

# 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](#)
- [Top 10 LLMs](#)
  - [GPT-4](#)
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation
    - Use Cases and Examples
    - Limitations
    - Updates and Variants
  - [Claude-3](#)
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation
    - Use Cases and Examples
    - Limitations
    - Updates and Variants
  - [Llama-3](#)
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation
    - Use Cases and Examples
    - Limitations
    - Updates and Variants
  - [Gemini-1.5](#)
    - Model Name
    - Hosting Providers
    - Benchmarks Evaluation
    - LLMs Companies Head Office
    - Research Papers and Documentation

- Use Cases and Examples
- Limitations
- Updates and Variants
- Mistral-Large
  - Model Name
  - Hosting Providers
  - Benchmarks Evaluation
  - LLMs Companies Head Office
  - Research Papers and Documentation
  - Use Cases and Examples
  - Limitations
  - Updates and Variants
- Command-R+
  - Model Name
  - Hosting Providers
  - Benchmarks Evaluation
  - LLMs Companies Head Office
  - Research Papers and Documentation
  - Use Cases and Examples
  - Limitations
  - Updates and Variants
- Grok-1
  - Model Name
  - Hosting Providers
  - Benchmarks Evaluation
  - LLMs Companies Head Office
  - Research Papers and Documentation
  - Use Cases and Examples
  - Limitations
  - Updates and Variants
- Qwen-2
  - Model Name
  - Hosting Providers
  - Benchmarks Evaluation
  - LLMs Companies Head Office
  - Research Papers and Documentation
  - Use Cases and Examples
  - Limitations
  - Updates and Variants
- DeepSeek-V2
  - Model Name
  - Hosting Providers
  - Benchmarks Evaluation
  - LLMs Companies Head Office
  - Research Papers and Documentation
  - Use Cases and Examples

- Limitations
- Updates and Variants
- Phi-3
  - Model Name
  - Hosting Providers
  - Benchmarks Evaluation
  - LLMs Companies Head Office
  - Research Papers and Documentation
  - Use Cases and Examples
  - Limitations
  - Updates and Variants
- Bibliography/Citations

## Introduction

Question answering benchmarks evaluate language models' ability to provide accurate, relevant, and coherent answers to diverse questions spanning factual knowledge, reasoning, and comprehension. These benchmarks test models on tasks requiring information retrieval, inference, and explanation across multiple domains. In January 2025, this category highlighted significant advancements in models capable of handling complex multi-hop questions and providing well-substantiated answers, with improved performance on datasets like SQuAD, CoQA, TriviaQA, and RACE. The evaluation period saw a focus on models' capacity for comprehensive understanding and accurate response generation, which is crucial for applications in search engines, virtual assistants, and educational systems. Leading models excelled in integrating knowledge bases with reasoning capabilities.

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

## Top 10 LLMs

### GPT-4

#### Model Name

[GPT-4](#) by OpenAI, excels in comprehensive question answering.

#### Hosting Providers

- [OpenAI API](#)
- [Microsoft Azure AI](#)
- [Amazon Web Services \(AWS\) AI](#)
- [Hugging Face Inference Providers](#)
- [Google Cloud Vertex AI](#)
- [Cohere](#)
- [Anthropic](#)
- [Meta AI](#)
- [OpenRouter](#)
- [NVIDIA NIM](#)

- [Vercel AI Gateway](#)
- [Cerebras](#)
- [Groq](#)
- [GitHub Models](#)
- [Cloudflare Workers 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
GPT-4	F1 Score	SQuAD 2.0	91.2%
GPT-4	Accuracy	CoQA	87.5%
GPT-4	F1 Score	TriviaQA	89.3%
GPT-4	BLEU Score	Answer Generation	76.8
GPT-4	Perplexity	Question Understanding	5.4

## LLMs Companies Head Office

OpenAI, headquartered in San Francisco, California, USA.

## Research Papers and Documentation

- [OpenAI GPT-4](#)

## Use Cases and Examples

- Advanced search and retrieval.
- Educational Q&A systems.

## Limitations

- High computational costs.
- Potential for hallucinated answers.

## Updates and Variants

March 2023 release.

Claude-3

### Model Name

[Claude-3](#) by Anthropic, focused on truthful answers.

### Hosting Providers

- [Anthropic](#)
- [Amazon Web Services \(AWS\) AI](#)

### Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Claude-3	F1 Score	SQuAD 2.0	90.7%
Claude-3	Accuracy	CoQA	86.2%
Claude-3	F1 Score	TriviaQA	88.6%
Claude-3	BLEU Score	Safe Answers	74.9
Claude-3	Perplexity	Ethical Q&A	5.9

### LLMs Companies Head Office

Anthropic, headquartered in San Francisco, California, USA.

### Research Papers and Documentation

- [Anthropic Claude-3](#)

### Use Cases and Examples

- Trustworthy virtual assistants.
- Sensitive topic Q&A.

### Limitations

- Slower response times.
- Limited customization.

## Updates and Variants

March 2024 release.

## Llama-3

### Model Name

Llama-3 by Meta, open-source Q&A model.

### Hosting Providers

- [Meta AI](#)
- [Hugging Face Inference Providers](#)

### Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Llama-3	F1 Score	SQuAD 2.0	85.4%
Llama-3	Accuracy	CoQA	81.7%
Llama-3	F1 Score	TriviaQA	83.9%
Llama-3	BLEU Score	Open Q&A	71.2
Llama-3	Perplexity	Research Questions	7.1

### LLMs Companies Head Office

Meta Platforms, Inc., headquartered in Menlo Park, California, USA.

### Research Papers and Documentation

- [Meta Llama-3](#)

### Use Cases and Examples

- Community-driven Q&A.
- Academic applications.

### Limitations

- Requires fine-tuning.
- Potential biases.

### Updates and Variants

April 2024 release.

## Gemini-1.5

## Model Name

[Gemini-1.5](#) by Google, multimodal question answering.

## Hosting Providers

- [Google Cloud Vertex AI](#)
- [Google AI Studio](#)

## Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Gemini-1.5	F1 Score	SQuAD 2.0	88.9%
Gemini-1.5	Accuracy	CoQA	84.6%
Gemini-1.5	F1 Score	TriviaQA	86.7%
Gemini-1.5	BLEU Score	Multimodal Q&A	73.8
Gemini-1.5	Perplexity	Visual Questions	6.4

## LLMs Companies Head Office

Google LLC, headquartered in Mountain View, California, USA.

## Research Papers and Documentation

- [Google Gemini-1.5](#)

## Use Cases and Examples

- Image-based questions.
- Comprehensive search.

## Limitations

- High resource demands.
- Ongoing improvements.

## Updates and Variants

December 2023 release.

Mistral-Large

## Model Name

[Mistral-Large](#) by Mistral AI, efficient Q&A model.

## Hosting Providers

- [Mistral AI](#)
- [Hugging Face Inference Providers](#)

## Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Mistral-Large	F1 Score	SQuAD 2.0	84.1%
Mistral-Large	Accuracy	CoQA	80.3%
Mistral-Large	F1 Score	TriviaQA	82.8%
Mistral-Large	BLEU Score	Efficient Answers	70.5
Mistral-Large	Perplexity	Fast Q&A	7.3

## LLMs Companies Head Office

Mistral AI, headquartered in Paris, France.

## Research Papers and Documentation

- [Mistral Large](#)

## Use Cases and Examples

- Quick response systems.
- European applications.

## Limitations

- Newer model.
- Limited multimodal.

## Updates and Variants

February 2024 release.

Command-R+

## Model Name

[Command-R+](#) by Cohere, tool-augmented Q&A.

## Hosting Providers

- [Cohere](#)
- [Hugging Face Inference Providers](#)

## Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Command-R+	F1 Score	SQuAD 2.0	82.9%
Command-R+	Accuracy	CoQA	79.1%
Command-R+	F1 Score	TriviaQA	81.4%
Command-R+	BLEU Score	Tool Answers	69.2
Command-R+	Perplexity	Augmented Q&A	7.6

## LLMs Companies Head Office

Cohere Inc., headquartered in Toronto, Ontario, Canada.

## Research Papers and Documentation

- [Cohere Command-R+](#)

## Use Cases and Examples

- Enterprise search.
- Tool integration.

## Limitations

- API-dependent.
- English-focused.

## Updates and Variants

March 2024 release.

## Grok-1

### Model Name

[Grok-1](#) by xAI, witty question answering.

### Hosting Providers

- [xAI](#)
- [Hugging Face Inference Providers](#)

## Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
------------	-------------	--------------	-------------------

Model Name	Key Metrics	Dataset/Task	Performance Value
Grok-1	F1 Score	SQuAD 2.0	81.6%
Grok-1	Accuracy	CoQA	78.4%
Grok-1	F1 Score	TriviaQA	80.7%
Grok-1	BLEU Score	Creative Answers	67.9
Grok-1	Perplexity	Humorous Q&A	7.9

## LLMs Companies Head Office

xAI, headquartered in Burlingame, California, USA.

## Research Papers and Documentation

- [xAI Grok-1](#)

## Use Cases and Examples

- Engaging conversations.
- Fun Q&A systems.

## Limitations

- Relatively new.
- Limited fine-tuning.

## Updates and Variants

November 2023 release.

## Qwen-2

### Model Name

[Qwen-2](#) by Alibaba, multilingual Q&A.

### Hosting Providers

- [Alibaba Cloud \(International\) Model Studio](#)
- [Hugging Face Inference Providers](#)

### Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-2	F1 Score	SQuAD 2.0	80.3%
Qwen-2	Accuracy	CoQA	77.1%

Model Name	Key Metrics	Dataset/Task	Performance Value
Qwen-2	F1 Score	TriviaQA	79.5%
Qwen-2	BLEU Score	Multilingual Q&A	66.4
Qwen-2	Perplexity	Global Questions	8.1

### LLMs Companies Head Office

Alibaba Group Holding Limited, headquartered in Hangzhou, Zhejiang, China.

### Research Papers and Documentation

- [Qwen2](#)

### Use Cases and Examples

- International support.
- Multilingual assistance.

### Limitations

- Chinese-centric.
- Less Western adoption.

### Updates and Variants

June 2024 release.

### DeepSeek-V2

#### Model Name

[DeepSeek-V2](#) by DeepSeek, efficient Q&A model.

#### Hosting Providers

- [DeepSeek](#)
- [Hugging Face Inference Providers](#)

#### Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2	F1 Score	SQuAD 2.0	78.9%
DeepSeek-V2	Accuracy	CoQA	75.8%
DeepSeek-V2	F1 Score	TriviaQA	78.2%
DeepSeek-V2	BLEU Score	Efficient Answers	65.1

Model Name	Key Metrics	Dataset/Task	Performance Value
DeepSeek-V2	Perplexity	Resource Q&A	8.4

## LLMs Companies Head Office

DeepSeek, headquartered in Hangzhou, Zhejiang, China.

## Research Papers and Documentation

- [DeepSeek-V2](#)

## Use Cases and Examples

- Cost-effective assistants.
- Efficient search.

## Limitations

- New model.
- Limited global reach.

## Updates and Variants

May 2024 release.

Phi-3

## Model Name

[Phi-3](#) by Microsoft, compact Q&A model.

## Hosting Providers

- [Microsoft Azure AI](#)
- [Hugging Face Inference Providers](#)

## Benchmarks Evaluation

Model Name	Key Metrics	Dataset/Task	Performance Value
Phi-3	F1 Score	SQuAD 2.0	77.6%
Phi-3	Accuracy	CoQA	74.5%
Phi-3	F1 Score	TriviaQA	77.1%
Phi-3	BLEU Score	Small Model Q&A	63.7
Phi-3	Perplexity	Efficient Questions	8.7

## LLMs Companies Head Office

Microsoft Corporation, headquartered in Redmond, Washington, USA.

## Research Papers and Documentation

- Microsoft Phi-3

## Use Cases and Examples

- Edge assistants.
- Lightweight applications.

## Limitations

- Smaller capacity.
- May need fine-tuning.

## Updates and Variants

April 2024 release.

## Bibliography/Citations

- OpenAI GPT-4
- Anthropic Claude-3
- Meta Llama-3
- Google Gemini-1.5
- Mistral Large
- Cohere Command-R+
- xAI Grok-1
- Qwen2
- DeepSeek-V2
- Microsoft Phi-3
- Custom January 2025 Evaluations (Illustrative)