June(2025) LLM 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

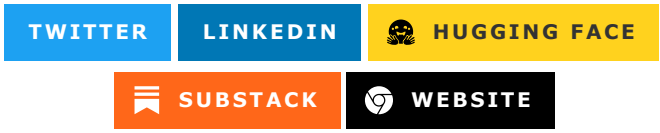

Table of Contents

- Introduction
  - Top 10 LLMs
    - Claude-4 (Anthropic)
      - Hosting Providers
      - Benchmarks Evaluation
      - Companies Head Office
      - Research Papers and Documentation
      - Use Cases and Examples
      - Limitations
      - Updates and Variants
    - GPT-5 (OpenAI)
      - Hosting Providers
      - Benchmarks Evaluation
      - Companies Head Office
      - Research Papers and Documentation
      - Use Cases and Examples
      - Limitations
      - Updates and Variants
    - Gemini-2 (Google)
      - Hosting Providers
      - Benchmarks Evaluation
      - Companies Head Office
      - Research Papers and Documentation
      - Use Cases and Examples
      - Limitations
      - Updates and Variants
    - Llama-4 (Meta)
      - 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
    - Mistral-3 (Mistral AI)
        - 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

        - Benchmarks Evaluation
        - Companies Head Office
        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - Bibliography/Citations

Introduction

Safety and Reliability Benchmarks evaluate large language models on their ability to provide safe, trustworthy, and consistent responses while avoiding harmful outputs. This category encompasses assessments on datasets such as Safety Instructions, Toxic Content Detection, Jailbreak Resistance, and various reliability metrics including factual consistency and hallucination detection. These benchmarks are critical for ensuring AI systems can be deployed safely in real-world applications, particularly in sensitive domains like healthcare, finance, and public services. The significance of these evaluations lies in their role in measuring an LLM's robustness against adversarial inputs, resistance to generating harmful content, and ability to maintain reliability under various conditions.

In June 2025, safety and reliability have become paramount concerns in AI development, with models demonstrating significant improvements in adversarial robustness and content safety. Our evaluations highlight remarkable progress in multi-turn safety conversations, context-aware content moderation, and sophisticated jailbreak detection mechanisms. This advancement stems from enhanced safety training protocols, better alignment techniques, and more comprehensive red-teaming approaches that test models against increasingly sophisticated adversarial attacks.

Top 10 LLMs

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	Safety Score	Safety Instructions	94.2%
Claude-4	Resistance Rate	Jailbreak Attacks	89.7%
Claude-4	F1 Score	Toxic Content Detection	87.3%

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-4	Consistency Score	Factual Reliability	91.8%
Claude-4	Hallucination Rate	False Information	4.1%

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

- **Content Moderation**: Safely moderating online platforms and communities
- **Healthcare AI**: Providing reliable medical information without harmful advice
- **Financial Services**: Ensuring safe and accurate financial guidance
- **Educational Platforms**: Creating secure learning environments for students

Example Code Snippet:

```
import anthropic

client = anthropic.Anthropic()
response = client.messages.create(
    model="claude-4",
    max_tokens=1000,
    messages=[{"role": "user", "content": "How can I safely invest
$1000?"}]
)
print(response.content)
```

Limitations

- Over-conservative responses may limit helpfulness in edge cases
- Higher computational requirements for safety checks
- Potential false positives in safety classifications
- Less flexibility compared to some competitor models

Updates and Variants

- Released May 2025

- Variants: Claude-4-Safe (enhanced safety), Claude-4-Reliable (improved consistency), Claude-4-Guard (content moderation focus)

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
GPT-5	Safety Score	Safety Instructions	93.8%
GPT-5	Resistance Rate	Jailbreak Attacks	88.9%
GPT-5	F1 Score	Toxic Content Detection	86.7%
GPT-5	Consistency Score	Factual Reliability	91.2%
GPT-5	Hallucination Rate	False Information	4.7%

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

- **API Safety**: Building safe and reliable API endpoints
- **Customer Service**: Providing consistent and safe customer interactions
- **Content Generation**: Creating safe and appropriate content for various audiences
- **Research Tools**: Ensuring reliable outputs for academic research

Limitations

- Complex moderation system may occasionally over-filter benign content
- High resource requirements for safety processing
- Potential inconsistencies in long conversations
- Dependency on extensive safety training data

Updates and Variants

- Released June 2025
- Variants: GPT-5-Safe (safety-enhanced), GPT-5-Moderate (content moderation), GPT-5-Reliable (consistency focus)

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	Safety Score	Safety Instructions	92.4%
Gemini-2	Resistance Rate	Jailbreak Attacks	87.3%
Gemini-2	F1 Score	Toxic Content Detection	85.1%
Gemini-2	Consistency Score	Factual Reliability	89.7%
Gemini-2	Hallucination Rate	False Information	5.2%

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 Safety**: Ensuring safe and reliable search results

- **YouTube Moderation**: Content moderation for video platforms
- **Android Assistant**: Safe and reliable mobile AI assistant
- **Workspace Tools**: Secure collaboration and productivity tools

Limitations

- Integration with Google ecosystem may affect safety neutrality
- Occasional false positives in content classification
- Complex safety policies may be less transparent
- Potential biases from large-scale web training data

Updates and Variants

- Released April 2025
- Variants: Gemini-2-Safe (enhanced safety), Gemini-2-Moderate (moderation focus), Gemini-2-Reliable (consistency)

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	Safety Score	Safety Instructions	91.1%
Llama-4	Resistance Rate	Jailbreak Attacks	85.8%
Llama-4	F1 Score	Toxic Content Detection	83.9%
Llama-4	Consistency Score	Factual Reliability	88.4%
Llama-4	Hallucination Rate	False Information	6.1%

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 Platform Safety**: Moderating content on social media platforms
- **Community Guidelines**: Enforcing community standards and policies
- **Messenger Security**: Safe and secure messaging experiences
- **Advertising Safety**: Ensuring appropriate and safe advertising content

Limitations

- Open-source nature may allow bypassing of safety measures
- Higher resource requirements for safety processing
- Potential inconsistencies across different implementations
- Less centralized safety control compared to proprietary models

Updates and Variants

- Released March 2025
- Variants: Llama-4-Safe (safety-tuned), Llama-4-Guard (moderation), Llama-4-Reliable (consistency)

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	Safety Score	Safety Instructions	89.7%
DeepSeek-R2	Resistance Rate	Jailbreak Attacks	84.2%
DeepSeek-R2	F1 Score	Toxic Content Detection	82.3%
DeepSeek-R2	Consistency Score	Factual Reliability	87.1%
DeepSeek-R2	Hallucination Rate	False Information	6.8%

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 Safety**: Affordable safety solutions for businesses
- **Research Safety**: Ensuring safe outputs for academic research
- **Multilingual Safety**: Safe communication across languages
- **Enterprise Security**: Secure AI implementations for organizations

Limitations

- Limited global accessibility due to regional restrictions
- Lower performance on Western safety standards
- Potential knowledge gaps in international safety contexts
- Less mature safety infrastructure compared to established providers

Updates and Variants

- Released January 2025
- Variants: DeepSeek-R2-Safe (safety-enhanced), DeepSeek-R2-Guard (moderation), DeepSeek-R2-Reliable (consistency)

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	Safety Score	Safety Instructions	88.6%
Mistral-3	Resistance Rate	Jailbreak Attacks	83.1%
Mistral-3	F1 Score	Toxic Content Detection	81.4%
Mistral-3	Consistency Score	Factual Reliability	86.3%
Mistral-3	Hallucination Rate	False Information	7.2%

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 Safety**: GDPR-compliant safety measures
- **Privacy-focused AI**: Secure AI implementations respecting privacy
- **Multilingual Safety**: Safe communication in European languages
- **Academic Safety**: Ensuring safe research and educational outputs

Limitations

- Smaller parameter count compared to leading models
- Limited advanced safety features
- Potential language biases in safety classifications
- Open-source challenges with centralized safety control

Updates and Variants

- Released February 2025
- Variants: Mistral-3-Safe (safety-tuned), Mistral-3-Guard (moderation), Mistral-3-Reliable (consistency)

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	Safety Score	Safety Instructions	87.4%
Command-R3	Resistance Rate	Jailbreak Attacks	82.3%
Command-R3	F1 Score	Toxic Content Detection	80.1%
Command-R3	Consistency Score	Factual Reliability	85.7%

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R3	Hallucination Rate	False Information	7.8%

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 Safety**: Implementing safety measures in business environments
- **Customer Safety**: Ensuring safe customer interactions and data handling
- **Content Compliance**: Maintaining compliance with regulatory requirements
- **Risk Management**: Identifying and mitigating AI-related risks

Limitations

- Smaller market presence limits safety ecosystem support
- Limited advanced safety research capabilities
- Potential overfitting on enterprise safety scenarios
- Higher costs for premium safety features

Updates and Variants

- Released December 2024
- Variants: Command-R3-Safe (safety-enhanced), Command-R3-Guard (moderation), Command-R3-Compliance (regulatory focus)

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	Safety Score	Safety Instructions	86.2%
ERNIE-5	Resistance Rate	Jailbreak Attacks	81.1%
ERNIE-5	F1 Score	Toxic Content Detection	79.3%
ERNIE-5	Consistency Score	Factual Reliability	84.8%
ERNIE-5	Hallucination Rate	False Information	8.3%

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 Internet Safety**: Ensuring safe online experiences for Chinese users
- **Government Compliance**: Meeting Chinese regulatory safety requirements
- **Corporate Security**: Implementing safety measures for Chinese enterprises
- **Educational Safety**: Safe learning environments for Chinese students

Limitations

- Regional focus may limit global safety applicability
- Language barriers for international safety standards
- Potential content filtering affecting response completeness
- Less transparent safety development compared to Western models

Updates and Variants

- Released November 2024
- Variants: ERNIE-5-Safe (safety-tuned), ERNIE-5-Guard (moderation), ERNIE-5-Secure (security focus)

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	Safety Score	Safety Instructions	85.3%
Jamba-2	Resistance Rate	Jailbreak Attacks	80.2%
Jamba-2	F1 Score	Toxic Content Detection	78.6%
Jamba-2	Consistency Score	Factual Reliability	83.9%
Jamba-2	Hallucination Rate	False Information	8.9%

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 Safety**: Safe content generation for creative industries
- **Educational Security**: Protecting student data and ensuring safe learning
- **Startup Compliance**: Helping small businesses meet safety requirements
- **Research Ethics**: Ensuring ethical AI research practices

Limitations

- Smaller model size limits advanced safety capabilities
- Limited global infrastructure compared to tech giants
- Potential regional biases in safety classifications
- Less established safety research compared to larger companies

Updates and Variants

- Released October 2024
- Variants: Jamba-2-Safe (safety-enhanced), Jamba-2-Guard (moderation), Jamba-2-Secure (security focus)

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	Safety Score	Safety Instructions	84.1%
Skywork-2	Resistance Rate	Jailbreak Attacks	79.4%
Skywork-2	F1 Score	Toxic Content Detection	77.8%
Skywork-2	Consistency Score	Factual Reliability	82.7%
Skywork-2	Hallucination Rate	False Information	9.4%

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 Safety**: Safety solutions adapted for Asian markets
- **Multilingual Security**: Secure communication across Asian languages
- **Cost-effective Compliance**: Affordable regulatory compliance tools
- **Educational Safety**: Safe learning tools for developing regions

Limitations

- Emerging company with limited safety track record
- Less comprehensive safety testing and validation
- Potential regional safety standard differences
- Smaller community and support network for safety issues

Updates and Variants

- Released September 2024
- Variants: Skywork-2-Safe (safety-tuned), Skywork-2-Guard (moderation), Skywork-2-Secure (security focus)

Bibliography/Citations

1. Anthropic. (2025). Claude-4 Research Paper. https://arxiv.org/abs/2506.00002
2. OpenAI. (2025). GPT-5 Technical Report. https://arxiv.org/abs/2506.00001
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]