BENCHMARK REPORT

# Independent Multi-Task Evaluation of Large Language Models

aiXplain

aiXplain

# Benchmark Report

## Independent Multi-Task Evaluation of Large Language Models

Arabic | 9 Models | 11 Tasks

Version 1.3

April 2025

aiXplain

aiXplain, Inc.
3031 Tisch Way, Suite 80
San Jose, CA 95128
United States
p. +1.408.601.0079
e. care@aixplain.com
w. www.aixplain.com

# Table of Content

## Disclaimer

The models presented in this report were assessed during April 2025, and it is important to note that developments or alterations may have occurred in the time elapsed since the evaluation. The performance of these models is contingent upon the extent of similarity between the data used for evaluation and the data employed in their training processes.

*This report is intended only for recipients who accessed it through their aiXplain subscription. To approve further distribution, please contact care@aixplain.com. We're happy to support your use of this report.*

## About aiXplain

aiXplain is the end-to-end agentic AI platform designed to help teams build, optimize, and deploy production-grade AI agents at scale. Whether you're automating workflows, enhancing customer experiences, or embedding AI into enterprise systems, aiXplain equips your team with everything needed—from asset selection to post-deployment monitoring—in one unified environment.

- **Extensive asset library**: Access over 38,000 AI models, tools, and agents—including more than 180 large language models—from 60+ global vendors. With one API key and a flexible pay-as-you-go model, you can test, integrate, and swap assets instantly. You can also onboard and manage your own models without vendor lock-in or infrastructure overhead.
- **Agentic framework**: Design intelligent, modular agents using role-based architectures. Leverage purpose-built micro-agents—such as the Orchestrator, Mentalist, Bodyguard, and Inspector—to handle multi-step planning, coordination, compliance, and output verification. Build everything from autonomous AI agents to complex multi-agent systems and flows, all built for transparency, reusability, and scale.
- **AI services**: Continuously improve your agents using integrated services for benchmarking, fine-tuning, auto-routing, and RAG indexing (text, image, and graph-based). Control usage and cost through rate limiting, and ensure relevance and reliability with real-time feedback loops. These services help keep your agents accurate, efficient, and responsive to evolving needs.
- **Production deployment**: Deploy securely across SaaS, hybrid, or on-prem environments. aiXplain handles infrastructure, scaling, and MLOps while giving you full visibility into agent behavior. Simplified auditing and continuous monitoring allow you to trace decisions, inspect model/tool usage, and enforce internal policies with confidence. Built-in trust mechanisms—including role-based access, guardrails, and output inspection—ensure your solutions meet business and operational standards long after deployment.

aiXplain helps teams move from experimentation to enterprise-grade execution—faster, safer, and with complete control over how agents operate in production.

# Executive Summary

This study presents the performance evaluation of 9 different Large Language Models (LLMs) across 11 diverse tasks: Question Answering, Reading Comprehension, Creative Writing, Information Extraction, Linguistic Processing, Logical Reasoning, Sequence Tagging, Summarization, Text Classification, Program Execution, and Translation. Nine LLMs were evaluated: Command R+, Gemma 2, Qwen 2.5, GPT-4o Mini, Llama 3.2, Llama 3.3, Deepseek V3, Llama 4 Maverick and Llama 4 Scout. All models, except GPT 4o Mini, are open-source and range in size from 3 billion to 671 billion parameters.

## Key Observations

GPT-4o Mini outperformed in most of the tasks such as Reading Comprehension, Linguistic Processing, Logical Reasoning, Sequence Tagging and Translation. Apart from it, the relatively smaller and open-source model, Command R+ with 104B parameters, delivered a robust performance in the remaining tasks like Question Answering, Program Execution and Text Classification.

The smaller models, Gemma 2 (9B) and Qwen 2.5 (32B), showed competitive performance in Creative Writing and Summarization tasks, respectively. It was interesting to observe the modest contribution of Llama 3.2(3B) and Llama 3.3(70B) in tasks like Information Extraction and Linguistic Processing. In the larger spectrum models, Deepseek V3 (671B) demonstrated decent output in terms of summaries and logical reasoning.

## Conclusion

Despite the varying size of these models, each has displayed strengths in certain tasks, often quite competitive with larger models. This demonstrates that effective task performance is not solely reliant on model size and encourages further exploration of specific training strategies or architectures for certain tasks.

The compelling performance of the open-source models, especially Command R+, reaffirms that the strategic direction of intelligent modelling is achieving desired results. Moving forward, fine-tuning these models, developing adaptations, and furthering customization based on specific tasks will be crucial.

# LLM Benchmarking Setup

In recent years, large language models (LLMs) have emerged as powerful tools in natural language processing (NLP), demonstrating remarkable capabilities across various tasks such as machine translation, sentiment analysis, question answering, and text generation. However, most benchmarking efforts have primarily focused on English and a few widely spoken languages, leaving gaps in evaluating LLM performance for languages with unique linguistic structures, such as Arabic.

Arabic presents distinct challenges for NLP due to its rich morphology, complex syntax, and diverse dialectal variations. As the adoption of LLMs expands in Arabic-speaking regions, there is an urgent need for rigorous evaluation tailored to the Arabic language. The performance of LLMs can vary significantly depending on the dataset, task, and evaluation metrics used, making it crucial to establish standardized benchmarks that accurately reflect real-world usage in Arabic.

This report focuses on the benchmarking of large language models specifically for Arabic. Our evaluation methodology follows a black-box approach, assessing models based solely on their outputs rather than their internal architectures. This approach enables a fair and objective comparison of different LLMs, independent of their underlying training strategies or architectures.

We evaluate various LLMs across a range of NLP tasks relevant to Arabic, including text classification, text generation, machine translation, and named entity recognition. Our benchmarking leverages Arabic NLP datasets and appropriate evaluation metrics to provide a comprehensive assessment of each model's strengths, weaknesses, and practical applicability.

By providing a dedicated benchmarking framework for Arabic LLMs, this report aims to equip researchers, developers, and industry practitioners with actionable insights to inform their model selection and deployment strategies. Through this effort, we contribute to the advancement of Arabic NLP and foster the development of more effective and inclusive AI models for Arabic-speaking users.

## Tasks

We evaluate Language Model Models (LLMs) based on the following tasks:

### Question Answering

LLMs' answers to questions are evaluated based on correctness and relevance. The model should select the most appropriate answer from the given options. This gives insight into the knowledge encoded in the LLM.

### Reading Comprehension

LLMs' answers should accurately reflect the information presented in the passage. Answers should be concise, relevant, and demonstrate comprehension of the text.

### Creative Writing

LLMs are assessed on their ability to generate original, engaging, and coherent creative text, such as stories, poems, or essays. The evaluation considers fluency, creativity, adherence to the given prompt, and overall readability.

### Information Extraction

LLMs are evaluated on their ability to identify and extract key pieces of information from structured or unstructured text. This includes named entity recognition, relation extraction, and fact retrieval, ensuring accuracy and consistency.

### Linguistic Processing

This task assesses the LLMs' ability to perform syntactic and semantic analysis, including tasks like part-of-speech tagging, parsing, and word sense disambiguation. The model's linguistic understanding and ability to process complex sentence structures are crucial evaluation factors.

## Logical Reasoning

LLMs are evaluated based on their ability to generate answers that demonstrate an understanding of common sense knowledge. Answers should reflect logical reasoning and a grasp of everyday situations.

## Sequence Tagging

LLMs are tested on their ability to assign labels to sequences of text, such as named entities, parts of speech, or syntactic roles. The evaluation focuses on accuracy, consistency, and adherence to linguistic patterns.

## Summarization

LLMs' generated summaries are evaluated based on their ability to capture the main points of the input text accurately while maintaining coherence and readability. Summaries should be concise and cover important information without losing key details.

## Text Classification

LLMs are evaluated on their ability to classify text into predefined categories, such as sentiment analysis, topic classification, or spam detection. The accuracy and robustness of the classifications are key performance metrics.

## Program Execution

LLMs are assessed on their ability to generate and execute code snippets correctly. This includes evaluating outputs against expected results, handling syntax and logical errors, and adhering to best programming practices.

## Translation

LLMs' translations are evaluated based on accuracy, fluency, and relevance. Translations should accurately convey the meaning of the source text in the target language while also being grammatically correct and natural-sounding.

## Evaluation Metrics

We measure the performance of models in different NLP tasks using the following metrics:

### a) ROUGE-L

Used for evaluating Text Summarization, ROUGE-L focuses on the longest common subsequence (LCS) between the generated and reference texts. This metric emphasizes fluency and coherence by capturing both the structure and meaning of the summarized content.

### b) BLEU

Bleu measures the overlap of n-grams (typically up to 4-grams) between the machine-translated text and human-translated references. It's widely used in machine translation tasks to assess the quality of translations.

## Datasets

To evaluate the performance of LLMs for each of the tasks, we use a number of widely-used benchmark test sets. The datasets used are either originally in Arabic or have been translated into Arabic. To ensure the quality and reliability of the test sets, the pipeline includes a filtering stage that eliminates poorly translated samples. A total of 61 test sets covering 11 tasks were used in the evaluation.

## Models

The benchmark covers a selection of LLMs, covering various aspects such as model size (in terms of parameters), accessibility (open vs. closed), and other relevant factors. Our selection encompasses LLMs of different sizes, from smaller to larger models, to evaluate their performance across a spectrum of scales. We also include both open and closed LLMs to ensure a comprehensive evaluation, considering the practicality and availability of models for different users and applications. Additionally, we consider factors like architecture, training data, and pre-training objectives to cover a wide range of LLM characteristics.

This approach allows us to provide a thorough and representative assessment of LLMs, considering their diverse characteristics. By benchmarking models across various sizes and accessibility levels, we offer insights into their performance and suitability for different NLP tasks and scenarios. The following table lists the LLMs under consideration.

| Model | Model Size | Context Length | Accessibility |
|---|---|---|---|
| Command R+ | 104B | 128k | Open |
| Gemma 2 | 9B | 8192 | Open |
| Qwen 2.5 | 32B | 128k | Open |
| GPT-4o Mini | Unknown | 128k | Closed |
| Llama 3.2 | 3B | 128k | Open |
| Llama 3.3 | 70B | 128k | Open |
| Deepseek V3 | 671B | 128k | Open |
| Llama 4 Maverick | 400B | 1M | Open |
| Llama 4 Scout | 109B | 10M | Open |

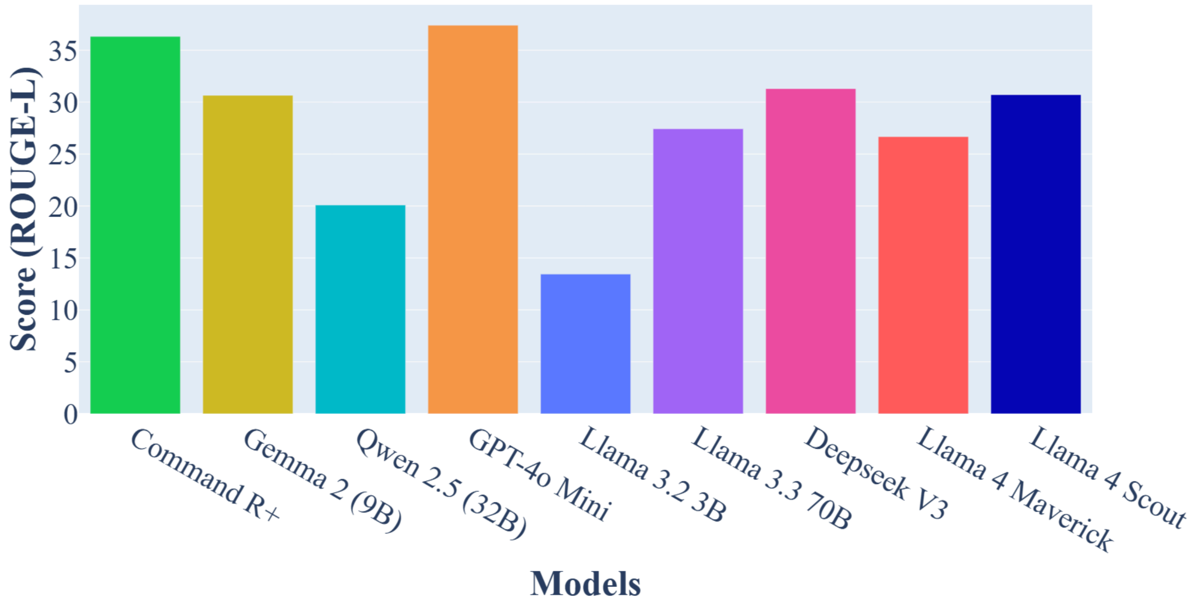
# Benchmark Results

This section presents results of the benchmarking of LLMs across different tasks. Results for the Translation task are reported in Bleu (the higher the better), and the results for other tasks are reported in ROUGE-L metric (the higher the better).

## All Results: An Overview

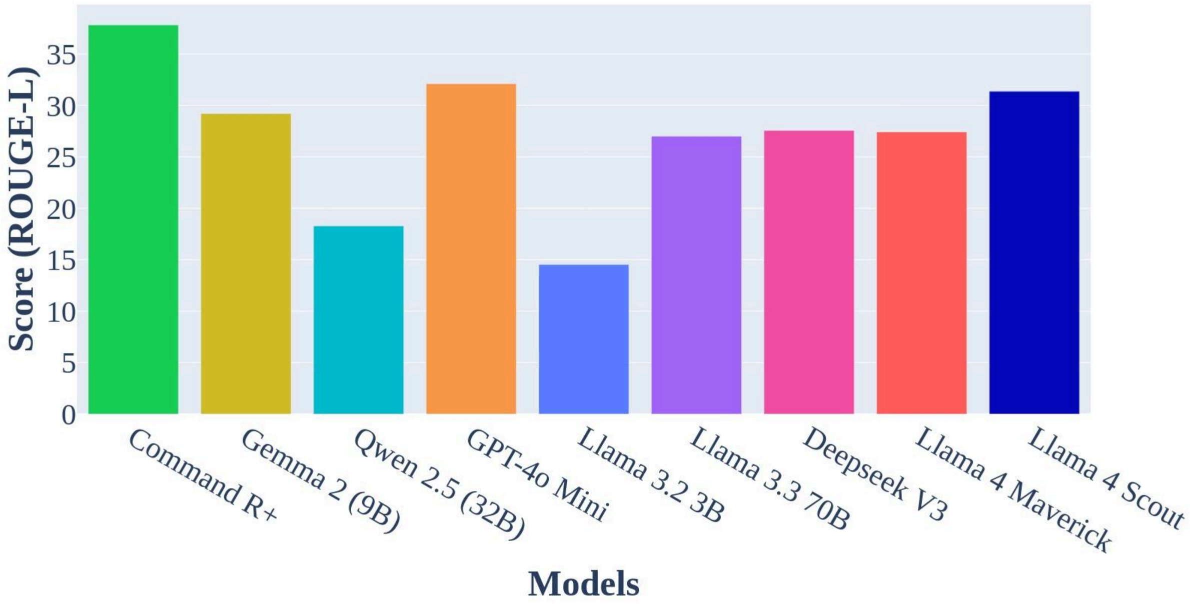
|  | Command R+ | Gemma 2 (9B) | Qwen 2.5 (32B) | GPT-4o Mini | Llama 3.2 3B | Llama 3.3 70B | Deepseek V3 | Llama 4 Maverick | Llama 4 Scout |
|---|---|---|---|---|---|---|---|---|---|
| **Question Answering** | **37.82** | 29.21 | 18.28 | 32.10 | 14.53 | 27.00 | 27.55 | 27.42 | 31.36 |
| **Reading Comprehension** | 41.61 | 39.13 | 20.02 | **51.81** | 24.83 | 36.66 | 35.02 | 37.17 | 41.01 |
| **Creative Writing** | 9.33 | 12.09 | 9.16 | 11.08 | 6.01 | 10.77 | 10.31 | **11.68** | 10.93 |
| **Information Extraction** | 22.05 | 19.43 | 12.29 | 20.10 | 9.77 | 20.08 | 18.05 | **23.24** | 23.74 |
| **Linguistic Processing** | 60.80 | 51.33 | 44.39 | **66.53** | 25.23 | 54.77 | 54.15 | 64.46 | 62.51 |
| **Logical Reasoning** | 32.35 | 18.30 | 11.45 | 40.38 | 7.98 | 22.92 | **33.84** | 16.55 | 22.67 |
| **Sequence Tagging** | 42.56 | 40.83 | 17.02 | **47.01** | 15.09 | 33.91 | 41.97 | 25.71 | 43.20 |
| **Summarization** | 23.69 | 23.44 | 21.34 | 24.64 | 14.32 | 19.80 | 21.77 | 25.39 | **25.82** |
| **Text Classification** | **25.99** | 15.70 | 11.26 | 17.96 | 7.26 | 15.38 | 17.26 | 19.90 | 22.91 |
| **Program Execution** | **68.24** | 55.76 | 32.43 | 56.31 | 13.62 | 38.71 | 54.30 | 16.95 | 29.85 |
| **Translation** | 35.18 | 31.93 | 23.34 | **43.53** | 9.17 | 21.81 | 30.13 | 24.92 | 23.78 |
| **Overall** | 36.33 | 30.65 | 20.09 | **37.40** | 13.44 | 27.44 | 31.30 | 26.67 | 30.71 |

## Overall Results



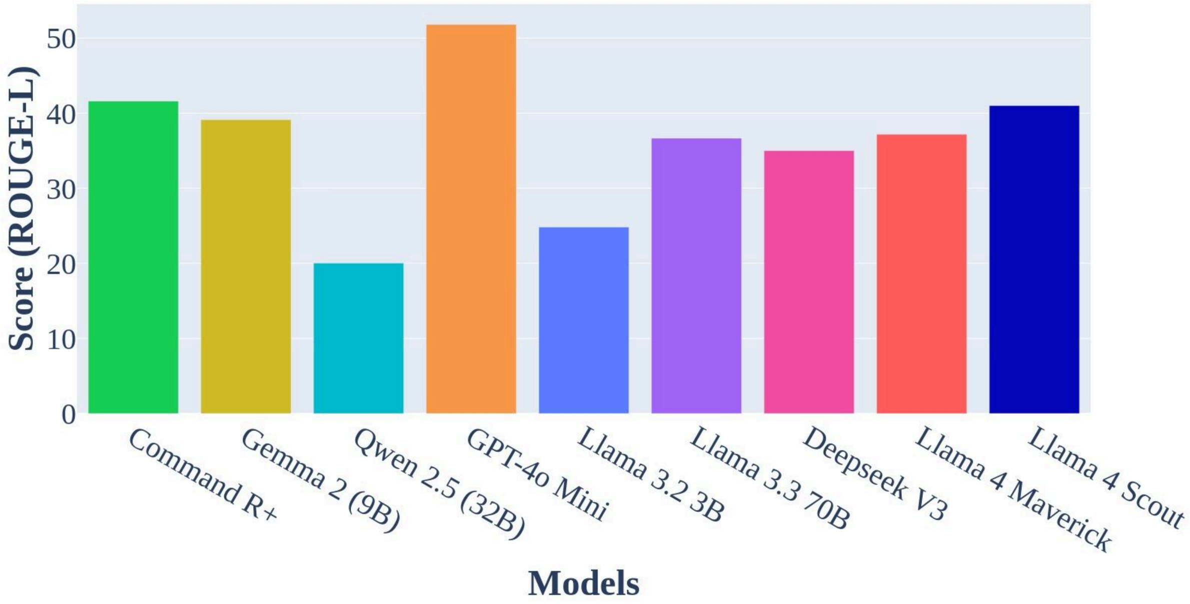
# Task Specific Performance

## Question Answering



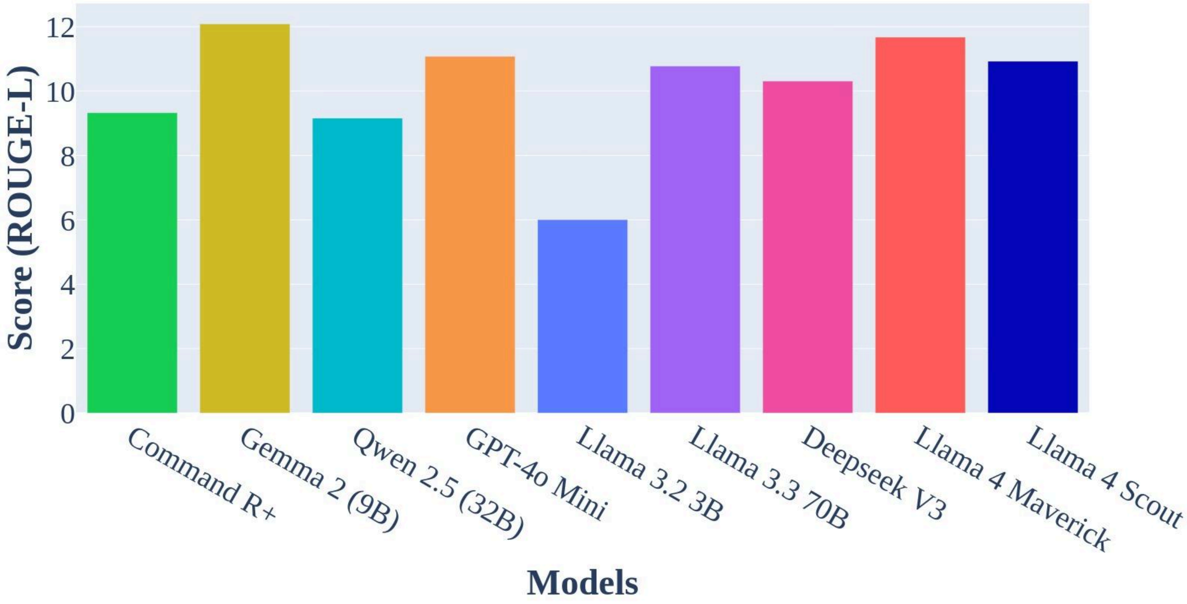
**Command R+** and **GPT-4o Mini** performs the best with ROUGE-L scores of 37.82 and 32.1 respectively. **Llama 3.2 3B i**s the lowest performer with a score of only 14.53.

## Reading Comprehension



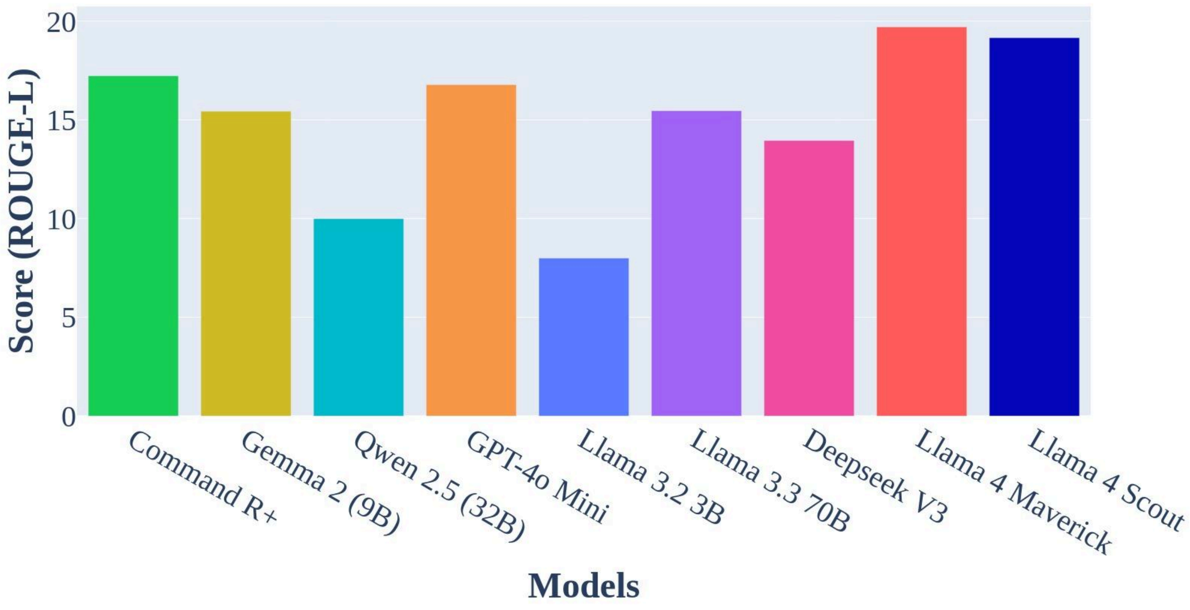
**GPT-4o Mini** shines here with the highest ROUGE-L score of 51.81. **Command R+** and **Llama 4 Scout** show similar performance, both scoring slightly above 40. **Llama 3.2 3B** is the least proficient model.

## Creative Writing



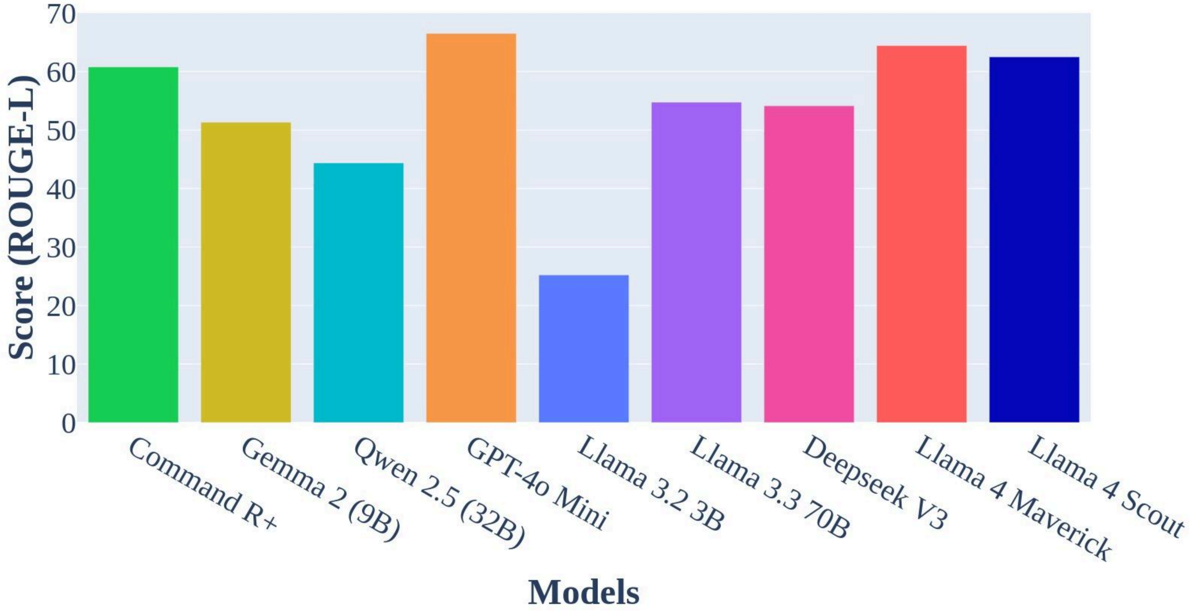
The performance of LLMs is relatively closer in this task. The score for **Gemma 2 (9B)** is the highest at 12.09, but **Llama 3.2 3B** comes last with a score of just 6.01.

## Information Extraction



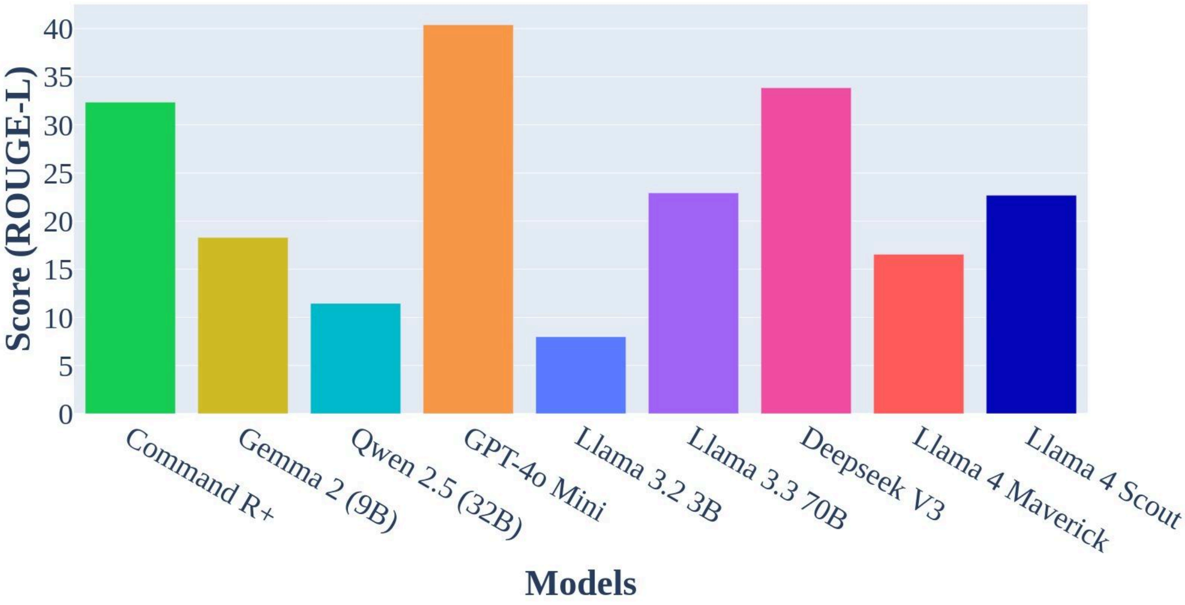
**Llama 4 Maverick** and **Scout** are leading models with nearly similar scores of 23.24 and 23.74 respectively. **Qwen 2.5 (32B**) and **Llama 3.2 3B** show lower performances.

## Linguistic Processing



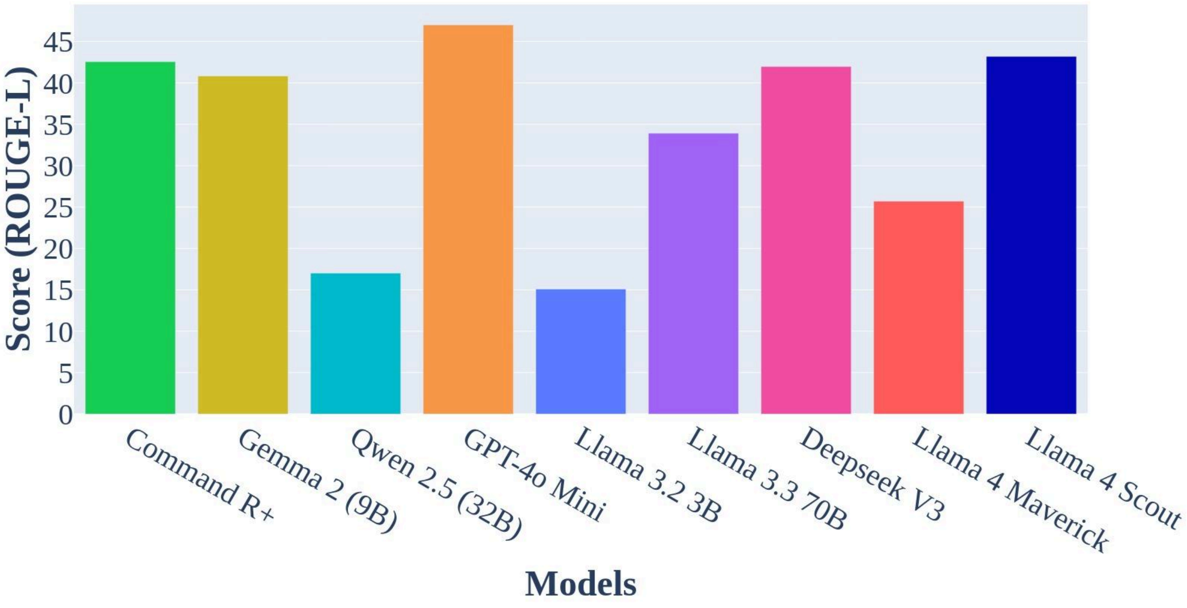
**GPT-4o Mini** holds the top spot with 66.53, followed closely by **Llama 4 Maverick** and **Scout**. **Llama 3.2 3B** gives a weaker performance, scoring only 25.23.

## Logical Reasoning



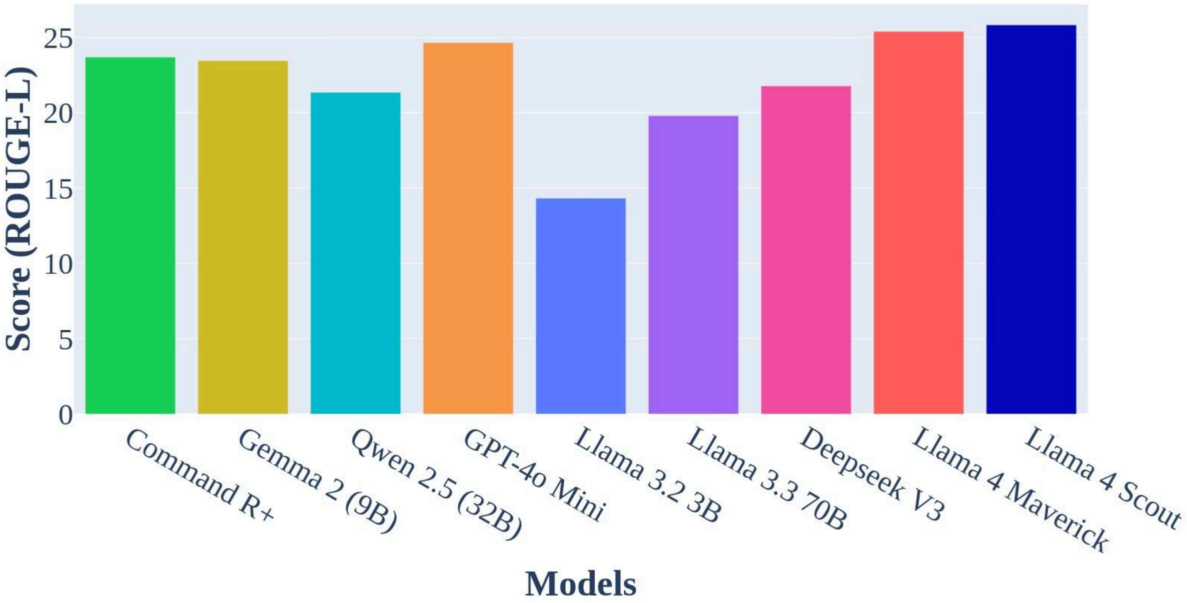
**GPT-4o** outperforms all, scoring 40.38. **Deepseek V3** also shows strong performance with 33.84. **Llama 3.2 3B** lacks performance in this task, scoring just 7.98.

## Sequence Tagging



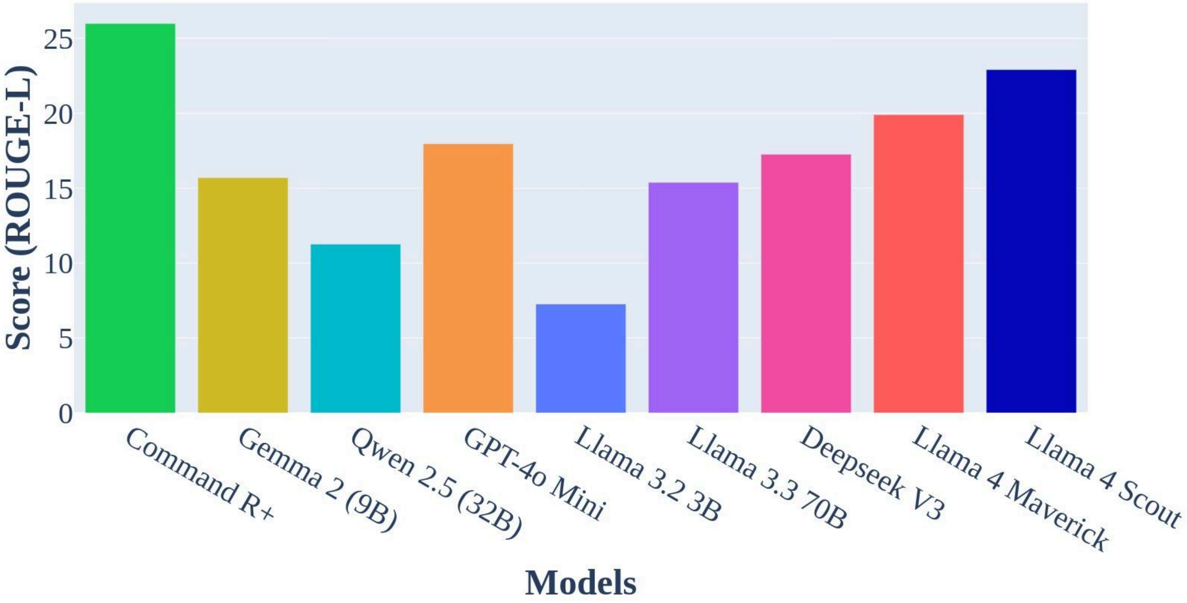
In this task, **GPT-4o Mini** reigns supreme with 47.01. **Command R+** and **Llama 4 Scout** hold similar scores around 42. **Llama 3.2 3B** is still on the lower end, scoring 15.09.

## Summarization



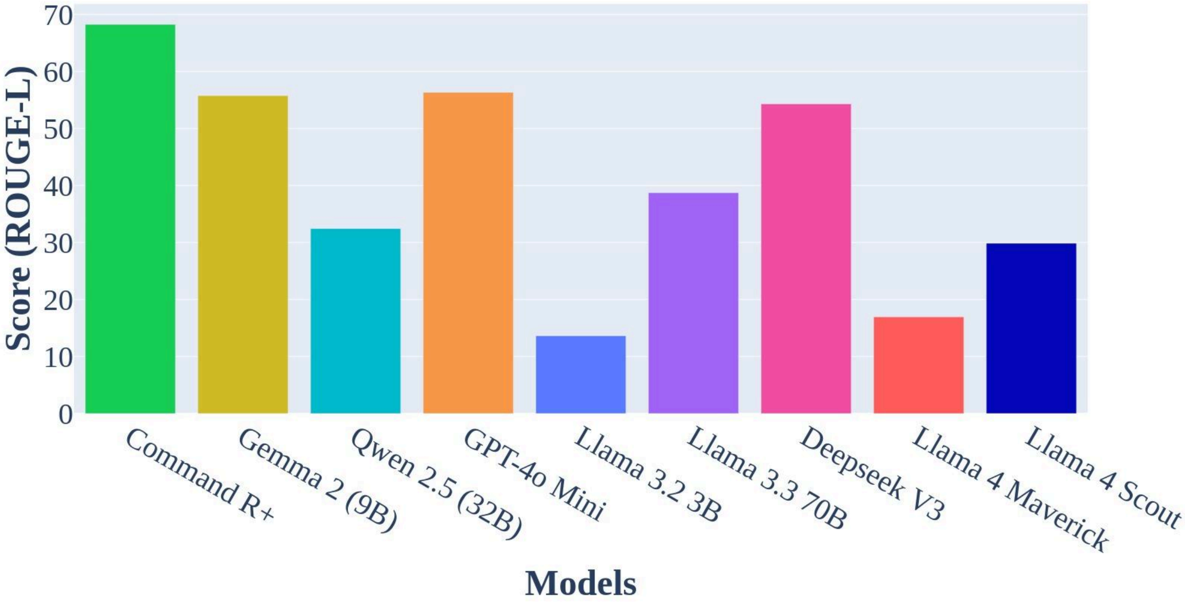
**Llama 4 models, Maverick and Scout**, outperform the others, scoring 25.39 and 25.82 respectively. **Llama 3.2 3B** again lags behind its peers with a score of 14.32.

## Text Classification



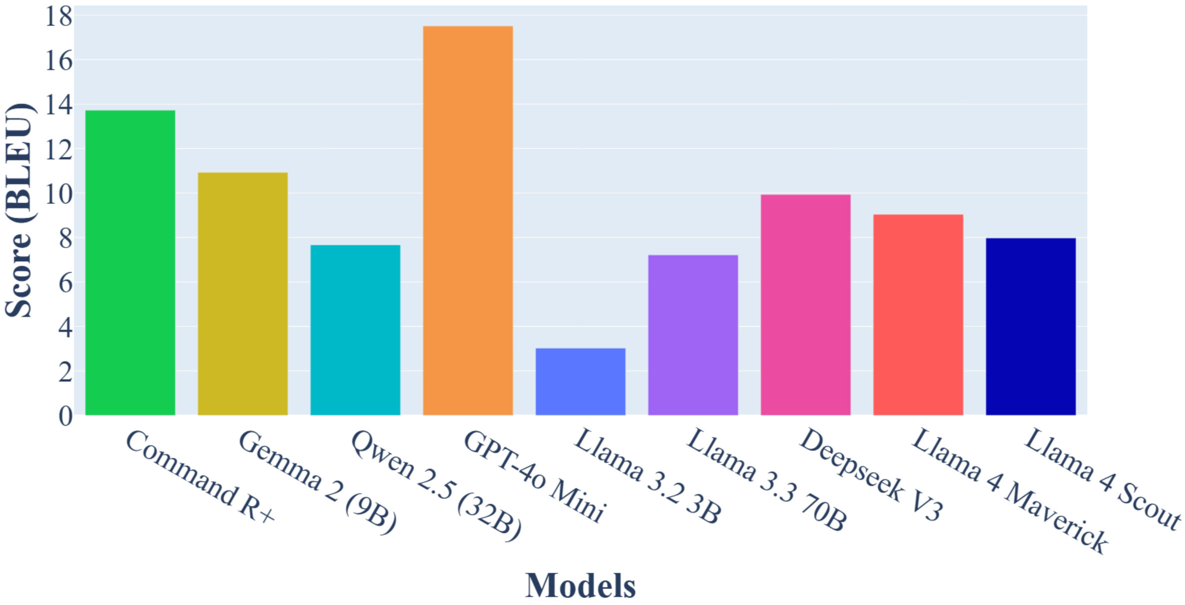
**Command R+** leads with 25.99. **Llama 4 Scout** also shows strong performance with 22.91. **Llama 3.2 3B** struggles to keep up in this task as well, scoring only 7.26.

## Program Execution



**Command R+** dominates with a soaring 68.24, followed by **Gemma 2 (9B)** and **GPT-4o Mini**. **Llama 3.2 3B** gives the least impressive performance, scoring 13.62.

## Translation



**GPT-4o Mini** excels in this task, scoring 17.53. **Command R+** also performs well with 13.73. **Llama 3.2 3B** struggles, with a BLEU of just 3.03.