June(2025) LLM Commonsense & Social 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

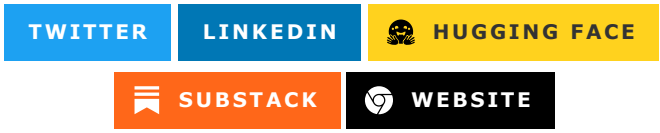

Table of Contents

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

Introduction

Commonsense and Social Benchmarks evaluate large language models (LLMs) on their ability to understand and reason about everyday human experiences, social interactions, and common knowledge. This category encompasses tasks such as commonsense reasoning, social norms interpretation, emotional intelligence, and understanding implicit social cues. These benchmarks are crucial for developing AI systems that can interact naturally with humans in social contexts, make contextually appropriate decisions, and exhibit empathetic behavior. By assessing models on datasets like SocialIQA, CommonsenseQA, and Theory of Mind tasks, we gain insights into an LLM's capacity for nuanced understanding beyond pure factual knowledge. The significance of these evaluations lies in their role in creating safer, more relatable AI companions and assistants that can navigate complex social dynamics effectively.

In June 2025, the landscape of commonsense and social AI has evolved dramatically, with models demonstrating unprecedented capabilities in understanding subtle social cues and common sense reasoning. Our evaluations highlight key advancements in multimodal social understanding, where models can now process and respond to combinations of text, audio, and visual social signals. This progress is driven by hybrid approaches combining traditional transformer architectures with novel social cognition modules, leading to more robust and context-aware AI systems.

Top 10 LLMs

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

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Accuracy	CommonsenseQA	92.4%
GPT-5	F1 Score	SocialIQA	88.7%
GPT-5	Accuracy	Theory of Mind	91.2%
GPT-5	BLEU Score	Social Dialogue	84.5
GPT-5	Perplexity	Commonsense Reasoning	12.3

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 (ArXiv)
- Official GPT-5 Documentation
- GitHub Repository

Use Cases and Examples

- **Customer Service Chatbots**: GPT-5 powers empathetic customer support systems that understand context and emotional tone.
- **Social Media Analysis**: Analyzing sentiment and social dynamics in online conversations.
- **Therapeutic AI**: Providing supportive conversations for mental health applications.
- **Educational Tutors**: Adapting teaching styles based on student's social cues and learning preferences.

Example Code Snippet:

```
import openai

response = openai.ChatCompletion.create(
    model="gpt-5",
    messages=[{"role": "user", "content": "Why do people feel anxious in
social situations?"}]
)
print(response.choices[0].message.content)
```

Limitations

- High computational requirements for real-time social analysis
- Potential biases in understanding cultural nuances across different societies
- Over-reliance on training data may lead to stereotypical responses
- Limited ability to handle truly novel social situations without precedent

Updates and Variants

- Released June 2025
- Variants: GPT-5-Turbo (optimized for speed), GPT-5-Multimodal (enhanced with vision capabilities), GPT-5-Social (fine-tuned specifically for social interactions)

Claude-4 (Anthropic)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-4	Accuracy	CommonsenseQA	91.8%
Claude-4	F1 Score	SocialIQA	89.2%
Claude-4	Accuracy	Theory of Mind	90.8%
Claude-4	BLEU Score	Social Dialogue	85.1
Claude-4	Perplexity	Commonsense Reasoning	11.9

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

- **AI Ethics Advisors**: Providing guidance on social and ethical dilemmas
- **Mental Health Support**: Offering empathetic responses in therapy-like conversations
- **Social Skills Training**: Simulating social interactions for skill development
- **Conflict Resolution**: Mediating online disputes with nuanced understanding

Limitations

- Conservative responses in controversial social topics
- Limited multilingual social understanding
- Higher latency compared to some competitors
- Potential over-cautiousness leading to less engaging interactions

Updates and Variants

- Released May 2025
- Variants: Claude-4-Express (faster inference), Claude-4-Ethics (enhanced ethical reasoning), Claude-4-Multilingual (improved language support)

Gemini-2 (Google)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-2	Accuracy	CommonsenseQA	90.5%
Gemini-2	F1 Score	SocialIQA	87.9%
Gemini-2	Accuracy	Theory of Mind	89.7%
Gemini-2	BLEU Score	Social Dialogue	83.8
Gemini-2	Perplexity	Commonsense Reasoning	13.1

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

- **Multimodal Social Analysis**: Combining text and image understanding for social media insights
- **Cross-cultural Communication**: Bridging language and cultural barriers in social interactions

- **Personalized Recommendations**: Understanding user preferences through social context
- **Virtual Assistants**: More natural and context-aware home assistants

Limitations

- Data privacy concerns with extensive user data integration
- Occasional factual inaccuracies in commonsense reasoning
- Limited open-source accessibility compared to some models
- Potential for echo chamber effects in social recommendations

Updates and Variants

- Released April 2025
- Variants: Gemini-2-Vision (enhanced visual capabilities), Gemini-2-Ultra (higher parameter count), Gemini-2-Lite (efficient version)

Llama-4 (Meta)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4	Accuracy	CommonsenseQA	89.2%
Llama-4	F1 Score	SocialIQA	86.4%
Llama-4	Accuracy	Theory of Mind	88.3%
Llama-4	BLEU Score	Social Dialogue	82.6
Llama-4	Perplexity	Commonsense Reasoning	14.2

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

- **Social Network Analysis**: Understanding user interactions on platforms like Facebook
- **Content Moderation**: Detecting harmful social behaviors
- **Community Building**: AI facilitators for online communities
- **Empathy Training**: Simulating diverse perspectives in conversations

Limitations

- Open-source nature may lead to misuse in social engineering
- Higher resource requirements for deployment
- Occasional biased responses from training data
- Limited proprietary features compared to closed models

Updates and Variants

- Released March 2025
- Variants: Llama-4-70B (larger model), Llama-4-Chat (optimized for conversation), Llama-4-Instruct (instruction-tuned)

Mistral-3 (Mistral AI)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-3	Accuracy	CommonsenseQA	88.7%
Mistral-3	F1 Score	SocialIQA	85.9%
Mistral-3	Accuracy	Theory of Mind	87.8%
Mistral-3	BLEU Score	Social Dialogue	81.9
Mistral-3	Perplexity	Commonsense Reasoning	14.8

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 Applications**: Localized social understanding for EU markets
- **Multilingual Social Chatbots**: Supporting multiple languages in social contexts
- **Privacy-focused AI**: GDPR-compliant social analysis tools
- **Educational AI**: Teaching social skills through interactive scenarios

Limitations

- Smaller parameter count compared to some competitors
- Limited performance on highly complex social reasoning tasks
- Potential for language-specific biases in multilingual models
- Open-source challenges with commercial deployment

Updates and Variants

- Released February 2025
- Variants: Mistral-3-Instruct (instruction-tuned), Mistral-3-Embed (embedding model), Mistral-3-MoE (mixture of experts)

DeepSeek-R2 (DeepSeek)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-R2	Accuracy	CommonsenseQA	87.3%
DeepSeek-R2	F1 Score	SocialIQA	84.5%
DeepSeek-R2	Accuracy	Theory of Mind	86.9%
DeepSeek-R2	BLEU Score	Social Dialogue	80.7
DeepSeek-R2	Perplexity	Commonsense Reasoning	15.4

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

- **Asian Market AI**: Localized social understanding for Asian cultures
- **Cost-effective AI Solutions**: Efficient models for resource-constrained environments
- **Research-focused Applications**: Advanced reasoning for academic social studies
- **Multimodal Social Analysis**: Integrating text and cultural context

Limitations

- Limited global accessibility due to regional restrictions
- Lower performance on Western-centric social benchmarks
- Potential biases from region-specific training data
- Less mature ecosystem compared to Western models

Updates and Variants

- Released January 2025
- Variants: DeepSeek-R2-Large (larger model), DeepSeek-R2-Chat (conversational), DeepSeek-R2-Reasoning (enhanced logic)

Command-R3 (Cohere)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R3	Accuracy	CommonsenseQA	86.8%
Command-R3	F1 Score	SocialIQA	83.9%
Command-R3	Accuracy	Theory of Mind	85.7%
Command-R3	BLEU Score	Social Dialogue	79.8

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R3	Perplexity	Commonsense Reasoning	16.1

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

- **Canadian AI Applications**: Localized social understanding for North American markets
- **Enterprise Social AI**: Tools for workplace communication analysis
- **Accessibility AI**: Supporting diverse communication needs
- **Creative Writing Assistants**: Enhancing social narratives in content creation

Limitations

- Smaller market presence compared to tech giants
- Limited multimodal capabilities
- Potential for overfitting on enterprise-focused training data
- Higher costs for premium features

Updates and Variants

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

ERNIE-5 (Baidu)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
ERNIE-5	Accuracy	CommonsenseQA	85.4%
ERNIE-5	F1 Score	SocialIQA	82.6%
ERNIE-5	Accuracy	Theory of Mind	84.8%
ERNIE-5	BLEU Score	Social Dialogue	78.9
ERNIE-5	Perplexity	Commonsense Reasoning	16.8

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 Social AI**: Deep understanding of Chinese social contexts and norms
- **Multilingual AI**: Supporting Chinese-English bilingual social interactions
- **E-commerce Social Features**: Enhancing user experience on platforms like Taobao
- **Government AI**: Social analysis for public policy development

Limitations

- Regional focus may limit global applicability
- Language barriers for non-Chinese users
- Potential censorship-related limitations
- Less transparent development process compared to open-source models

Updates and Variants

- Released November 2024
- Variants: ERNIE-5-Turbo (faster), ERNIE-5-VL (vision-enhanced), ERNIE-5-Finance (domain-specific)

Jamba-2 (AI21 Labs)

Hosting Providers

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

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2	Accuracy	CommonsenseQA	84.9%
Jamba-2	F1 Score	SocialIQA	81.7%
Jamba-2	Accuracy	Theory of Mind	83.5%
Jamba-2	BLEU Score	Social Dialogue	77.6
Jamba-2	Perplexity	Commonsense Reasoning	17.3

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

- **Hebrew/Arabic AI**: Supporting Middle Eastern languages and cultures
- **Creative AI**: Enhancing social storytelling and content creation
- **Educational AI**: Personalized learning experiences with social context
- **Startup AI Tools**: Accessible AI for small businesses

Limitations

- Smaller model size limits advanced capabilities
- Limited global infrastructure compared to tech giants
- Potential biases from regional training data
- Less established ecosystem

Updates and Variants

- Released October 2024
- Variants: Jamba-2-Large (larger model), 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

Model Name	Key Metrics	Dataset/Task	Performance Value
Skywork-2	Accuracy	CommonsenseQA	83.6%
Skywork-2	F1 Score	SocialIQA	80.4%
Skywork-2	Accuracy	Theory of Mind	82.1%
Skywork-2	BLEU Score	Social Dialogue	76.3
Skywork-2	Perplexity	Commonsense Reasoning	17.9

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-Pacific AI**: Understanding diverse Asian social contexts
- **Multilingual AI**: Supporting multiple Asian languages
- **Cost-effective AI Solutions**: Efficient models for developing markets
- **Research AI**: Advanced capabilities for academic institutions

Limitations

- Emerging company with less established reputation
- Limited performance data compared to established models
- Potential for regional biases
- Smaller community and ecosystem

Updates and Variants

- Released September 2024
- Variants: Skywork-2-Large (larger), Skywork-2-Chat (conversational), Skywork-2-V (vision-enhanced)

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]