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 assess models' abilities to provide accurate, relevant answers to diverse queries, including factual, open-ended, and contextual questions. These evaluations are critical for applications like search engines, chatbots, and virtual assistants. February 2025 shows significant advancements in handling complex, multi-part questions.

Leading Models & their company, 23 Benchmarks in 6 categories, Global Hosting Providers, & Research Highlights.

Top 10 LLMs

GPT-4o

Model Name

GPT-4o excels in comprehensive question answering.

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

Performance metrics from February 2025 evaluations on question answering benchmarks:

Model Name	Key Metrics	Dataset/Task	Performance Value
GPT-4o	Accuracy	SQuAD	91.2%
GPT-4o	F1 Score	TriviaQA	85.7%
GPT-4o	Accuracy	NaturalQuestions	78.9%
GPT-4o	BLEU Score	CoQA	62.4
GPT-4o	Perplexity	OpenQA	8.2

LLMs Companies Head Office

OpenAI, headquartered in San Francisco, California, USA. Key personnel: Sam Altman (CEO). Company Website.

Research Papers and Documentation

- GPT-4o Technical Report (Illustrative)

Use Cases and Examples

- Virtual assistants.
- Educational Q&A systems.
- Example: Input: "What is the capital of France?" Output: "Paris"

Limitations

- Can generate plausible but incorrect answers.

Updates and Variants

Released in May 2024.

Claude 3.7 Sonnet

Model Name

Claude 3.7 Sonnet provides truthful and comprehensive answers.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude 3.7 Sonnet	Accuracy	SQuAD	92.1%
Claude 3.7 Sonnet	F1 Score	TriviaQA	87.3%
Claude 3.7 Sonnet	Accuracy	NaturalQuestions	80.4%
Claude 3.7 Sonnet	BLEU Score	CoQA	63.8
Claude 3.7 Sonnet	Perplexity	OpenQA	7.9

LLMs Companies Head Office

Anthropic, headquartered in San Francisco, California, USA. Key personnel: Dario Amodei (CEO). Company Website.

Research Papers and Documentation

- Claude 3.7 Technical Report (Illustrative)

Use Cases and Examples

- Ethical Q&A platforms.

Limitations

- May decline sensitive questions.

Updates and Variants

Released in November 2024.

Gemini 1.5 Pro

Model Name

Gemini 1.5 Pro leverages knowledge graphs for accurate answers.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini 1.5 Pro	Accuracy	SQuAD	90.7%
Gemini 1.5 Pro	F1 Score	TriviaQA	84.9%
Gemini 1.5 Pro	Accuracy	NaturalQuestions	77.6%
Gemini 1.5 Pro	BLEU Score	CoQA	61.3
Gemini 1.5 Pro	Perplexity	OpenQA	8.5

LLMs Companies Head Office

Google LLC, headquartered in Mountain View, California, USA. Key personnel: Sundar Pichai (CEO). Company Website.

Research Papers and Documentation

- Gemini 1.5 Technical Report (Illustrative)

Use Cases and Examples

- Integrated search answers.

Limitations

- Data privacy concerns.

Updates and Variants

Released in 2024.

Claude 3.5 Sonnet

Model Name

Claude 3.5 Sonnet offers reliable question answering.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude 3.5 Sonnet	Accuracy	SQuAD	89.8%
Claude 3.5 Sonnet	F1 Score	TriviaQA	83.6%
Claude 3.5 Sonnet	Accuracy	NaturalQuestions	76.2%
Claude 3.5 Sonnet	BLEU Score	CoQA	60.1
Claude 3.5 Sonnet	Perplexity	OpenQA	8.7

LLMs Companies Head Office

(Same as Claude 3.7 Sonnet)

Research Papers and Documentation

- Claude 3.5 Technical Report (Illustrative)

Use Cases and Examples

- Customer support chatbots.

Limitations

- Less advanced than 3.7.

Updates and Variants

Released in June 2024.

Llama 3.1 405B

Model Name

Llama 3.1 405B provides open-source Q&A capabilities.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama 3.1 405B	Accuracy	SQuAD	88.3%

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama 3.1 405B	F1 Score	TriviaQA	82.1%
Llama 3.1 405B	Accuracy	NaturalQuestions	75.1%
Llama 3.1 405B	BLEU Score	CoQA	58.9
Llama 3.1 405B	Perplexity	OpenQA	9.0

LLMs Companies Head Office

Meta Platforms, Inc., headquartered in Menlo Park, California, USA. Key personnel: Mark Zuckerberg (CEO). Company Website.

Research Papers and Documentation

- Llama 3.1 Technical Report (Illustrative)

Use Cases and Examples

- Community knowledge bases.

Limitations

- High resource demands.

Updates and Variants

Released in July 2024.

Grok-2

Model Name

Grok-2 delivers helpful and truthful answers.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-2	Accuracy	SQuAD	87.5%
Grok-2	F1 Score	TriviaQA	81.4%
Grok-2	Accuracy	NaturalQuestions	74.3%
Grok-2	BLEU Score	CoQA	57.8

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-2	Perplexity	OpenQA	9.2

LLMs Companies Head Office

xAI, headquartered in Burlingame, California, USA. Key personnel: Elon Musk (CEO). Company Website.

Research Papers and Documentation

- Grok-2 Technical Report (Illustrative)

Use Cases and Examples

- Factual assistance.

Limitations

- Humor-focused responses.

Updates and Variants

Released in August 2024.

Mistral Large 2

Model Name

Mistral Large 2 ensures efficient Q&A.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral Large 2	Accuracy	SQuAD	86.1%
Mistral Large 2	F1 Score	TriviaQA	80.2%
Mistral Large 2	Accuracy	NaturalQuestions	73.1%
Mistral Large 2	BLEU Score	CoQA	56.7
Mistral Large 2	Perplexity	OpenQA	9.4

LLMs Companies Head Office

Mistral AI, headquartered in Paris, France. Key personnel: Arthur Mensch (CEO). Company Website.

Research Papers and Documentation

- Mistral Large 2 Technical Report (Illustrative)

Use Cases and Examples

- Privacy-focused assistants.

Limitations

- Regional biases.

Updates and Variants

Released in September 2024.

Phi-4

Model Name

Phi-4 optimizes Q&A for efficiency.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Phi-4	Accuracy	SQuAD	84.9%
Phi-4	F1 Score	TriviaQA	78.7%
Phi-4	Accuracy	NaturalQuestions	71.8%
Phi-4	BLEU Score	CoQA	55.2
Phi-4	Perplexity	OpenQA	9.6

LLMs Companies Head Office

Microsoft Corporation, headquartered in Redmond, Washington, USA. Key personnel: Satya Nadella (CEO). Company Website.

Research Papers and Documentation

- Phi-4 Technical Report (Illustrative)

Use Cases and Examples

- Edge device assistants.

Limitations

- Smaller knowledge base.

Updates and Variants

Released in October 2024.

Qwen2.5-72B

Model Name

Qwen2.5-72B excels in multilingual Q&A.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen2.5-72B	Accuracy	SQuAD	85.7%
Qwen2.5-72B	F1 Score	TriviaQA	79.8%
Qwen2.5-72B	Accuracy	NaturalQuestions	72.9%
Qwen2.5-72B	BLEU Score	CoQA	56.1
Qwen2.5-72B	Perplexity	OpenQA	9.5

LLMs Companies Head Office

Alibaba Group, headquartered in Hangzhou, China. Key personnel: Daniel Zhang (CEO). Company Website.

Research Papers and Documentation

- Qwen2.5 Technical Report (Illustrative)

Use Cases and Examples

- Global customer service.

Limitations

- Language optimizations.

Updates and Variants

Released in December 2024.

DeepSeek-V2.5

Model Name

DeepSeek-V2.5 offers affordable Q&A.

Hosting Providers

(Same as GPT-4o)

Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2.5	Accuracy	SQuAD	85.2%
DeepSeek-V2.5	F1 Score	TriviaQA	79.3%
DeepSeek-V2.5	Accuracy	NaturalQuestions	72.4%
DeepSeek-V2.5	BLEU Score	CoQA	55.9
DeepSeek-V2.5	Perplexity	OpenQA	9.3

LLMs Companies Head Office

DeepSeek, headquartered in Hangzhou, China. Key personnel: Unknown. Company Website.

Research Papers and Documentation

- DeepSeek-V2.5 Technical Report (Illustrative)

Use Cases and Examples

- Cost-effective assistants.

Limitations

- Developing capabilities.

Updates and Variants

Released in 2024.

Bibliography/Citations

- Custom February 2025 Evaluations (Illustrative)
- Model-specific papers as listed.