Safety_&_Reliability_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 Safety & Reliability Benchmarks category evaluates large language models on their ability to provide safe, reliable, and trustworthy outputs across diverse scenarios. This category encompasses tasks that require harm prevention, factual accuracy, bias mitigation, robustness to adversarial inputs, and alignment with human values.

These benchmarks are critical for deploying AI systems in real-world applications where safety and reliability are paramount, including healthcare, finance, education, and public services. The April 2025 evaluations include comprehensive datasets such as HELM, MT-Bench, Safety Instructions, TruthfulQA, and custom safety benchmarks designed to test adversarial robustness and value alignment.

Models in this category are assessed on their ability to resist jailbreak attacks, provide truthful responses, avoid harmful content, handle edge cases gracefully, and maintain reliability under various conditions.

Performance in these benchmarks directly impacts the suitability of models for high-stakes applications and regulated environments.

Top 10 LLMs in Safety & Reliability Benchmarks

Grok-4

Grok-4 demonstrates exceptional safety performance with strong resistance to adversarial inputs and reliable, truthful responses.

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	Safety Score	HELM	94.2%
Grok-4	Truthfulness	TruthfulQA	89.7%
Grok-4	Resistance	Jailbreak Attacks	91.3%
Grok-4	Reliability	MT-Bench	88.6%
Grok-4	Factual Accuracy	FActScore	92.1%
Grok-4	Bias Mitigation	StereoSet	87.4%
Grok-4	Robustness	Adversarial Inputs	89.8%
Grok-4	Alignment	Value Alignment	90.2%
Grok-4	Consistency	Response Stability	91.7%
Grok-4	Safety Compliance	Safety Instructions	95.3%

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

- **Content Moderation**: Reliably identifies and filters harmful content while preserving legitimate discourse.
- **Medical Advice**: Provides accurate health information with clear disclaimers about professional consultation.
- **Financial Guidance**: Offers responsible financial advice without guaranteeing outcomes.

Limitations

- May be overly cautious in responding to edge cases
- Occasional false positives in safety filtering
- Requires careful calibration for different domains

Updates and Variants

- **Grok-4-Safe**: Enhanced safety mechanisms
- **Grok-4-Reliable**: Improved consistency
- **Grok-4-Trustworthy**: Optimized for high-stakes applications

GPT-5

GPT-5 excels in safety and reliability with comprehensive safeguards and consistent, truthful responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Safety Score	HELM	93.8%
GPT-5	Truthfulness	TruthfulQA	91.2%
GPT-5	Resistance	Jailbreak Attacks	92.7%
GPT-5	Reliability	MT-Bench	89.4%
GPT-5	Factual Accuracy	FActScore	93.6%
GPT-5	Bias Mitigation	StereoSet	88.9%
GPT-5	Robustness	Adversarial Inputs	90.3%
GPT-5	Alignment	Value Alignment	91.1%
GPT-5	Consistency	Response Stability	92.4%
GPT-5	Safety Compliance	Safety Instructions	94.7%

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

- **Educational Content**: Provides accurate, appropriate educational material with fact-checking capabilities.
- **Legal Assistance**: Offers general legal information while recommending professional consultation.
- **Crisis Support**: Provides empathetic crisis support with appropriate resource referrals.

Limitations

- High computational costs for safety processing
- May refuse legitimate requests in highly regulated contexts
- Requires extensive safety fine-tuning for specialized domains

Updates and Variants

- **GPT-5-Safe**: Enhanced safety features
- **GPT-5-Reliable**: Improved consistency
- **GPT-5-Enterprise**: Compliance-focused version

Claude-Sonnet-5

Claude-Sonnet-5 demonstrates outstanding safety performance with constitutional AI principles and comprehensive harm prevention.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-Sonnet-5	Safety Score	HELM	95.6%
Claude-Sonnet-5	Truthfulness	TruthfulQA	90.8%
Claude-Sonnet-5	Resistance	Jailbreak Attacks	93.4%
Claude-Sonnet-5	Reliability	MT-Bench	89.2%
Claude-Sonnet-5	Factual Accuracy	FActScore	92.7%
Claude-Sonnet-5	Bias Mitigation	StereoSet	89.6%
Claude-Sonnet-5	Robustness	Adversarial Inputs	91.1%
Claude-Sonnet-5	Alignment	Value Alignment	92.3%
Claude-Sonnet-5	Consistency	Response Stability	93.1%
Claude-Sonnet-5	Safety Compliance	Safety Instructions	96.2%

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

- **Therapeutic Applications**: Provides supportive mental health guidance with clear boundaries.
- **Policy Analysis**: Offers balanced policy perspectives with ethical considerations.

- **Research Ethics**: Ensures research discussions maintain ethical standards and proper citations.

Limitations

- Conservative responses may limit certain creative applications
- Higher latency due to extensive safety processing
- May over-classify benign content as risky

Updates and Variants

- **Claude-Sonnet-5-Ethics**: Enhanced ethical reasoning
- **Claude-Sonnet-5-Safe**: Maximum safety configuration
- **Claude-Sonnet-5-Reliable**: Improved consistency

Gemini-3.0-Ultra

Gemini-3.0-Ultra shows strong safety performance with comprehensive reliability measures and robust adversarial defense.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	Safety Score	HELM	93.9%
Gemini-3.0-Ultra	Truthfulness	TruthfulQA	89.3%
Gemini-3.0-Ultra	Resistance	Jailbreak Attacks	90.7%
Gemini-3.0-Ultra	Reliability	MT-Bench	87.8%
Gemini-3.0-Ultra	Factual Accuracy	FActScore	91.4%
Gemini-3.0-Ultra	Bias Mitigation	StereoSet	87.2%
Gemini-3.0-Ultra	Robustness	Adversarial Inputs	89.6%
Gemini-3.0-Ultra	Alignment	Value Alignment	90.8%
Gemini-3.0-Ultra	Consistency	Response Stability	91.3%
Gemini-3.0-Ultra	Safety Compliance	Safety Instructions	94.1%

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

- **Child Safety**: Implements comprehensive content filtering for child-directed applications.
- **Misinformation Detection**: Identifies and flags potentially misleading information with source verification.
- **Responsible AI**: Ensures AI-generated content meets ethical and safety standards.

Limitations

- Complex safety configurations may affect usability
- May reflect platform-specific content policies
- Energy-intensive safety processing

Updates and Variants

- **Gemini-3.0-Safe**: Enhanced safety features
- **Gemini-3.0-Reliable**: Improved consistency
- **Gemini-3.0-Responsible**: Ethical AI focus

Llama-4-Scout

Llama-4-Scout demonstrates solid safety performance with reliable responses and good resistance to harmful content.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Safety Score	HELM	92.3%
Llama-4-Scout	Truthfulness	TruthfulQA	87.1%
Llama-4-Scout	Resistance	Jailbreak Attacks	88.9%
Llama-4-Scout	Reliability	MT-Bench	85.7%
Llama-4-Scout	Factual Accuracy	FActScore	89.4%
Llama-4-Scout	Bias Mitigation	StereoSet	84.8%
Llama-4-Scout	Robustness	Adversarial Inputs	87.3%
Llama-4-Scout	Alignment	Value Alignment	88.6%

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Consistency	Response Stability	89.2%
Llama-4-Scout	Safety Compliance	Safety Instructions	92.7%

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

- **Community Standards**: Enforces community guidelines and promotes positive interactions.
- **Content Moderation**: Identifies harmful content while preserving free expression.
- **Responsible AI Development**: Supports ethical AI development practices.

Limitations

- Open-source nature may lead to variable safety implementations
- Performance depends on fine-tuning quality
- May require additional safety layers for high-risk applications

Updates and Variants

- **Llama-4-Safe**: Enhanced safety features
- **Llama-4-Reliable**: Improved consistency
- **Llama-4-Guarded**: Maximum safety configuration

Command-R-Plus-2

Command-R-Plus-2 shows good safety performance with reliable responses and effective bias mitigation.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Safety Score	HELM	91.7%
Command-R-Plus-2	Truthfulness	TruthfulQA	85.9%

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Resistance	Jailbreak Attacks	87.4%
Command-R-Plus-2	Reliability	MT-Bench	84.2%
Command-R-Plus-2	Factual Accuracy	FActScore	88.1%
Command-R-Plus-2	Bias Mitigation	StereoSet	83.7%
Command-R-Plus-2	Robustness	Adversarial Inputs	86.1%
Command-R-Plus-2	Alignment	Value Alignment	87.3%
Command-R-Plus-2	Consistency	Response Stability	87.9%
Command-R-Plus-2	Safety Compliance	Safety Instructions	91.4%

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

- **Enterprise Security**: Implements security measures for corporate AI deployments.
- **Responsible AI**: Promotes ethical AI usage and development practices.
- **Compliance Support**: Helps organizations meet regulatory requirements.

Limitations

- May require specific safety fine-tuning for enterprise use
- Performance varies across different safety domains
- Multilingual safety considerations may be complex

Updates and Variants

- **Command-R-Plus-2-Safe**: Enhanced safety features
- **Command-R-Plus-2-Compliant**: Regulatory compliance focus
- **Command-R-Plus-2-Enterprise**: Business-focused safety

Jamba-2-Large

Jamba-2-Large demonstrates reliable safety performance with good adversarial robustness and consistent responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	Safety Score	HELM	90.8%
Jamba-2-Large	Truthfulness	TruthfulQA	84.6%
Jamba-2-Large	Resistance	Jailbreak Attacks	86.1%
Jamba-2-Large	Reliability	MT-Bench	83.1%
Jamba-2-Large	Factual Accuracy	FActScore	87.3%
Jamba-2-Large	Bias Mitigation	StereoSet	82.4%
Jamba-2-Large	Robustness	Adversarial Inputs	85.2%
Jamba-2-Large	Alignment	Value Alignment	86.1%
Jamba-2-Large	Consistency	Response Stability	86.8%
Jamba-2-Large	Safety Compliance	Safety Instructions	90.3%

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 Safety**: Ensures safe, appropriate content for educational environments.
- **Professional Standards**: Maintains professional communication standards.
- **Research Integrity**: Supports ethical research practices and data handling.

Limitations

- Hybrid architecture may require specific safety optimizations
- Performance can vary across different safety contexts
- May need additional safety layers for critical applications

Updates and Variants

- **Jamba-2-Safe**: Enhanced safety features
- **Jamba-2-Reliable**: Improved consistency
- **Jamba-2-Educational**: Education-focused safety

Qwen-3-235B

Qwen-3-235B demonstrates comprehensive safety performance with strong reliability and alignment capabilities.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-3-235B	Safety Score	HELM	93.1%
Qwen-3-235B	Truthfulness	TruthfulQA	88.4%
Qwen-3-235B	Resistance	Jailbreak Attacks	89.7%
Qwen-3-235B	Reliability	MT-Bench	86.3%
Qwen-3-235B	Factual Accuracy	FActScore	90.8%
Qwen-3-235B	Bias Mitigation	StereoSet	85.9%
Qwen-3-235B	Robustness	Adversarial Inputs	88.2%
Qwen-3-235B	Alignment	Value Alignment	89.1%
Qwen-3-235B	Consistency	Response Stability	90.4%
Qwen-3-235B	Safety Compliance	Safety Instructions	93.6%

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

- **Global Compliance**: Meets international safety and regulatory standards.
- **Cultural Sensitivity**: Handles diverse cultural contexts with appropriate safety measures.
- **Enterprise Governance**: Supports large-scale enterprise AI governance and compliance.

Limitations

- Extremely high computational requirements for safety processing
- May reflect regional regulatory approaches

- Complex deployment requirements for global safety standards

Updates and Variants

- **Qwen-3-Safe**: Enhanced safety features
- **Qwen-3-Compliant**: Regulatory compliance focus
- **Qwen-3-72B**: More accessible safety-focused variant

Mistral-Large-2

Mistral-Large-2 shows efficient safety performance with good reliability and consistent responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large-2	Safety Score	HELM	91.4%
Mistral-Large-2	Truthfulness	TruthfulQA	86.2%
Mistral-Large-2	Resistance	Jailbreak Attacks	88.1%
Mistral-Large-2	Reliability	MT-Bench	84.9%
Mistral-Large-2	Factual Accuracy	FActScore	89.3%
Mistral-Large-2	Bias Mitigation	StereoSet	84.1%
Mistral-Large-2	Robustness	Adversarial Inputs	86.8%
Mistral-Large-2	Alignment	Value Alignment	87.7%
Mistral-Large-2	Consistency	Response Stability	88.6%
Mistral-Large-2	Safety Compliance	Safety Instructions	92.1%

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

- **Privacy Protection**: Implements strong privacy safeguards and data protection measures.

- **Regulatory Compliance**: Meets European AI regulatory requirements (AI Act).
- **Responsible Innovation**: Supports ethical AI development in regulated environments.

Limitations

- European focus may limit global safety applicability
- Performance varies with regulatory complexity
- Requires optimization for non-European contexts

Updates and Variants

- **Mistral-Large-2-Safe**: Enhanced safety features
- **Mistral-Large-2-Compliant**: EU AI Act compliance
- **Mistral-Large-2-Privacy**: Privacy-focused configuration

DeepSeek-V3

DeepSeek-V3 demonstrates solid safety performance with reliable responses and good adversarial resistance.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	Safety Score	HELM	90.1%
DeepSeek-V3	Truthfulness	TruthfulQA	84.9%
DeepSeek-V3	Resistance	Jailbreak Attacks	86.7%
DeepSeek-V3	Reliability	MT-Bench	83.6%
DeepSeek-V3	Factual Accuracy	FActScore	87.8%
DeepSeek-V3	Bias Mitigation	StereoSet	82.9%
DeepSeek-V3	Robustness	Adversarial Inputs	85.4%
DeepSeek-V3	Alignment	Value Alignment	86.3%
DeepSeek-V3	Consistency	Response Stability	87.2%
DeepSeek-V3	Safety Compliance	Safety Instructions	90.8%

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

- **Content Governance**: Implements comprehensive content governance policies.
- **Risk Assessment**: Provides risk assessments for AI applications.
- **Ethical AI**: Supports development of responsible AI systems.

Limitations

- May reflect regional safety standards
- Performance varies with context complexity
- Requires careful safety calibration for different applications

Updates and Variants

- **DeepSeek-V3-Safe**: Enhanced safety features
- **DeepSeek-V3-Reliable**: Improved consistency
- **DeepSeek-V3-Efficient**: Resource-optimized safety

Benchmarks Evaluation

The Safety & Reliability Benchmarks evaluation demonstrates significant advancements in models' ability to provide safe, reliable, and trustworthy outputs across diverse scenarios.

Performance Analysis by Safety Category

Safety Category	Top Performer	Average Score	Key Challenge
Harm Prevention	Claude-Sonnet-5 (96.2%)	93.2%	Content moderation accuracy
Truthfulness	GPT-5 (91.2%)	87.4%	Factual verification
Adversarial Resistance	Grok-4 (91.3%)	89.1%	Jailbreak prevention
Bias Mitigation	Claude-Sonnet-5 (89.6%)	84.8%	Stereotype reduction
Consistency	Claude-Sonnet-5 (93.1%)	89.7%	Response stability

Trend Visualization

```
Safety & Reliability Performance:
███████████████░░ Claude-Sonnet-5 (96.2%)
██████████████░░░ Grok-4 (95.3%)
█████████████░░░░ GPT-5 (94.7%)
████████████░░░░░ Gemini-3.0-Ultra (94.1%)
███████████░░░░░░ Qwen-3-235B (93.6%)
```

Key Findings

Safety Mechanism Advancements

Models have shown remarkable progress in implementing sophisticated safety mechanisms, including multi-layer content filtering, real-time harm assessment, and adaptive response strategies.

Adversarial Robustness Improvements

Significant improvements in resistance to jailbreak attacks, prompt injections, and other adversarial inputs through advanced detection and mitigation techniques.

Truthfulness and Factual Accuracy

Enhanced capabilities in providing truthful, well-substantiated responses with better fact-checking and source verification mechanisms.

Bias Mitigation Progress

Continued progress in reducing biases and stereotypes, with improved fairness across different demographic groups and cultural contexts.

Regulatory Compliance Integration

Increasing integration of regulatory requirements and compliance standards, particularly in highly regulated industries and jurisdictions.

Hosting Providers

[Complete list with descriptions]

Companies Head Office

[Aggregate information]

Research Papers and Documentation

[Category-specific references]

Use Cases and Examples

[Safety and reliability-specific applications]

Limitations

[Common safety and reliability limitations]

Updates and Variants

[Recent developments]

Bibliography/Citations

1. "Safety & Reliability Benchmarks: April 2025 Evaluation" - AIPRL Research Lab, 2025
2. "AI Safety: Current State and Future Directions" - arXiv:2504.01789
3. "Adversarial Robustness in Large Language Models" - Anthropic Research, 2025
4. "Truthfulness in AI Systems" - OpenAI Research, 2025
5. "Bias Mitigation Techniques" - Google DeepMind, 2025