Utrecht University

2026 SIGdial Atlanta, GA, USA

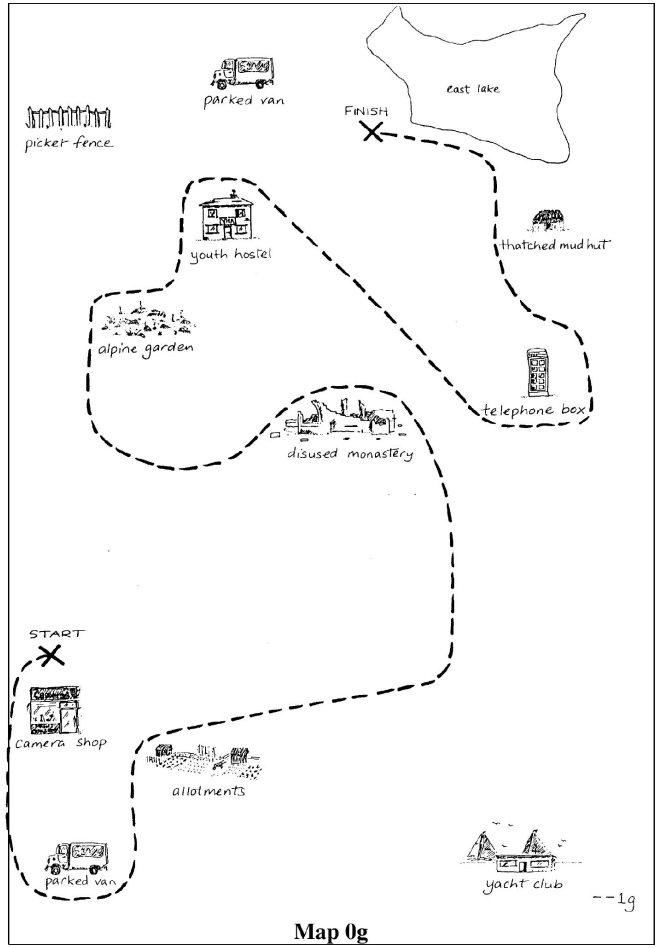
SIGDIAL 2026 · Atlanta · August 5th

Seeing Is Not Sharing

Some Vision-Language Models Overestimate Common Ground in Asymmetric Dialogue

Nan Li · Albert Gatt · Massimo Poesio

Utrecht University, The Netherlands



Map 0g

What could be shared ≠ what has been shared

MOTIVATION

Shared perception is not shared interpretation

POTENTIAL COMMON GROUND

Both participants could have access to the same kind of landmark.

Co-presence in the scene.

≠

ESTABLISHED COMMON GROUND

Built incrementally: the speaker refers, the addressee recognizes or accommodates, both confirm or repair.

Clark & Wilkes-Gibbs 1986; Clark & Brennan 1991

A MODEL IS AN OVERHEARER

It sees the words, and sometimes the maps. It cannot ask a clarification question and it cannot repair.

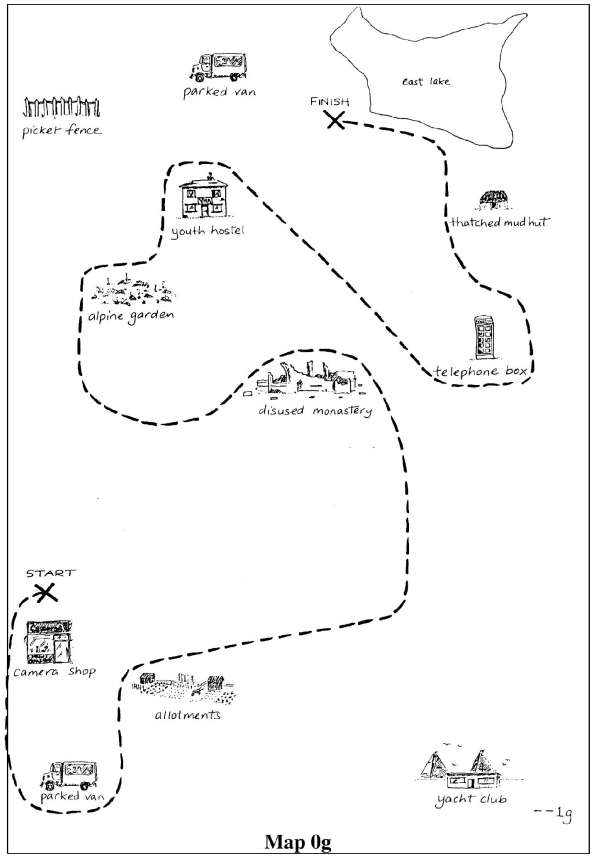
Overhearers overestimate their own understanding (Schober & Clark 1989).

Can a vision-language model tell what could be shared from what has actually become shared?

In **asymmetric dialogue**, participants hold different private information by design. So what looks shared from the outside may not be shared at all.

TESTBED

MapTask: two private maps that differ by design



Map 0g

- **Lexical variant** — the same landmark is named differently — “white water” vs. “rapids”
- **Existence** — a landmark is on one participant's map only
- **Multiplicity** — e.g., two parked vans on the giver’s map, one on the follower’s

Perspectivist annotation (Li et al., LREC 2026): for every reference expression we record the giver's intended landmark and the follower's interpreted landmark, separately.

UNDERSTANDING STATE

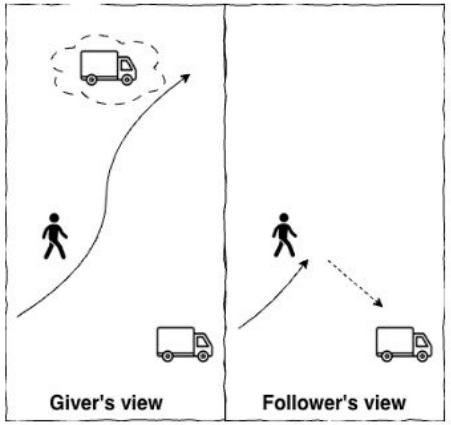
Aligned 72.1% n = 9,435 **Pending 26.0%** n = 3,403 **Misunderstood 1.8%** n = 239

13,077 reference expressions · 128 dialogues · 16 map pairs. HCRC MapTask (Anderson et al., 1991); GMMT annotation, Li et al., LREC 2026.

THE TASK

Interpretation matching: do both participants mean the same landmark?

1 · MapTask Dialogue Example



G Go to **«the parked van»**.

F Alright, to the bottom.

G **No!**

2 · Interpretations

Perspectivist annotations (Li et al., 2026)

G Giver's interpretation

upper parked van

landmark ID: m0_parked_van#1@g

≠

F Follower's interpretation

lower parked van

landmark ID: m0_parked_van#0@f

The landmark ID encodes map (m0), side (@g / @f), and ordinal index (#0 = lower, #1 = upper).

3 · Our task

Interpretation matching

Do the two participants interpret the target RE the same way?

Yes — same landmark

No — different landmarks

Gold Label: No

Model role: *overhearer (third-party analyzer) — observes but cannot ask questions, or repair.*

Experiment Conditions: *dialogue context window and map information modality.*

DESIGN

Two independent variables, fully crossed, zero-shot

1 · DIALOGUE CONTEXT

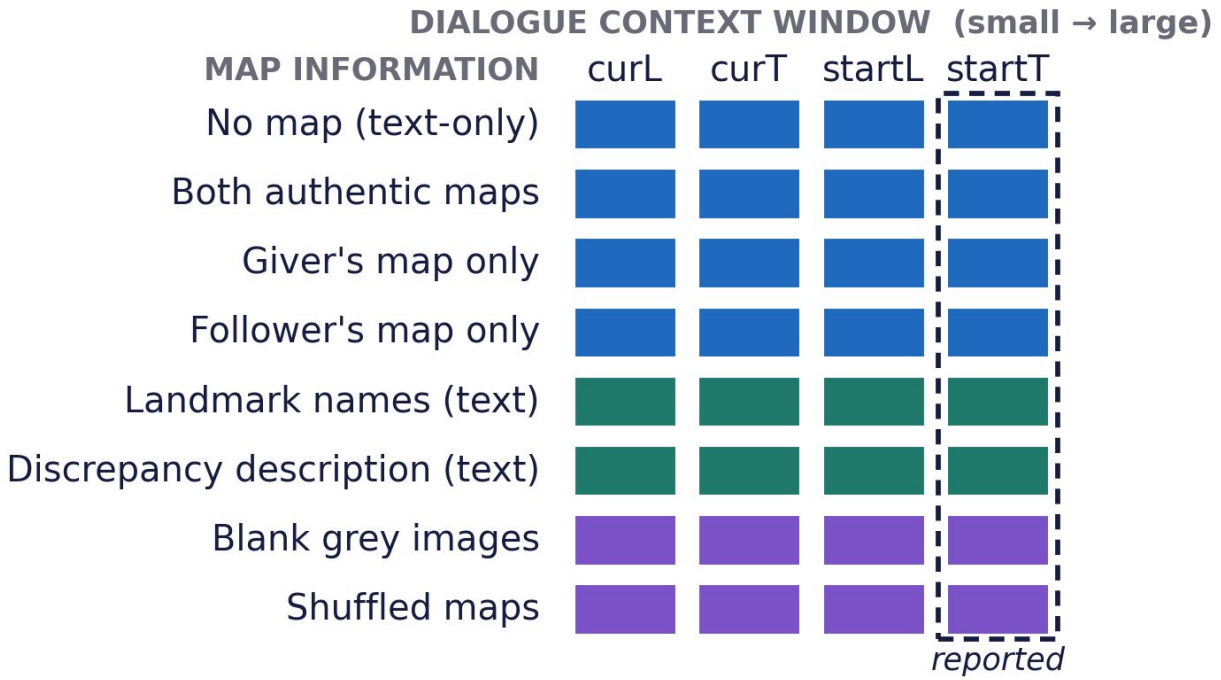
Four windows, from the current transaction up to the target line (curL) to the whole dialogue through the end of that transaction (startT).

Later turns may carry confirmation and/or repair.

2 · MAP INFORMATION

Authentic maps (both / giver / follower); the same content delivered as text (landmark names, discrepancy descriptions); two content-free visual controls (blank grey, shuffled maps).

The controls are what separate content from channel.

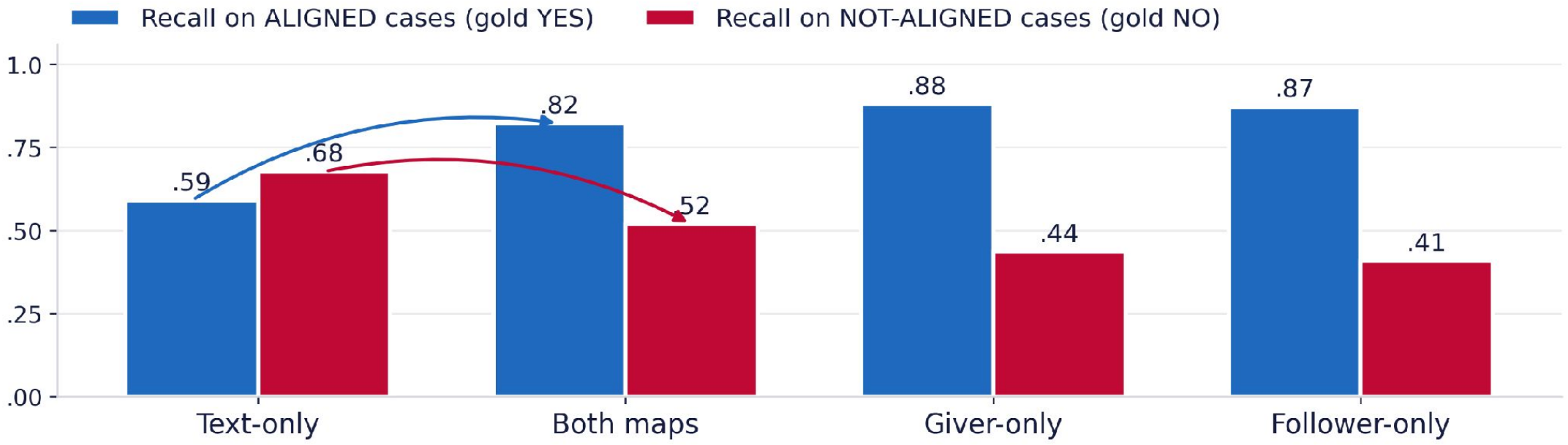


Five open VLMs: Qwen3-VL 2B / 4B / 8B and Gemma-3 4B / 12B.
Qwen3-VL-8B runs the full grid; the other four run the four primary map conditions, also across all four windows.
Decoding is constrained to a single YES / NO token — so the token probability is directly the model's confidence.

FINDING 1

Maps help — by pushing the model toward YES

MACRO-F1	RECALL · ALIGNED	RECALL · NOT ALIGNED	YES-RATE
.591 → **.671**	.590 → **.822**	.677 → **.518**	.515 → **.727**

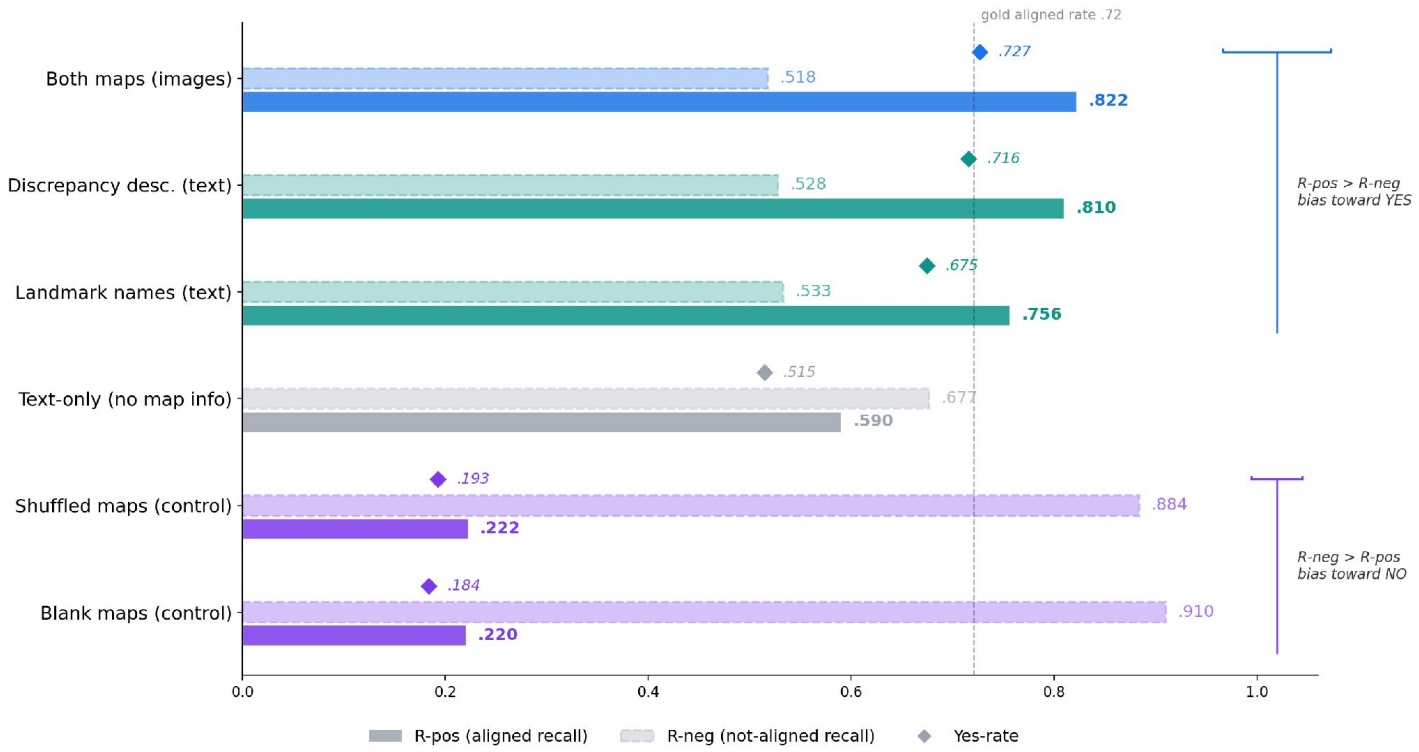


Qwen3-VL-8B at startT. The gold aligned rate is .721 — the yes-rate alone is not the test; the direction of the recall trade-off is.

Either single map triggers it more strongly; both maps moderate it, via cross-map discrepancy evidence.

FINDING 2

The bias travels with content, not with the channel



CONTENT-RICH CONDITIONS

Images and text both push R-pos up, R-neg down — the same trade-off direction as Finding 1.

CONTENT-FREE CONTROLS

Blank and shuffled maps flip the pattern: R-neg dominates. The model turns hyper-conservative.

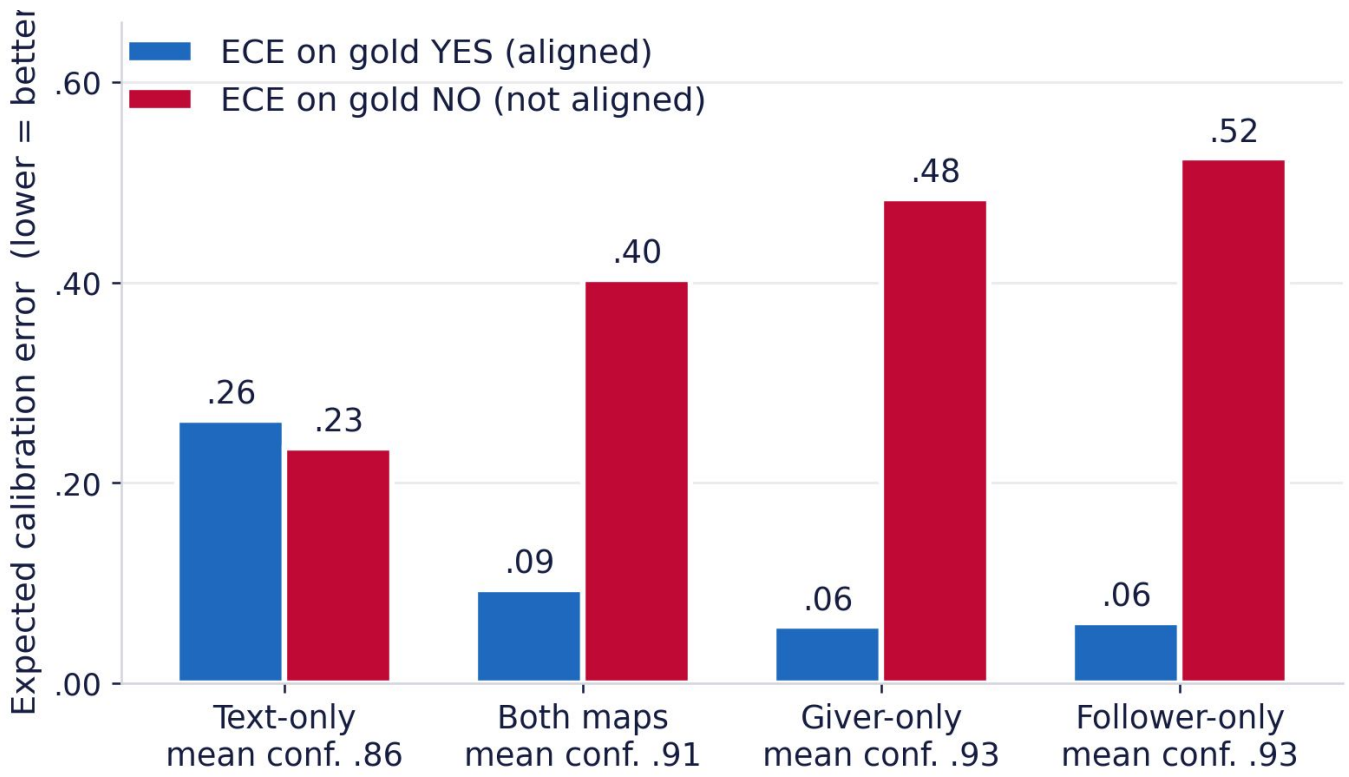
SO THE CUE IS MAP CONTENT

in pixels or in words.

All conditions: Qwen3-VL-8B at startT. R-pos = recall on aligned; R-neg = recall on not-aligned.

MECHANISM

Under map access, the model is confidently wrong on non-aligned cases



WELL CALIBRATED ON GOLD YES

ECE .263 → .094 with both maps.

CONFIDENTLY WRONG ON GOLD NO

ECE .235 → .403, and .524 with the follower's map alone.

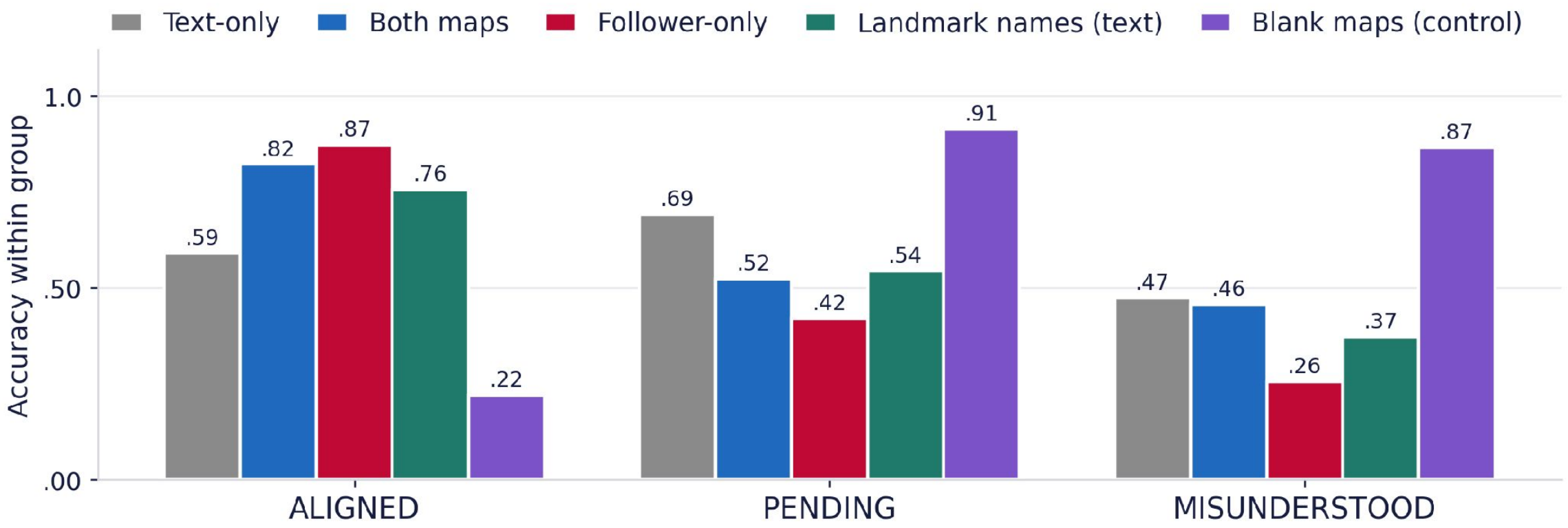
THE ASYMMETRY IS THE EVIDENCE

The class gap widens from .028 to .463.

ECE = weighted mean of |accuracy − confidence| per probability bin (Naeini et al., 2015). Lower = better calibrated.
Confidence := exp(logprob) of the single YES/NO token.
n = 13,077, Qwen3-VL-8B at startT.

WHERE IT COSTS

The aggregate gain is paid for by pending cases



n = 9,435
+23.2 pp .590 → .822

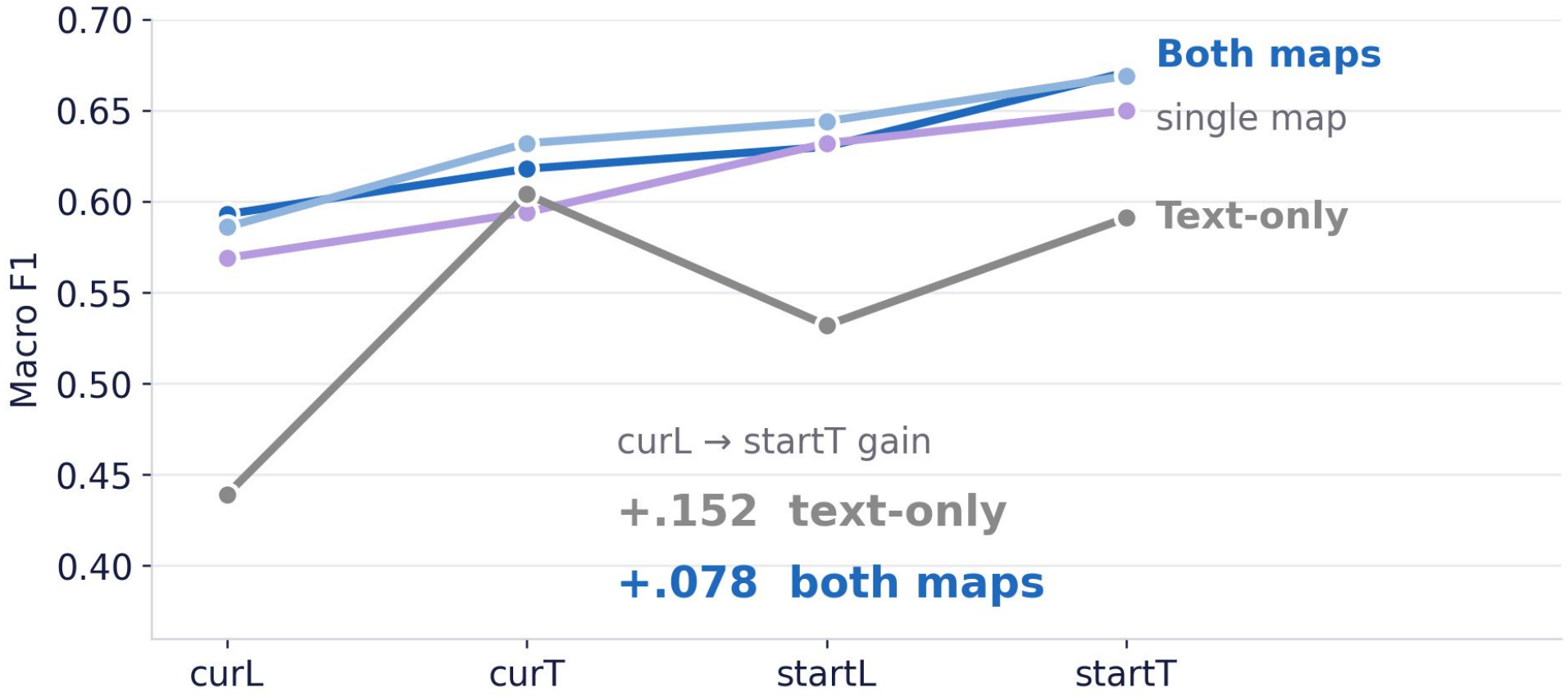
n = 3,403
−16.8 pp $p < 10^{-6}$

n = 239
−1.7 pp n.s. p = .724

Text-only → both maps. Blank maps look strong on pending and misunderstood only because they answer NO to almost everything.

DYNAMICS

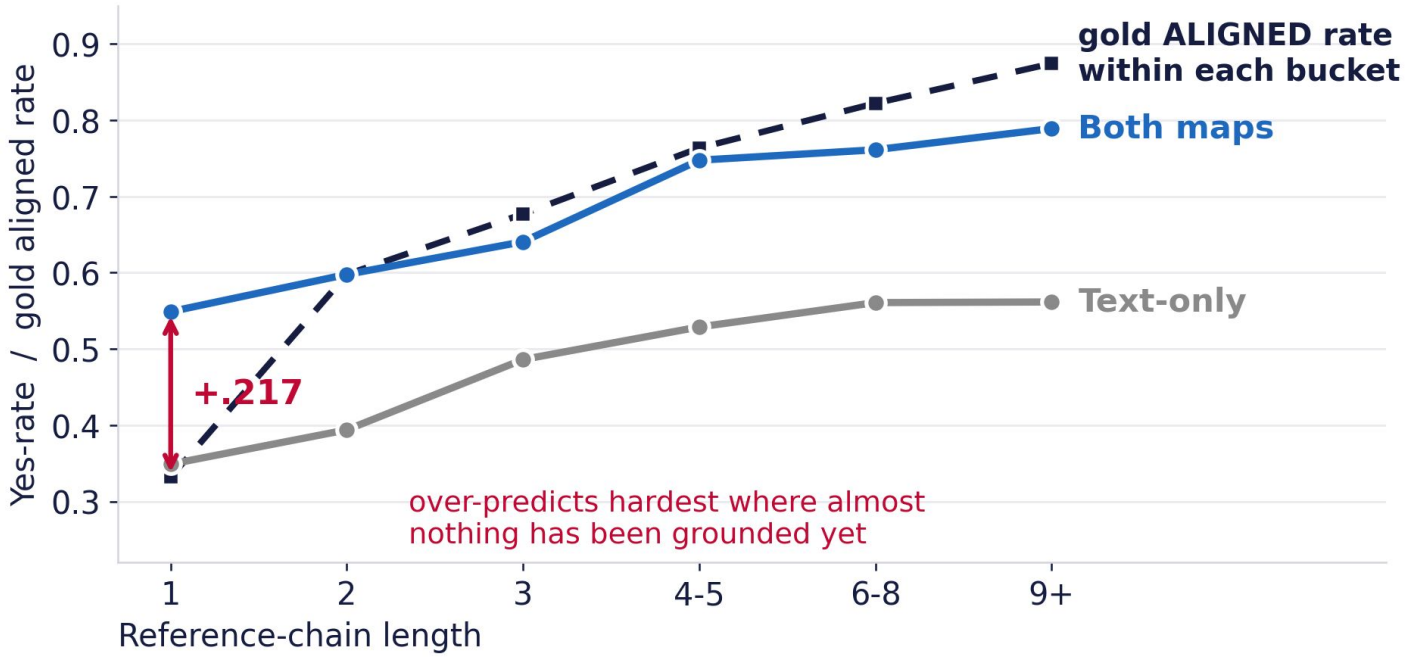
Dialogue evidence carries less weight once maps are in



Qwen3-VL-8B macro-F1 across the four context windows.

EVIDENCE

Over-prediction is strongest before grounding accumulates



FIRST MENTION

Yes-rate .55 against a gold rate of .33 — a 22-point over-prediction, exactly where grounding has barely begun.

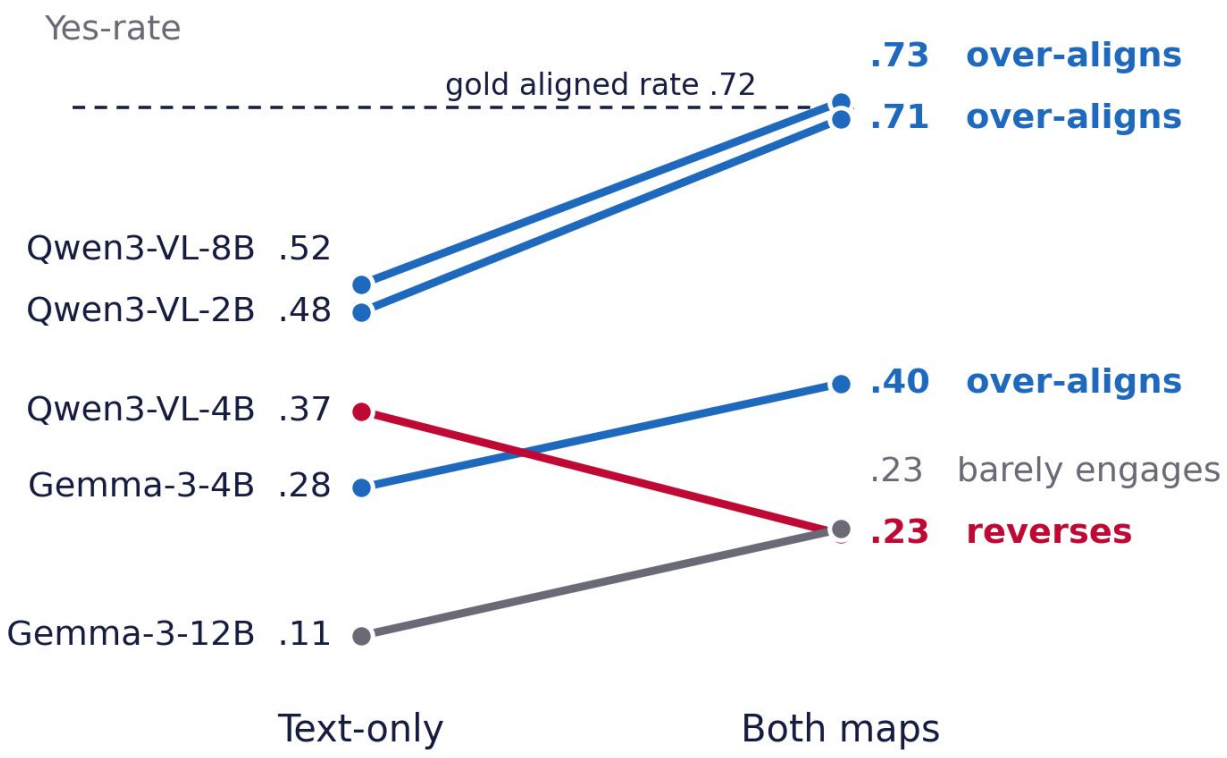
REPEATED MENTION

By 9+ mentions the model sits slightly below the local gold rate. Over-prediction fades as grounding evidence accumulates.

1,665 chains covering 11,473 of the 13,077 REs.

GENERALITY

Model-dependent — and scale does not fix it



"SOME" IS DOING REAL WORK

Four of five shift upward, but only two land near the base rate. Gemma-3-12B stays extremely conservative; Qwen3-VL-4B reverses.

LARGER ≠ MORE DISCERNING

2B beats 4B (F1 .566 vs .411); Gemma-3-4B beats 12B (.510 vs .416).

ONE LIKELY CONTRIBUTOR

A model that cannot read the maps cannot show a content-driven bias.

Baseline grid at startT, n = 13,077.
Gemma's SigLIP encoder pools every image to 256 tokens; in a map-reading check Gemma scored 81–82 F1 on landmark names against 88–90 for Qwen3-VL.

INTERPRETATION

What the failure tells us about VLM grounding

SCENE EVIDENCE

what COULD be shared — landmarks co-present on both maps

DIALOGUE EVIDENCE

what HAS been grounded — recognition, confirmation, repair

co-presence ⇒ alignment

under-weighted

THE MODEL'S JUDGEMENT

“ALIGNED” — because the landmark is on both maps.

Potential common ground is read as established common ground: the computational form of the overhearer's illusion.

Evaluation should separate visual reference recognition from interactional common-ground tracking.

Seeing is not sharing

1 Maps improve macro-F1 by reading potential referential overlap as established common ground.

2 The bias travels with content, not channel — text reproduces it, content-free images reverse it.

3 The cost lands where grounding matters — confident errors on pending cases.

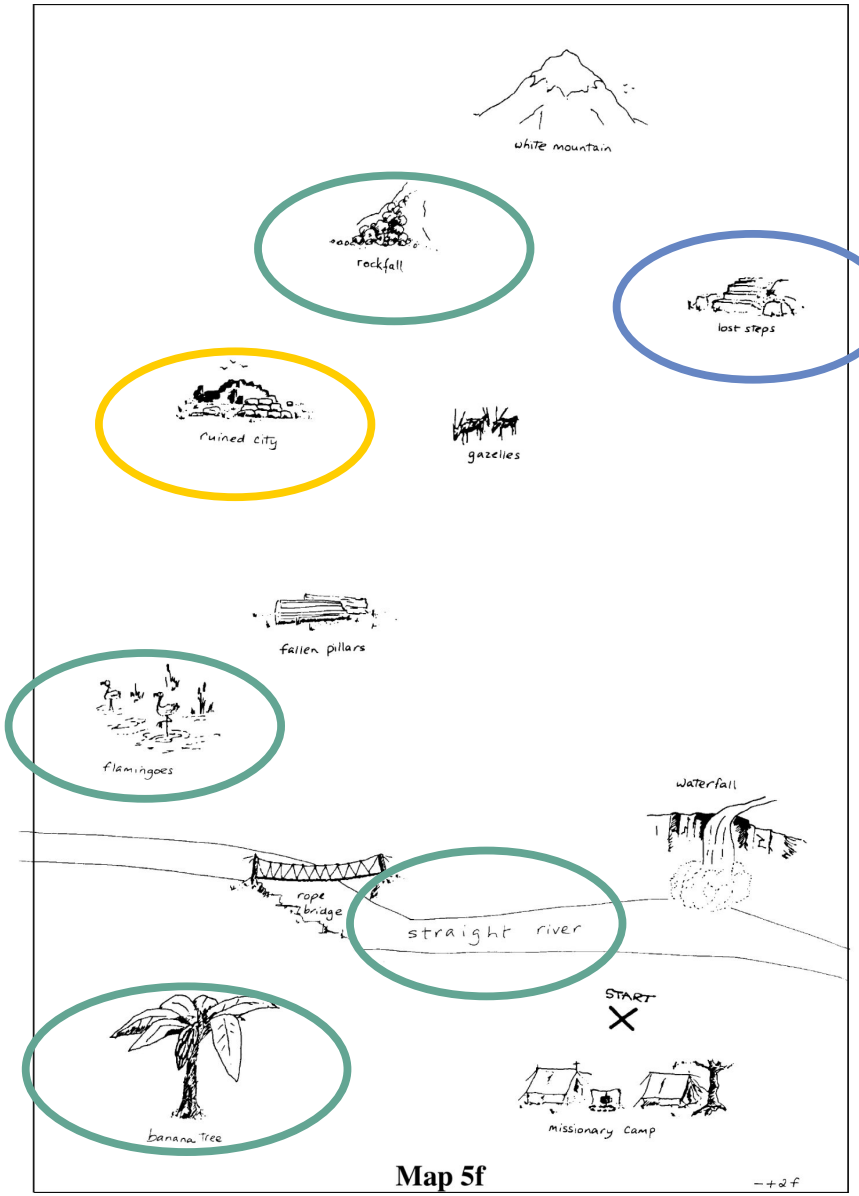
These models infer what the interlocutors could be talking about — not what they have established together.

Paper
aclanthology.org/2026.sigdial-1.49

Code + prompts
github.com/chnln/seeing-is-not-sharing

SINS dataset
https://huggingface.co/datasets/chnln/seeing-is-not-sharing

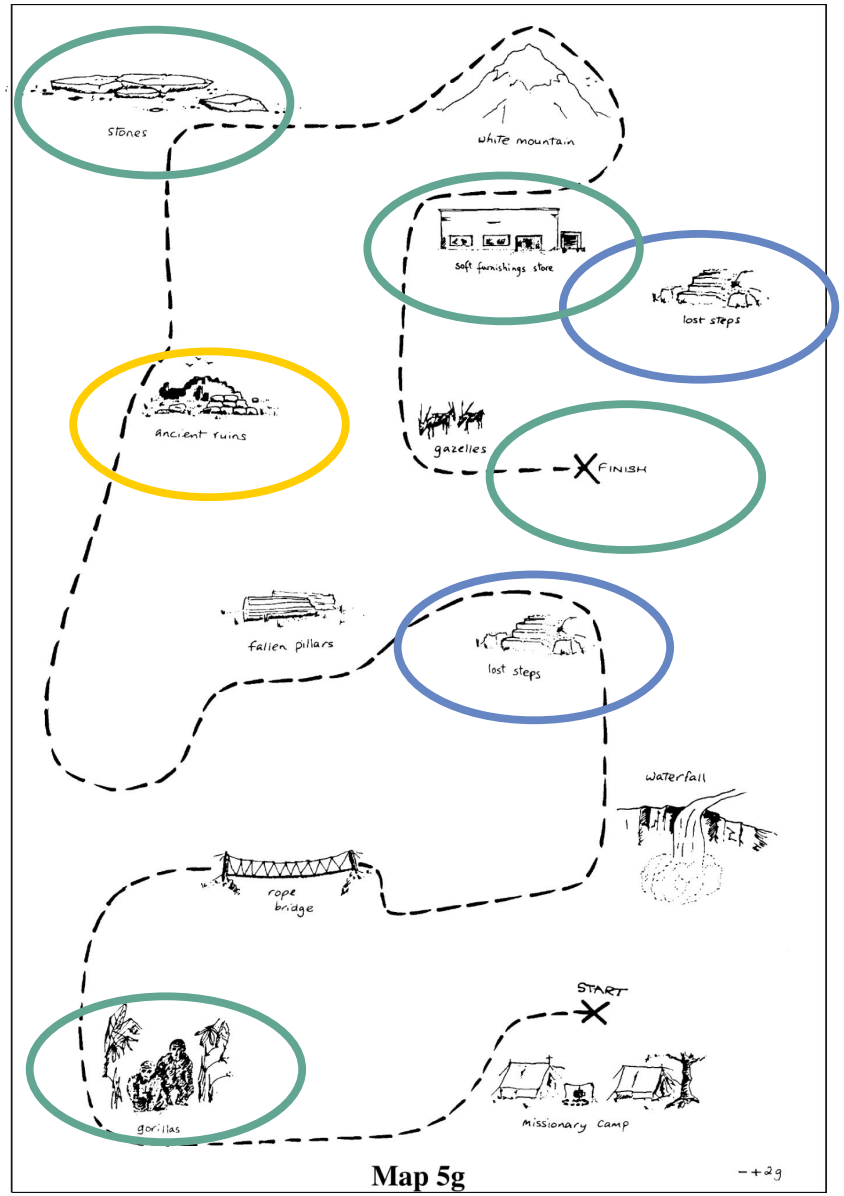
Thank you — questions welcome. Funded by NWO, AiNed Fellowship NGF.1607.22.002 — Dealing with Meaning Variation in NLP.



Lexical

Existence

Multiplicity



Baseline and modality results (Qwen3-VL-8B, startT)

Map information	F1-macro	R-pos	R-neg	Yes-rate
Text-only	.591	.590	.677	.515
Both maps	.671	.822	.518	.727
Giver-only	.669	.879	.436	.791
Follower-only	.650	.872	.408	.794
Landmark names (text)	.636	.756	.533	.675
Discrepancy description (text)	.668	.810	.528	.716
Blank maps	.407	.220	.910	.184
Shuffled maps	.402	.222	.884	.193

READING THE TABLE

R-pos = recall on aligned; R-neg = recall on not-aligned.

Gold aligned rate: .721.

n = 13,077 for every row.

Tables 1 and 2 of the paper.