BENCHMARK REPORT

Independent Multi-Task Evaluation of Large Language Models

aixplain

aixplain

Benchmark Report

Independent Multi-Task Evaluation of Large Language Models

Arabic | 18 Models | 11 Tasks

Version 2.1

May 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 May 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.

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 18 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 addition to the 9 models evaluated in the previous evaluation (Command R+, Deepseek V3, Gemma 2, GPT-4o Mini, Llama 3.2, Llama 3.3, Llama 4 Maverick and Llama 4 Scout, Qwen 2.5), we added 9 additional models: ALLaM 7B, ALLaM 13B, Fanar C-1, GPT-4.1, GPT-4.1 mini, GPT-4.1 nano, LFM 40B, Qwen3 14B, Qwen3 32B.

Key Observations

Among the models evaluated, GPT-4.1 (closed-source) achieves the highest overall score, closely followed by GPT-4o mini and then Command R+. Top-performing open models such as Command R+, ALLaM 7B, and LFM 40B deliver results that are nearly comparable to the leading closed alternatives on most language tasks. Notably, open models are advancing rapidly and have demonstrated strong performance, particularly in reasoning, extraction, and summarization tasks.

Interestingly, larger model size does not always guarantee better performance. Smaller models like ALLaM 7B frequently outperform much larger models, suggesting that factors like training data quality and architecture outweigh sheer parameter count. Effective model training with relevant data and design are proving to be as important as size when it comes to real-world results.

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.

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.

For organizations that require the very best out-of-the-box performance—especially at enterprise scale—closed models such as GPT-4.1 are still the first recommendation. However, for research, development, and customization needs, top open-source models like Command R+, ALLaM 7B, and LFM 40B offer strong, competitive results along with benefits in transparency and flexibility. 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.

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
ALLaM 7B	7B	4096	Open
ALLaM 13B	13B	4096	Open
Command R+	104B	128k	Open
Deepseek V3	671B	128k	Open
Fanar C-1	9B	4k	Open
Gemma 2	9B	8192	Open
GPT-4o mini	Unknown	128k	Closed
GPT-4.1	Unknown	1M	Closed
GPT-4.1 mini	Unknown	1M	Closed
GPT-4.1 nano	Unknown	1M	Closed
LFM 40B	40B	32k	Open
Llama 3.2	3B	128k	Open
Llama 3.3	70B	128k	Open
Llama 4 Maverick	400B	1M	Open
Llama 4 Scout	109B	10M	Open
Qwen2.5	32B	128k	Open
Qwen3 14B	14B	32k	Open
Qwen3 32B	32B	32k	Open

Benchmark Results

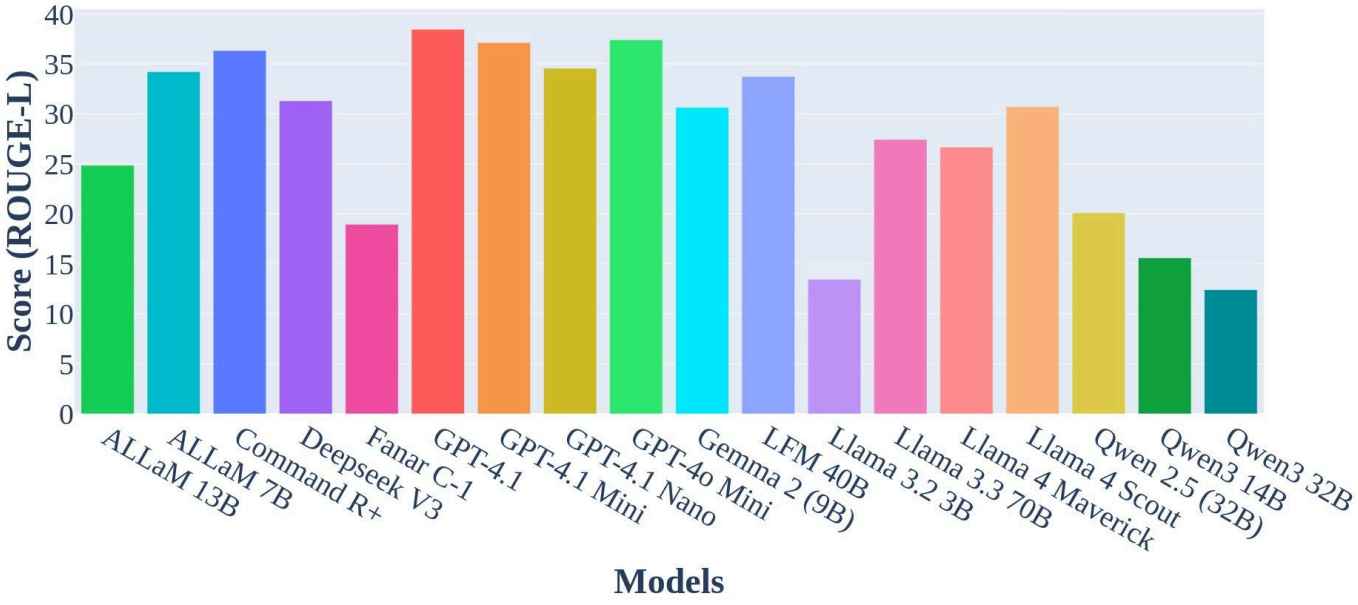
This section presents results of the benchmarking of LLMs across different tasks. Results for the Translation task are reported in Bleu (the higher the better), and the results for other tasks are reported in ROUGE-L metric (the higher the better).

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*
ALLaM 13B	24.85	24.36	28.63	12.94	21.65	54.72	19.65	34.35	24.13	12.48	16.27	10.50
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
Gemma 2 (9B)	30.65	29.21	39.13	12.09	19.43	51.33	18.30	40.83	23.44	15.70	55.76	11.50
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
GPT-4.1 mini	37.12	32.11	43.35	12.28	30.40	64.70	31.34	44.80	24.72	20.88	63.72	**19.99**
GPT-4.1 nano	34.56	30.27	48.50	11.85	24.49	58.50	31.48	37.42	23.96	16.98	65.58	15.86
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 3.2 3B	13.44	14.53	24.83	6.01	9.77	25.23	7.98	15.09	14.32	7.26	13.62	2.95
Llama 3.3 70B	27.44	27.00	36.66	10.77	20.08	54.77	22.92	33.91	19.80	15.38	38.71	7.50
Llama 4 Maverick	26.67	27.42	37.17	11.68	23.24	64.46	16.55	25.71	25.39	19.90	16.95	9.20
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
Qwen2.5 (32B)	20.09	18.28	20.02	9.16	12.29	44.39	11.45	17.02	21.34	11.26	32.43	11.50
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
Qwen3 32B	12.40	6.52	5.79	6.21	6.90	14.79	1.42	1.89	13.62	4.34	59.47	14.40

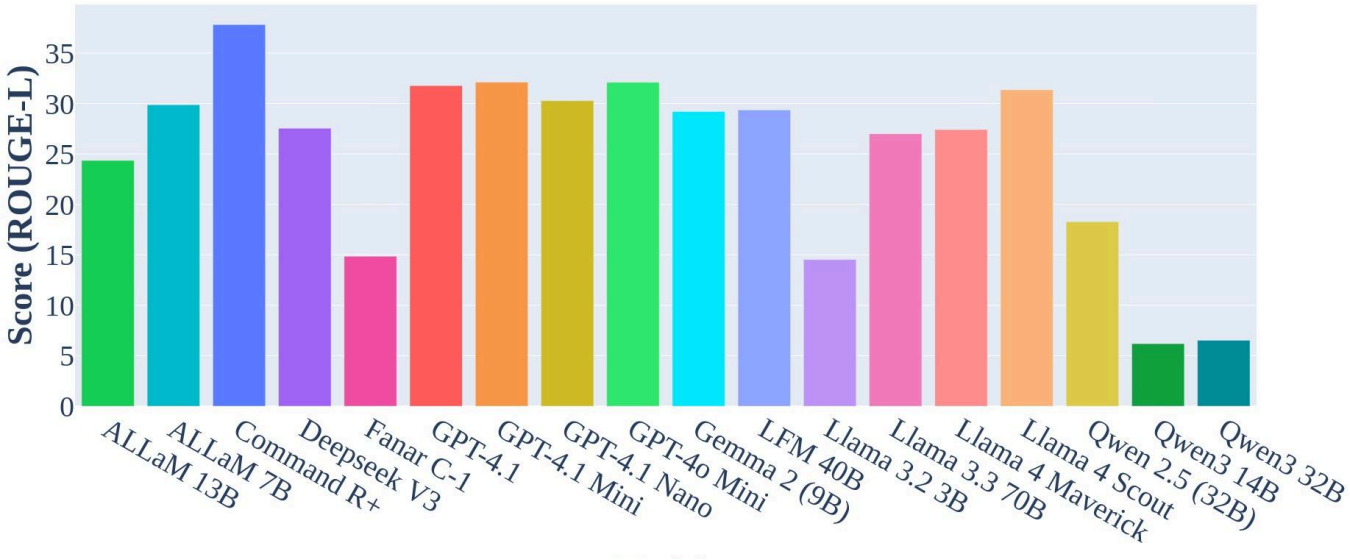
*Blue metric

Overall Results



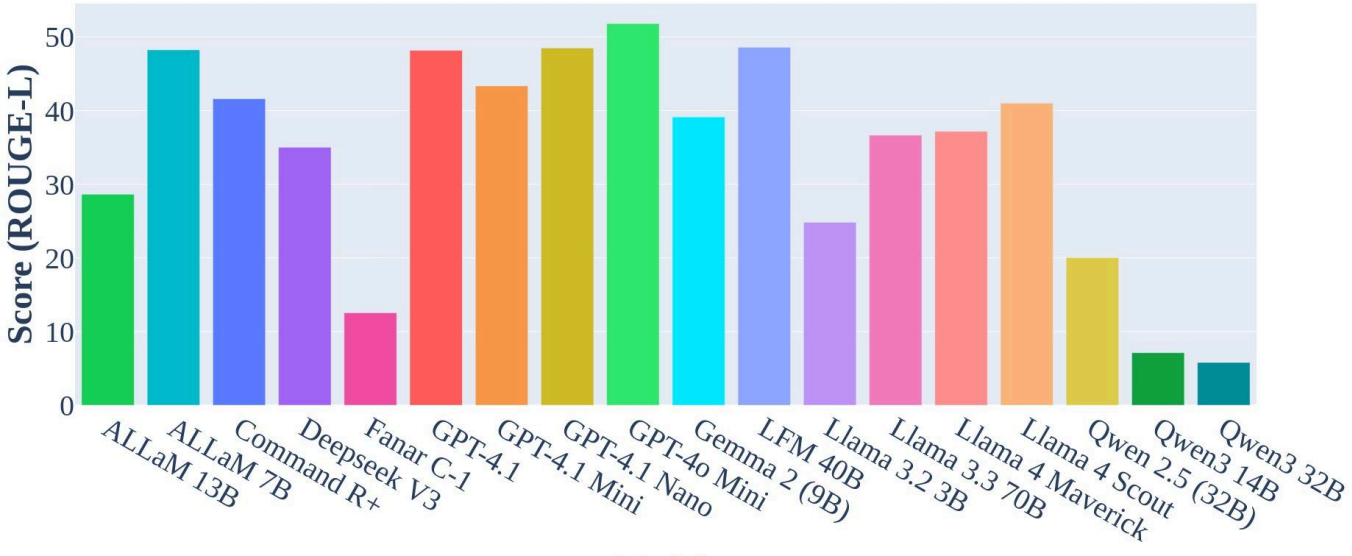
Task Specific Performance

Question Answering



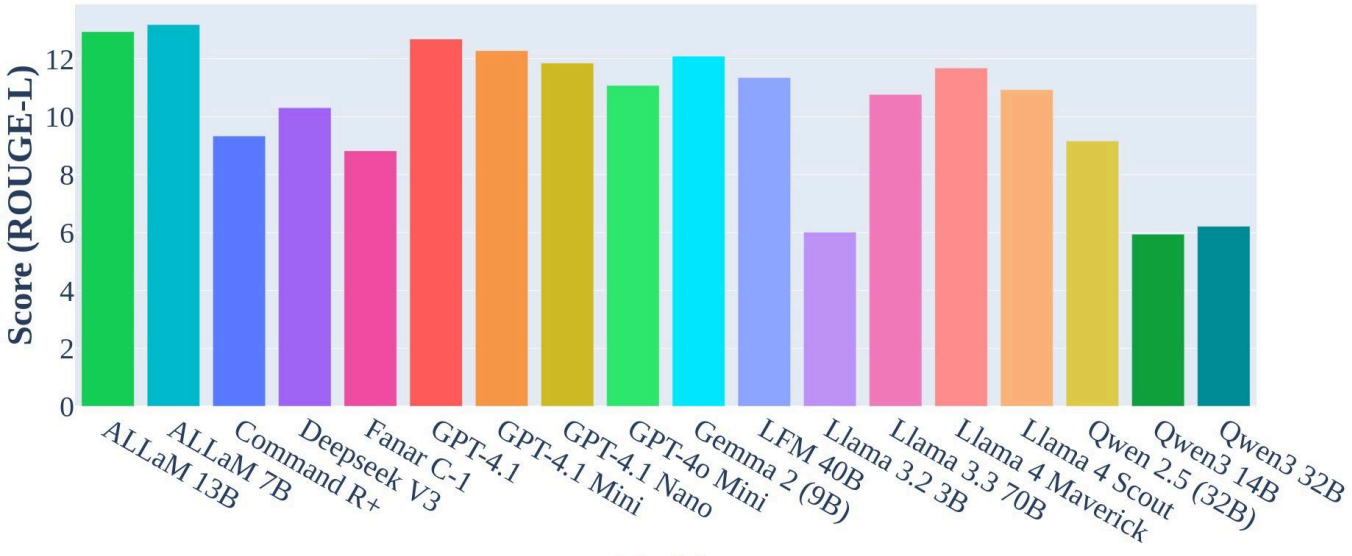
Command R+ leads with a ROUGE-L score of 37.82, closely followed by **GPT-4o mini** and the **GPT-4.1 mini** family. **Llama 3.2 3B** and **Qwen3 models (14B/32B)** perform much lower, scoring under 15.

Reading Comprehension



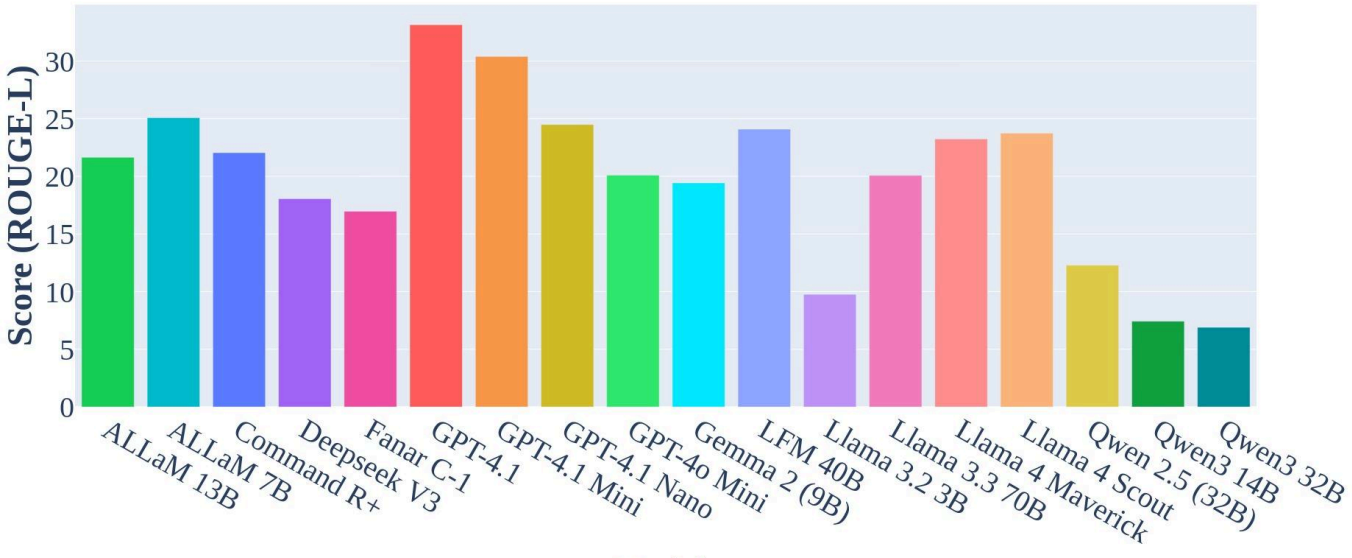
GPT-4o mini and **LFM 40B** are top performers (51.81 and 48.6), with **ALLaM 7B** and **GPT-4.1 variants** also strong. **Qwen3 14B/32B** and **Fanar C-1** lag far behind, under 13.

Creative Writing



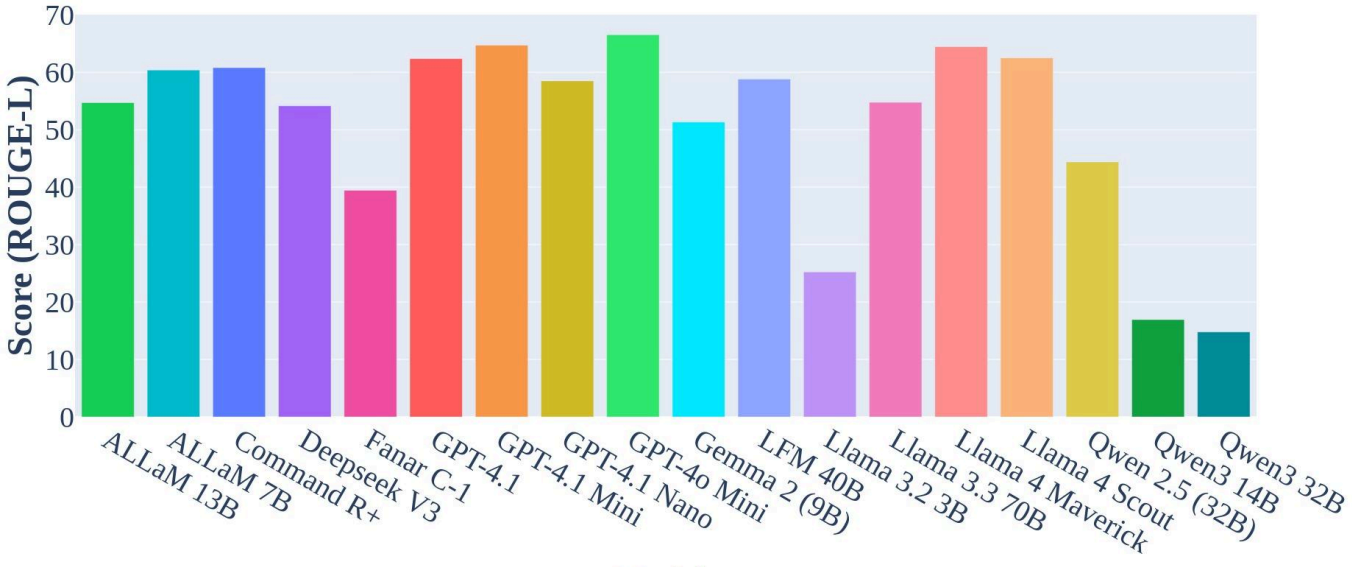
ALLaM 7B/13B, GPT-4.1, and **Gemma 2 (9B)** are best, all around 12–13 points. **Llama 3.2 3B** and **Qwen3 14B/32B** are weakest, below 6.5.

Information Extraction



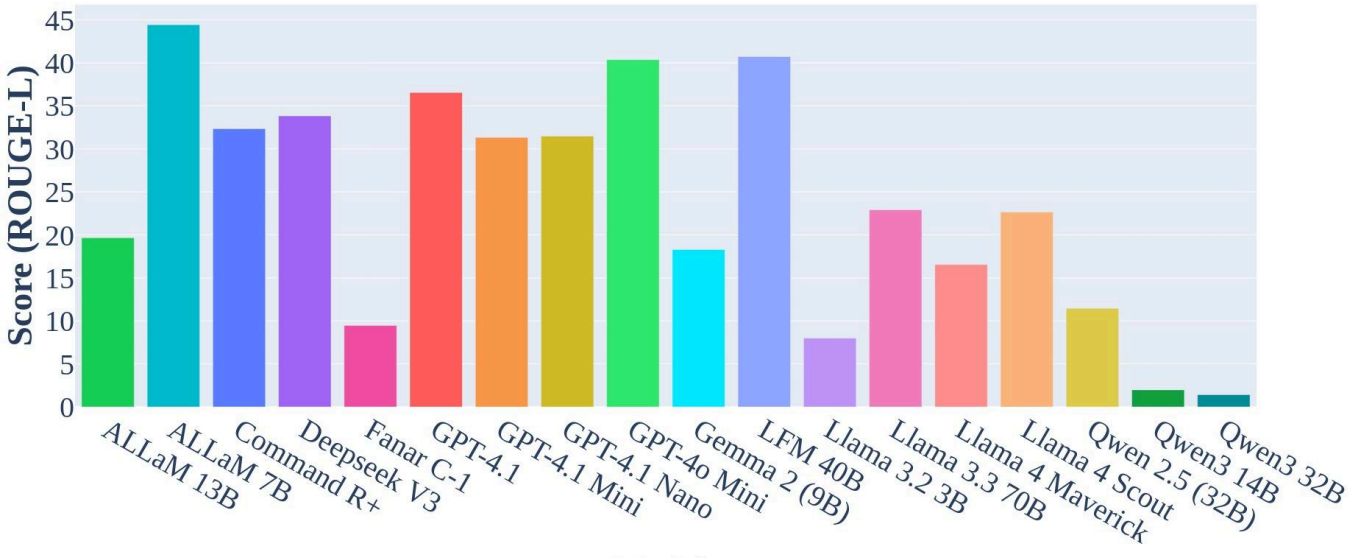
GPT-4.1 (33.15) is out front, with **ALLaM 7B, LFM 40B,** and **the GPT-4.1 mini/nano** close behind. **Qwen3 14B/32B** are lowest at around 7.

Linguistic Processing



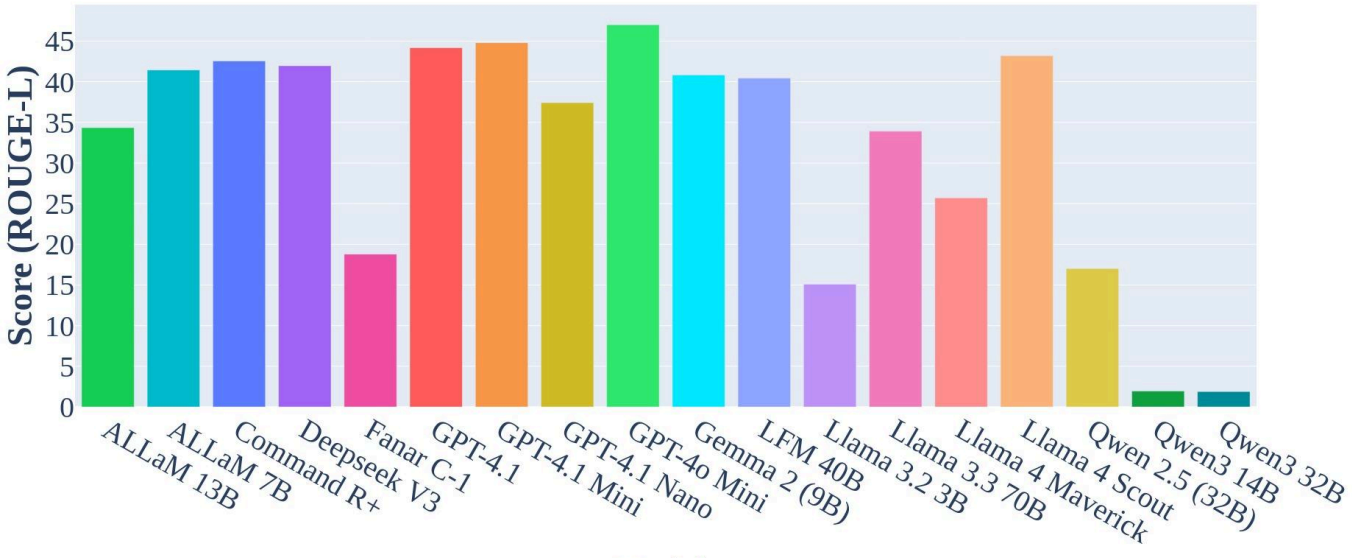
GPT-4o mini, Llama 4 Maverick, and **GPT-4.1 mini** shine above 64, while **Llama 3.2 3B** (25.23) and **Qwen3 14B/32B** (~15) underperform.

Logical Reasoning



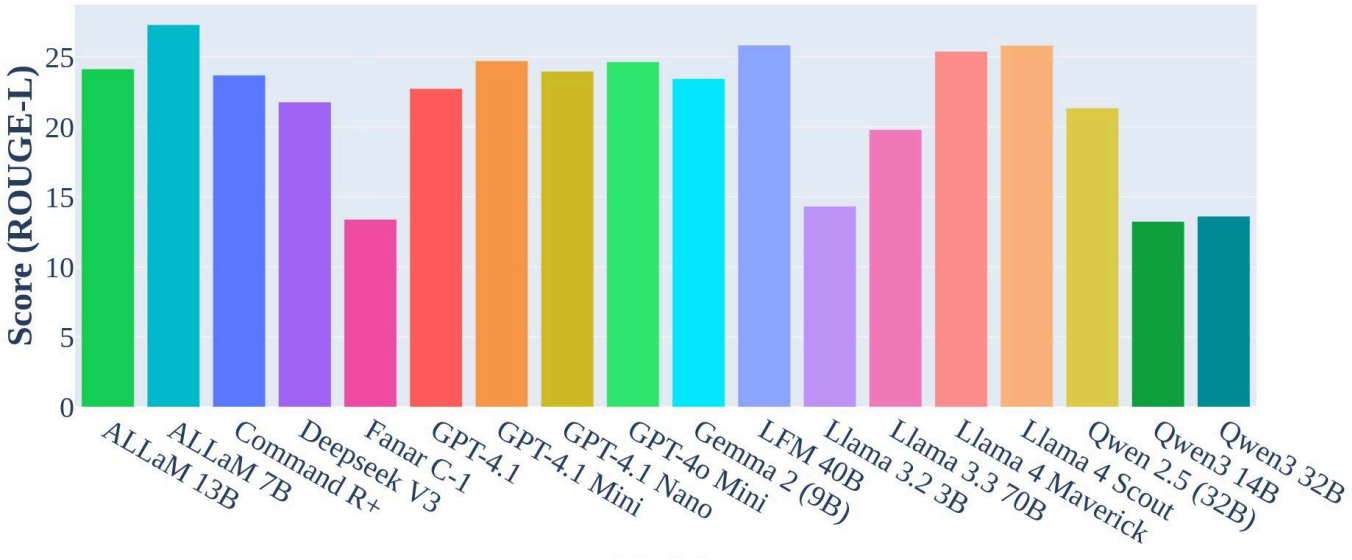
ALLaM 7B (44.44) and **LFM 40B** (40.74) lead, with **GPT-4o mini** and **Deepseek V3** also strong. **Qwen3 models** are lowest, near 2, with **Llama 3.2 3B** not far ahead.

Sequence Tagging



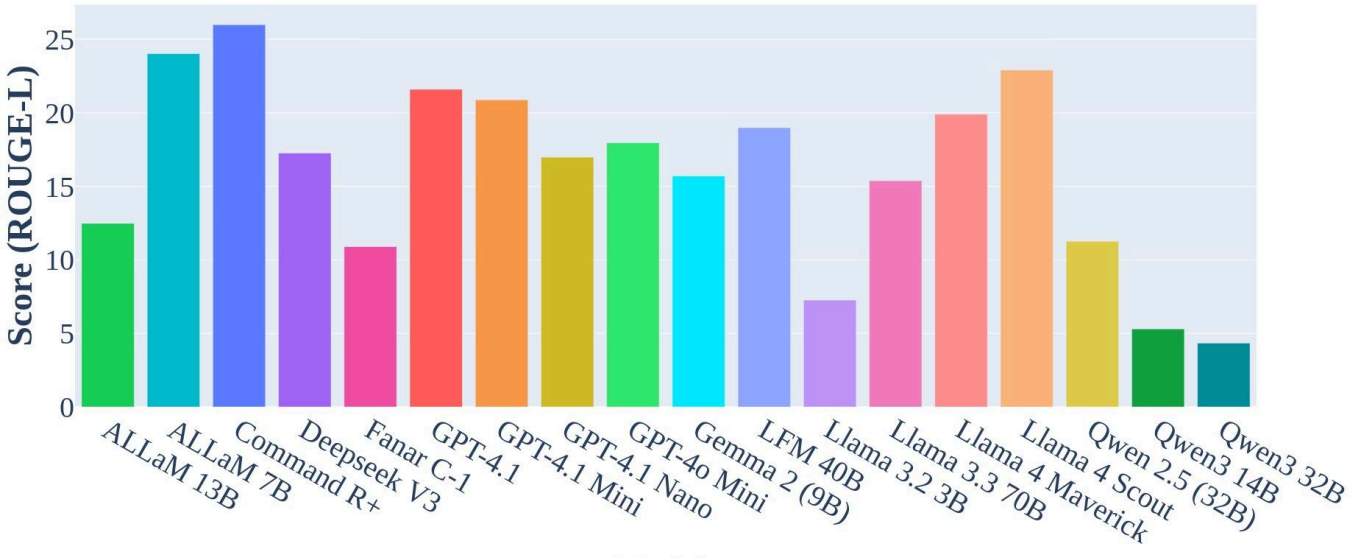
GPT-4o mini (47.01), **Llama 4 Scout** (43.2), and **GPT-4.1/mini** dominate, while **Qwen3 14B/32B** and **Llama 3.2 3B** lags behind.

Summarization



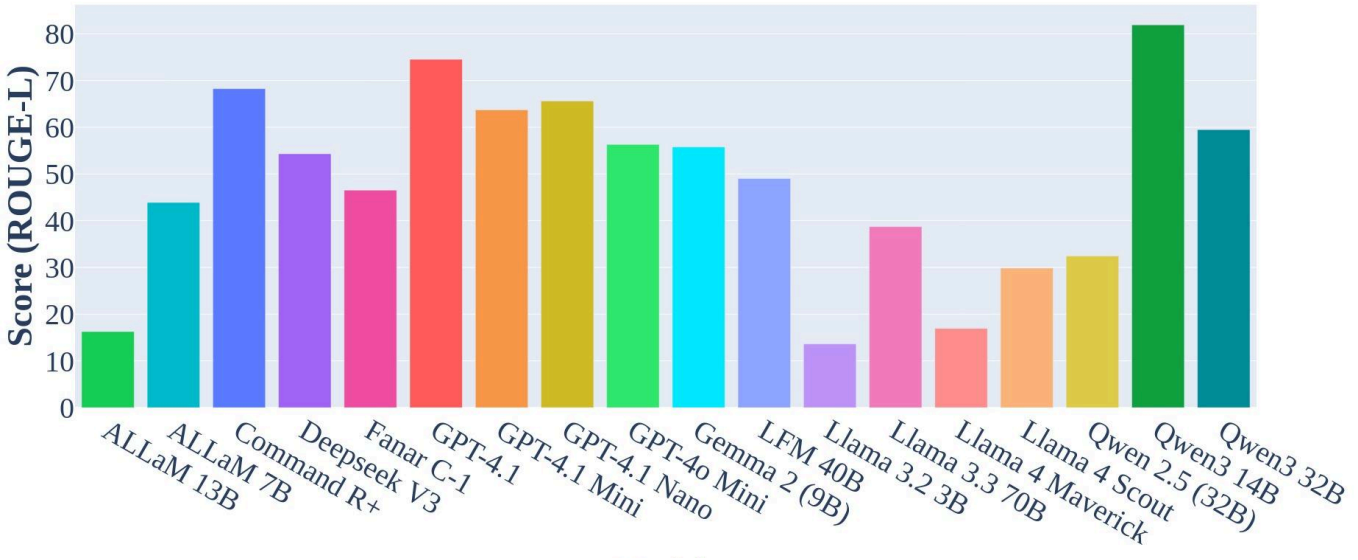
LFM 40B (25.83) and **ALLaM 7B** (27.29) perform best. Most models cluster between 20–25, with **Llama 3.2 3B** and **Fanar C-1** at the bottom (13–14).

Text Classification



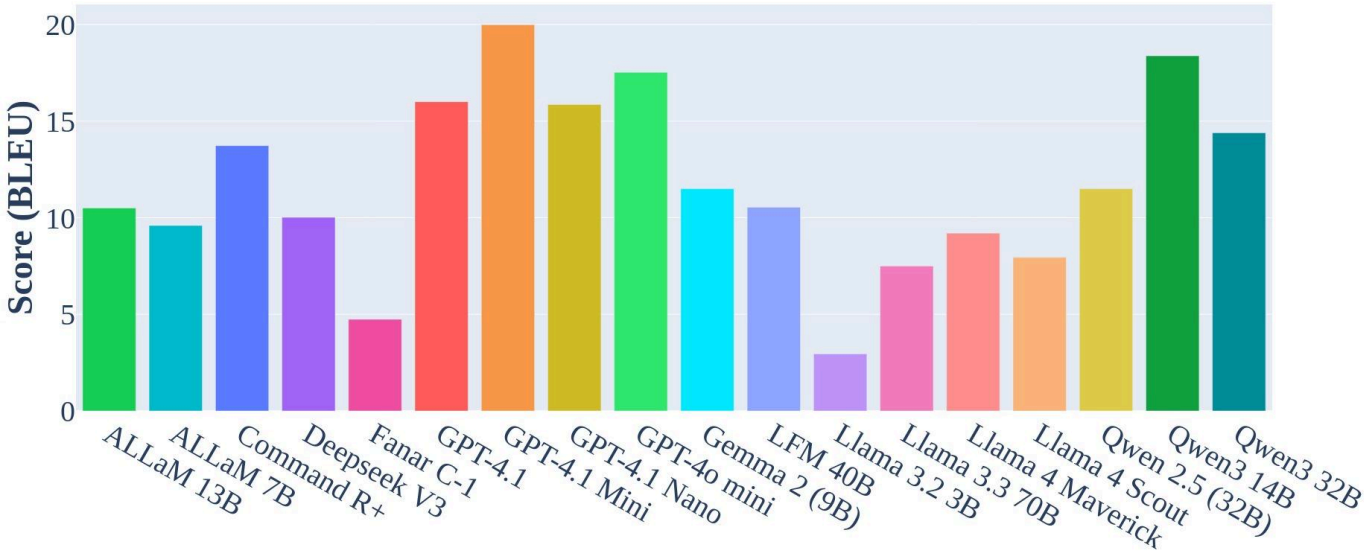
Command R+ (25.99) performs best, with **ALLaM 7B** and **Llama 4 Scou**t close. **Qwen3 14B/32B** have the weakest results.

Program Execution



Qwen3 14B tops with 81.89—much higher than others—while **Command R+, GPT-4.1, nano/mini,** and **Fanar C-1** also score well (46–74). **Llama 3.2 3B** trails at 13.62.

Translation



GPT-4.1 mini (19.99) stands out, followed by **Qwen3 14B** (18.38) **GPT-4o mini** (17.53). **Llama 3.2 3B** and **Fanar C-1** are ranked the lowest (~2–5).