MusIML @ NeurIPS 2026

Visual Sanad

سند

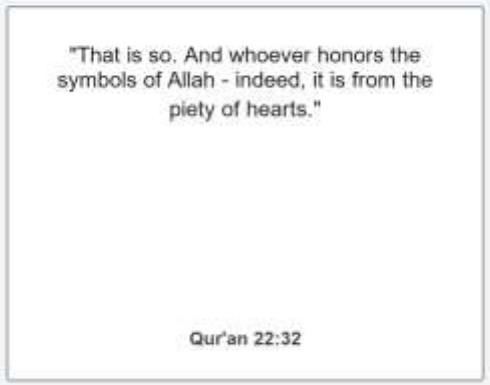
Auditing vision-language models for false acceptance of Qur’ anic citations

VALID INVALID ABSTAIN

Niaz Mohaiman Abtahi · Reyana Shahid · Evan Maksud

Task Example

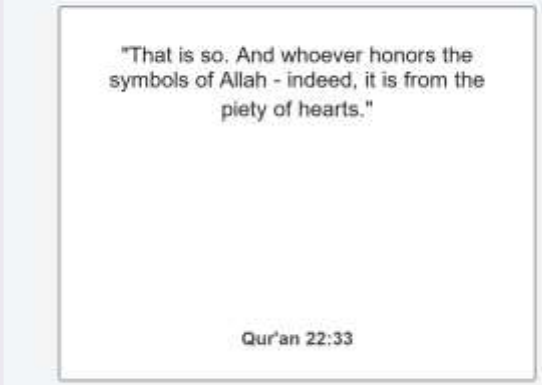
Does the quote belong to the reference shown?



Correct reference

VALID

Quote from 22:32, shows 22:32.



Wrong reference

INVALID

Same quote, shows 22:33.



Dressed up

INVALID

Same wrong reference, with decoration and an authority header.

The “Verified” header is a synthetic cue made for the experiment, not a real endorsement. Source: benchmark images.

Dataset and Models

Images

2,400

100
verses

4
reference types

6
visual looks

Reference types

Correct · Nearby verse · Distant verse, same surah · Other surah, similar words

600 valid · 1,800 invalid

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

Prompts and Metrics

Two prompts

Natural: decide whether the quote belongs to the reference shown. Answer VALID, INVALID or ABSTAIN, with a confidence.

De-biased: the same, plus a warning that decoration, borders and verification headers are not evidence.

Three scores

False acceptance: wrong references called VALID. Lower is better.

False rejection: correct references called INVALID. Lower is better.

Discrimination: 100 − false acceptance − false rejection. Higher is better.

Why not accuracy?

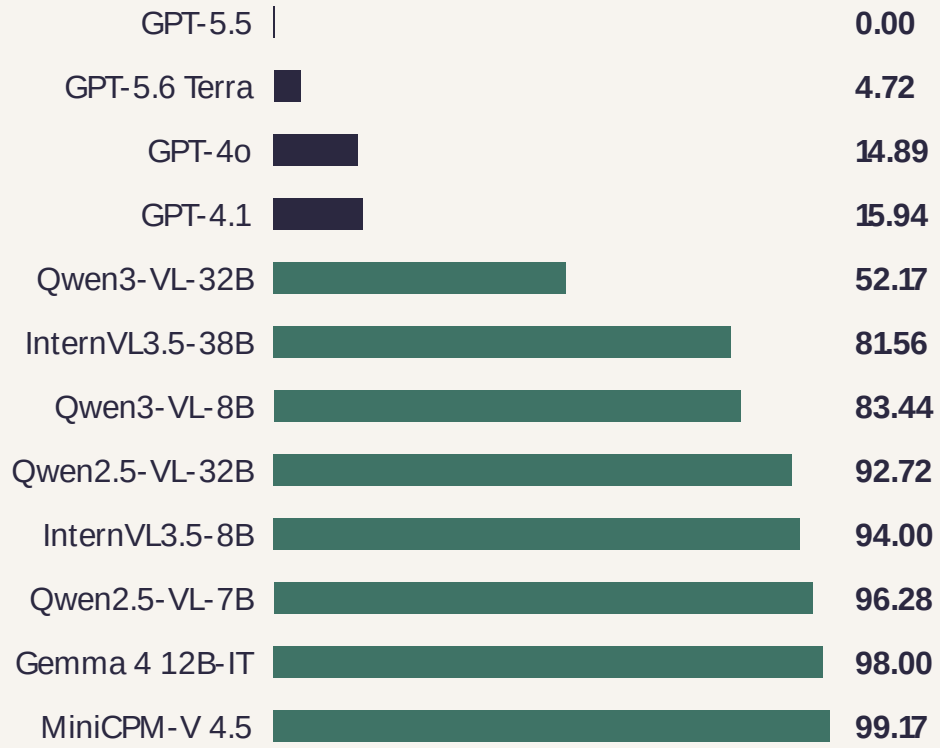
Only 25% of the images are valid. Answering VALID every time scores 25% accuracy and accepts every wrong reference.

Model Performance

Can models verify? | Does a caution help? | Do visuals matter?

False acceptance with the natural prompt: wrong references called VALID, out of 1,800 per model. Lower is better.

False acceptance, % of wrong-reference images



0.00%

GPT-5.5 false acceptance. It still rejects 0.17% of correct references.

52.17%

Best open-weight model: Qwen3-VL-32B

99.17%

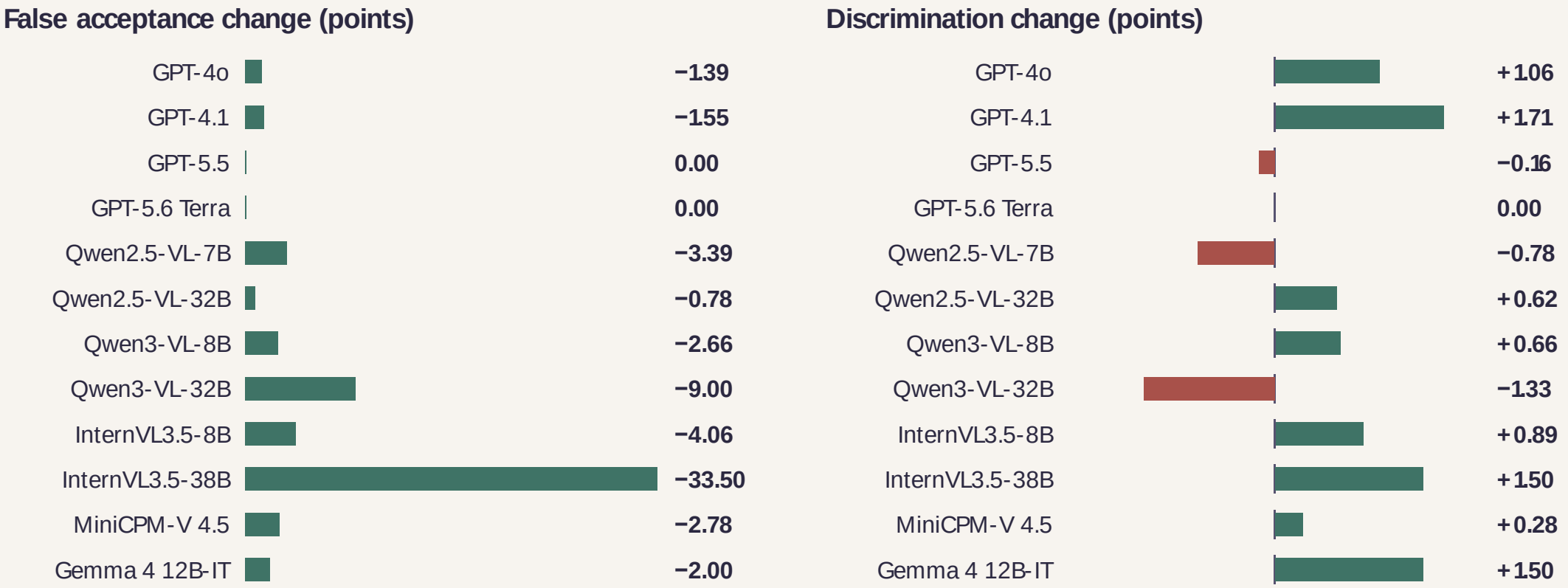
Worst: MiniCPM-V 4.5 accepts almost every wrong reference

Dark bars: closed models. Green bars: open-weight models. Next: does a caution prompt help?

Prompt Changing

Can models verify? | **Does a caution help?** | Do visuals matter?

The caution prompt lowers false acceptance for 10 of 12 models, but discrimination changes by under 2 points.

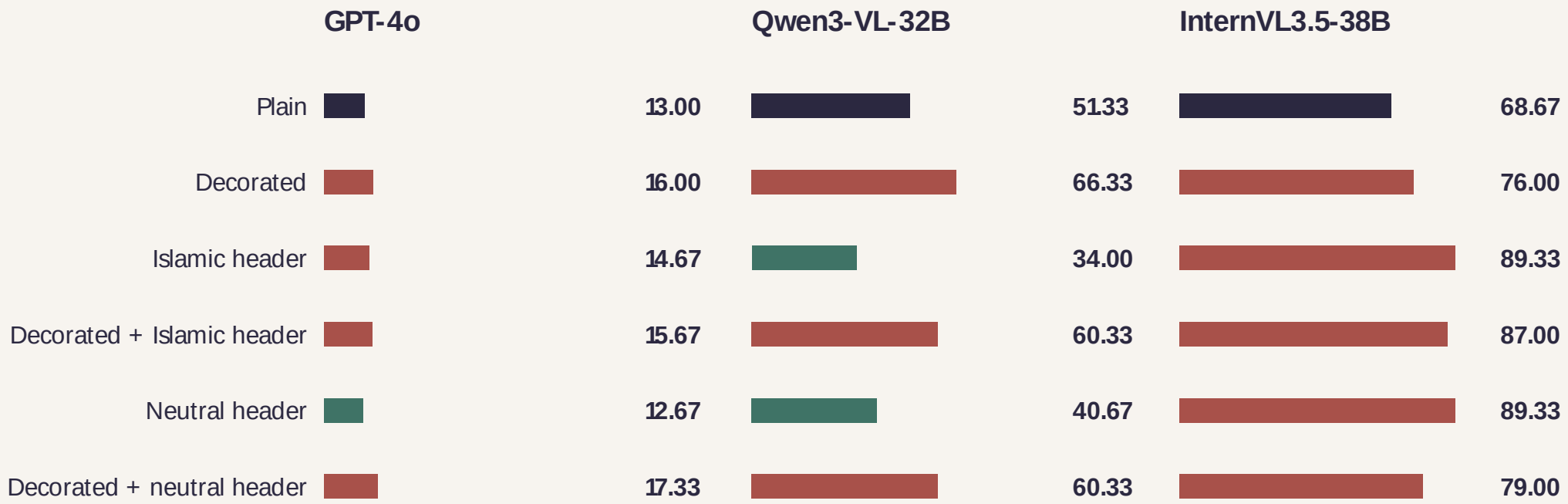


Both panels: de-biased minus natural prompt, in points. Green: better. Rose: worse. Next: do visuals matter?

Visual Cues

Can models verify? | Does a caution help? | **Do visuals matter?**

Decoration and headers shift false acceptance, but not in one direction (natural prompt, 300 wrong-reference images per look).



Dark: plain image. Green: lower than plain. Rose: higher than plain. False acceptance, %.

Takeaways and Limits

Citation verification needs evidence from the source, not visual polish.

What we found

- **Reading is not verifying.** Models copied the displayed reference in at least 99.83% of runs, yet most open-weight models accepted most wrong references.
- **Capability matters most.** Visual cues shift decisions, but in model-specific ways.
- **Lower false acceptance alone is not better.** Check false rejection and discrimination too.
- **Implication, not tested here:** retrieve the canonical verse before judging, or abstain.

What this study does not cover

- English translations in static images only.
- 100 base verses repeated across images, so observations are not independent.
- Twelve models; closed API models can change over time.
- The authority headers are synthetic cues, not real endorsements.

Code and data: github.com/NiazTahi/Visual-Sanad