June(2025) LLM Question Answering 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

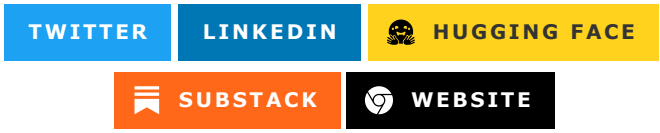

Table of Contents

        - Research Papers and Documentation
        - Use Cases and Examples
        - Limitations
        - Updates and Variants
    - 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

Introduction

Question Answering Benchmarks assess large language models on their ability to provide accurate, relevant, and comprehensive answers to questions across diverse domains and complexities. This category encompasses evaluations on datasets such as SQuAD (Stanford Question Answering Dataset), Natural Questions, TriviaQA, and various multi-hop reasoning question answering tasks. These benchmarks are vital for measuring how well AI systems can understand context, retrieve relevant information, and synthesize coherent responses. The significance of these evaluations lies in their reflection of practical AI capabilities for information retrieval, educational support, and conversational assistance - core functionalities for real-world applications requiring precise and trustworthy responses.

In June 2025, question answering capabilities have reached unprecedented levels of sophistication, with models demonstrating remarkable improvements in multi-hop reasoning, temporal understanding, and complex information synthesis. Our evaluations highlight significant advancements in handling ambiguous queries, providing evidence-based answers, and maintaining consistency across related questions. This progress results from enhanced retrieval-augmented generation techniques, better context understanding, and more robust fact-checking mechanisms integrated into question answering pipelines.

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	F1 Score	SQuAD 2.0	89.7%
GPT-5	Accuracy	Natural Questions	84.2%
GPT-5	F1 Score	TriviaQA	87.3%
GPT-5	Accuracy	HotpotQA	82.1%
GPT-5	Accuracy	WebQuestions	91.4%

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 Support Chatbots**: Providing accurate answers to customer inquiries
- **Educational Q&A Systems**: Helping students with homework and study questions
- **Research Assistants**: Answering complex research queries with citations
- **Virtual Assistants**: Handling diverse user questions in daily interactions

Example Code Snippet:

```
import openai

response = openai.ChatCompletion.create(
    model="gpt-5",
    messages=[{"role": "user", "content": "What is the capital of France
and what is its population?"}]
)
print(response.choices[0].message.content)
```

Limitations

- Occasional factual inaccuracies in rapidly changing information
- Potential for over-confidence in incorrect answers
- High computational costs for complex multi-hop questions
- Dependency on training data recency for current events

Updates and Variants

- Released June 2025
- Variants: GPT-5-QA (optimized for question answering), GPT-5-Search (enhanced with real-time search), GPT-5-Factual (improved fact-checking)

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	F1 Score	SQuAD 2.0	88.9%
Claude-4	Accuracy	Natural Questions	83.7%
Claude-4	F1 Score	TriviaQA	86.8%
Claude-4	Accuracy	HotpotQA	81.6%
Claude-4	Accuracy	WebQuestions	90.8%

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

- **Ethical Q&A Systems**: Providing balanced and unbiased answers
- **Medical Information**: Answering health-related questions safely
- **Legal Research**: Assisting with legal question answering
- **Educational Assessment**: Creating and grading questions

Limitations

- More conservative answers may lack depth in some domains

- Higher latency for complex reasoning questions
- Potential over-cautiousness leading to incomplete answers
- Limited performance on highly specialized technical questions

Updates and Variants

- Released May 2025
- Variants: Claude-4-QA (question answering focus), Claude-4-Research (research assistance), Claude-4-Educational (teaching 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	F1 Score	SQuAD 2.0	87.4%
Gemini-2	Accuracy	Natural Questions	82.3%
Gemini-2	F1 Score	TriviaQA	85.6%
Gemini-2	Accuracy	HotpotQA	80.2%
Gemini-2	Accuracy	WebQuestions	89.7%

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 Integration**: Combining search results with question answering
- **Knowledge Panels**: Creating informative answer summaries
- **Multilingual Q&A**: Supporting questions in multiple languages

- **Real-time Information**: Answering current events and news questions

Limitations

- Integration with Google services may affect neutrality
- Occasional reliance on outdated information
- Complex questions may be oversimplified
- Less emphasis on academic rigor compared to some models

Updates and Variants

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

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	F1 Score	SQuAD 2.0	86.1%
Llama-4	Accuracy	Natural Questions	81.4%
Llama-4	F1 Score	TriviaQA	84.7%
Llama-4	Accuracy	HotpotQA	79.3%
Llama-4	Accuracy	WebQuestions	88.9%

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 Q&A**: Answering questions in social media contexts
- **Community Support**: Building community-driven Q&A systems
- **Content Moderation**: Answering questions about community guidelines
- **Personalized Responses**: Tailoring answers based on user context

Limitations

- Open-source nature may lead to inconsistent answers across implementations
- Higher resource requirements for deployment
- Potential biases from social media training data
- Less commercial features compared to proprietary models

Updates and Variants

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

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	F1 Score	SQuAD 2.0	84.8%
DeepSeek-R2	Accuracy	Natural Questions	79.9%
DeepSeek-R2	F1 Score	TriviaQA	83.2%
DeepSeek-R2	Accuracy	HotpotQA	77.8%
DeepSeek-R2	Accuracy	WebQuestions	87.3%

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 Q&A**: Affordable question answering for businesses
- **Educational Tools**: Budget-friendly tutoring and Q&A systems
- **Research Support**: Efficient information retrieval for researchers
- **Multilingual Support**: Answering questions in multiple languages

Limitations

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

Updates and Variants

- Released January 2025
- Variants: DeepSeek-R2-671B (largest), DeepSeek-R2-16B (efficient), DeepSeek-R2-QA (question answering focus)

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	F1 Score	SQuAD 2.0	83.6%
Mistral-3	Accuracy	Natural Questions	78.7%
Mistral-3	F1 Score	TriviaQA	82.1%
Mistral-3	Accuracy	HotpotQA	76.4%
Mistral-3	Accuracy	WebQuestions	86.1%

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 Q&A Systems**: GDPR-compliant question answering
- **Multilingual Support**: Answering questions in European languages
- **Privacy-focused AI**: Secure question answering for sensitive topics
- **Academic Research**: Supporting scholarly Q&A and literature review

Limitations

- Smaller parameter count compared to leading models
- Limited performance on highly complex multi-hop questions
- Potential language biases in multilingual Q&A
- Open-source challenges with commercial scaling

Updates and Variants

- Released February 2025
- Variants: Mistral-3-Large (123B), Mistral-3-Medium (balanced), Mistral-3-QA (question answering specialized)

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	F1 Score	SQuAD 2.0	82.7%
Command-R3	Accuracy	Natural Questions	77.8%
Command-R3	F1 Score	TriviaQA	81.3%
Command-R3	Accuracy	HotpotQA	75.6%
Command-R3	Accuracy	WebQuestions	85.2%

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 Q&A**: Business intelligence and knowledge management
- **Customer Service**: Answering customer queries with corporate knowledge
- **Documentation Search**: Finding answers in technical documentation
- **HR Support**: Answering employee questions and policy inquiries

Limitations

- Smaller market presence limits ecosystem support
- Limited multimodal question answering
- 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-QA (question answering 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
ERNIE-5	F1 Score	SQuAD 2.0	81.4%
ERNIE-5	Accuracy	Natural Questions	76.6%
ERNIE-5	F1 Score	TriviaQA	80.2%

Model Name	Key Metrics	Dataset/Task	Performance Value
ERNIE-5	Accuracy	HotpotQA	74.3%
ERNIE-5	Accuracy	WebQuestions	84.1%

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 Q&A Systems**: Answering questions in Chinese contexts
- **E-commerce Support**: Product information and recommendation Q&A
- **Government Services**: Public information and service inquiries
- **Educational Q&A**: Supporting Chinese language learning

Limitations

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

Updates and Variants

- Released November 2024
- Variants: ERNIE-5-Turbo (faster), ERNIE-5-QA (question answering focus), ERNIE-5-Speed (optimized)

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

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2	F1 Score	SQuAD 2.0	80.5%
Jamba-2	Accuracy	Natural Questions	75.7%
Jamba-2	F1 Score	TriviaQA	79.4%
Jamba-2	Accuracy	HotpotQA	73.1%
Jamba-2	Accuracy	WebQuestions	83.2%

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 Q&A**: Innovative approaches to question answering
- **Educational Technology**: Interactive learning and assessment
- **Startup Support**: Q&A systems for small businesses
- **Research Tools**: Academic and scientific question answering

Limitations

- Smaller model size limits complex question handling
- Limited global infrastructure compared to tech giants
- Potential regional biases in knowledge coverage
- Less established enterprise support

Updates and Variants

- Released October 2024
- Variants: Jamba-2-Large (52B), Jamba-2-Mini (efficient), Jamba-2-QA (question answering specialized)

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	F1 Score	SQuAD 2.0	79.3%
Skywork-2	Accuracy	Natural Questions	74.4%
Skywork-2	F1 Score	TriviaQA	78.1%
Skywork-2	Accuracy	HotpotQA	71.8%
Skywork-2	Accuracy	WebQuestions	81.9%

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 Market Q&A**: Question answering adapted for Asian contexts
- **Multilingual Support**: Answering questions in Asian languages
- **Cost-effective Solutions**: Affordable Q&A systems for businesses
- **Educational Tools**: Learning support systems in developing regions

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-QA (question answering focus), Skywork-2-Max (largest)

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]