MusIML @ NeurIPS 2026

When Models Misread Ink

Auditing vision-language models on Arabic calligraphy recognition

Niaz Mohaiman Abtahi · Evan Maksud · Reyana Shahid

Dataset and Models

Images

1,195

كوفي Kufic · 106	نسخ Naskh · 218	ثلث Thuluth · 216
ديواني Diwani · 217	رقعة Ruq’ah · 225	نستعليق Nasta’liq · 213

Models

12

4 closed (API)

GPT-4o · GPT-4.1 · GPT-5.5 · GPT-5.6 Terra

8 open-weight

Qwen2.5-VL-7B · Qwen2.5-VL-32B · Qwen3-VL-8B · Qwen3-VL-32B · InternVL3.5-8B · InternVL3.5-38B · MiniCPM-V 4.5 · Gemma 4 12B-IT

Prompt Description

Numeric

Sees: the six styles, numbered 1 to 6.

Answers: one digit.

Visual-cue

Sees: the same list, plus a hint for each style, such as “Diwani: dense, ornamental curves”.

Answers: one digit.

Diwani-first (diagnostic)

Sees: the hints, with Diwani listed first.

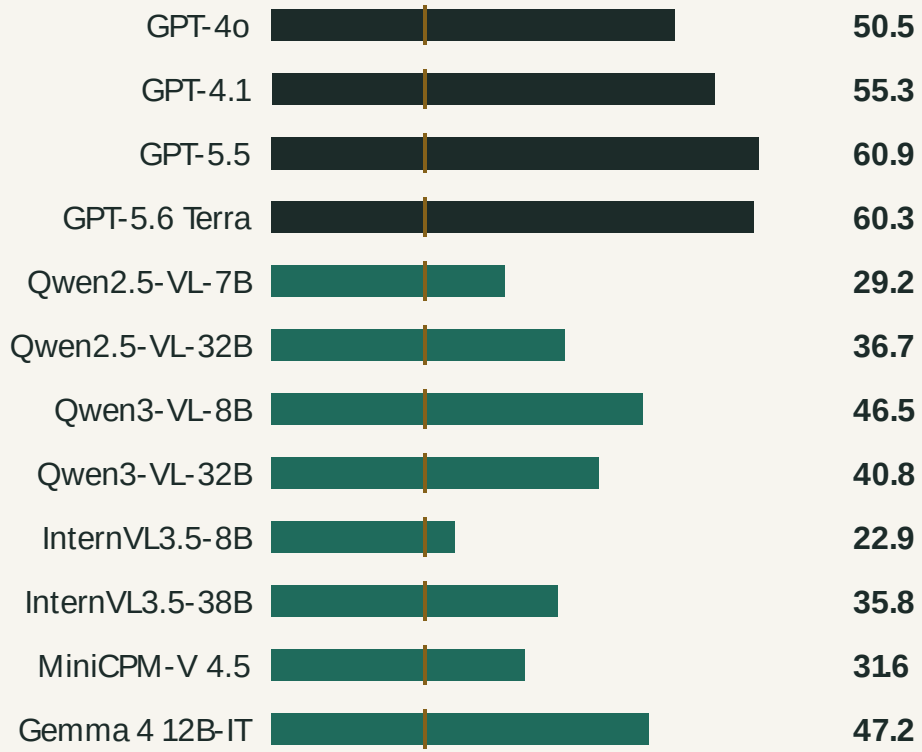
Answers: the style name.

Example: the Numeric prompt (shortened)

“You are shown an Arabic calligraphy image. Below are six possible styles, each with a number: 1. Kufic, 2. Naskh, 3. Thuluth, 4. Diwani, 5. Ruq’ah, 6. Nasta’liq. Respond with only the number of the style that best matches this image.”

Model Performance

How accurate? | Where do errors go? | Can prompts fix it?



60.9%
Best closed model: GPT-5.5

47.2%
Best open-weight model: Gemma 4 12B-IT

18.8%
Always guessing Ruq’ah, the most common style

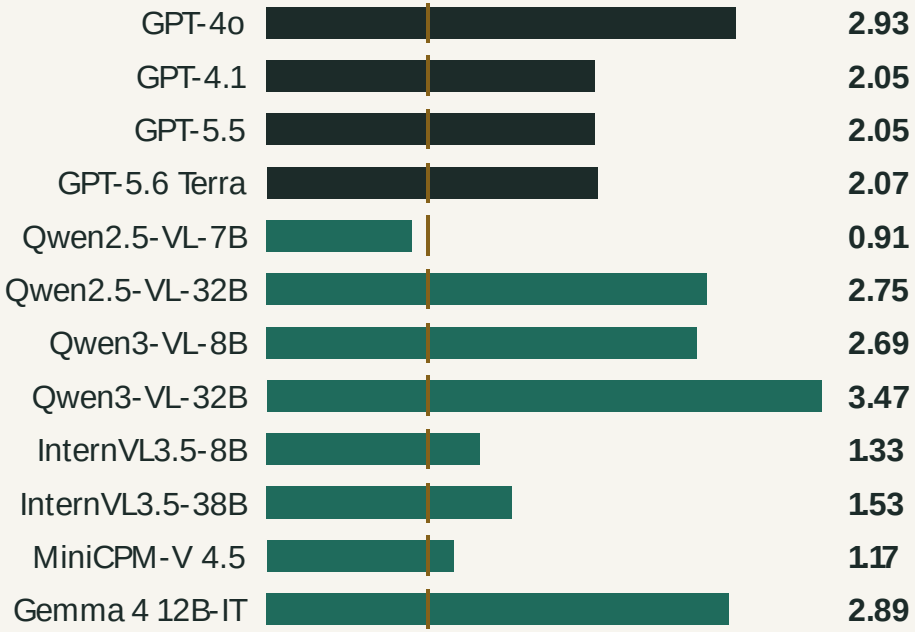
23.0%
Diwani found on average, the most-missed style

Where Errors Go

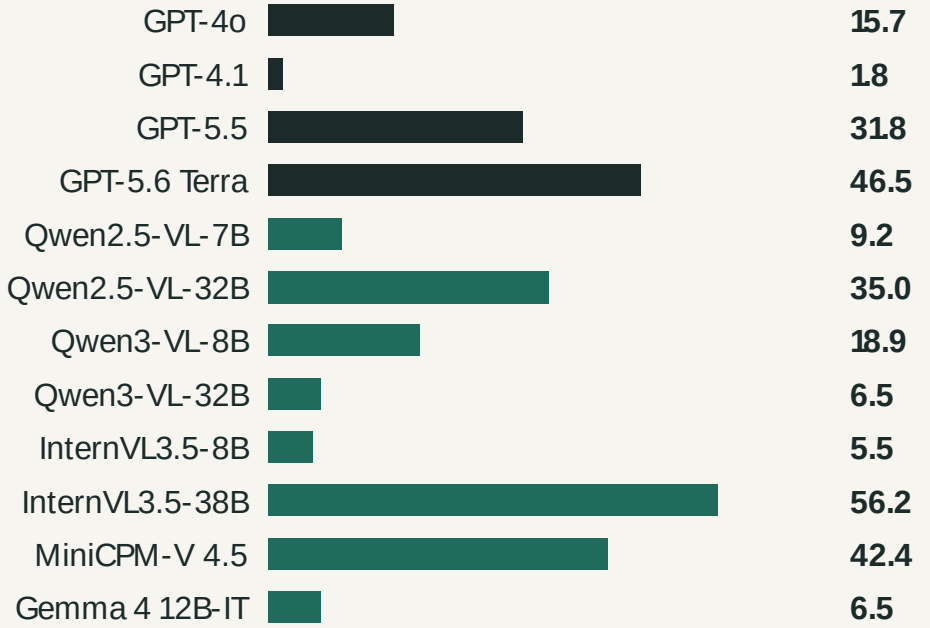
How accurate? | **Where do errors go?** | Can prompts fix it?

The errors follow a pattern, so the paper tracks two signals: Thuluth over-prediction and Diwani recall. GPT-4o calls 52.9% of all images Thuluth; only 18.1% truly are.

Thuluth: times predicted ÷ true count (gold line = 1.00, the right amount)



Diwani images found (recall, %)

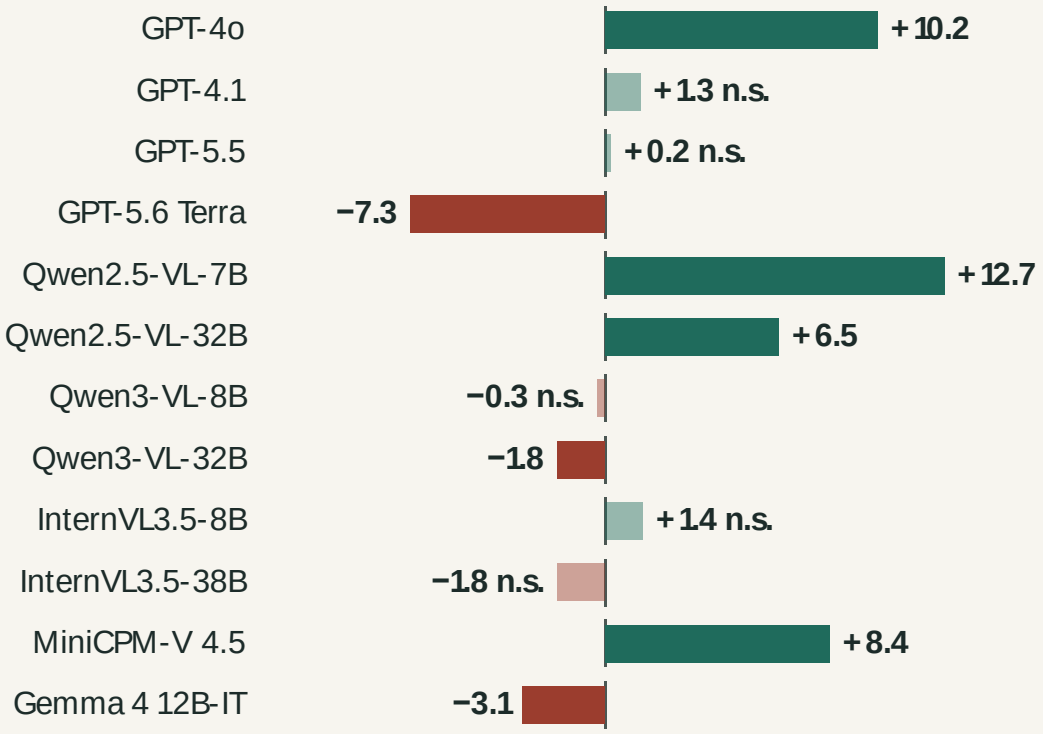


Prompt Changing

How accurate? | Where do errors go? | **Can prompts fix it?**

Not reliably. Adding the style hints from slide 3 raised accuracy for some models and lowered it for others.

Accuracy change from adding style descriptions (points)



Colour key

Green: accuracy goes up

Orange-red: accuracy goes down

Faded: change is not statistically significant (n.s.), so no direction can be claimed

Each bar is the change in accuracy, in points, compared with the numeric prompt. Right of the line is up, left is down.

Takeaways and Limits

Overall accuracy hides which calligraphy traditions go unseen.

What tools should do

- **Report accuracy for each style,** not only one overall number.
- **Show uncertainty** when a tool labels or searches calligraphy.
- **Do not rely on prompt descriptions alone.** Test the exact model you plan to use.
- **Why it matters:** museums, archives, teaching and search all use these labels. A Diwani piece called Thuluth becomes harder to find.

What this study does not cover

- Six broad styles only: no regional schools, mixed styles, dating or handwriting.
- Some images come from other sources, so style and source are partly mixed.
- The Diwani-first prompt changes label order and answer format together.
- Closed models are served through an API and can change over time.