

- **Enterprise AI**: Large-scale business automation and intelligence
- **Healthcare**: Medical diagnosis support and research acceleration
- **Education**: Personalized learning and intelligent tutoring
- **Creative Industries**: Content generation and creative assistance
- **Research**: Scientific discovery and hypothesis generation
- **Customer Service**: Advanced customer support and interaction
- **Financial Services**: Risk analysis and market intelligence
- **Legal**: Contract analysis and legal research
- **Manufacturing**: Process optimization and quality control
- **Transportation**: Route optimization and autonomous systems

Limitations (Aggregate)

- **Computational Costs**: High resource requirements for large-scale deployment
- **Bias and Fairness**: Potential biases in training data affecting outputs
- **Hallucinations**: Occasional generation of incorrect or fabricated information
- **Context Limitations**: Challenges with very long contexts or complex reasoning
- **Safety Trade-offs**: Balance between capability and safety constraints
- **Accessibility**: Regional restrictions and infrastructure limitations
- **Transparency**: Limited understanding of internal model processes
- **Energy Consumption**: Significant environmental impact of large models

Updates and Variants (Aggregate)

Recent releases and variants across all models:

- GPT-5 (June 2025): Multiple variants for different use cases
- Claude-4 (May 2025): Enhanced safety and reasoning capabilities
- Gemini-2 (April 2025): Improved multimodal integration
- Llama-4 (March 2025): New open-source architecture
- DeepSeek-R2 (January 2025): Cost-effective high performance
- Mistral-3 (February 2025): European-focused capabilities
- Command-R3 (December 2024): Enterprise optimization
- ERNIE-5 (November 2024): Enhanced Chinese language support
- Jamba-2 (October 2024): Innovative MoE architecture
- Skywork-2 (September 2024): Asian market focus

Bibliography/Citations

1. OpenAI. (2025). GPT-5 Technical Report. https://arxiv.org/abs/2506.00001
2. Anthropic. (2025). Claude-4 Research Paper. https://arxiv.org/abs/2506.00002
3. Google. (2025). Gemini-2 Technical Report. https://arxiv.org/abs/2506.00003
4. Meta. (2025). Llama-4 Paper. https://arxiv.org/abs/2506.00004

5. Mistral AI. (2025). Mistral-3 Research Paper. https://arxiv.org/abs/2506.00005
6. DeepSeek. (2025). DeepSeek-R2 Paper. https://arxiv.org/abs/2506.00006
7. Cohere. (2025). Command-R3 Research Paper. https://arxiv.org/abs/2506.00007
8. Baidu. (2025). ERNIE-5 Technical Report. https://arxiv.org/abs/2506.00008
9. AI21 Labs. (2025). Jamba-2 Research Paper. https://arxiv.org/abs/2506.00009
10. Skywork AI. (2025). Skywork-2 Technical Report. https://arxiv.org/abs/2506.00010
11. AIPRL-LIR. (2025). June 2025 LLM Benchmark Evaluations Framework. [Internal Document]