The Impact of Stated Source Reliability on Face Verification with Multimodal Large Language Models

Artūras Nakvosas Virginijus Marcinkevičius
Institute of Data Science and Digital Technologies, Vilnius University arturas.nakvosas@mif.stud.vu.lt

None of the tested Multi-modal Large Language Models consistently weights textual claims according to the stated reliability of their source; how the source is described and a shift shared by opposing claims also shape their decisions.

Question and method

Our previous work examined adversarial and cooperative context effects in MLLM face verification [1]. Here we ask: **do MLLMs weight a textual claim by the stated reliability of its source?**

- **Data and models:** 300 LFW pairs (150 genuine, 150 impostor), 3 verification prompts; Gemma-3-12B, Qwen3-VL-8B, InternVL3.5-8B. 200 further pairs set the temperature and the security thresholds.
- **Claims** ("same person"/"different people", $r = N/100 \in \{.5, .6, .75, .9, .95, .99\}$):

witness	*"A witness who is right about N% of the time says the two photos show the same person."*
matcher	*"An automated face matcher with N% accuracy reports that these images show different people."*
source tag	*"Source reliability: N%. Statement: the same person."*
analyst	*"An analyst whose past calls were correct in N% of cases concluded: different people."*

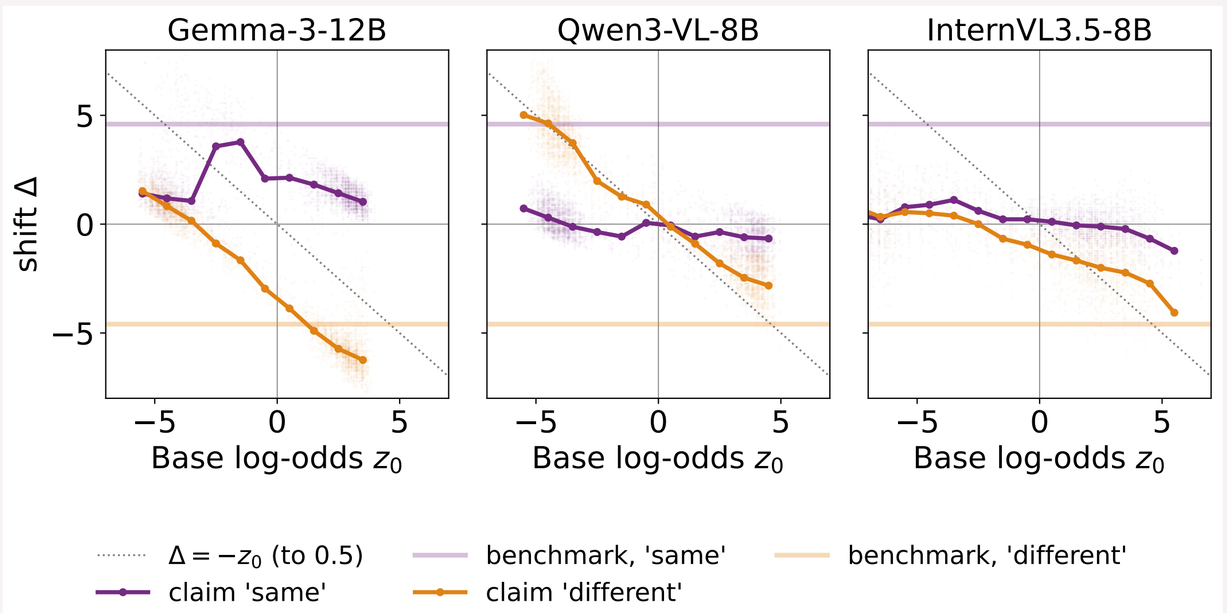
- **Query:** claim + images + *"Are these two images showing the same person? Answer only 'yes' or 'no'."*
- **Paired design:** every pair, prompt, description and r is queried with *both* claims and without a claim (image-only log-odds z_0).

With $\Delta = z_1 - z_0$ the shift of the temperature-scaled log-odds $z = \log p(\text{yes}) - \log p(\text{no})$:

$$\text{claim effect} = \tfrac{1}{2}[\Delta(\text{same}) - \Delta(\text{diff})], \qquad \text{common part} = \tfrac{1}{2}[\Delta(\text{same}) + \Delta(\text{diff})].$$

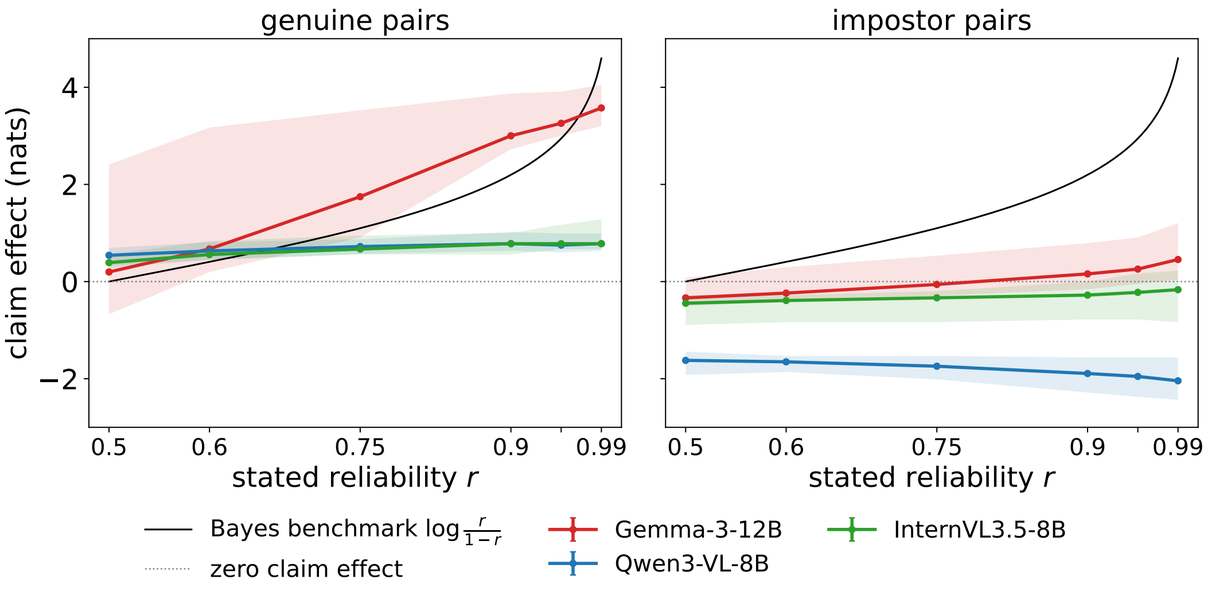
Bayesian benchmark: if the source is equally reliable on genuine and impostor pairs and independent of the images, claim effect $= \log \frac{r}{1-r}$ and common part $= 0$.

A shift shared by opposing claims ($r = 0.99$)



Benchmark: flat horizontal lines. In all three models the common part (the mean of the two curves) pulls decisions toward 0.5 (dotted line). **Control:** a neutral sentence does not reproduce it, whereas a source-attributed statement that makes no identity claim reproduces much of it in Qwen3-VL and InternVL3.5.

Claim effect vs. benchmark (four source descriptions)



Median with 95 % bootstrap intervals; bands: range across source descriptions. **Gemma-3** rises with r on genuine pairs only; **Qwen3-VL** is nearly flat and negative on impostor pairs; **InternVL3.5** stays small.

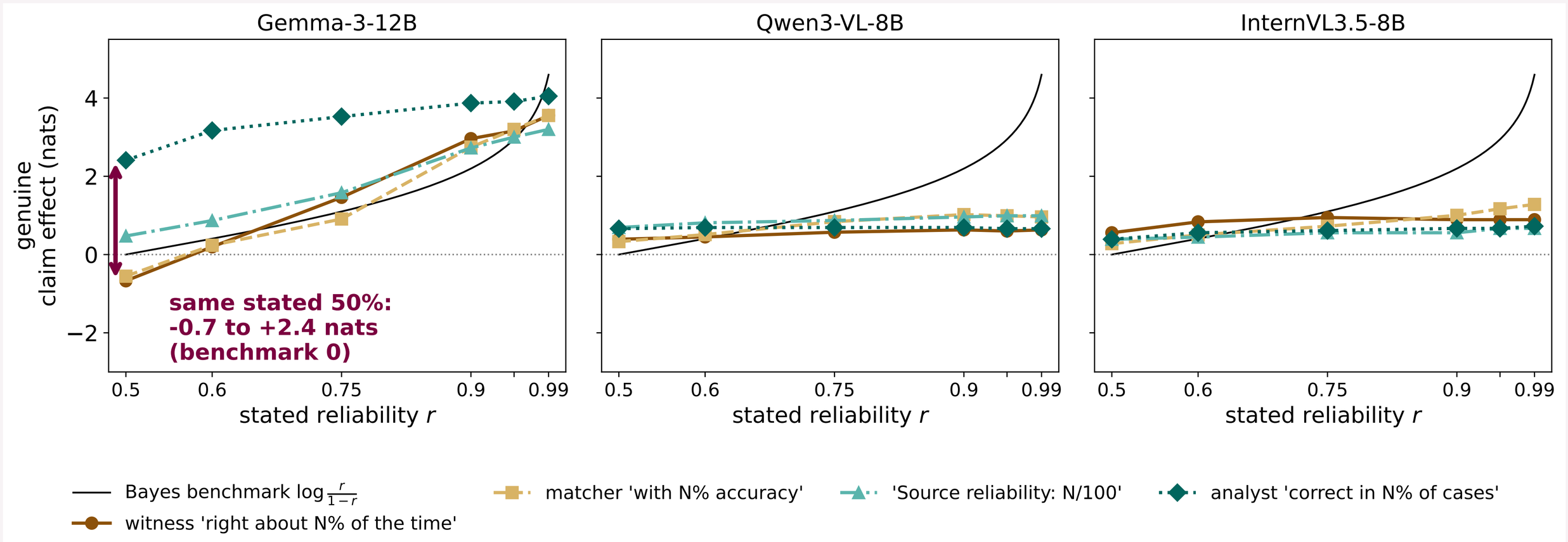
Security view: false "same" claims

Measured: FAR, the % of the 150 impostor pairs accepted ($z \geq t$, z = yes/no log-odds). **Threshold t:** one per row, set on image-only answers of 200 separate held-out pairs at FAR 1 % (strict) or 10 % (lenient), then kept fixed while a false "same person" claim is added. **Threshold check:** near-target image-only values show that t transfers to unseen pairs. Cells: FAR [95 % bootstrap interval, %].

threshold		image	+ false "same" claim, $r =$		
(FAR on images)		only	50 %	90 %	99 %
strict, 1 %	Gemma-3	1.6 % [0.0, 4.0]	0.7 % [0.1, 2.5]	9.4 % [5.7, 15.1]	15.8 % [9.9, 23.0]
	Qwen3-VL	1.1 % [0.0, 4.7]	23.7 % [19.1, 32.7]	8.6 % [2.5, 21.6]	2.5 % [0.6, 13.6]
	InternVL3.5	0.2 % [0.0, 6.0]	0.2 % [0.0, 7.3]	0.7 % [0.0, 7.6]	0.8 % [0.0, 8.1]
lenient, 10 %	Gemma-3	12.0 % [4.0, 24.0]	5.2 % [2.5, 12.8]	20.2 % [13.9, 31.7]	27.4 % [20.3, 38.6]
	Qwen3-VL	10.0 % [4.9, 16.2]	44.4 % [34.6, 56.6]	27.8 % [22.5, 34.4]	23.5 % [15.9, 31.2]
	InternVL3.5	8.9 % [4.4, 15.6]	11.5 % [5.7, 19.5]	12.4 % [6.2, 21.6]	12.9 % [6.4, 22.4]

Gemma-3's FAR rises with r; Qwen3-VL's is highest at the lowest stated reliability. For InternVL3.5, changes in error rates are not consistently supported across the two thresholds, given bootstrap uncertainty.

Same stated reliability, different source descriptions: claim effect on genuine pairs



Conclusion

At the same stated reliability, the source description alone moved Gemma-3's claim effect on genuine pairs from -0.7 to $+2.4$ nats at 50 % (benchmark: 0); Qwen3-VL and InternVL3.5 barely respond to r, and in all three models opposing claims share a shift toward 0.5. False "same" claims raised the share of accepted impostors from 1–2 % to as much as 16 % (Gemma-3, 99 %) and 24 % (Qwen3-VL, 50 %) at a 1 % target FAR. The paired protocol, which separates the claim effect from the common part and compares it with an explicit benchmark, provides a basis for evaluating progress toward calibrated evidence use. Generalization beyond these three models and LFW face verification remains untested.

[1] A. Nakvosas, P. Daniušis, V. Marcinkevičius, "Benchmarking multimodal large language models for face verification under adversarial and cooperative contexts," IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3753521.