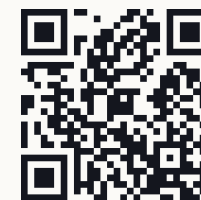


Token-Set Choice Confounds POPE

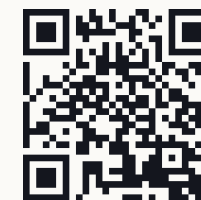
A Systematic Audit of Yes/No Extraction in VLM Hallucination Evaluation

Kesav Kumar Jayakumar¹ · Karthigeyan Thilak²

¹Sri Krishna College of Technology, Coimbatore · ²Sri Eshwar College of Engineering, Coimbatore



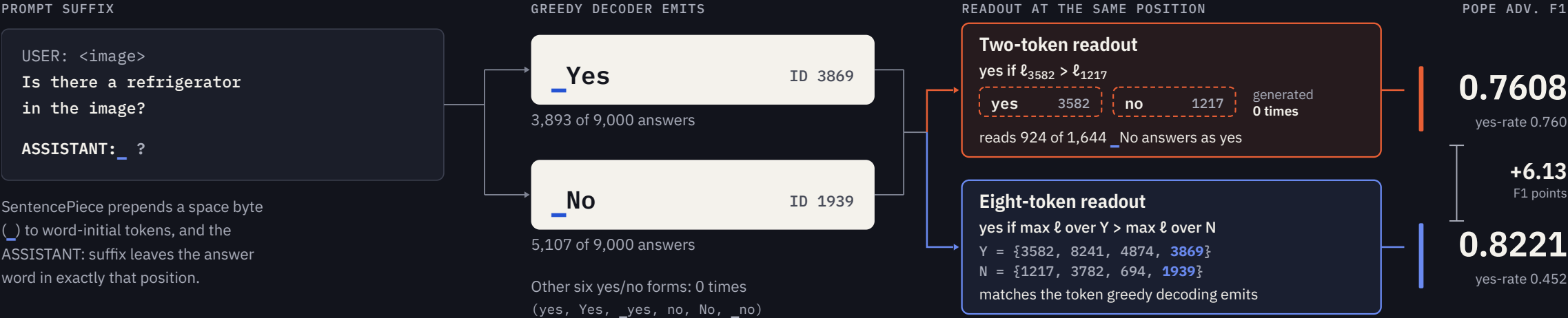
Paper
arXiv:2610.25964



Code and data
ugaa-research

The readout compared two tokens the model never generates.

A two-token POPE readout compares the logits of **yes (3582)** and **no (1217)**. In 9,000 greedy answers LLaVA-1.5-7B generated neither ID once; it always emits **_Yes (3869)** or **_No (1939)**.



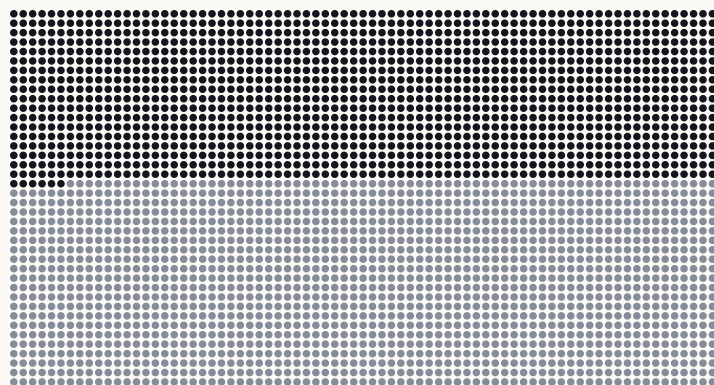
0 of 9,000

answers have ID 3582 or 1217 as the generated token

Two IDs carry every answer: Yes 3869 and No 1939; the other six yes/no forms also appear zero times. SentencePiece prepends a space byte to word-initial tokens, and the ASSISTANT: suffix leaves the answer word exactly there.

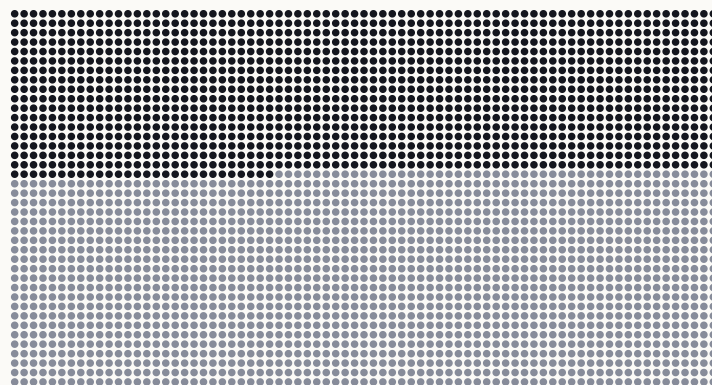
Adversarial

1,356 Yes · 1,644 No



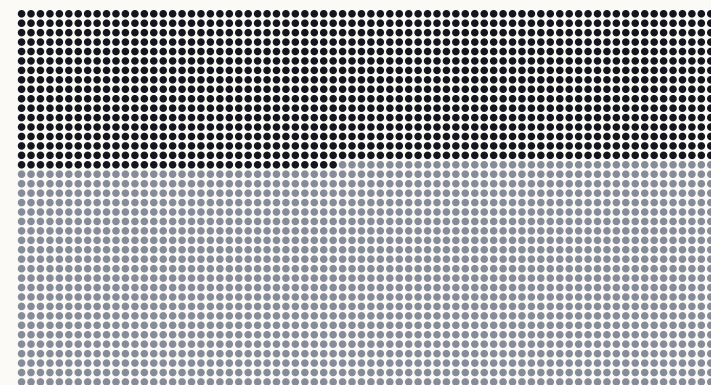
Popular

1,303 Yes · 1,697 No



Random

1,234 Yes · 1,766 No

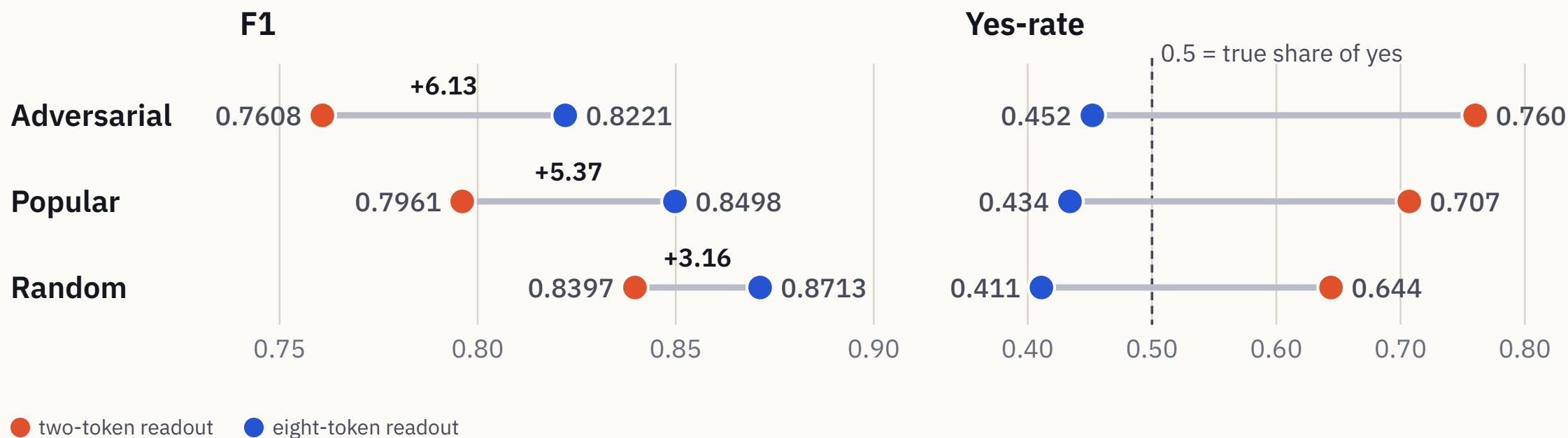


Yes (3869)
 No (1939)
 yes (3582), no (1217): 0 dots

points on POPE adversarial, with nothing in the model's behaviour changed

+6.13 F1

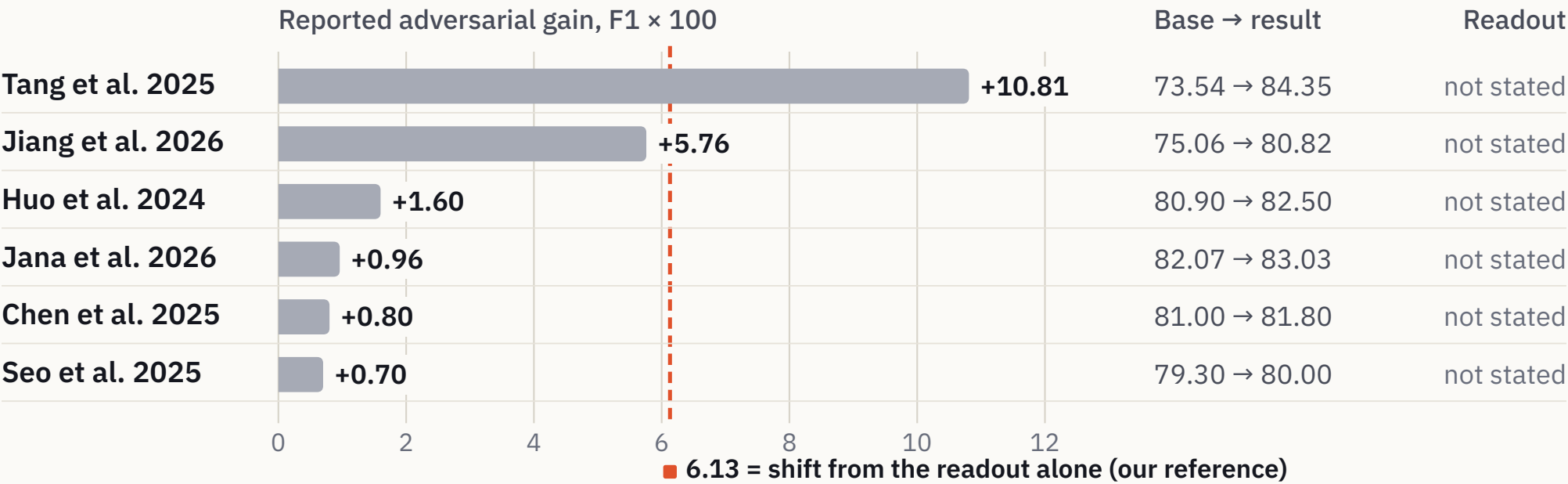
The two-token rule reads **924 of the 1,644** No answers as yes: 660 true negatives become false positives and 264 false negatives become true positives. Recall climbs to 0.9587 while precision falls to 0.6307.



5 of 6

surveyed papers report a smaller adversarial gain than the 6.13-point shift. None states its readout.

Each gain is measured against a baseline from the same paper. Our 82.21 reference ran at 4-bit NF4, so we compute no corrected gains and show no method is broken.



0.00 F1

when the LLaVA-1.5 IDs are ported to the Mistral and Qwen vocabularies

A readout derived from each model's own tokenizer holds between **0.80 and 0.92** on every model and split. The confound belongs to the protocol, not to one model.

Model tokenizer	legacy_2tok LLaVA-1.5 IDs, 2 tokens	legacy_8tok LLaVA-1.5 IDs, 8 tokens	dynamic_single IDs from own tokenizer	string_parse parse decoded text
LLaVA-1.5-7B LLaMA-2 · reference	0.7608 0.7961 · 0.8397	0.8221 0.8498 · 0.8713	0.8221 0.8498 · 0.8713	0.8221 0.8498 · 0.8713
LLaVA-1.6-Mistral-7B Mistral	0.6667 0.6667 · 0.6667	0.0000 0.0000 · 0.0000	0.8521 0.8953 · 0.9163	0.8521 0.8953 · 0.9163
InstructBLIP-Vicuna-7B LLaMA	0.7074 0.7277 · 0.7613	0.8183 0.8553 · 0.8815	0.8183 0.8553 · 0.8815	0.6667 0.6667 · 0.6667
mPLUG-Owl2-LLaMA2-7B LLaMA-2	0.7754 0.8060 · 0.8466	0.8001 0.8312 · 0.8707	0.8001 0.8312 · 0.8707	0.8001 0.8312 · 0.8707
Qwen2-VL-7B-Instruct Qwen	0.6674 0.6671 · 0.6667	0.0000 0.0000 · 0.0000	0.8457 0.8578 · 0.8671	0.8457 0.8578 · 0.8671

Large: adversarial · small: popular · random 0.00: every answer read as no ≈0.667: at the all-yes floor

Derive the yes/no IDs from the tokenizer

```
tok = processor.tokenizer
enc = lambda s: tok.encode(s, add_special_tokens=False)
yes_forms = ["yes", "Yes", " yes", " Yes"]
no_forms = ["no", "No", " no", " No"]
yes_ids = [enc(s)[0] for s in yes_forms if len(enc(s)) == 1]
no_ids = [enc(s)[0] for s in no_forms if len(enc(s)) == 1]
out = model.generate(**inputs, max_new_tokens=1,
                    do_sample=False,
                    return_dict_in_generate=True,
                    output_scores=True)

logit = out.scores[0][0]
y = max(logit[i] for i in yes_ids)
n = max(logit[j] for j in no_ids)
pred = "yes" if y > n else "no"
```

- 1 **Log the generated token ID and its string** for every question.
- 2 **Verify the yes/no token sets** under the exact prompt template.
- 3 **Prefer the tokenizer-derived logit readout.** String parsing matches it only when it reads the first generated token.
- 4 **Release prediction-level records** so any aggregate can be recomputed.

SECONDARY FINDINGS

Why logit-level fixes stall on adversarial

Nine inference-time corrections of our own, six grounding-based and three contrastive. None beats the 0.8221 baseline; the closest, CLIP-B/32 certainty at $\beta = 1.0$, sits 0.005 below it.

Wrong answers draw more image attention than correct ones: 0.0801 vs 0.0738 (layers 14 to 20, $n = 510$ vs 2,490), Cohen's $d = -0.46$, $p = 1.4 \times 10^{-24}$. Grounding gates assume the opposite.

Eight-token baseline

CLIP-B/32 certainty ($\beta = 1.0$)

VCD-noise perturbation

Object-noun CLIP

Entropy gating

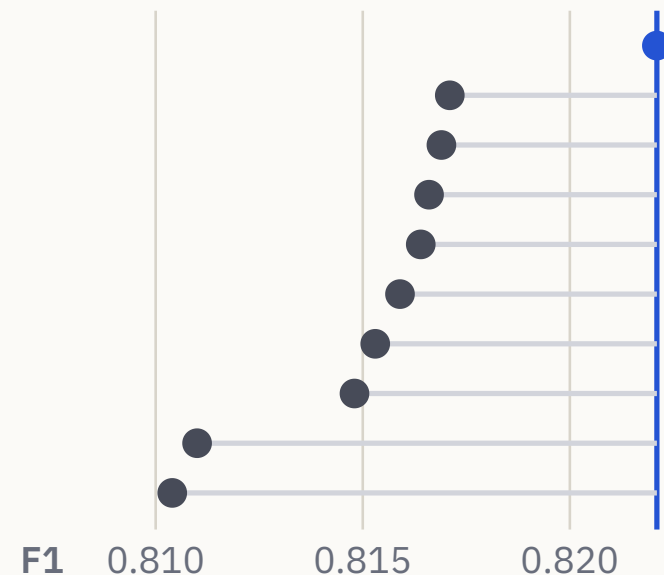
CLIP-L/14-336 certainty

CLS feature similarity

Magnitude gating

Uniform suppression

CLIP-B/32 certainty ($\beta = 1.5$)



SUMMARY

State the readout. Log the token.

The 6.13-point shift comes from one logit comparison, and the correction is a one-line change in the evaluation code. Published POPE numbers cannot be compared at face value until the readout and the precision are matched and reported.

0 of 9,000

answers generate ID 3582 or 1217

+6.13 F1

from the readout alone, POPE
adversarial

5 of 6

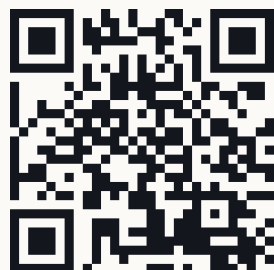
surveyed papers report a smaller
gain

0.80 to 0.92

F1 with tokenizer-derived IDs, 5
VLMs



Paper
arxiv.org/abs/2610.25964



Code, 9,000 prediction records,
diagnostic script
github.com/Kesav2k04/ugaa-research

Kesav Kumar Jayakumar · kesavk659@gmail.com
Karthigeyan Thilak · karthik.suthraye@gmail.com