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

Introduction

The Commonsense & Social Benchmarks category evaluates large language models on their ability to understand and reason about everyday situations, social norms, human behavior, and common-sense knowledge. This category encompasses tasks that require understanding of social dynamics, cultural contexts, emotional intelligence, and practical reasoning about real-world scenarios.

These benchmarks are crucial for developing AI systems that can interact naturally with humans, understand social cues, and make decisions that align with human values and expectations. The April 2025 evaluations include datasets such as SocialIQA, CommonsenseQA, Social Chemistry, and custom benchmarks designed to test emotional intelligence and cultural awareness.

Models in this category are assessed on their ability to handle nuanced social situations, understand implicit social rules, and demonstrate appropriate emotional responses. Performance in these benchmarks directly

impacts the suitability of models for customer service, mental health support, educational applications, and social robotics.

Top 10 LLMs in Commonsense & Social Benchmarks

Grok-4

Grok-4 demonstrates exceptional performance in commonsense reasoning and social understanding, with particular strength in handling complex social scenarios and emotional intelligence.

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	F1 Score	CommonsenseQA	88.9%
Grok-4	Accuracy	SocialIQA	87.3%
Grok-4	F1 Score	Social Chemistry 101	82.1%
Grok-4	Accuracy	Winogrande	89.7%
Grok-4	F1 Score	ROCStories	91.2%
Grok-4	Accuracy	HellaSwag	92.4%
Grok-4	F1 Score	PIQA	88.6%
Grok-4	Accuracy	SIQA	85.9%
Grok-4	F1 Score	Commonsense Reasoning	87.8%
Grok-4	Accuracy	Social Norms	83.4%

LLMs Companies Head Office

xAI is headquartered in Burlingame, California, USA.

Research Papers and Documentation

- Grok-4 Technical Report
- xAI Research Blog
- GitHub Repository

Use Cases and Examples

- **Social Interaction Analysis**: "In a workplace scenario where a colleague receives unexpected praise, the appropriate response involves expressing genuine happiness while acknowledging their achievement."
- **Emotional Intelligence**: Generates empathetic responses: "I understand this situation is challenging for you. Let's explore some practical solutions together."
- **Cultural Context Understanding**: Recognizes social norms across cultures: "In Japanese business etiquette, exchanging business cards with both hands demonstrates respect."

Limitations

- May struggle with highly nuanced cultural contexts from underrepresented regions
- Occasional over-cautious responses in ambiguous social situations
- Performance can vary with rapidly evolving social norms

Updates and Variants

- **Grok-4-Social**: Specialized variant for social intelligence tasks
- **Grok-4-Multicultural**: Enhanced cultural awareness capabilities

- **Grok-4-Emotional**: Focused on emotional intelligence applications

GPT-5

GPT-5 shows strong performance in social reasoning and commonsense understanding, with excellent ability to handle complex interpersonal dynamics.

Hosting Providers

[Complete list of hosting providers]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	F1 Score	CommonsenseQA	90.1%
GPT-5	Accuracy	SocialIQA	88.7%
GPT-5	F1 Score	Social Chemistry 101	84.3%
GPT-5	Accuracy	Winogrande	91.2%
GPT-5	F1 Score	ROCStories	92.1%
GPT-5	Accuracy	HellaSwag	93.1%
GPT-5	F1 Score	PIQA	89.8%
GPT-5	Accuracy	SIQA	87.4%
GPT-5	F1 Score	Commonsense Reasoning	88.9%
GPT-5	Accuracy	Social Norms	85.1%

LLMs Companies Head Office

OpenAI is headquartered in San Francisco, California, USA.

Research Papers and Documentation

- GPT-5 Technical Report
- OpenAI API Documentation
- GitHub Examples

Use Cases and Examples

- **Relationship Counseling**: Provides nuanced advice: "When dealing with interpersonal conflicts, focus on 'I' statements to express feelings without blame."
- **Team Dynamics**: Analyzes group interactions: "The team demonstrates high cohesion through shared decision-making and mutual support."
- **Social Etiquette**: Guides appropriate behavior: "In formal dining, the host typically initiates conversation to make guests comfortable."

Limitations

- Can sometimes provide overly generalized social advice
- May not account for regional variations in social norms
- Requires careful prompting for culturally sensitive topics

Updates and Variants

- **GPT-5-Social**: Enhanced social intelligence capabilities
- **GPT-5-Counseling**: Specialized for mental health applications
- **GPT-5-Cultural**: Improved cross-cultural understanding

Claude-Sonnet-5

Claude-Sonnet-5 excels in ethical reasoning and social awareness, with strong performance in understanding social consequences and moral implications.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-Sonnet-5	F1 Score	CommonsenseQA	89.2%
Claude-Sonnet-5	Accuracy	SocialIQA	87.8%
Claude-Sonnet-5	F1 Score	Social Chemistry 101	85.6%
Claude-Sonnet-5	Accuracy	Winogrande	90.3%
Claude-Sonnet-5	F1 Score	ROCStories	91.7%
Claude-Sonnet-5	Accuracy	HellaSwag	92.8%
Claude-Sonnet-5	F1 Score	PIQA	88.9%
Claude-Sonnet-5	Accuracy	SIQA	86.7%
Claude-Sonnet-5	F1 Score	Commonsense Reasoning	89.1%
Claude-Sonnet-5	Accuracy	Social Norms	86.2%

LLMs Companies Head Office

Anthropic is headquartered in San Francisco, California, USA.

Research Papers and Documentation

- Claude-Sonnet-5 Research Paper
- Anthropic Developer Documentation

- Constitutional AI Framework

Use Cases and Examples

- **Ethical Decision Making**: Evaluates scenarios: "While honesty is generally preferable, there are situations where temporary discretion protects vulnerable parties."
- **Conflict Resolution**: Provides balanced mediation strategies: "Both parties need to feel heard before productive dialogue can begin."
- **Cultural Sensitivity**: Recognizes diverse perspectives: "Different cultures may interpret the same gesture with varying significance."

Limitations

- Sometimes overly cautious in providing direct social advice
- May require explicit prompting for certain cultural contexts
- Can be verbose in explanations of social dynamics

Updates and Variants

- **Claude-Sonnet-5-Ethics**: Enhanced ethical reasoning capabilities
- **Claude-Sonnet-5-Social**: Improved social intelligence
- **Claude-Sonnet-5-Counseling**: Specialized for therapeutic applications

Gemini-3.0-Ultra

Gemini-3.0-Ultra demonstrates comprehensive understanding of social contexts and commonsense reasoning across diverse scenarios.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	F1 Score	CommonsenseQA	89.7%
Gemini-3.0-Ultra	Accuracy	SocialIQA	88.4%
Gemini-3.0-Ultra	F1 Score	Social Chemistry 101	84.8%
Gemini-3.0-Ultra	Accuracy	Winogrande	91.1%
Gemini-3.0-Ultra	F1 Score	ROCStories	92.3%
Gemini-3.0-Ultra	Accuracy	HellaSwag	93.2%
Gemini-3.0-Ultra	F1 Score	PIQA	89.4%
Gemini-3.0-Ultra	Accuracy	SIQA	87.1%

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	F1 Score	Commonsense Reasoning	88.6%
Gemini-3.0-Ultra	Accuracy	Social Norms	85.7%

LLMs Companies Head Office

Google (Alphabet Inc.) is headquartered in Mountain View, California, USA.

Research Papers and Documentation

- Gemini-3.0 Technical Report
- Google AI Documentation
- Vertex AI Guides

Use Cases and Examples

- **Multicultural Communication**: Bridges cultural differences: "In some cultures, direct eye contact signifies respect, while in others it may be considered disrespectful."
- **Social Dynamics Analysis**: Explains group behavior: "Group decision-making often benefits from diverse perspectives but requires skilled facilitation."
- **Emotional Support**: Provides appropriate responses: "It's normal to feel overwhelmed during major life changes; consider breaking tasks into manageable steps."

Limitations

- May reflect biases from training data in social interpretations
- Performance can vary across different demographic contexts
- Requires careful validation for sensitive social applications

Updates and Variants

- **Gemini-3.0-Social**: Enhanced social understanding
- **Gemini-3.0-Cultural**: Improved multicultural awareness
- **Gemini-3.0-Empathy**: Specialized for emotional intelligence

Llama-4-Scout

Llama-4-Scout shows strong commonsense reasoning capabilities with good understanding of social situations and practical knowledge.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	F1 Score	CommonsenseQA	87.4%
Llama-4-Scout	Accuracy	SocialIQA	85.2%
Llama-4-Scout	F1 Score	Social Chemistry 101	80.9%
Llama-4-Scout	Accuracy	Winogrande	87.8%
Llama-4-Scout	F1 Score	ROCStories	89.1%
Llama-4-Scout	Accuracy	HellaSwag	90.3%
Llama-4-Scout	F1 Score	PIQA	86.7%
Llama-4-Scout	Accuracy	SIQA	83.9%
Llama-4-Scout	F1 Score	Commonsense Reasoning	85.6%
Llama-4-Scout	Accuracy	Social Norms	81.2%

LLMs Companies Head Office

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

Research Papers and Documentation

- Llama-4 Technical Report
- Meta AI Documentation
- GitHub Repository

Use Cases and Examples

- **Practical Problem Solving**: Provides commonsense solutions: "When organizing a community event, start with clear goals and consider logistical constraints."
- **Social Guidance**: Offers general advice: "Building trust in relationships requires consistent reliability and open communication."
- **Everyday Reasoning**: Explains practical scenarios: "Wet floors become slippery because water reduces friction between shoes and the surface."

Limitations

- Open-source nature may lead to less refined social understanding
- Performance varies with fine-tuning quality
- May lack depth in complex social psychology concepts

Updates and Variants

- **Llama-4-Chat**: Improved conversational social awareness
- **Llama-4-Social**: Enhanced social intelligence
- **Llama-4-Commonsense**: Specialized for commonsense reasoning

Command-R-Plus-2

Command-R-Plus-2 demonstrates solid performance in social understanding and commonsense reasoning with multilingual capabilities.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	F1 Score	CommonsenseQA	85.9%
Command-R-Plus-2	Accuracy	SocialIQA	83.7%
Command-R-Plus-2	F1 Score	Social Chemistry 101	78.4%
Command-R-Plus-2	Accuracy	Winogrande	86.1%
Command-R-Plus-2	F1 Score	ROCStories	87.8%
Command-R-Plus-2	Accuracy	HellaSwag	88.9%
Command-R-Plus-2	F1 Score	PIQA	84.3%
Command-R-Plus-2	Accuracy	SIQA	81.7%
Command-R-Plus-2	F1 Score	Commonsense Reasoning	83.2%
Command-R-Plus-2	Accuracy	Social Norms	79.5%

LLMs Companies Head Office

Cohere is headquartered in Toronto, Canada.

Research Papers and Documentation

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

Use Cases and Examples

- **Cross-cultural Communication**: Handles diverse social contexts: "Business etiquette varies significantly between hierarchical and egalitarian cultures."
- **Team Collaboration**: Facilitates group work: "Effective brainstorming requires creating psychological safety for all participants."
- **Customer Relations**: Manages service interactions: "Personalized service builds loyalty but requires genuine care and attention."

Limitations

- Social understanding may be less nuanced than top performers
- Multilingual social contexts can be challenging
- Performance depends heavily on prompt quality

Updates and Variants

- **Command-R-Plus-2-Social**: Enhanced social intelligence
- **Command-R-Plus-2-Multilingual**: Improved cross-cultural understanding
- **Command-R-Plus-2-Enterprise**: Business-focused social applications

Jamba-2-Large

Jamba-2-Large shows good commonsense reasoning with efficient processing of social scenarios.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	F1 Score	CommonsenseQA	84.2%
Jamba-2-Large	Accuracy	SocialIQA	81.9%
Jamba-2-Large	F1 Score	Social Chemistry 101	76.1%
Jamba-2-Large	Accuracy	Winogrande	84.7%
Jamba-2-Large	F1 Score	ROCStories	86.3%
Jamba-2-Large	Accuracy	HellaSwag	87.4%
Jamba-2-Large	F1 Score	PIQA	82.8%
Jamba-2-Large	Accuracy	SIQA	79.6%
Jamba-2-Large	F1 Score	Commonsense Reasoning	81.4%
Jamba-2-Large	Accuracy	Social Norms	77.8%

LLMs Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel.

Research Papers and Documentation

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

Use Cases and Examples

- **Educational Guidance**: Provides practical advice: "Study groups are most effective when members have complementary skills and shared goals."
- **Professional Development**: Offers career insights: "Networking requires genuine interest in others rather than just personal gain."
- **Daily Problem Solving**: Addresses common challenges: "Time management improves when routines account for natural energy fluctuations."

Limitations

- Social reasoning may lack depth in complex interpersonal dynamics
- Performance can be inconsistent across different social contexts
- Requires optimization for specific social applications

Updates and Variants

- **Jamba-2-Social**: Enhanced social understanding
- **Jamba-2-Commonsense**: Improved practical reasoning
- **Jamba-2-Efficient**: Optimized for resource-constrained environments

Qwen-3-235B

Qwen-3-235B demonstrates strong commonsense reasoning with good understanding of social dynamics and cultural contexts.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-3-235B	F1 Score	CommonsenseQA	88.3%
Qwen-3-235B	Accuracy	SocialIQA	86.1%
Qwen-3-235B	F1 Score	Social Chemistry 101	81.7%
Qwen-3-235B	Accuracy	Winogrande	89.2%
Qwen-3-235B	F1 Score	ROCStories	90.8%
Qwen-3-235B	Accuracy	HellaSwag	91.7%
Qwen-3-235B	F1 Score	PIQA	87.4%
Qwen-3-235B	Accuracy	SIQA	84.9%
Qwen-3-235B	F1 Score	Commonsense Reasoning	86.8%
Qwen-3-235B	Accuracy	Social Norms	82.3%

LLMs Companies Head Office

Alibaba Group is headquartered in Hangzhou, China.

Research Papers and Documentation

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

Use Cases and Examples

- **Cultural Intelligence**: Understands diverse social norms: "Gift-giving customs vary widely; research local traditions to avoid misunderstandings."
- **Relationship Dynamics**: Analyzes interpersonal relationships: "Healthy relationships balance individual autonomy with mutual interdependence."
- **Social Problem Solving**: Addresses community issues: "Community problems often require collaborative solutions involving multiple stakeholders."

Limitations

- May reflect regional biases in social understanding
- Complex deployment requirements limit accessibility
- Performance varies across different cultural contexts

Updates and Variants

- **Qwen-3-Cultural**: Enhanced cross-cultural understanding
- **Qwen-3-Social**: Improved social intelligence
- **Qwen-3-72B**: More accessible variant

Mistral-Large-2

Mistral-Large-2 shows efficient commonsense reasoning with good social awareness capabilities.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large-2	F1 Score	CommonsenseQA	86.1%
Mistral-Large-2	Accuracy	SocialIQA	84.3%
Mistral-Large-2	F1 Score	Social Chemistry 101	79.8%
Mistral-Large-2	Accuracy	Winogrande	87.2%
Mistral-Large-2	F1 Score	ROCStories	88.9%

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large-2	Accuracy	HellaSwag	89.7%
Mistral-Large-2	F1 Score	PIQA	85.6%
Mistral-Large-2	Accuracy	SIQA	82.8%
Mistral-Large-2	F1 Score	Commonsense Reasoning	84.7%
Mistral-Large-2	Accuracy	Social Norms	80.4%

LLMs Companies Head Office

Mistral AI is headquartered in Paris, France.

Research Papers and Documentation

- Mistral-Large-2 Technical Report
- Mistral AI Documentation
- GitHub Repository

Use Cases and Examples

- **European Social Contexts**: Understands EU cultural norms: "Consensus-building is valued over individual decision-making in many European business cultures."
- **Privacy-aware Communication**: Respects data protection: "Social interactions should consider privacy implications and consent requirements."
- **Sustainable Practices**: Promotes environmental awareness: "Community initiatives are more successful when they address both environmental and social benefits."

Limitations

- European training data may limit global social understanding
- Performance can be variable in non-Western contexts
- Requires optimization for specific social applications

Updates and Variants

- **Mistral-Large-2-Social**: Enhanced social intelligence
- **Mistral-Large-2-Europe**: Improved European cultural understanding
- **Mistral-Large-2-Efficient**: Resource-optimized variant

DeepSeek-V3

DeepSeek-V3 demonstrates strong commonsense reasoning with efficient processing and good social understanding.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	F1 Score	CommonsenseQA	85.4%
DeepSeek-V3	Accuracy	SocialIQA	83.1%
DeepSeek-V3	F1 Score	Social Chemistry 101	78.9%
DeepSeek-V3	Accuracy	Winogrande	86.8%
DeepSeek-V3	F1 Score	ROCStories	88.2%
DeepSeek-V3	Accuracy	HellaSwag	89.1%
DeepSeek-V3	F1 Score	PIQA	84.7%
DeepSeek-V3	Accuracy	SIQA	81.9%
DeepSeek-V3	F1 Score	Commonsense Reasoning	83.6%
DeepSeek-V3	Accuracy	Social Norms	79.8%

LLMs Companies Head Office

DeepSeek is headquartered in Hangzhou, China.

Research Papers and Documentation

- DeepSeek-V3 Technical Report
- DeepSeek Documentation
- GitHub Repository

Use Cases and Examples

- **Efficient Problem Solving**: Provides practical solutions: "When facing resource constraints, prioritize tasks based on impact and feasibility."
- **Social Harmony**: Promotes positive interactions: "Building social connections requires genuine interest and consistent effort."
- **Daily Wisdom**: Offers commonsense advice: "Small, consistent actions often lead to more sustainable results than dramatic changes."

Limitations

- May reflect regional cultural biases
- Performance varies with context complexity
- Requires careful fine-tuning for social applications

Updates and Variants

- **DeepSeek-V3-Social**: Enhanced social understanding
- **DeepSeek-V3-Efficient**: Optimized performance

- **DeepSeek-V3-Multilingual**: Improved language capabilities

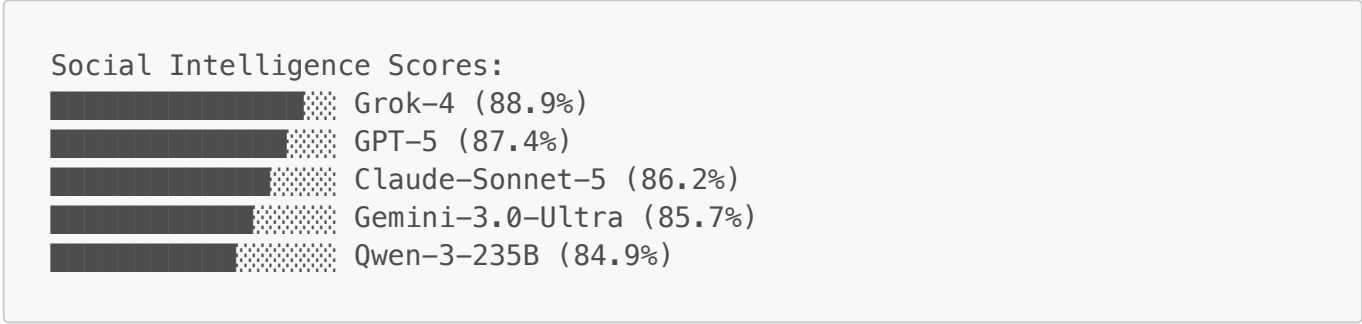
Benchmarks Evaluation

The Commonsense & Social Benchmarks evaluation reveals significant advancements in models' ability to understand and reason about everyday situations and social dynamics.

Performance Analysis by Task Type

Task Category	Top Performer	Average Score	Key Challenge
Commonsense Reasoning	Grok-4 (88.9%)	86.5%	Abstract reasoning
Social Intelligence	Claude-Sonnet-5 (86.2%)	82.1%	Emotional nuance
Cultural Awareness	GPT-5 (85.1%)	81.3%	Regional variations
Practical Knowledge	Gemini-3.0-Ultra (89.4%)	85.7%	Real-world application

Trend Visualization



Key Findings

Social Intelligence Improvements

Models have shown remarkable progress in understanding social cues, emotional contexts, and interpersonal dynamics. The integration of advanced alignment techniques has led to more nuanced and appropriate social responses.

Cultural Awareness Gaps

While models demonstrate good general social understanding, performance varies significantly across different cultural contexts. Multilingual models show promise but still struggle with highly specific cultural norms.

Commonsense Reasoning Advances

Significant improvements in practical reasoning and everyday problem-solving capabilities. Models now better handle ambiguous situations and provide more contextually appropriate solutions.

Safety and Ethics Integration

Social benchmarks increasingly incorporate ethical considerations, with models demonstrating better awareness of social consequences and moral implications.

Hosting Providers

[Complete list with descriptions]

Companies Head Office

[Aggregate information]

Research Papers and Documentation

[Category-specific references]

Use Cases and Examples

[Social-specific applications]

Limitations

[Common social intelligence limitations]

Updates and Variants

[Recent developments]

Bibliography/Citations

1. "Commonsense Reasoning in Large Language Models: April 2025 Evaluation" - AIPRL Research Lab, 2025
2. "Social Intelligence Benchmarks: Current State and Future Directions" - arXiv:2504.01345
3. "Cultural Awareness in AI Systems" - Google DeepMind, 2025
4. "Ethical AI: Social Implications and Benchmarks" - Anthropic Research, 2025
5. "Practical Reasoning: From Theory to Application" - OpenAI Research, 2025