BENCHMARK REPORT

Independent Multi-Task Evaluation of Large Language Models

aixplain

aiXplain

Benchmark Report

Independent Multi-Task Evaluation of Large Language Models

Arabic | 12 Models | 11 Tasks

Version 3.2

June 2025

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 June 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 12 different Large Language Models (LLMs) across 11 diverse tasks in Arabic: Question Answering, Reading Comprehension, Creative Writing, Information Extraction, Linguistic Processing, Logical Reasoning, Sequence Tagging, Summarization, Text Classification, Program Execution, and Translation. In this round of evaluation, we added 3 LLMs that are specifically developed for Arabic: AceGPT 7B, Jais 13B, SILMA 9B. We also retained a representative sample of LLMs from the previous evaluation: ALLaM 7B, Command R+, Deepseek V3, Fanar C-1, GPT-4o mini, GPT-4.1, FLM 40B, Llama 4 Scout, and Quen3 14B.

In addition, we introduce a new metric "LLM as a Judge", which evaluates the outputs using an LLM. We present results for several tasks: Reading Comprehension, Information Extraction and Logical Reasoning

Key Observations

Across the tasks, GPT-4.1 and SILMA 9B emerge as top performers, with GPT-4.1 scoring particularly high in Program Execution and Linguistic Processing. SILMA 9B excels in Reading Comprehension and demonstrates consistent strength overall. Command R+ shows robust performance in Program Execution and performs competitively in most other tasks, signaling its versatility. Among smaller models, ALLaM 7B performs notably well in Logical Reasoning and Sequence Tagging. Notably, open models are advancing rapidly and have demonstrated strong performance, particularly in reasoning, extraction, and summarization tasks.

The models exhibit substantial variation in size, from compact 7B models to expansive 671B models like Deepseek V3. Interestingly, larger models do not always guarantee superior performance, although they generally perform well in tasks requiring extensive contextual comprehension.

In terms of strengths and weaknesses, closed-source models still hold a slight edge in complex tasks and sophisticated linguistic processing, but open-source models are catching up quickly. All models, regardless of their size or source, find creative writing and nuanced text classification to be especially challenging, with noticeably lower scores across the board in those areas.

“LLM as a Judge” metric (which used Gemini 2.5 Flash as the judge LLM) , while broadly confirming the performance of automatic metrics, also highlights some differences. High ROUGE scores don't always align with high LLM Judge scores, indicating differences between content overlap and qualitative understanding. For instance, SILMA 9B in Reading Comprehension scores high on ROUGE but not on the LLM Judge metric. Similarly, ALLaM 7B in Logical Reasoning is an outlier with a high ROUGE but a lower LLM Judge score. Qwen3 14B exhibits an inverse trend with relatively low ROUGE scores but higher LLM Judge scores across tasks.

Conclusion

The performance evaluation reveals the dynamic landscape of LLM capabilities, where both model size and accessibility influence outcomes across tasks. Models like GPT-4.1 and SILMA 9B lead in overall performance due to their balanced strengths across a spectrum of tasks. Open-access models, despite their competitive edge in certain areas, are outperformed in tasks requiring advanced linguistic and logical processing by models leveraging closed-access optimizations.

As the field of AI continues to evolve, future developments should consider the nuanced balance between model size, accessibility, and specific enhancements to address a diverse range of applications effectively. Ultimately, model selection should be guided by the requirements of each specific task rather than by model size or vendor reputation alone.

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.

Included Tasks: Answer Verification, Answerability Classification, Multiple Choice Q&A, Question Answering, Question Categorization, Question Decomposition, Question Duplication Detection, Question Generation, Question Rewriting, and Question Understanding

Reading Comprehension

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

Included Tasks: Reading Comprehension, and Sentence Perturbation

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.

Included Tasks: News Article Generation, News Article Writing, Poem Generation, Story Composition, Story Generation, and Title Generation

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.

Included Tasks: Diacritization, Grammar Correction, and Transliteration

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.

Included Tasks: Cause Effect Classification, Commonsense Validation, Evidence Evaluation, Explanation, Fact Verification, Inference Detection, Logical Reasoning, Natural Language Inference, Semantic Matching, Semantic Similarity, and Textual Entailment

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.

Included Tasks: Named Entity Recognition, Entity Recognition & Gender ID, Entity Relation Classification, and Relation Extraction

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.

Included Tasks: Text Summarization, Sentence Compression, and Text Simplification

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.

Included Tasks: Text Classification, Coherence Classification, Emotion Analysis, Emotion Detection, Entity Categorization, Intent Classification, Intent Identification, Query Classification, Review Rating Prediction, Section Classification, Sentiment Analysis, Spam Detection, Text Categorization, Topic Classification, and Topic Prediction

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.

c) LLM as a Judge

LLM-as-a-judge leverages advanced LLMs to evaluate other LLMs, offering a scalable and consistent approach to assessing the quality of generative models as an alternative to human assessment. The core principle involves prompting a "judge" LLM with the input query, the generated output from the "candidate" LLM, and clear evaluation criteria. The judge LLM then scores, critiques, or compares outputs, often providing explanations for its judgments.

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
AceGPT 7B	7B	2k	Open
ALLaM 7B	7B	4096	Open
Command R+	104B	128k	Open
Deepseek V3	671B	128k	Open
Fanar C-1	9B	4k	Open
GPT-4o mini	Unknown	128k	Closed
GPT-4.1	Unknown	1M	Closed
Jais 13B	13B	2048	Open
LFM 40B	40B	32k	Open
Llama 4 Scout	109B	10M	Open
Qwen3 14B	14B	32k	Open
SILMA 9B	9B	12k	Open

Benchmark Results

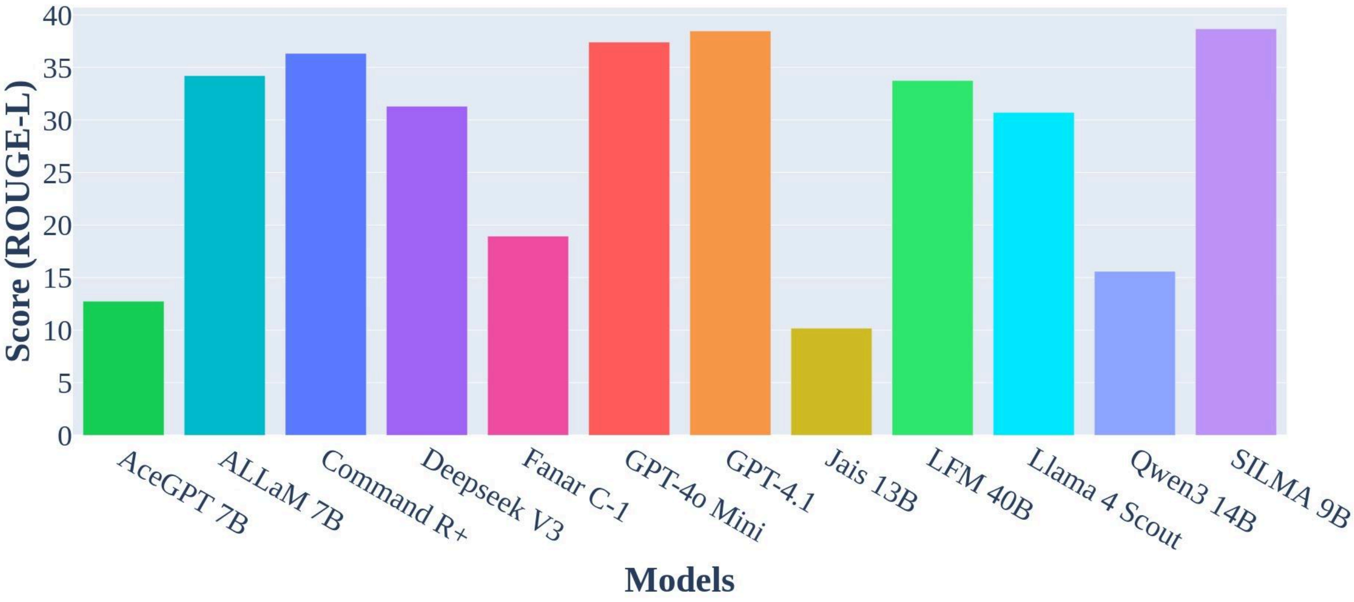
This section presents results of the benchmarking of LLMs across different tasks. Results for the Translation task are reported in Bleu (the higher the better), and the results for other tasks are reported in ROUGE-L metric (the higher the better).

All Results: An Overview

	Overall	Question Answering	Reading Compre-hension	Creative Writing	Info Extraction	Linguistic Processing	Logical Reasoning	Sequence Tagging	Summari-zation	Text Classifi-cation	Program Execution	Translation*
AceGPT 7B	12.75	8.00	9.88	9.50	9.57	35.30	7.75	10.35	11.75	11.76	6.59	11.31
ALLaM 7B	34.21	29.87	48.25	13.18	25.09	60.39	**44.44**	41.45	**27.29**	24.02	43.90	9.60
Command R+	36.33	**37.82**	41.61	9.33	22.05	60.80	32.35	42.56	23.69	25.99	68.24	13.73
Deepseek V3	31.30	27.55	35.02	10.31	18.05	54.15	33.84	41.97	21.77	17.26	54.30	10.02
Fanar C-1	18.94	14.86	12.54	8.82	16.97	39.45	9.46	18.79	13.39	10.90	46.52	4.74
GPT-4o mini	37.40	32.10	51.81	11.08	20.10	**66.53**	40.38	**47.01**	24.64	17.96	56.31	17.53
GPT-4.1	38.46	31.78	48.18	12.68	33.15	62.38	36.56	44.18	22.73	21.60	74.52	16.00
Jais 13B	10.18	9.92	7.81	8.28	6.30	16.16	9.91	7.21	7.52	5.23	25.73	2.41
LFM 40B	33.74	29.36	48.60	11.35	24.09	58.80	40.74	40.44	25.83	18.99	49.02	10.54
Llama 4 Scout	30.71	31.36	41.01	10.93	23.74	62.51	22.67	43.20	25.82	22.91	29.85	7.95
Qwen3 14B	15.59	6.19	7.11	5.94	7.43	16.94	1.98	1.94	13.24	5.30	**81.89**	18.38
SILMA 9B	**38.66**	31.23	**55.49**	**22.92**	**33.21**	64.22	29.48	37.32	20.96	**32.76**	56.67	**20.68**

*Blue metric

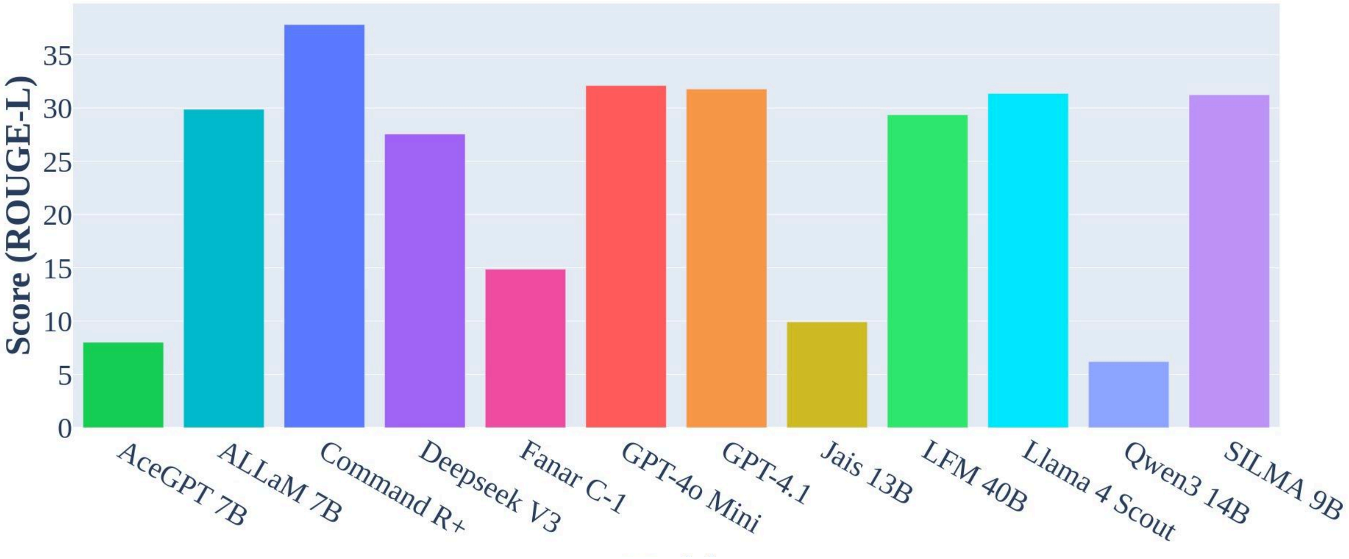
Overall Results



SILMA 9B tops with 38.66, followed closely by **GPT-4.1** and **GPT-4o Mini**. **Qwen3 14B, AceGPT 7B,** and **Jais 13B** remain at the lower end.

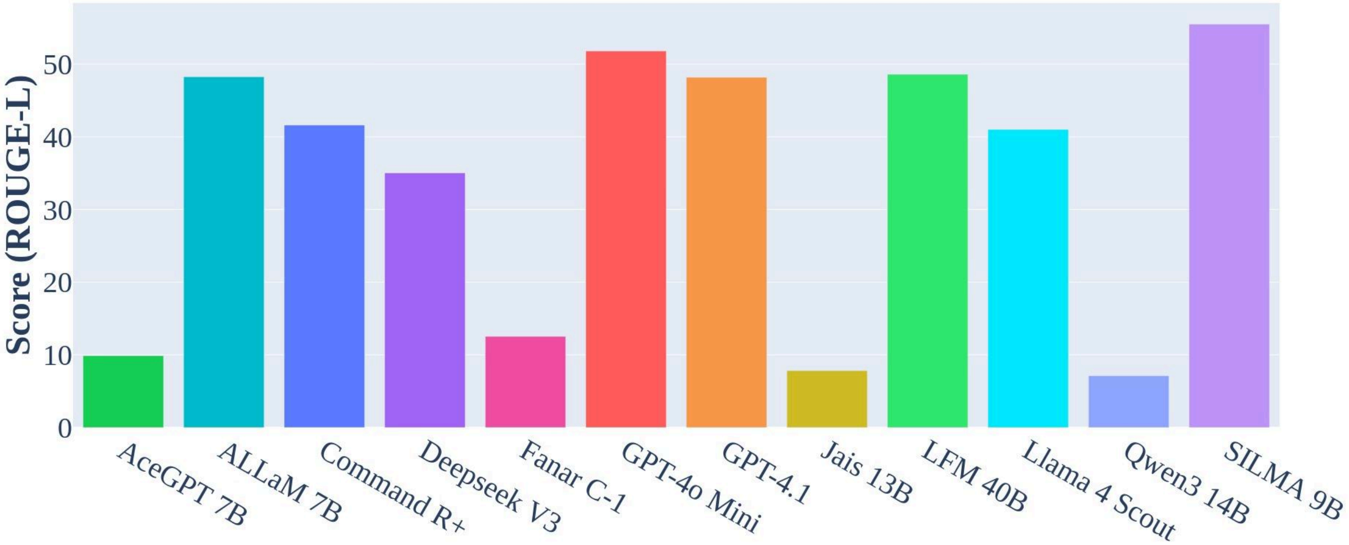
Task Specific Performance

Question Answering



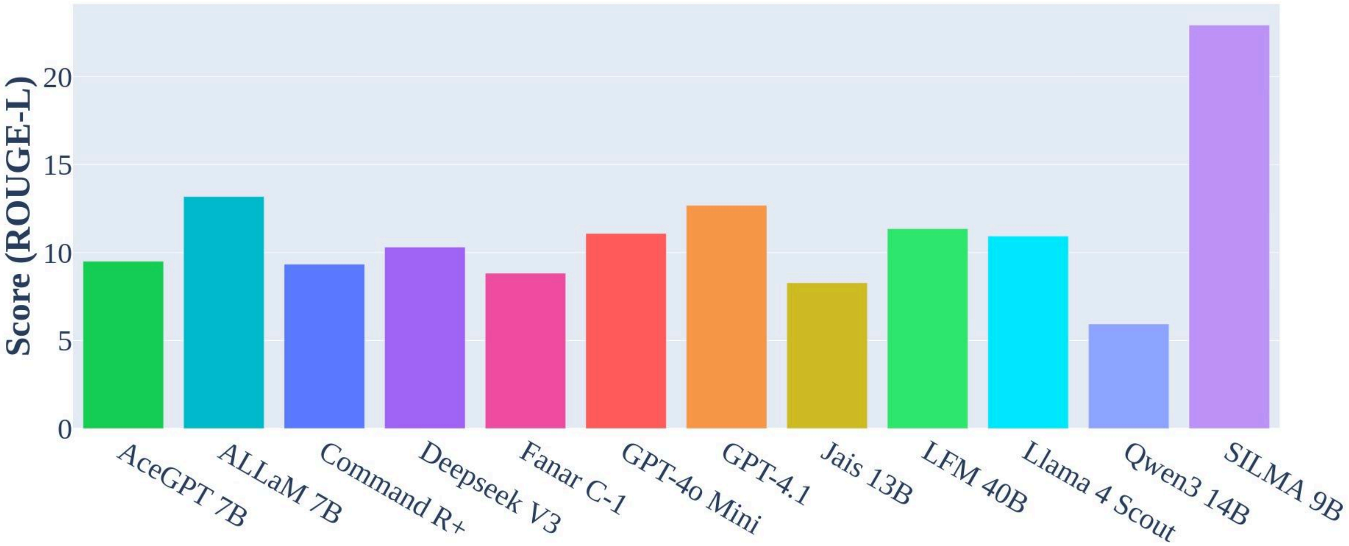
Command R+ leads with a ROUGE-L score of 37.82, closely followed by **SILMA 9B**, **GPT-4o Mini** and **GPT-4.1. Qwen3 14B** and **AceGPT 7B** perform much lower, scoring under 10

Reading Comprehension



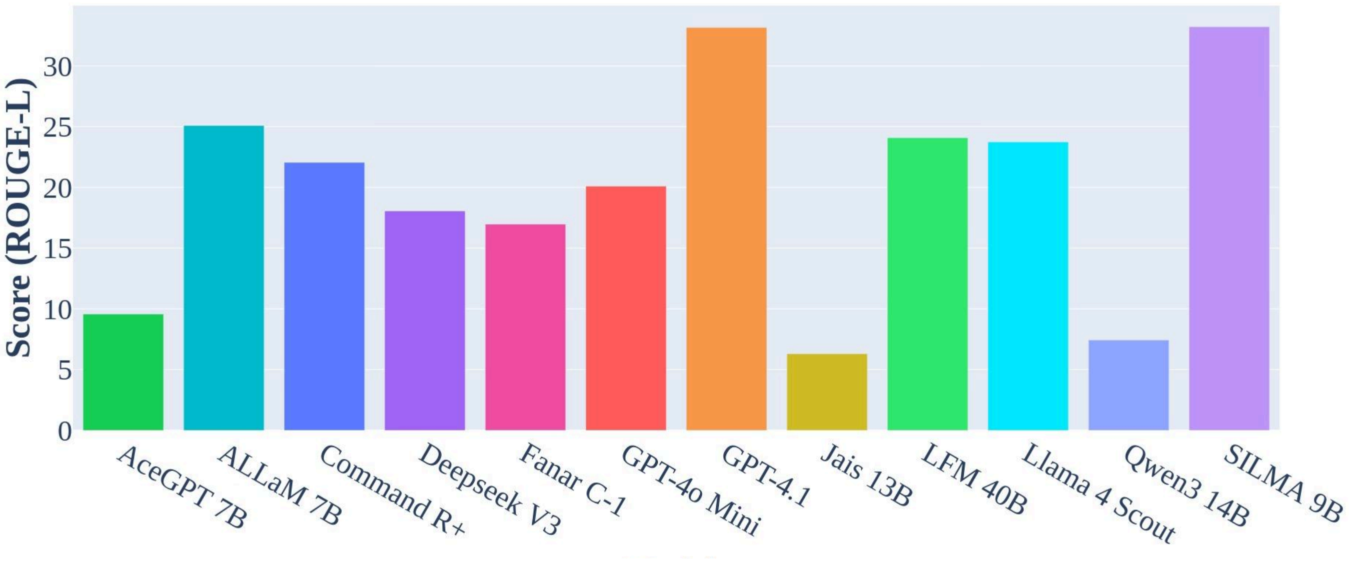
SILMA 9B excels with 55.49, while **GPT-4o Mini** and **LFM 40B** also show strong performance. **Qwen3 14B** and **Jais 13B** lag behind, scoring below 8.

Creative Writing



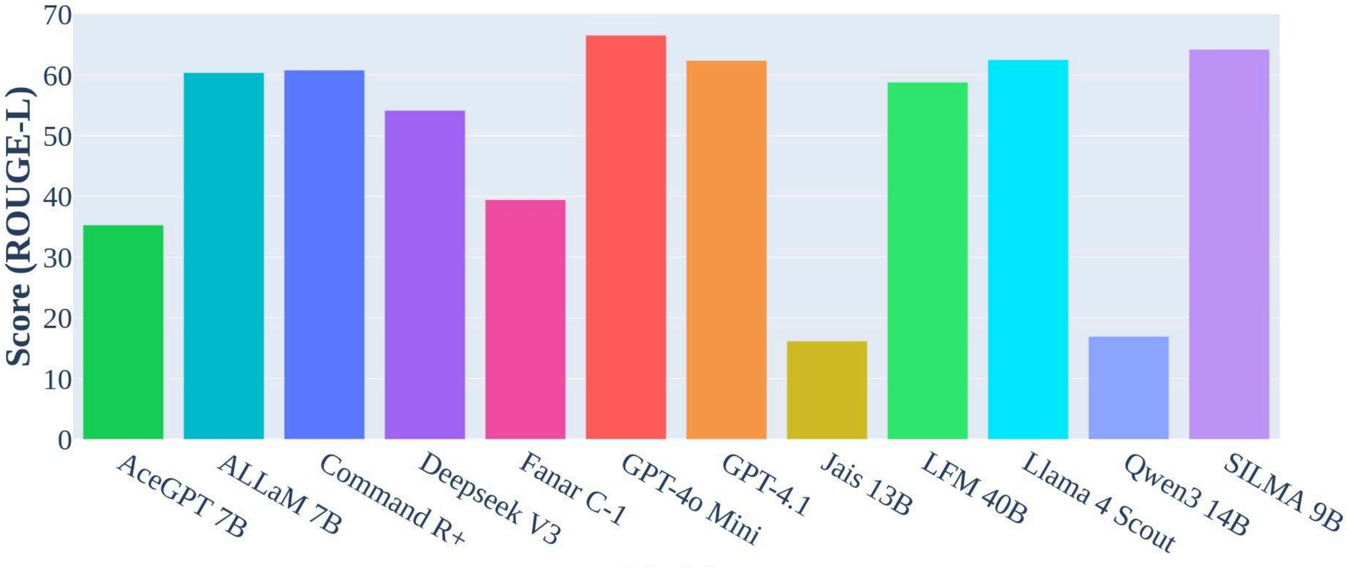
SILMA 9B is the front-runner at 22.92. Other models, like **ALLaM 7B** and **GPT-4.1,** perform similarly around 12-13, with **Qwen3 14B** at a low of 5.94.

Information Extraction



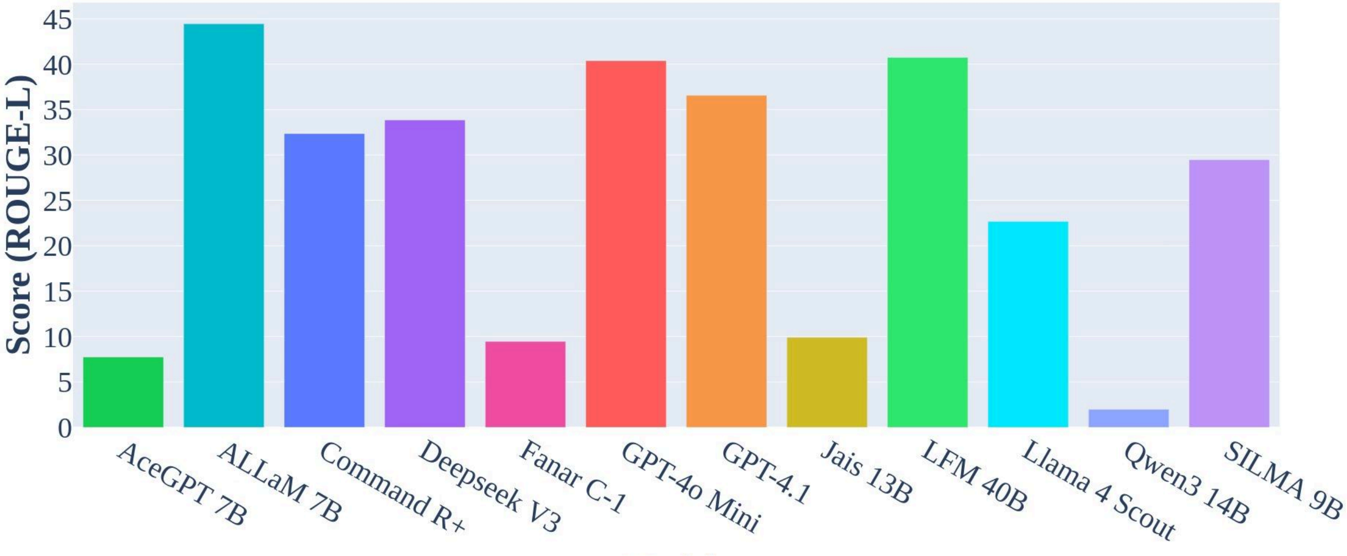
SILMA 9B and **GPT-4.1** lead**,** both scoring above 33, while **Jais 13B** and **Qwen3 14B** perform poorly below 8.

Linguistic Processing



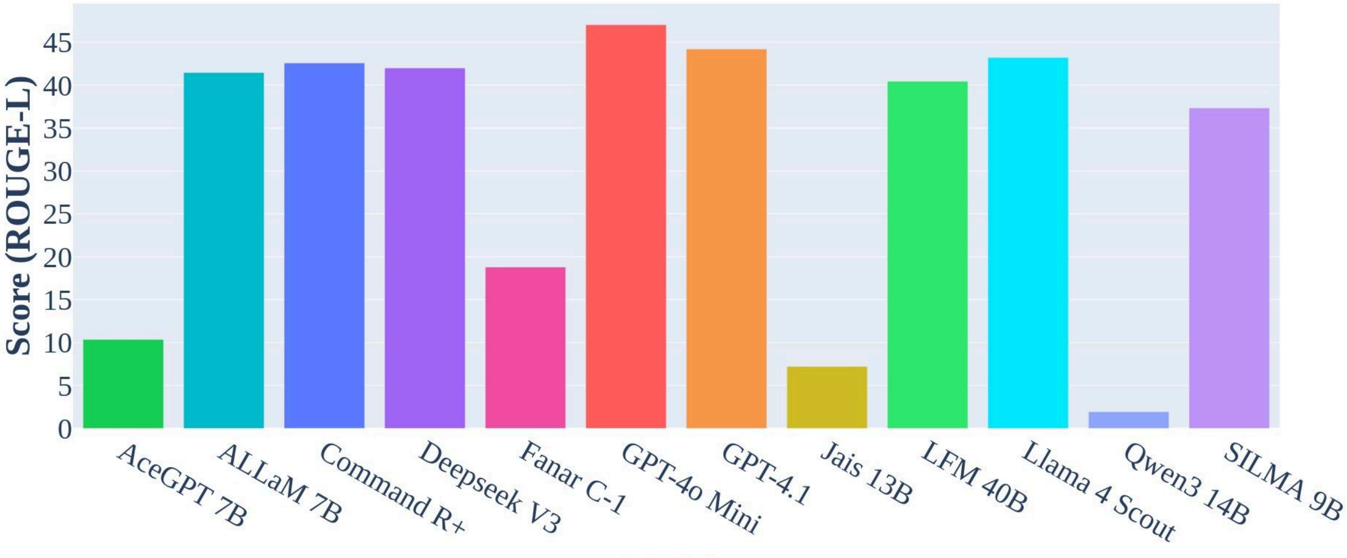
GPT-4o Mini scores highest at 66.53, with **SILMA 9B** close behind. **Jais 13B** and **Qwen3 14B** perform significantly lower, with scores under 17.

Logical Reasoning



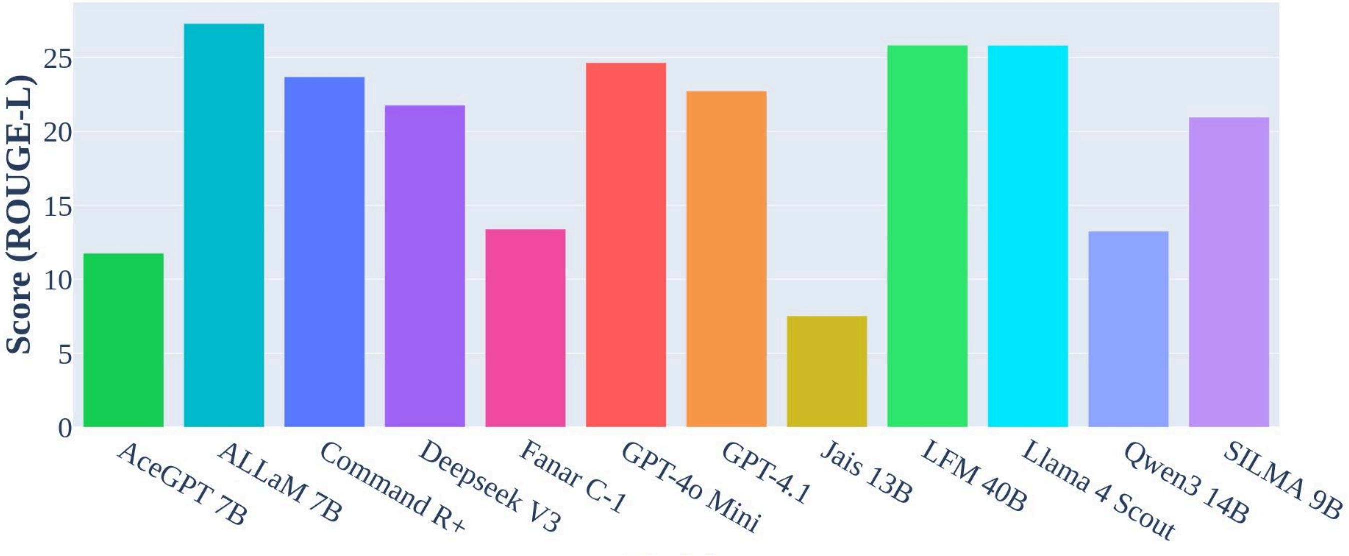
ALLaM 7B excels at 44.44, with **LFM 40B** close behind. **Qwen3 14B** scores the lowest, at 1.98.

Sequence Tagging



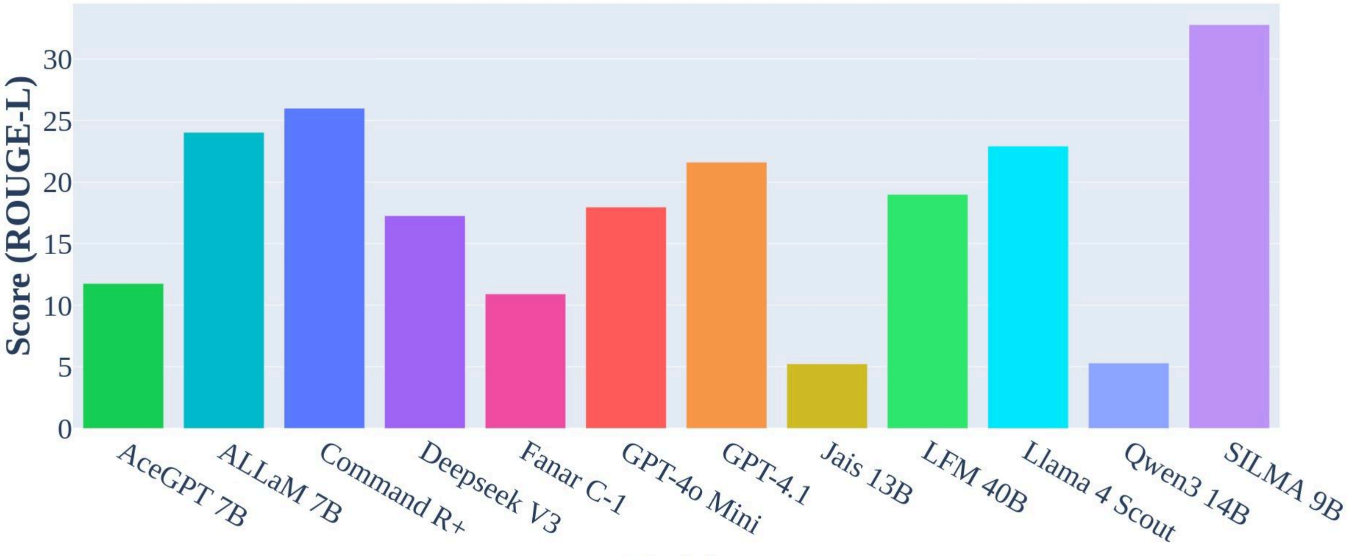
GPT-4o Mini leads with 47.01, followed by models like **Command R+** and **Llama 4 Scout. Qwen3 14B** performs the worst, scoring just 1.94.

Summarization



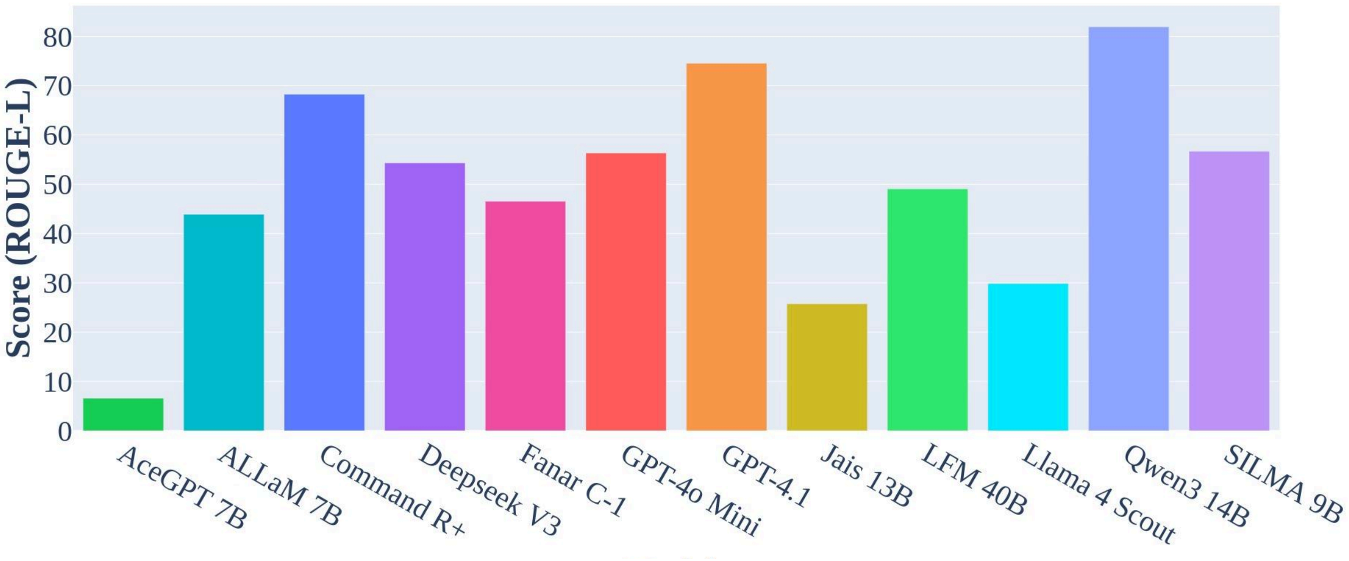
LFM 40B and **Llama 4 Scout** are top performers around **25.83. Qwen3 14B** and **Jais 13B** struggle, both under 14.

Text Classification



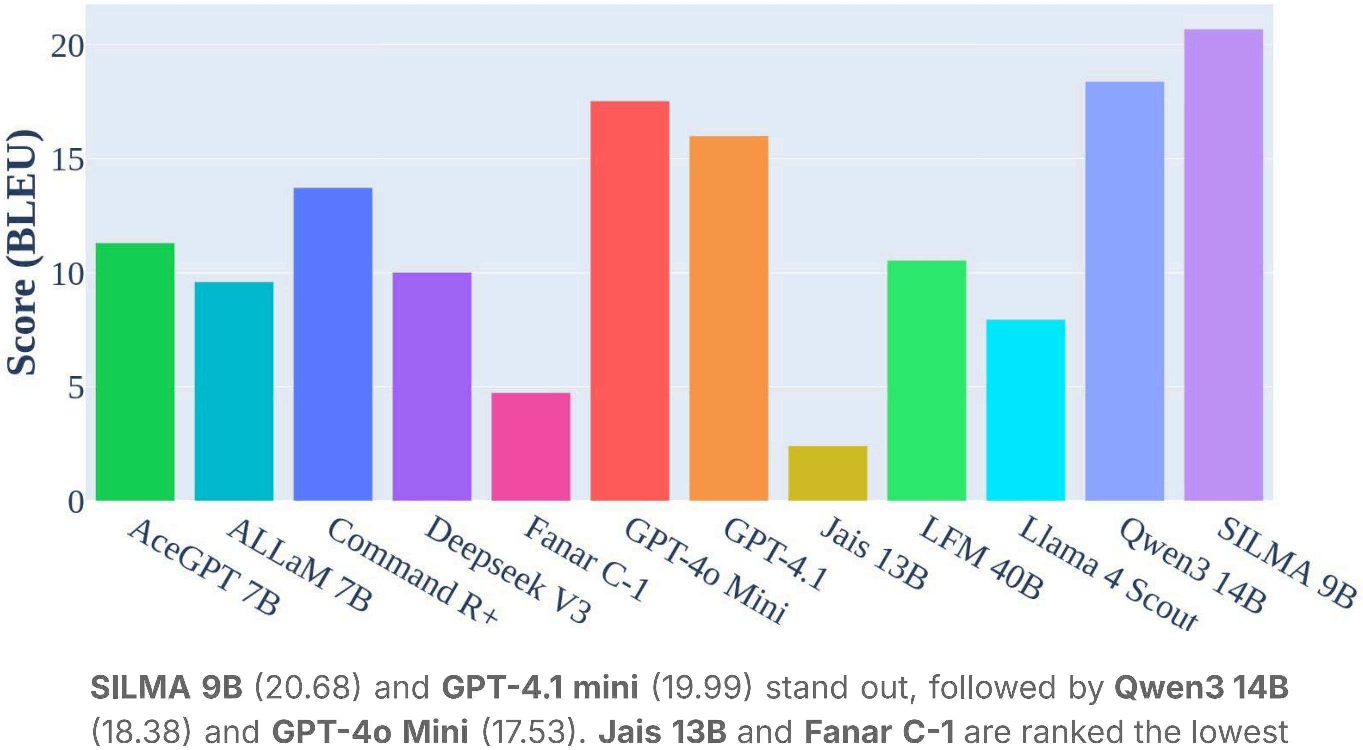
SILMA 9B stands out with **32.76,** while models like **Command R+** and **ALLaM 7B** show decent performance. **Qwen3 14B** and **Jais 13B** are at the bottom, below 6.

Program Execution



Qwen3 14B leads with 81.89, followed by **GPT-4.1** at 74.52. **AceGPT 7B** and **Jais 13B** show the weakest performance, under 26.

Translation



SILMA 9B (20.68) and **GPT-4.1 mini** (19.99) stand out, followed by **Qwen3 14B** (18.38) and **GPT-4o Mini** (17.53). **Jais 13B** and **Fanar C-1** are ranked the lowest (~2–5).

LLM as a Judge

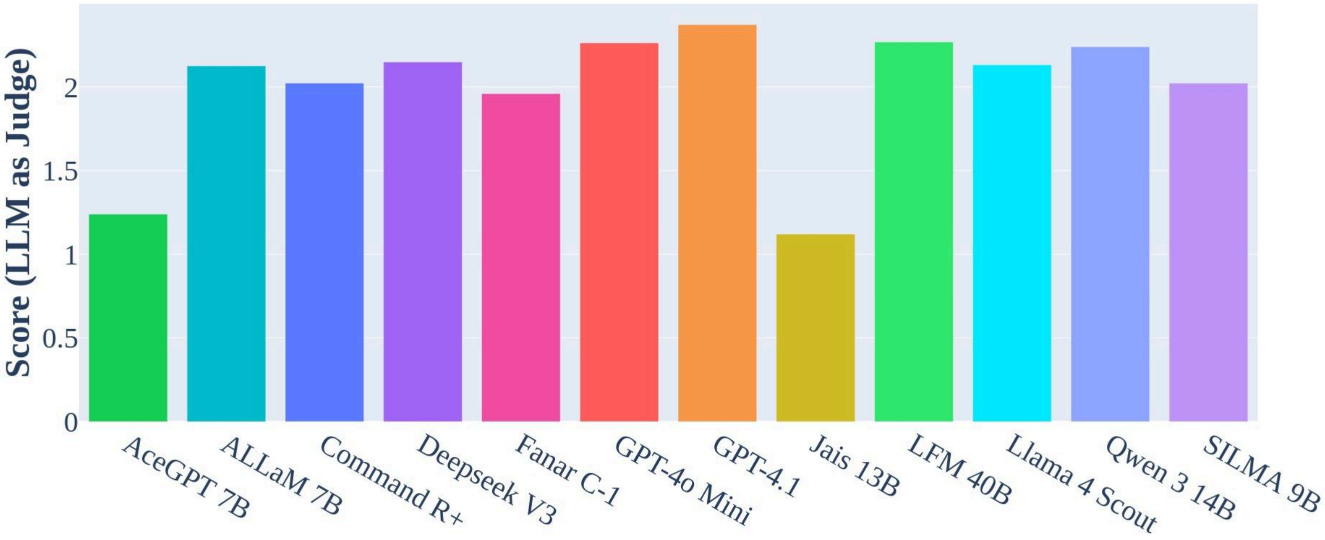
The rise of Large Language Models (LLMs) has necessitated more sophisticated evaluation methods beyond traditional metrics like BLEU or ROUGE, which sometimes fail to capture the nuanced quality of human-like text generation. This has led to the development of "LLM-as-a-judge" – a paradigm where a powerful LLM acts as an evaluator, leveraging its advanced linguistic understanding to assess the outputs of other LLMs.

The core principle involves prompting a "judge" LLM with the input query, the generated output from the "candidate" LLM, and clear evaluation criteria. The judge LLM then scores, critiques, or compares outputs, providing explanations for its judgments. In our experiments, the scores for each candidate LLM ranged from 0 to 3, 0 being the worst and 3 being the best. After testing several candidate models for the Judge, we selected the Gemini 2.5 Flash model, which offered the best performance-to-cost ratio for the Judge LLM. This model was not included among the benchmarking candidates.

This method offers significant advantages, primarily scalability and cost-effectiveness, as LLMs can evaluate vast amounts of text rapidly and consistently - which is not possible via Human Evaluation. They also excel at capturing linguistic nuances like coherence and tone, which traditional metrics often miss.

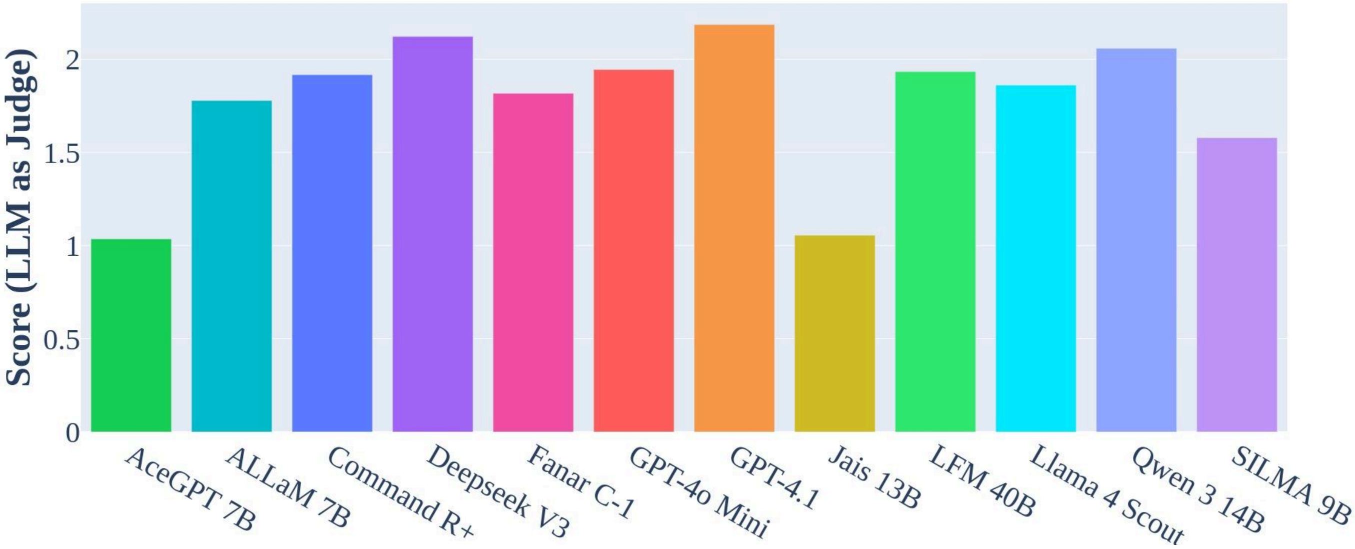
In the following plots, we provide scores for 3 tasks: Reading Comprehension, Information Extraction and Logical Reasoning. A more comprehensive evaluation will be provided later.

Reading Comprehension



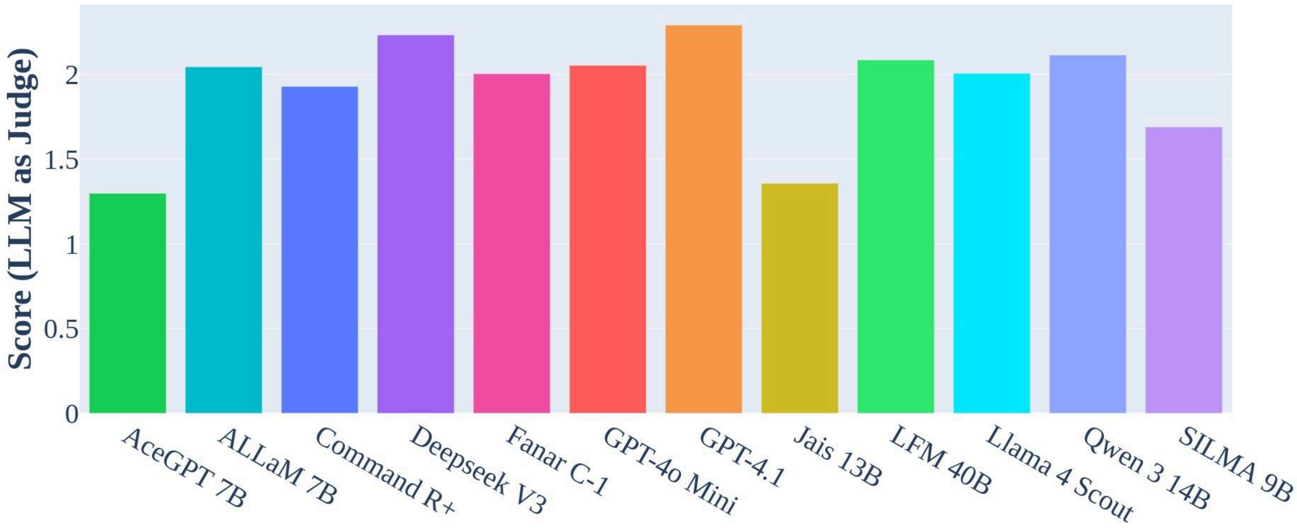
LFM 40B and **GPT 4.1** stand out. **AceGPT 7B** and **Jais 13B** are having the worst scores

Information Extraction



Deepseek V3 and **GPT 4.1** have leading scores. **AceGPT 7B** and **Jais 13B** are ranked the lowest.

Logical Reasoning



Deepseek V3 and **GPT 4.1** lead the fray with **AceGPT 7B** and **Jais 13B** having the worst scores

Comparing Traditional Metrics vs. LLM as a Judge

When comparing traditional metrics like ROUGE to "LLM as a Judge," distinct differences emerge in evaluating LLM performance. ROUGE emphasizes surface-level lexical similarity through n-gram overlap, while "LLM as a Judge" assesses semantic understanding and coherence, leveraging deeper model capabilities.

In **Information Extraction**, models like GPT-4o mini and Deepseek V3 score highly on "LLM as a Judge," indicating strong semantic alignment despite moderate ROUGE scores, while SILMA 9B shows high ROUGE but lower semantic alignment. Both metrics agree on models like Command R+, where scores are consistently moderate.

In **Logical Reasoning**, GPT-4o mini and Deepseek V3 excel in logical consistency with strong scores in both metrics. Conversely, ALLaM 7B's high ROUGE doesn't reflect in "LLM as a Judge" scores. Both metrics align for Llama 4 Scout, suggesting a balance between lexical and logical strengths

In **Reading Comprehension**, models such as GPT-4o mini and LFM 40B score well on both metrics, indicating a good balance of lexical fidelity and understanding. Meanwhile, SILMA 9B's high ROUGE highlights limitations in deeper comprehension. Consistently high scores for models like GPT-4o mini reflect agreement in evaluating comprehensive performance.

Overall, discrepancies between ROUGE and LLM as a Judge scores highlight that traditional metrics like ROUGE may not fully capture the nuanced aspects of language understanding that AI models can now achieve. LLM as a Judge scores offer an additional layer of evaluation, focusing on semantic accuracy, coherence, and logical reasoning, which are crucial for tasks requiring deeper language comprehension. Where both metrics agree, they reinforce the balanced strengths of a model, and combining them provides a comprehensive evaluation of modern AI capabilities.