

The Acacus Arabic test

Version 1.1, 15 October 2026

Version 1.1: 6 models added on 15 October 2026

Summary

A small test of twelve Arabic tasks, run once on 21 models on 15 October 2026.

We show scores task by task and do not put the models in one ranking. The answers were judged automatically and have not been reviewed by a person yet.

The full answers, the method and the open data are on the page:

<https://acacus.ly/research/arabic-test-2026-10>

How we ran the test

- We wrote twelve tasks ourselves: formal writing in standard Arabic, understanding a question in Libyan dialect, translation both ways, a summary, a lesson explained, a CV section, a maths word problem, code with Arabic comments, following format rules, and a text that mixes Arabic, English and numbers.
- We sent each task to each model once through the Acacus API, the way any customer does. No hidden instructions beyond one neutral sentence, and each model on its own default settings, temperature and thinking included. The answer limit was 8000 tokens for every model.
- We scored each answer two ways. First, mechanical checks: counting points and words, finding given text verbatim, and actually running the code on test cases. Second, two automatic judges, two different AI models, each reading only the task, the criteria and one anonymous answer in random order, then giving a score from 1 to 5.
- The score in the table is the mean of the two judges' scores, or the score a person corrected if the reader changed it. When the judges differ by more than one point we say so next to the answer, and those answers have not been reviewed by a person yet. The judges are DeepSeek V4 Pro and Xiaomi MiMo V2.6 Pro. DeepSeek V4 Pro is also one of the tested models, so in the first pass it judged its own answers; that is a possible bias we disclose, and the second pass uses a model from another maker.

What the scores mean

- 5: excellent, nothing worth a remark.
- 4: good, with a small remark.
- 3: acceptable, with visible errors.
- 2: weak, meets part of the task.
- 1: does not meet the task, is wrong, or has no answer.

The limits of this test: it is small, with few tasks, one run per model, on one day. Models and the way they are served change, and an automatic judge can be wrong or lean toward one style. That is why we do not put the models in one ranking or crown one model; look at the task closest to yours and try the model on your real work.

The scores in charts

These charts draw the same numbers as the tables below. Models are in alphabetical order, not by score, and we do not add the scores into one figure per model. It is a small test, one run per model.

Average score in each kind of task

Each bar is the model's average score in the tasks of that kind, from 1 to 5. A striped bar includes an answer the judges differed on by more than a point.

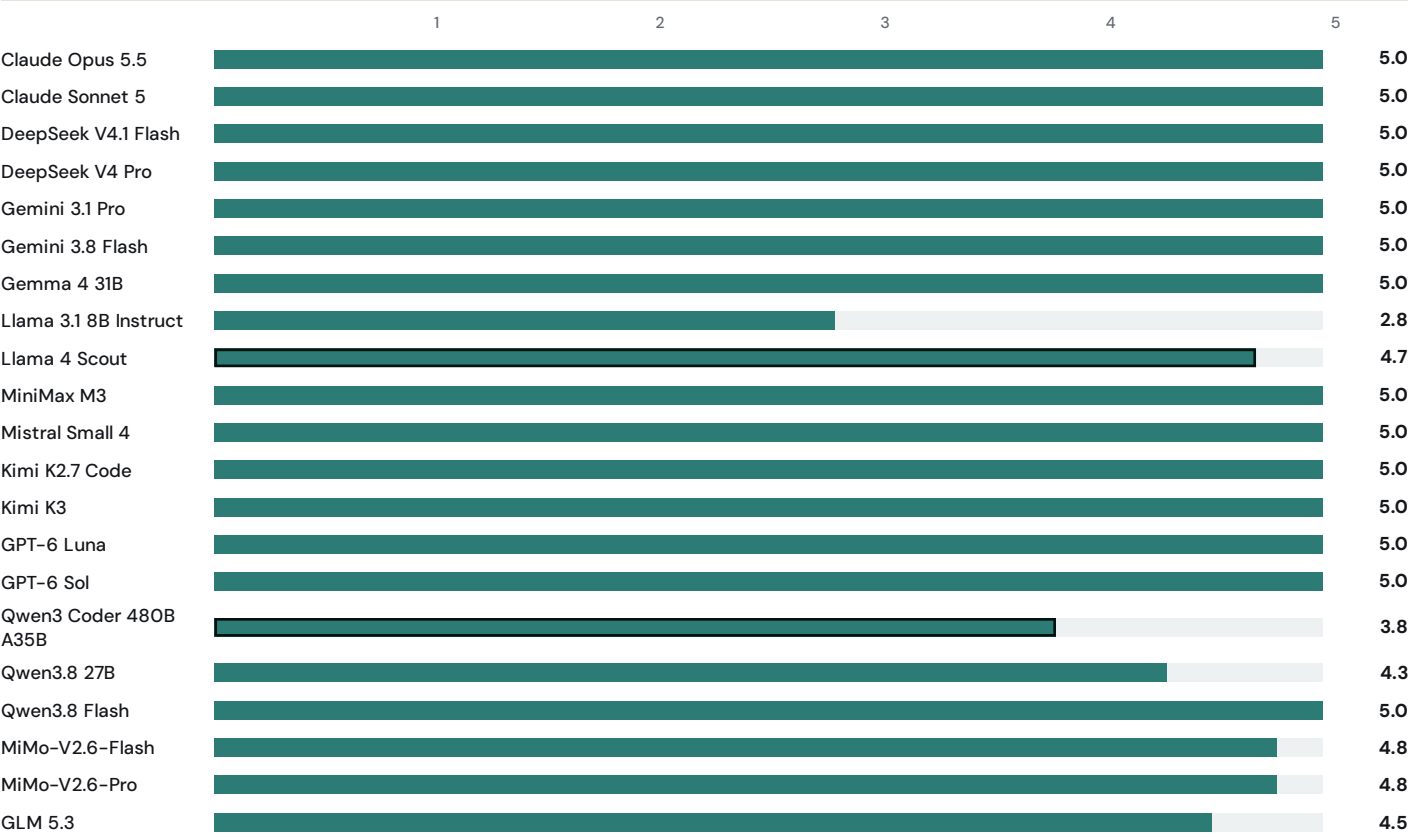
Writing in standard Arabic



Libyan dialect



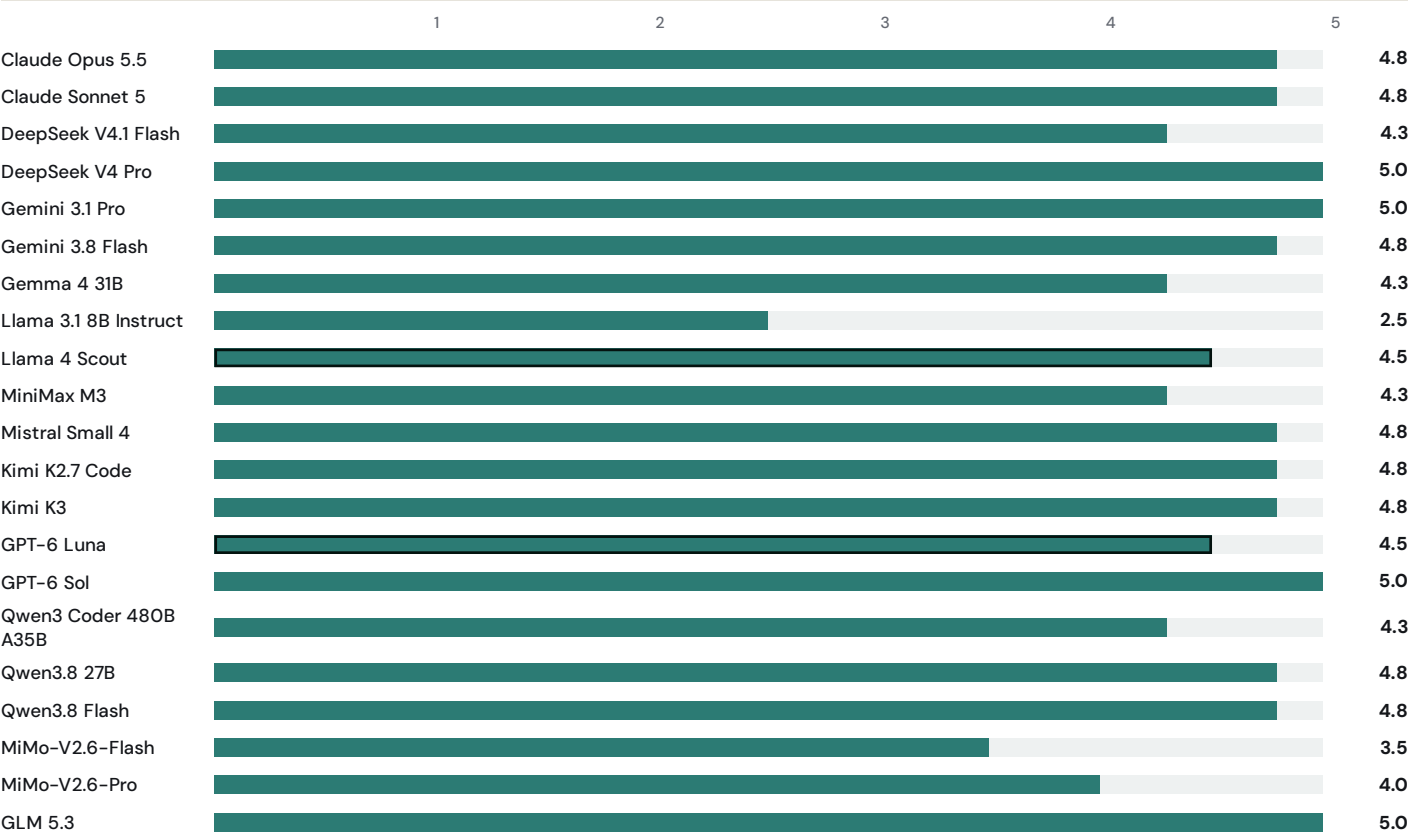
Translation



Reading and summarising



Study and arithmetic



Work and CVs



Code



Format and text direction



Every model on every task

Each cell is one model's score on one task; the darker the cell, the higher the score. On a phone, scroll the chart sideways.

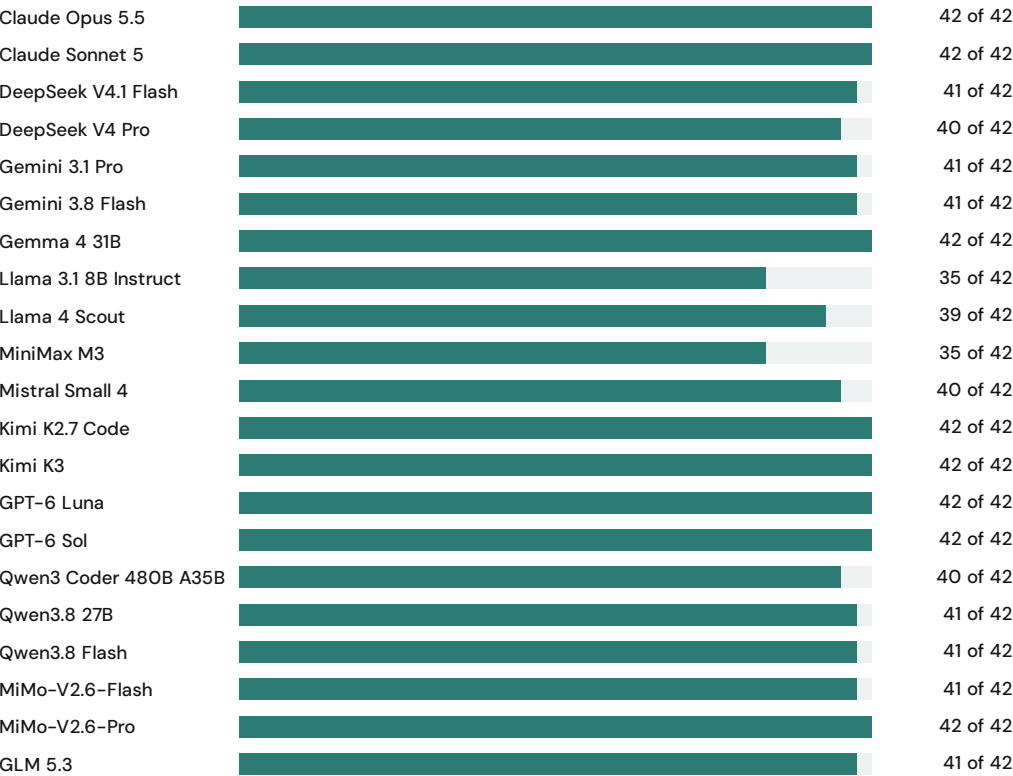
	Writing in standard Arabic		Libyan dialect	Translation		Reading and summarising	Study and arithmetic		Work and CVs	Code	For...
Model	Formal letter in Modern Standard Arabic	Short report with correct percentages	A question in Libyan dialect	Arabic to English translation	English to Arabic translation	Summarising an Arabic text	Explaining a lesson step by step	A maths word problem in Arabic	A section of a CV	A function with Arabic comments	Followi...
Claude Opus 5.5	5.0	5.0	5.0	5.0	5.0	4.5	4.5	5.0	5.0	4.5	5.0
Claude Sonnet 5	4.5	4.0	4.5	5.0	5.0	5.0	4.5	5.0	4.5	4.0	5.0
DeepSeek V4.1 Flash	5.0	4.5	4.5	5.0	5.0	4.0	3.5	5.0	5.0	4.0	5.0
DeepSeek V4 Pro	5.0	4.0	3.3	5.0	5.0	5.0	5.0	5.0	5.0	4.5	5.0
Gemini 3.1 Pro	4.5	4.5	5.0	5.0	5.0	4.5	5.0	5.0	5.0	5.0	4.0
Gemini 3.8 Flash	4.5	5.0	4.0	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.5
Gemma 4 31B	4.5	5.0	4.5	5.0	5.0	3.0	3.5	5.0	5.0	3.5	5.0
Llama 3.1 8B Instruct	2.0	1.5	1.0	3.5	2.0	1.0	1.0	4.0	4.5	2.0	2.0
Llama 4 Scout	4.5	4.0	3.0	5.0	4.3	3.0	4.0	5.0	4.0	4.0	5.0
MiniMax M3	4.5	4.0	4.0	5.0	5.0	4.0	4.0	4.5	5.0	4.5	4.0
Mistral Small 4	3.5	4.5	2.5	5.0	5.0	4.0	4.5	5.0	4.0	4.0	3.5
Kimi K2.7 Code	5.0	5.0	3.0	5.0	5.0	4.5	4.5	5.0	5.0	4.5	5.0
Kimi K3	5.0	4.5	4.0	5.0	5.0	5.0	4.5	5.0	5.0	4.0	5.0
GPT-6 Luna	5.0	5.0	5.0	5.0	5.0	4.5	4.0	5.0	5.0	5.0	5.0

Model	Writing in standard Arabic		Libyan dialect	Translation		Reading and summarising	Study and arithmetic		Work and CVs	Code	Forr c
	Formal letter in Modern Standard Arabic	Short report with correct percentages	A question in Libyan dialect	Arabic to English translation	English to Arabic translation	Summarising an Arabic text	Explaining a lesson step by step	A maths word problem in Arabic	A section of a CV	A function with Arabic comments	Followi forme rules
GPT-6 Sol	5.0	5.0	5.0	5.0	5.0	4.5	5.0	5.0	5.0	5.0	5.0
Qwen3 Coder 480B A35B	3.0	2.0	1.5	3.0	4.5	4.0	3.5	5.0	3.0	3.0	4.5
Qwen3.8 27B	4.5	4.5	2.5	5.0	3.5	4.5	4.5	5.0	4.5	3.5	5.0
Qwen3.8 Flash	4.5	4.0	3.5	5.0	5.0	5.0	4.5	5.0	4.5	3.5	5.0
MiMo-V2.6-Flash	4.0	4.5	3.0	5.0	4.5	3.0	2.0	5.0	3.5	4.0	3.0
MiMo-V2.6-Pro	4.0	4.0	3.0	5.0	4.5	3.5	3.0	5.0	3.0	2.0	5.0
GLM 5.3	4.0	4.5	4.5	5.0	4.0	5.0	5.0	5.0	4.0	5.0	4.5

Score: 1 2 3 4 5 Dashed outline: the judges differed by more than a point

Mechanical checks each model passed

Mechanical checks are counts, exact-text tests and running code, 42 in all across the twelve tasks for each model; no judge is involved.



How much the two judges agreed

Each cell counts the answers that the first judge scored as in the row and the second judge as in the column. The diagonal is exact agreement.

First judge's score	Second judge's score				
	1	2	3	4	5
1	3	1	0	0	0

First judge's score	Second judge's score				
	1	2	3	4	5
2	1	8	1	2	0
3	0	1	11	8	7
4	0	2	7	20	31
5	0	0	7	25	117

The judges gave the same score on 159 of 252 answers and differed by more than a point on 20 (dashed outline).

Scores by kind of task

Each cell shows the score out of 5 on top and, below it, how many of the mechanical checks passed. Models are in alphabetical order, not by score.

Writing in standard Arabic

Model	Formal letter in Modern Standard Arabic	Short report with correct percentages
Claude Opus 5.5	5.0 4 of 4	5.0 4 of 4
Claude Sonnet 5	4.5 4 of 4	4.0 4 of 4
DeepSeek V4.1 Flash	5.0 4 of 4	4.5 4 of 4
DeepSeek V4 Pro	5.0 4 of 4	4.0 3 of 4
Gemini 3.1 Pro	4.5 4 of 4	4.5 4 of 4
Gemini 3.8 Flash	4.5 4 of 4	5.0 4 of 4
Gemma 4 31B	4.5 4 of 4	5.0 4 of 4
Llama 3.1 8B Instruct	2.0 4 of 4	1.5 3 of 4
Llama 4 Scout	4.5 4 of 4	4.0 4 of 4
MiniMax M3	4.5 4 of 4	4.0 4 of 4
Mistral Small 4	3.5 4 of 4	4.5 4 of 4
Kimi K2.7 Code	5.0 4 of 4	5.0 4 of 4
Kimi K3	5.0 4 of 4	4.5 4 of 4
GPT-6 Luna	5.0 4 of 4	5.0 4 of 4
GPT-6 Sol	5.0 4 of 4	5.0 4 of 4
Qwen3 Coder 480B A35B	3.0 4 of 4	2.0 3 of 4
Qwen3.8 27B	4.5 4 of 4	4.5 4 of 4
Qwen3.8 Flash	4.5 4 of 4	4.0 4 of 4
MiMo-V2.6-Flash	4.0 4 of 4	4.5 4 of 4
MiMo-V2.6-Pro	4.0 4 of 4	4.0 4 of 4

Model	Formal letter in Modern Standard Arabic	Short report with correct percentages
<u>GLM 5.3</u>	4.0 4 of 4	4.5 4 of 4

Libyan dialect

Model	A question in Libyan dialect
Claude Opus 5.5	5.0 2 of 2
Claude Sonnet 5	4.5 2 of 2
<u>DeepSeek V4.1 Flash</u>	4.5 2 of 2
<u>DeepSeek V4 Pro</u>	3.3 2 of 2
Gemini 3.1 Pro	5.0 2 of 2
Gemini 3.8 Flash	4.0 2 of 2
<u>Gemma 4 31B</u>	4.5 2 of 2
<u>Llama 3.1 8B Instruct</u>	1.0 1 of 2
<u>Llama 4 Scout</u>	3.0 2 of 2
<u>MiniMax M3</u>	4.0 2 of 2
<u>Mistral Small 4</u>	2.5 2 of 2
<u>Kimi K2.7 Code</u>	3.0 2 of 2
<u>Kimi K3</u>	4.0 2 of 2
GPT-6 Luna	5.0 2 of 2
GPT-6 Sol	5.0 2 of 2
<u>Qwen3 Coder 480B A35B</u>	1.5 2 of 2
<u>Qwen3.8 27B</u>	2.5 2 of 2
<u>Qwen3.8 Flash</u>	3.5 2 of 2
<u>MiMo-V2.6-Flash</u>	3.0 2 of 2
<u>MiMo-V2.6-Pro</u>	3.0 2 of 2
<u>GLM 5.3</u>	4.5 2 of 2

Translation

Model	Arabic to English translation	English to Arabic translation
Claude Opus 5.5	5.0 4 of 4	5.0 3 of 3
Claude Sonnet 5	5.0 4 of 4	5.0 3 of 3
<u>DeepSeek V4.1 Flash</u>	5.0 4 of 4	5.0 3 of 3
<u>DeepSeek V4 Pro</u>	5.0 4 of 4	5.0 3 of 3

Model	Arabic to English translation	English to Arabic translation
Gemini 3.1 Pro	5.0 4 of 4	5.0 3 of 3
Gemini 3.8 Flash	5.0 4 of 4	5.0 3 of 3
Gemma 4 31B	5.0 4 of 4	5.0 3 of 3
Llama 3.1 8B Instruct	3.5 4 of 4	2.0 3 of 3
Llama 4 Scout	5.0 4 of 4	4.3 3 of 3
MiniMax M3	5.0 4 of 4	5.0 1 of 3
Mistral Small 4	5.0 4 of 4	5.0 3 of 3
Kimi K2.7 Code	5.0 4 of 4	5.0 3 of 3
Kimi K3	5.0 4 of 4	5.0 3 of 3
GPT-6 Luna	5.0 4 of 4	5.0 3 of 3
GPT-6 Sol	5.0 4 of 4	5.0 3 of 3
Qwen3 Coder 480B A35B	3.0 4 of 4	4.5 3 of 3
Qwen3.8 27B	5.0 4 of 4	3.5 3 of 3
Qwen3.8 Flash	5.0 4 of 4	5.0 3 of 3
MiMo-V2.6-Flash	5.0 4 of 4	4.5 3 of 3
MiMo-V2.6-Pro	5.0 4 of 4	4.5 3 of 3
GLM 5.3	5.0 4 of 4	4.0 3 of 3

Reading and summarising

Model	Summarising an Arabic text
Claude Opus 5.5	4.5 3 of 3
Claude Sonnet 5	5.0 3 of 3
DeepSeek V4.1 Flash	4.0 3 of 3
DeepSeek V4 Pro	5.0 3 of 3
Gemini 3.1 Pro	4.5 3 of 3
Gemini 3.8 Flash	5.0 3 of 3
Gemma 4 31B	3.0 3 of 3
Llama 3.1 8B Instruct	1.0 2 of 3
Llama 4 Scout	3.0 2 of 3
MiniMax M3	4.0 3 of 3
Mistral Small 4	4.0 3 of 3
Kimi K2.7 Code	4.5 3 of 3

Model	Summarising an Arabic text
Kimi K3	5.0 3 of 3
GPT-6 Luna	4.5 3 of 3
GPT-6 Sol	4.5 3 of 3
Qwen3 Coder 480B A35B	4.0 3 of 3
Qwen3.8 27B	4.5 3 of 3
Qwen3.8 Flash	5.0 3 of 3
MiMo-V2.6-Flash	3.0 3 of 3
MiMo-V2.6-Pro	3.5 3 of 3
GLM 5.3	5.0 3 of 3

Study and arithmetic

Model	Explaining a lesson step by step	A maths word problem in Arabic
Claude Opus 5.5	4.5 3 of 3	5.0 2 of 2
Claude Sonnet 5	4.5 3 of 3	5.0 2 of 2
DeepSeek V4.1 Flash	3.5 2 of 3	5.0 2 of 2
DeepSeek V4 Pro	5.0 2 of 3	5.0 2 of 2
Gemini 3.1 Pro	5.0 2 of 3	5.0 2 of 2
Gemini 3.8 Flash	4.5 3 of 3	5.0 2 of 2
Gemma 4 31B	3.5 3 of 3	5.0 2 of 2
Llama 3.1 8B Instruct	1.0 2 of 3	4.0 2 of 2
Llama 4 Scout	4.0 2 of 3	5.0 2 of 2
MiniMax M3	4.0 1 of 3	4.5 2 of 2
Mistral Small 4	4.5 3 of 3	5.0 2 of 2
Kimi K2.7 Code	4.5 3 of 3	5.0 2 of 2
Kimi K3	4.5 3 of 3	5.0 2 of 2
GPT-6 Luna	4.0 3 of 3	5.0 2 of 2
GPT-6 Sol	5.0 3 of 3	5.0 2 of 2
Qwen3 Coder 480B A35B	3.5 2 of 3	5.0 2 of 2
Qwen3.8 27B	4.5 2 of 3	5.0 2 of 2
Qwen3.8 Flash	4.5 2 of 3	5.0 2 of 2
MiMo-V2.6-Flash	2.0 3 of 3	5.0 2 of 2
MiMo-V2.6-Pro	3.0 3 of 3	5.0 2 of 2

Model	Explaining a lesson step by step	A maths word problem in Arabic
GLM 5.3	5.0 2 of 3	5.0 2 of 2

Work and CVs

Model	A section of a CV
Claude Opus 5.5	5.0 4 of 4
Claude Sonnet 5	4.5 4 of 4
DeepSeek V4.1 Flash	5.0 4 of 4
DeepSeek V4 Pro	5.0 4 of 4
Gemini 3.1 Pro	5.0 4 of 4
Gemini 3.8 Flash	5.0 3 of 4
Gemma 4 31B	5.0 4 of 4
Llama 3.1 8B Instruct	4.5 4 of 4
Llama 4 Scout	4.0 3 of 4
MiniMax M3	5.0 4 of 4
Mistral Small 4	4.0 3 of 4
Kimi K2.7 Code	5.0 4 of 4
Kimi K3	5.0 4 of 4
GPT-6 Luna	5.0 4 of 4
GPT-6 Sol	5.0 4 of 4
Qwen3 Coder 480B A35B	3.0 4 of 4
Qwen3.8 27B	4.5 4 of 4
Qwen3.8 Flash	4.5 4 of 4
MiMo-V2.6-Flash	3.5 3 of 4
MiMo-V2.6-Pro	3.0 4 of 4
GLM 5.3	4.0 4 of 4

Code

Model	A function with Arabic comments
Claude Opus 5.5	4.5 3 of 3
Claude Sonnet 5	4.0 3 of 3
DeepSeek V4.1 Flash	4.0 3 of 3
DeepSeek V4 Pro	4.5 3 of 3

Model	A function with Arabic comments
Gemini 3.1 Pro	5.0 3 of 3
Gemini 3.8 Flash	5.0 3 of 3
Gemma 4 31B	3.5 3 of 3
Llama 3.1 8B Instruct	2.0 2 of 3
Llama 4 Scout	4.0 3 of 3
MiniMax M3	4.5 3 of 3
Mistral Small 4	4.0 3 of 3
Kimi K2.7 Code	4.5 3 of 3
Kimi K3	4.0 3 of 3
GPT-6 Luna	5.0 3 of 3
GPT-6 Sol	5.0 3 of 3
Qwen3 Coder 480B A35B	3.0 3 of 3
Qwen3.8 27B	3.5 3 of 3
Qwen3.8 Flash	3.5 3 of 3
MiMo-V2.6-Flash	4.0 3 of 3
MiMo-V2.6-Pro	2.0 3 of 3
GLM 5.3	5.0 3 of 3

Format and text direction

Model	Following format rules	Mixed Arabic, English and numbers
Claude Opus 5.5	5.0 5 of 5	5.0 5 of 5
Claude Sonnet 5	5.0 5 of 5	4.0 5 of 5
DeepSeek V4.1 Flash	5.0 5 of 5	5.0 5 of 5
DeepSeek V4 Pro	5.0 5 of 5	5.0 5 of 5
Gemini 3.1 Pro	4.0 5 of 5	4.5 5 of 5
Gemini 3.8 Flash	4.5 5 of 5	5.0 5 of 5
Gemma 4 31B	5.0 5 of 5	3.5 5 of 5
Llama 3.1 8B Instruct	2.0 4 of 5	2.0 4 of 5
Llama 4 Scout	5.0 5 of 5	4.5 5 of 5
MiniMax M3	4.0 2 of 5	3.0 5 of 5
Mistral Small 4	3.5 4 of 5	4.5 5 of 5
Kimi K2.7 Code	5.0 5 of 5	5.0 5 of 5

Model	Following format rules	Mixed Arabic, English and numbers
Kimi K3	5.0 5 of 5	5.0 5 of 5
GPT-6 Luna	5.0 5 of 5	5.0 5 of 5
GPT-6 Sol	5.0 5 of 5	5.0 5 of 5
Qwen3 Coder 480B A35B	4.5 5 of 5	4.5 5 of 5
Qwen3.8 27B	5.0 5 of 5	4.5 5 of 5
Qwen3.8 Flash	5.0 5 of 5	5.0 5 of 5
MiMo-V2.6-Flash	3.0 5 of 5	4.0 5 of 5
MiMo-V2.6-Pro	5.0 5 of 5	4.5 5 of 5
GLM 5.3	4.5 5 of 5	4.5 5 of 5

How to cite this test

مختبر أكاكوس (2026). اختبار أكاكوس العربي، الإصدار 1.1. شفرة التكوين. <https://acacus.ly/research/arabic-test-2026-10>

Acacus Lab (2026). Acacus Arabic Test, v1.1. Shafra Al-Takween. <https://acacus.ly/research/arabic-test-2026-10>