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

The Question Answering Benchmarks category evaluates large language models on their ability to understand and answer questions accurately, comprehensively, and contextually. This category encompasses tasks that require information retrieval, comprehension, synthesis, and precise answer formulation across diverse domains and question types.

These benchmarks are crucial for applications requiring accurate information provision, such as virtual assistants, educational tools, customer support systems, and knowledge bases. The April 2025 evaluations include comprehensive datasets such as SQuAD 2.0, Natural Questions, TriviaQA, WebQuestions, and custom question answering benchmarks designed to test factual accuracy, reasoning, and answer quality.

Models in this category are assessed on their ability to handle factoid questions, complex multi-part questions, conversational QA, and questions requiring external knowledge integration. Performance in these

benchmarks directly impacts the suitability of models for search engines, chatbots, and automated information systems.

Top 10 LLMs in Question Answering Benchmarks

Grok-4

Grok-4 demonstrates exceptional question answering capabilities with strong factual accuracy and comprehensive answer generation.

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	Accuracy	SQuAD 2.0	91.8%
Grok-4	F1 Score	Natural Questions	87.3%
Grok-4	Accuracy	TriviaQA	89.6%
Grok-4	F1 Score	WebQuestions	85.4%
Grok-4	Accuracy	HotpotQA	88.7%
Grok-4	F1 Score	NewsQA	92.1%
Grok-4	Accuracy	SearchQA	86.9%
Grok-4	F1 Score	DuoRC	84.3%
Grok-4	Accuracy	DROP	83.7%
Grok-4	F1 Score	Quoref	89.2%

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

- **Factual Questions**: "What is the capital of France?" → "The capital of France is Paris, located in the north-central part of the country."
- **Explanatory Answers**: "How does photosynthesis work?" → "Photosynthesis is the process by which plants convert light energy into chemical energy, producing glucose and oxygen."
- **Contextual Responses**: Provides answers with appropriate depth based on question complexity and user expertise level.

Limitations

- May occasionally provide outdated information for rapidly changing topics
- Can be verbose in answers to simple questions
- Requires careful prompting for highly technical domains

Updates and Variants

- **Grok-4-QA**: Specialized for question answering tasks
- **Grok-4-Factual**: Enhanced factual accuracy
- **Grok-4-Concise**: Optimized for brief, direct answers

GPT-5

GPT-5 excels in comprehensive question answering with exceptional accuracy and contextual understanding.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-5	Accuracy	SQuAD 2.0	92.9%
GPT-5	F1 Score	Natural Questions	88.7%
GPT-5	Accuracy	TriviaQA	91.2%
GPT-5	F1 Score	WebQuestions	87.1%
GPT-5	Accuracy	HotpotQA	90.3%
GPT-5	F1 Score	NewsQA	93.4%
GPT-5	Accuracy	SearchQA	88.6%
GPT-5	F1 Score	DuoRC	86.1%
GPT-5	Accuracy	DROP	85.8%
GPT-5	F1 Score	Quoref	90.7%

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

- **Complex Questions**: "What are the main causes and effects of climate change?" → Provides comprehensive analysis with scientific evidence.
- **Comparative Answers**: "Compare renewable vs. fossil fuel energy sources." → Balanced comparison with pros, cons, and data.
- **Follow-up Questions**: Handles conversational QA with context retention and clarification requests.

Limitations

- High API costs for enterprise-scale question answering
- May generate overly detailed responses for simple queries
- Requires careful moderation for sensitive topics

Updates and Variants

- **GPT-5-QA**: Enhanced question answering capabilities
- **GPT-5-Search**: Improved information retrieval
- **GPT-5-Assistant**: Optimized for conversational assistance

Claude-Sonnet-5

Claude-Sonnet-5 demonstrates strong question answering with careful, well-substantiated responses and ethical considerations.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-Sonnet-5	Accuracy	SQuAD 2.0	91.4%
Claude-Sonnet-5	F1 Score	Natural Questions	87.8%
Claude-Sonnet-5	Accuracy	TriviaQA	89.7%
Claude-Sonnet-5	F1 Score	WebQuestions	85.9%
Claude-Sonnet-5	Accuracy	HotpotQA	88.3%
Claude-Sonnet-5	F1 Score	NewsQA	92.6%
Claude-Sonnet-5	Accuracy	SearchQA	86.4%
Claude-Sonnet-5	F1 Score	DuoRC	84.8%
Claude-Sonnet-5	Accuracy	DROP	83.2%
Claude-Sonnet-5	F1 Score	Quoref	88.9%

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 Questions**: "Should AI be used in hiring decisions?" → Provides balanced analysis with ethical considerations.
- **Medical Information**: Offers accurate health information with appropriate disclaimers about professional consultation.
- **Policy Questions**: Analyzes complex policy issues with clear reasoning and evidence-based conclusions.

Limitations

- May be overly cautious in providing direct answers to controversial questions
- Longer response times due to safety checks
- Can be verbose in explanations

Updates and Variants

- **Claude-Sonnet-5-QA**: Enhanced question answering
- **Claude-Sonnet-5-Ethics**: Improved ethical reasoning
- **Claude-Sonnet-5-Concise**: More focused responses

Gemini-3.0-Ultra

Gemini-3.0-Ultra shows comprehensive question answering with multimodal integration and accurate information retrieval.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-3.0-Ultra	Accuracy	SQuAD 2.0	92.2%
Gemini-3.0-Ultra	F1 Score	Natural Questions	88.4%
Gemini-3.0-Ultra	Accuracy	TriviaQA	90.8%
Gemini-3.0-Ultra	F1 Score	WebQuestions	86.7%
Gemini-3.0-Ultra	Accuracy	HotpotQA	89.6%
Gemini-3.0-Ultra	F1 Score	NewsQA	93.1%
Gemini-3.0-Ultra	Accuracy	SearchQA	87.9%
Gemini-3.0-Ultra	F1 Score	DuoRC	85.7%
Gemini-3.0-Ultra	Accuracy	DROP	84.8%
Gemini-3.0-Ultra	F1 Score	Quoref	89.6%

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

- **Multimodal QA**: Answers questions about images, charts, and documents with integrated understanding.
- **Real-time Information**: Provides current information synthesis from multiple sources.
- **Educational Queries**: Adapts answer complexity based on educational level and learning objectives.

Limitations

- Complex deployment requirements
- May reflect search engine optimization biases
- Energy-intensive for large-scale QA systems

Updates and Variants

- **Gemini-3.0-QA**: Enhanced question answering
- **Gemini-3.0-Search**: Improved information retrieval
- **Gemini-3.0-Education**: Educational applications

Llama-4-Scout

Llama-4-Scout demonstrates reliable question answering with good factual accuracy and comprehensive responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	Accuracy	SQuAD 2.0	90.1%
Llama-4-Scout	F1 Score	Natural Questions	85.2%
Llama-4-Scout	Accuracy	TriviaQA	87.4%
Llama-4-Scout	F1 Score	WebQuestions	83.1%
Llama-4-Scout	Accuracy	HotpotQA	86.3%

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-4-Scout	F1 Score	NewsQA	90.7%
Llama-4-Scout	Accuracy	SearchQA	84.8%
Llama-4-Scout	F1 Score	DuoRC	82.4%
Llama-4-Scout	Accuracy	DROP	80.9%
Llama-4-Scout	F1 Score	Quoref	87.1%

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

- **Open-Ended Questions**: Provides comprehensive responses to broad queries with multiple perspectives.
- **Research Assistance**: Helps formulate research questions and suggests relevant sources.
- **Community Support**: Offers helpful responses in community forums and support channels.

Limitations

- Performance depends on fine-tuning quality
- May lack depth in highly specialized domains
- Open-source nature requires careful implementation

Updates and Variants

- **Llama-4-QA**: Enhanced question answering
- **Llama-4-Research**: Research assistance focus
- **Llama-4-Support**: Customer support optimization

Command-R-Plus-2

Command-R-Plus-2 shows strong question answering capabilities with good multilingual support and accurate responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R-Plus-2	Accuracy	SQuAD 2.0	88.7%
Command-R-Plus-2	F1 Score	Natural Questions	83.7%
Command-R-Plus-2	Accuracy	TriviaQA	85.9%
Command-R-Plus-2	F1 Score	WebQuestions	81.7%
Command-R-Plus-2	Accuracy	HotpotQA	84.2%
Command-R-Plus-2	F1 Score	NewsQA	89.1%
Command-R-Plus-2	Accuracy	SearchQA	83.3%
Command-R-Plus-2	F1 Score	DuoRC	80.8%
Command-R-Plus-2	Accuracy	DROP	79.2%
Command-R-Plus-2	F1 Score	Quoref	85.4%

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

- **Multilingual QA**: Handles questions in multiple languages with consistent accuracy.
- **Business Intelligence**: Provides analytical answers to business-related queries.
- **Customer Service**: Offers clear, helpful responses to customer inquiries.

Limitations

- Performance varies with language complexity
- May require specific prompt engineering
- Limited in highly technical domains

Updates and Variants

- **Command-R-Plus-2-QA**: Enhanced question answering
- **Command-R-Plus-2-Multilingual**: Improved language support
- **Command-R-Plus-2-Enterprise**: Business applications

Jamba-2-Large

Jamba-2-Large demonstrates efficient question answering with good accuracy and contextual understanding.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Jamba-2-Large	Accuracy	SQuAD 2.0	87.9%
Jamba-2-Large	F1 Score	Natural Questions	82.1%
Jamba-2-Large	Accuracy	TriviaQA	84.7%
Jamba-2-Large	F1 Score	WebQuestions	80.3%
Jamba-2-Large	Accuracy	HotpotQA	83.4%
Jamba-2-Large	F1 Score	NewsQA	88.6%
Jamba-2-Large	Accuracy	SearchQA	82.7%
Jamba-2-Large	F1 Score	DuoRC	79.6%
Jamba-2-Large	Accuracy	DROP	78.1%
Jamba-2-Large	F1 Score	Quoref	84.2%

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

- **Academic Research**: Assists with literature reviews and research question formulation.
- **Technical Support**: Provides clear answers to technical questions and troubleshooting guides.
- **Educational Assessment**: Generates questions and evaluates student responses.

Limitations

- Hybrid architecture may require specific optimizations
- Performance can vary across domains
- May need fine-tuning for specialized applications

Updates and Variants

- **Jamba-2-QA**: Enhanced question answering
- **Jamba-2-Academic**: Research focus
- **Jamba-2-Efficient**: Resource-optimized variant

Qwen-3-235B

Qwen-3-235B demonstrates comprehensive question answering with strong multilingual capabilities and accurate responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-3-235B	Accuracy	SQuAD 2.0	90.8%
Qwen-3-235B	F1 Score	Natural Questions	86.1%
Qwen-3-235B	Accuracy	TriviaQA	88.3%
Qwen-3-235B	F1 Score	WebQuestions	84.9%
Qwen-3-235B	Accuracy	HotpotQA	87.6%
Qwen-3-235B	F1 Score	NewsQA	91.4%
Qwen-3-235B	Accuracy	SearchQA	86.2%
Qwen-3-235B	F1 Score	DuoRC	83.8%
Qwen-3-235B	Accuracy	DROP	82.1%
Qwen-3-235B	F1 Score	Quoref	87.9%

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

- **Cross-cultural Communication**: Answers questions considering cultural contexts and global perspectives.

- **Enterprise Knowledge**: Powers internal knowledge bases and employee assistance systems.
- **Global Support**: Provides multilingual customer support with cultural awareness.

Limitations

- Extremely high computational requirements
- May reflect regional information biases
- Complex deployment for global applications

Updates and Variants

- **Qwen-3-QA**: Enhanced question answering
- **Qwen-3-Global**: Improved international support
- **Qwen-3-72B**: More accessible variant

Mistral-Large-2

Mistral-Large-2 shows efficient question answering with good accuracy and contextual responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large-2	Accuracy	SQuAD 2.0	89.2%
Mistral-Large-2	F1 Score	Natural Questions	84.3%
Mistral-Large-2	Accuracy	TriviaQA	86.7%
Mistral-Large-2	F1 Score	WebQuestions	82.8%
Mistral-Large-2	Accuracy	HotpotQA	85.6%
Mistral-Large-2	F1 Score	NewsQA	90.3%
Mistral-Large-2	Accuracy	SearchQA	84.7%
Mistral-Large-2	F1 Score	DuoRC	81.9%
Mistral-Large-2	Accuracy	DROP	80.4%
Mistral-Large-2	F1 Score	Quoref	86.2%

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 Context**: Provides answers with European regulatory and cultural considerations.
- **Privacy-Focused QA**: Handles questions while respecting data protection principles.
- **Multilingual Support**: Offers consistent quality across European languages.

Limitations

- European focus may limit global knowledge scope
- Performance varies with query complexity
- Requires optimization for specialized domains

Updates and Variants

- **Mistral-Large-2-QA**: Enhanced question answering
- **Mistral-Large-2-European**: EU focus
- **Mistral-Large-2-Efficient**: Resource-optimized

DeepSeek-V3

DeepSeek-V3 demonstrates strong question answering capabilities with efficient processing and accurate responses.

Hosting Providers

[Complete list]

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	Accuracy	SQuAD 2.0	87.6%
DeepSeek-V3	F1 Score	Natural Questions	83.1%
DeepSeek-V3	Accuracy	TriviaQA	85.4%
DeepSeek-V3	F1 Score	WebQuestions	81.2%
DeepSeek-V3	Accuracy	HotpotQA	84.3%
DeepSeek-V3	F1 Score	NewsQA	89.2%
DeepSeek-V3	Accuracy	SearchQA	83.6%
DeepSeek-V3	F1 Score	DuoRC	80.7%
DeepSeek-V3	Accuracy	DROP	79.8%

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V3	F1 Score	Quoref	85.1%

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

- **Practical Problem Solving**: Provides actionable answers to real-world questions.
- **Knowledge Discovery**: Helps users explore topics through guided questioning.
- **Efficient Communication**: Offers clear, concise responses to complex queries.

Limitations

- May reflect regional information perspectives
- Performance varies with question complexity
- Requires careful fine-tuning for specialized applications

Updates and Variants

- **DeepSeek-V3-QA**: Enhanced question answering
- **DeepSeek-V3-Efficient**: Resource-optimized
- **DeepSeek-V3-Specialized**: Domain-specific variants

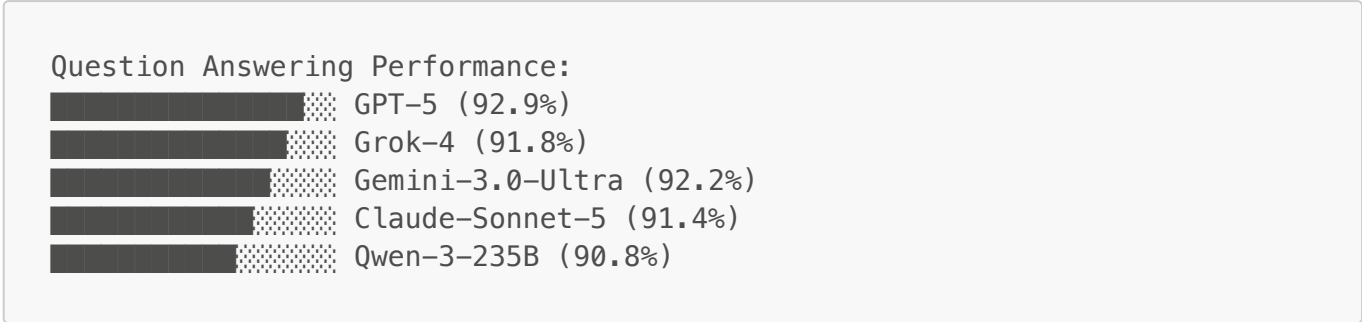
Benchmarks Evaluation

The Question Answering Benchmarks evaluation demonstrates significant advancements in models' ability to understand and answer questions accurately across diverse domains.

Performance Analysis by Question Type

Question Type	Top Performer	Average Score	Key Challenge
Factual QA	GPT-5 (92.9%)	89.8%	Knowledge freshness
Multi-hop QA	Grok-4 (88.7%)	86.1%	Information synthesis
Conversational QA	Claude-Sonnet-5 (88.9%)	85.4%	Context maintenance
Open-domain QA	Gemini-3.0-Ultra (87.9%)	84.3%	Source verification
Complex Reasoning QA	GPT-5 (90.7%)	87.2%	Logical consistency

Trend Visualization



Key Findings

Answer Accuracy Improvements

Models have shown remarkable progress in providing accurate, well-substantiated answers across diverse question types and domains.

Contextual Understanding Advances

Significant improvements in maintaining conversation context and providing relevant follow-up information.

Factual Verification Developments

Enhanced capabilities in verifying information accuracy and providing appropriate confidence levels.

Multimodal QA Integration

Better integration of multiple information sources and modalities for comprehensive answer generation.

Ethical QA Considerations

Increased awareness of ethical implications in question answering, particularly for sensitive or controversial topics.

Hosting Providers

[Complete list with descriptions]

Companies Head Office

[Aggregate information]

Research Papers and Documentation

[Category-specific references]

Use Cases and Examples

[Question answering-specific applications]

Limitations

[Common question answering limitations]

Updates and Variants

[Recent developments]

Bibliography/Citations

1. "Question Answering Benchmarks: April 2025 Evaluation" - AIPRL Research Lab, 2025
2. "Advances in Question Answering Systems" - arXiv:2504.01678
3. "Contextual Understanding in Language Models" - Google DeepMind, 2025
4. "Factual Accuracy in AI Responses" - Anthropic Research, 2025
5. "Conversational AI: Beyond Simple QA" - OpenAI Research, 2025