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

Introduction

Question answering benchmarks evaluate language models' ability to comprehend questions, retrieve relevant information, and provide accurate, concise answers across diverse domains and question types. These evaluations are essential for assessing AI systems' factual knowledge, reading comprehension, and information synthesis capabilities. The March 2025 evaluations reveal substantial improvements in multi-hop reasoning, factual accuracy, and the ability to handle complex, multi-part questions, though challenges persist with temporal reasoning and very specialized knowledge domains.

This category includes benchmarks like SQuAD 2.0, Natural Questions, TriviaQA, HotpotQA, and various reading comprehension datasets. Performance in these areas is critical for applications requiring reliable information retrieval and synthesis.

Top 10 LLMs

1. GPT-4o (OpenAI)

Model Name

GPT-4o

Hosting Providers

- OpenAI API
- Microsoft Azure AI
- Amazon Web Services (AWS) AI
- Hugging Face Inference Providers
- Vercel AI Gateway
- NVIDIA NIM
- Together AI

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	95.2%
Natural Questions	F1 Score	92.8%
TriviaQA	Accuracy	87.4%
HotpotQA	F1 Score	89.1%
NewsQA	F1 Score	83.7%
Multi-hop QA	Accuracy	76.3%

LLMs Companies Head Office

OpenAI is headquartered in San Francisco, California, USA. Key personnel include Sam Altman (CEO) and Mira Murati (CTO).

Research Papers and Documentation

- GPT-4o Technical Report
- Question Answering in GPT-4
- OpenAI API Documentation

Use Cases and Examples

- Advanced search and information retrieval
- Customer support automation
- Educational question answering
- Research literature analysis

Limitations

- Occasional factual inconsistencies
- Context window limitations for long documents
- Potential bias in answer generation
- Dependency on training data recency

Updates and Variants

Latest update: March 2025 - Enhanced factual accuracy. Variants include GPT-4o-mini and GPT-4o-turbo.

2. Claude-3.5-Sonnet (Anthropic)

Model Name

Claude-3.5-Sonnet

Hosting Providers

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

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	94.7%
Natural Questions	F1 Score	92.1%
TriviaQA	Accuracy	86.8%
HotpotQA	F1 Score	88.6%
NewsQA	F1 Score	83.2%
Multi-hop QA	Accuracy	75.7%

LLMs Companies Head Office

Anthropic is headquartered in San Francisco, California, USA. Key personnel include Dario Amodei (CEO) and Daniela Amodei (President).

Research Papers and Documentation

- Claude 3.5 Model Card
- Constitutional AI and QA
- Claude API Documentation

Use Cases and Examples

- Safe and reliable information retrieval
- Ethical question answering
- Content moderation assistance
- Educational tutoring systems

Limitations

- Conservative responses in controversial topics
- Occasional verbosity in answers
- Limited real-time information access
- Slower processing for complex queries

Updates and Variants

Latest update: February 2025 - Improved answer accuracy. Variants include Claude-3.5-Haiku and Claude-3.5-Opus.

3. Gemini-1.5-Pro (Google)

Model Name

Gemini-1.5-Pro

Hosting Providers

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

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	94.1%
Natural Questions	F1 Score	91.6%
TriviaQA	Accuracy	86.2%
HotpotQA	F1 Score	87.9%
NewsQA	F1 Score	82.8%
Multi-hop QA	Accuracy	75.1%

LLMs Companies Head Office

Google DeepMind is headquartered in London, UK. Parent company Google/Alphabet headquartered in Mountain View, California, USA.

Research Papers and Documentation

- Gemini 1.5 Technical Report
- Question Answering in Multimodal Systems
- Gemini API Documentation

Use Cases and Examples

- Multimodal question answering
- Real-time information synthesis
- Cross-language QA systems
- Scientific literature search

Limitations

- Search integration limitations
- Occasional multimodal confusion
- Privacy concerns with queries
- Variable performance across languages

Updates and Variants

Latest update: January 2025 - Enhanced search integration. Variants include Gemini-1.5-Flash and Gemini-1.5-Ultra.

4. Llama-3.3-70B (Meta)

Model Name

Llama-3.3-70B

Hosting Providers

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

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	92.8%
Natural Questions	F1 Score	89.7%
TriviaQA	Accuracy	84.6%
HotpotQA	F1 Score	86.1%
NewsQA	F1 Score	81.2%
Multi-hop QA	Accuracy	73.4%

LLMs Companies Head Office

Meta Platforms, Inc. is headquartered in Menlo Park, California, USA. AI division led by Yann LeCun.

Research Papers and Documentation

- Llama 3.3 Technical Report
- Open Question Answering Research
- Hugging Face Model Page

Use Cases and Examples

- Open-source QA development
- Social media content analysis
- Academic research assistance
- Community knowledge sharing

Limitations

- Requires significant fine-tuning
- Community-driven quality variations
- Limited real-time knowledge updates
- Potential misinformation amplification

Updates and Variants

Latest update: December 2024 - Enhanced knowledge retrieval. Variants include Llama-3.3-8B and Llama-3.3-405B.

5. Mistral-Large-2.1 (Mistral AI)

Model Name

Mistral-Large-2.1

Hosting Providers

- Mistral AI
- Hugging Face Inference Providers
- Together AI
- Fireworks

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	92.3%
Natural Questions	F1 Score	89.2%
TriviaQA	Accuracy	84.1%

Dataset/Task	Key Metrics	Performance Value
HotpotQA	F1 Score	85.7%
NewsQA	F1 Score	80.8%
Multi-hop QA	Accuracy	72.9%

LLMs Companies Head Office

Mistral AI is headquartered in Paris, France. Founded by former DeepMind researchers.

Research Papers and Documentation

- Mistral Large 2.1 Release Notes
- European AI QA Research
- Hugging Face Model Page

Use Cases and Examples

- European regulatory compliance QA
- Multilingual question answering
- Cultural knowledge preservation
- Privacy-focused information retrieval

Limitations

- European training data limitations
- Smaller knowledge base compared to global models
- Limited real-time information access
- Regulatory constraints on certain topics

Updates and Variants

Latest update: November 2024 - Enhanced multilingual QA. Variants include Mistral-Medium and Mistral-Small.

6. Qwen2.5-72B (Alibaba)

Model Name

Qwen2.5-72B

Hosting Providers

- Alibaba Cloud Model Studio
- Hugging Face Inference Providers
- Together AI
- NVIDIA NIM

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	91.7%
Natural Questions	F1 Score	88.6%
TriviaQA	Accuracy	83.4%
HotpotQA	F1 Score	85.1%
NewsQA	F1 Score	80.3%
Multi-hop QA	Accuracy	72.3%

LLMs Companies Head Office

Alibaba Group is headquartered in Hangzhou, China. AI division led by Wang Xiaoyun.

Research Papers and Documentation

- Qwen2.5 Technical Report
- Chinese AI Question Answering
- Hugging Face Model Page

Use Cases and Examples

- Chinese language QA systems
- Cross-cultural information retrieval
- E-commerce customer support
- Traditional knowledge QA

Limitations

- Chinese language optimization
- Limited global knowledge depth
- International regulatory constraints
- Variable performance outside Chinese contexts

Updates and Variants

Latest update: October 2024 - Enhanced cross-lingual QA. Variants include Qwen2.5-7B and Qwen2.5-32B.

7. DeepSeek-V3.1 (DeepSeek)

Model Name

DeepSeek-V3.1

Hosting Providers

- DeepSeek Platform
- Hugging Face Inference Providers
- Together AI
- SiliconCloud

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	91.2%
Natural Questions	F1 Score	88.1%
TriviaQA	Accuracy	82.9%
HotpotQA	F1 Score	84.6%
NewsQA	F1 Score	79.8%
Multi-hop QA	Accuracy	71.8%

LLMs Companies Head Office

DeepSeek is headquartered in Hangzhou, China. Founded by former Alibaba researchers.

Research Papers and Documentation

- DeepSeek-V3.1 Technical Report
- Efficient Question Answering
- Hugging Face Model Page

Use Cases and Examples

- Cost-effective QA systems
- Chinese information retrieval
- Efficient knowledge base querying
- Resource-constrained applications

Limitations

- New architecture with limited validation
- Primarily Chinese-focused training
- Smaller community for feedback
- Potential optimization issues

Updates and Variants

Latest update: September 2024 - Improved QA efficiency. Variants include DeepSeek-V2 and DeepSeek-Chat.

8. Grok-2 (xAI)

Model Name

Grok-2

Hosting Providers

- xAI Platform
- Hugging Face Inference Providers
- Together AI

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	90.8%
Natural Questions	F1 Score	87.6%
TriviaQA	Accuracy	82.4%
HotpotQA	F1 Score	84.1%
NewsQA	F1 Score	79.3%
Multi-hop QA	Accuracy	71.3%

LLMs Companies Head Office

xAI is headquartered in Burlingame, California, USA. Founded by Elon Musk.

Research Papers and Documentation

- Grok-2 Release Notes
- Truth-seeking QA
- Hugging Face Model Page

Use Cases and Examples

- Honest information retrieval
- Bias-free question answering
- Educational fact-checking
- Real-time knowledge validation

Limitations

- New model with limited validation
- Smaller knowledge base
- Experimental approach
- Limited third-party testing

Updates and Variants

Latest update: August 2024 - Enhanced factual accuracy. Variants include Grok-1 and Grok-2-Mini.

9. Yi-1.5-34B (01.AI)

Model Name

Yi-1.5-34B

Hosting Providers

- 01.AI Platform
- Hugging Face Inference Providers
- Together AI

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	90.3%
Natural Questions	F1 Score	87.1%
TriviaQA	Accuracy	81.9%
HotpotQA	F1 Score	83.7%
NewsQA	F1 Score	78.9%
Multi-hop QA	Accuracy	70.9%

LLMs Companies Head Office

01.AI is headquartered in Beijing, China. Founded by Kai-Fu Lee.

Research Papers and Documentation

- Yi-1.5 Technical Report
- Chinese AI QA Research
- Hugging Face Model Page

Use Cases and Examples

- Chinese QA systems
- Educational question answering
- Cultural knowledge retrieval
- Cross-lingual information access

Limitations

- Chinese language focus
- Limited international presence
- Smaller ecosystem
- Cultural context dependencies

Updates and Variants

Latest update: July 2024 - Enhanced QA capabilities. Variants include Yi-6B and Yi-9B.

10. Jamba-1.7-Large (AI21 Labs)

Model Name

Jamba-1.7-Large

Hosting Providers

- AI21 Labs
- Hugging Face Inference Providers
- Together AI
- Amazon Web Services (AWS) AI

Benchmarks Evaluation

Dataset/Task	Key Metrics	Performance Value
SQuAD 2.0	F1 Score	89.9%
Natural Questions	F1 Score	86.7%
TriviaQA	Accuracy	81.4%
HotpotQA	F1 Score	83.2%
NewsQA	F1 Score	78.4%
Multi-hop QA	Accuracy	70.4%

LLMs Companies Head Office

AI21 Labs is headquartered in Tel Aviv, Israel. Led by Ori Goshen and Yoav Shoham.

Research Papers and Documentation

- Jamba Model Paper
- Hybrid QA Architectures
- Hugging Face Model Page

Use Cases and Examples

- Long-context document QA

- Legal document analysis
- Complex contract understanding
- Research literature synthesis

Limitations

- Complex architecture deployment challenges
- Higher computational costs
- Limited community adoption
- New model performance variability

Updates and Variants

Latest update: June 2024 - Improved QA efficiency. Variants include Jamba-Mini and Jamba-Large.

Bibliography/Citations

1. Rajpurkar, P., et al. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv preprint arXiv:1806.03822.
2. Kwiatkowski, T., et al. (2019). Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics.
3. Joshi, M., et al. (2017). TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv preprint arXiv:1705.03551.
4. Yang, Z., et al. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv preprint arXiv:1809.09600.
5. OpenAI. (2025). GPT-4o Question Answering Evaluation. Retrieved from https://openai.com/research/gpt-4o
6. Anthropic. (2025). Claude 3.5 QA Assessment. Retrieved from https://www.anthropic.com/research