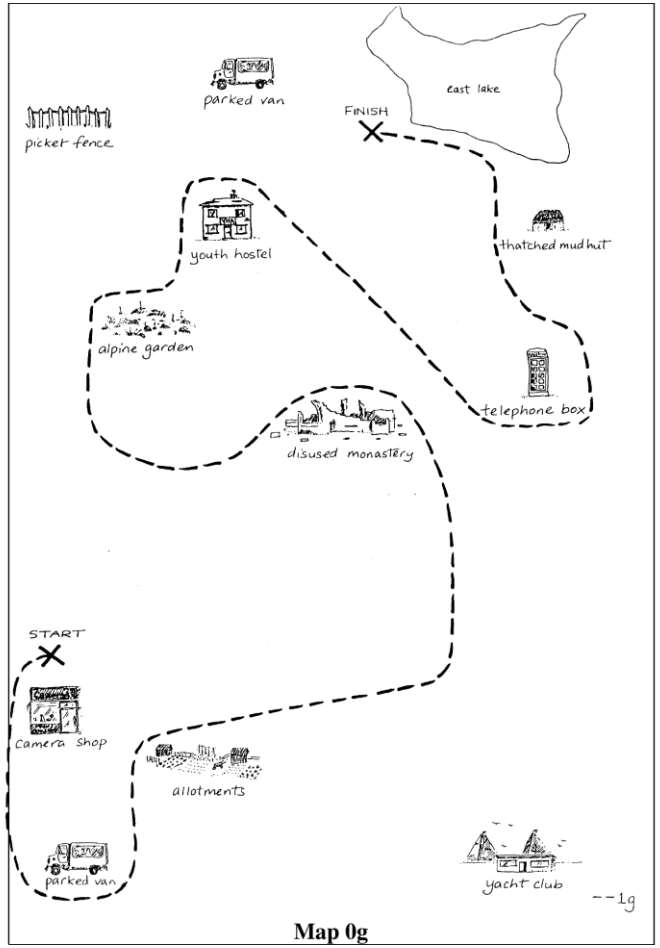
Personal Interpretations as Pragmatic Label Variation

Perspectivist Grounding Annotation and (V)LLM Evaluation



Nan Li

Utrecht University

n.li@uu.nl

Li, Gatt & Poesio. Grounded Misunderstandings in Asymmetric Dialogue: A Perspectivist Annotation Scheme for MapTask. LREC 2026.

Li, Gatt & Poesio. Seeing Is Not Sharing: Some Vision-Language Models Overestimate Common Ground in Asymmetric Dialogue. SIGDIAL 2026.

THE QUESTION

When two interlocutors appear to use the same expression, do they actually interpret it in the same way?

One expression, two interpretations. The variation is perspective-conditioned: it comes from the participants' asymmetric information, not from the annotators.

PART 1

LREC 2026

A perspectivist annotation resource

Record both participants' interpretations for every reference expression in a map navigation task.

PART 2

SIGDIAL 2026

A vision-language model evaluation

Can VLMs tell what could be shared from what has been shared in common ground?

PART 3

HVL at the discourse level

Conditioned ground truth

MOTIVATION

Common ground is built incrementally

GROUNDING IS INCREMENTAL

I present → you accept → we ground

An utterance joins the common ground only once the addressee recognises and acknowledges it.

≠

WHAT CORPORA ASSUME

one expression = one gold referent

For each referring expression, both interlocutors are always referring to the same entity.

UNDER ASYMMETRY

grounding can fail silently

Both sides believe they agreed. The dialogue sounds smooth. They mean different objects.

A one-gold-label format cannot represent this.

There is no slot in which the giver's referent and the follower's referent can differ — so the phenomenon is invisible before analysis even starts.

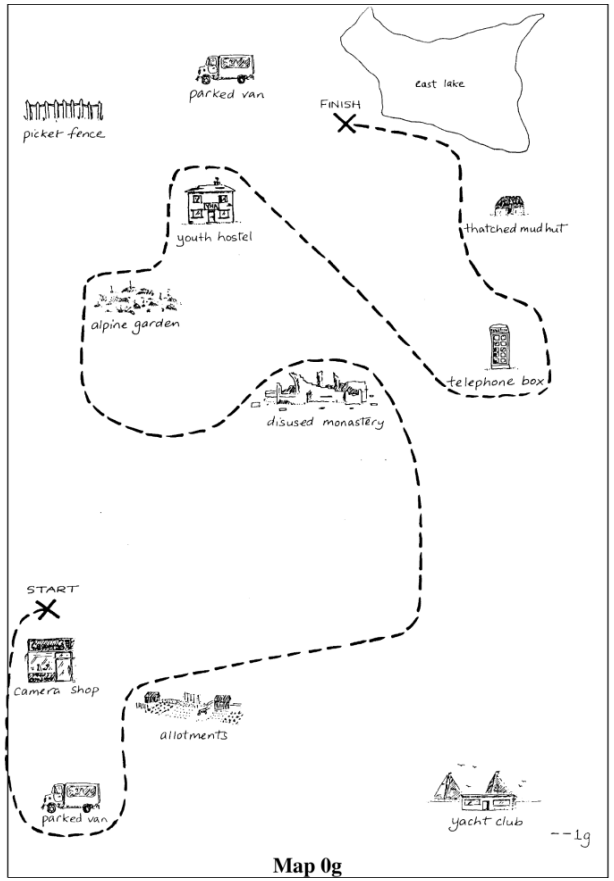
TESTBED

The HCRC MapTask corpus (Anderson et al., 1991)

128	16	267
dialogues	map pairs	landmarks

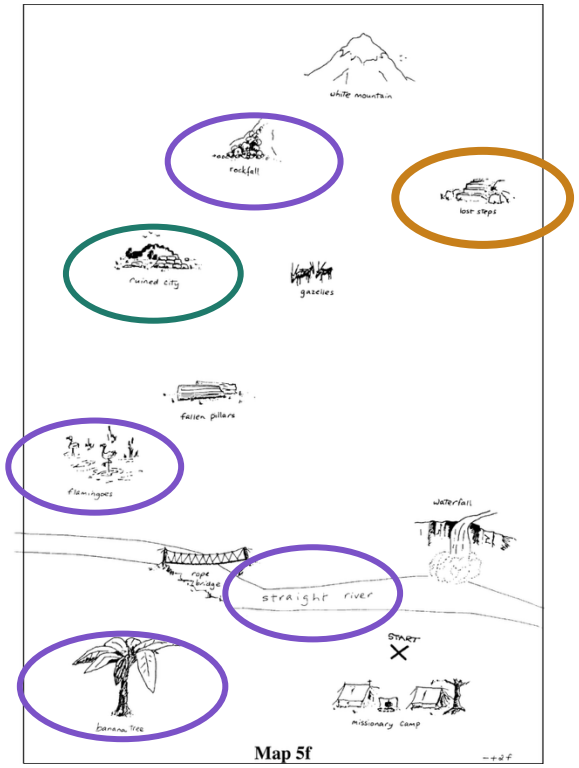
The instruction giver has a route; the follower has to reproduce it. Verbally only — they cannot see each other's maps.

The maps are similar but deliberately not identical, which makes MapTask a controlled environment for eliciting referential misalignment.

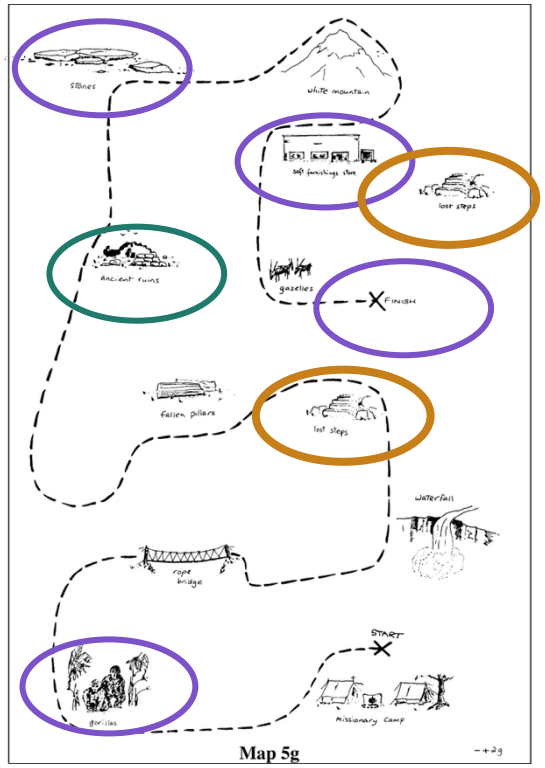


TESTBED

Three ways the maps differ



Map 5f · follower



Map 5g · giver

Lexical

One place, two names — “ruined city” on one map, “ancient ruins” on the other.

20 landmarks (7.5%) · 914 REs

Existence

On one map only — flamingoes, banana tree, gorillas, stones, straight river, …

99 landmarks (37.1%) · 2,878 REs

Multiplicity

“Lost steps” — twice on the giver’s map, once on the follower’s.

16 landmarks (6.0%) · 956 REs

Multiplicity is 6% of landmarks and 7.3% of mentions. It will carry half of all the misunderstandings.

Map pair m5, every discrepancy circled. 267 landmarks total, half identical; 36.3% of REs target a discrepant one.

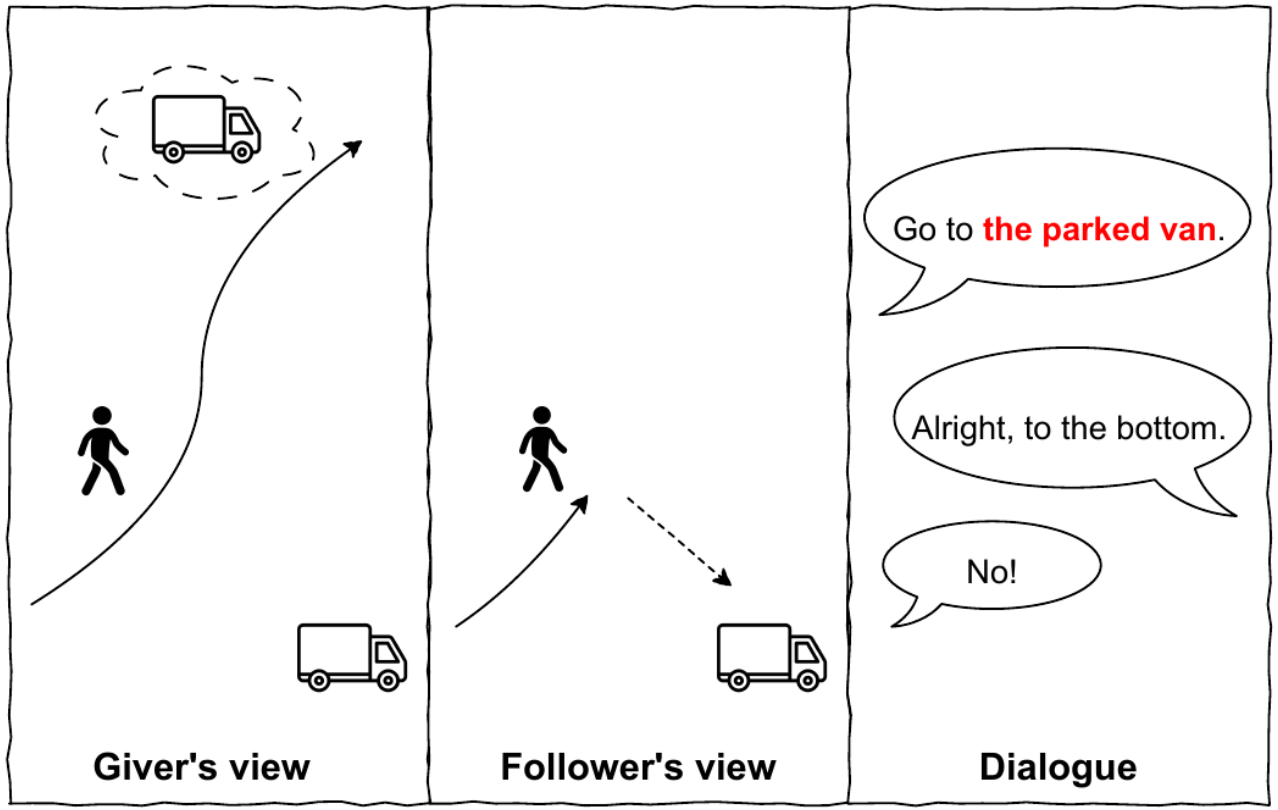
PART 1 · LREC 2026

Grounded Misunderstandings in Asymmetric Dialogue

A perspectivist annotation scheme for MapTask

THE PHENOMENON

“Go to the parked van.”



Different private information

Two parked vans on the giver’s map vs. one on the follower’s map.

Confirmation signals grounding

“Alright” says the follower thinks they are referring to the same parked van.

But common ground tacitly breaks here

If without the spatial description “to the bottom”, the dialogue would move on.

Two labels, one expression.

THE GAP

Why the existing annotation cannot see it

ORIGINAL MAPTASK IDS · NAME-BASED

upper parked van vs lower parked van

SAME id — different places, different referents

“cliffs” vs “sandstone cliffs”

DIFFERENT ids — the same landmark for any participant

WHAT WE NEED INSTEAD

Each participant’s personal interpretation —

recorded separately,
for every reference expression,
tracked as it evolves.

SCHEME · LANDMARK ID

Perspectivist landmark IDs

m0_parked_van#1@g

map pair | landmark name | instance 0 = lower | map side g / f

FIXES BOTH DEFECTS

The multiplicity twins get distinct IDs; lexical variants can be explicitly unified where participants treat them as one place.

EVERY RE GETS TWO INTERPRETATION SLOTS

speaker — **intended landmark**

addressee — **interpreted landmark**

These are not two guesses about one truth.

They are facts about two different mental states. Both are gold.

Five attributes, one decision cascade

STEP 1

is_quantificational

Is this even a referring use?

“Do you have a parked van?” checks existence.

STEP 2

is_specified

Any uptake at all?

Or was the expression simply ignored?

STEP 3

is_accommodated

Did the signal get through?

No “What?”, no “Pardon?”

STEP 4

is_grounded

Linked to a concrete landmark?

Prerequisite for assigning an addressee ID.

STEP 5

is_imagined

Projected from the description?

Not on my map, but I can place it. Still aligned.

Each step is evaluated only if the previous one succeeded.

That mirrors incremental reference resolution, and it operates as a built-in chain-of-thought scaffold when the task is handed to a language model.

“Go to the parked van.” grounded → assign both IDs: giver **`m0_parked_van#1@g`** follower **`m0_parked_van#0@f`**

Structured Annotation Prompt for LLM

BACKGROUND

MapTask's design and the three discrepancy types.

TASK & LANDMARK ID SCHEME

The task itself, and the unified landmark ID format.

ANNOTATION RULE

The five-step cascade in order, plus a required one-sentence reason citing dialogue evidence.

OUTPUT FORMAT

A JSON schema with required fields, enforced at decoding.

DIALOGUE CONTEXT

From the start of the conversation through the end of the current transaction, every RE bracketed.

DIALOGUE ACTS

MapTask's own move tags — instruct, acknowledge, query.

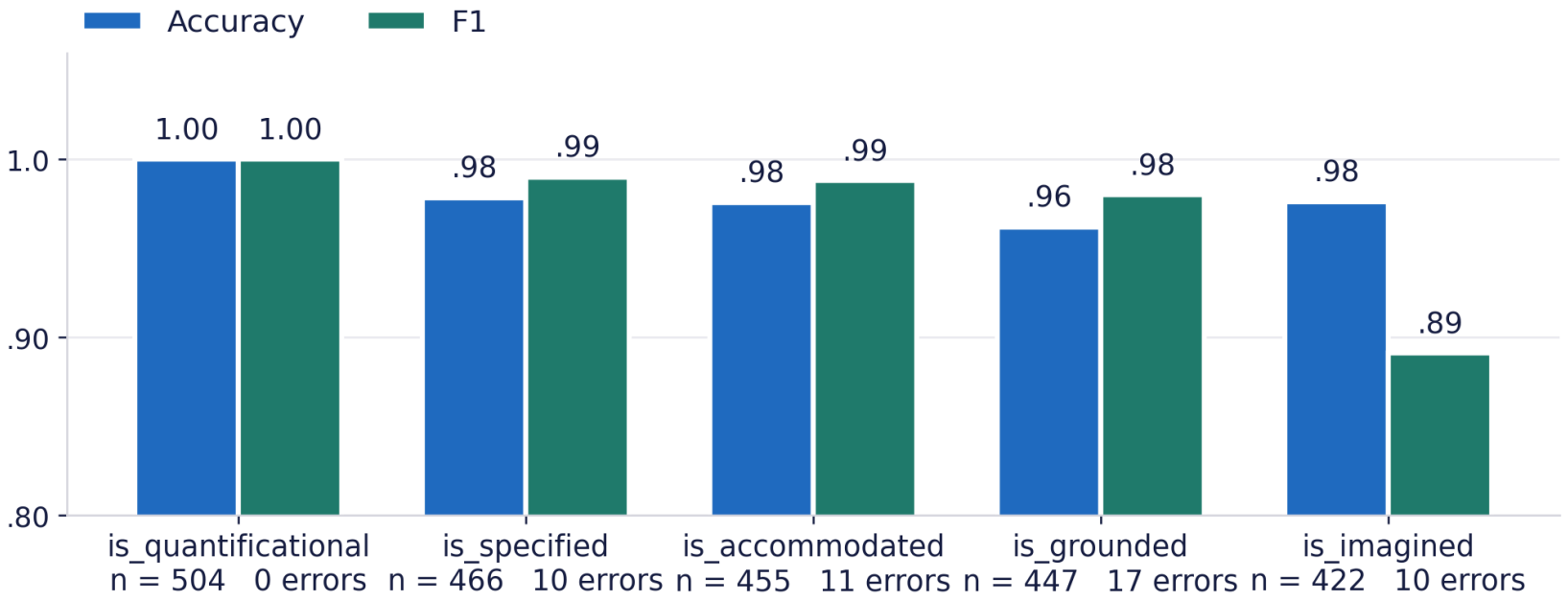
LANDMARK CANDIDATES

Every landmark on BOTH maps, with its unified ID.

PIPELINE & RELIABILITY

Scaling it, and verifying it

GPT-5, BATCH API	CONTEXT	CANDIDATES	OUTPUT
JSON-schema-enforced output	dialogue start → end of current transaction	all landmarks, both maps	2 IDs + 5 attributes + a reason



13,077
reference expressions
all 128 dialogues

94.4%
of 504 verified REs
completely correct

Human gold standard: one expert annotator, ambiguous cases discussed among the authors.
3 dialogues · 504 REs · 28 with ≥1 attribute error.

UNDERSTANDING DISTRIBUTION

Three understanding states

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

ALIGNED

Both ground the expression to the same landmark.

PENDING

Still under way, or someone signalled trouble.

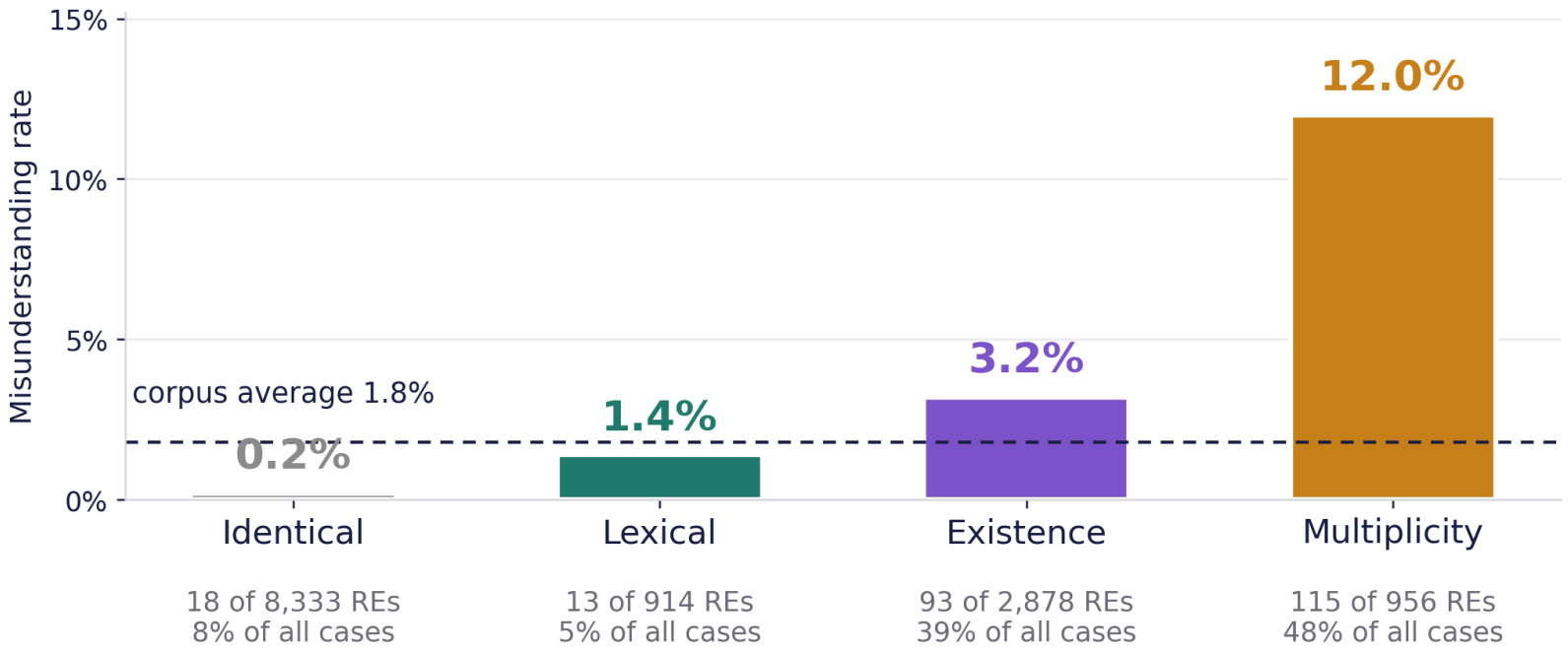
MISUNDERSTOOD

Both believe reference succeeded, but to different landmarks.

Raw misunderstanding is 7.07%. Most of it is lexical variants — “old mill” vs “mill wheel” — that participants quickly agree name one place. After unifying the 10 pairs: **1.82%.**

THE HAZARD

Rare — but not randomly distributed



THE MECHANISM

“the parked van” presupposes uniqueness.

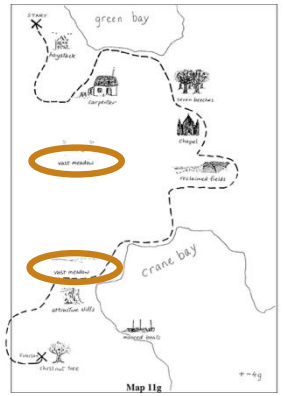
Saying "the parked van" claims there is only one. Both sides resolve it instantly — to different vans — and neither has any reason to ask.

Misunderstood: 7.3% of reference expressions produce 48% of all misunderstandings.

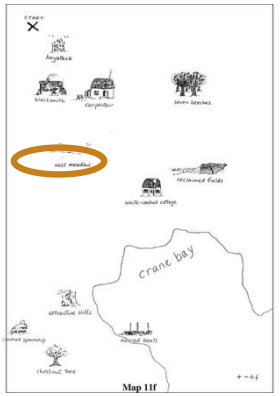
A DYNAMIC EXAMPLE

A misunderstanding, live

q7ec2 · “vast meadow” — twice on the giver’s map, once on the follower’s.



Map 11g · giver
(two meadows)



Map 11f · follower
(one meadow)

Ref	Spk	Expression	Giver	Follower	State
112	G	“a meadow”	#0@g	*ungrounded*	Pending
114	F	“a vast meadow”	#0@g	*ungrounded*	Pending
115	G	“a vast meadow”	#0@g	#1@f	**Misunderstood**
116	F	“a vast meadow”	#1@g	#1@f	**Aligned**
119	F	“vast meadow”	#1@g	#1@f	**Aligned**
127	G	“a vast meadow”	#0@g	*ungrounded*	Pending

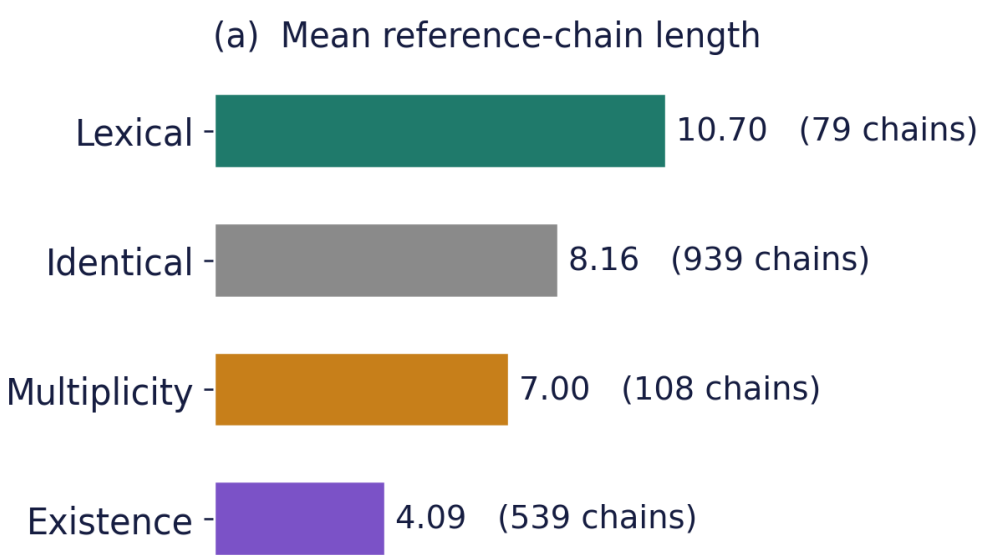
THE DIALOGUE 115 misunderstood → 116 repaired by spatial detail → 127 resurfaces

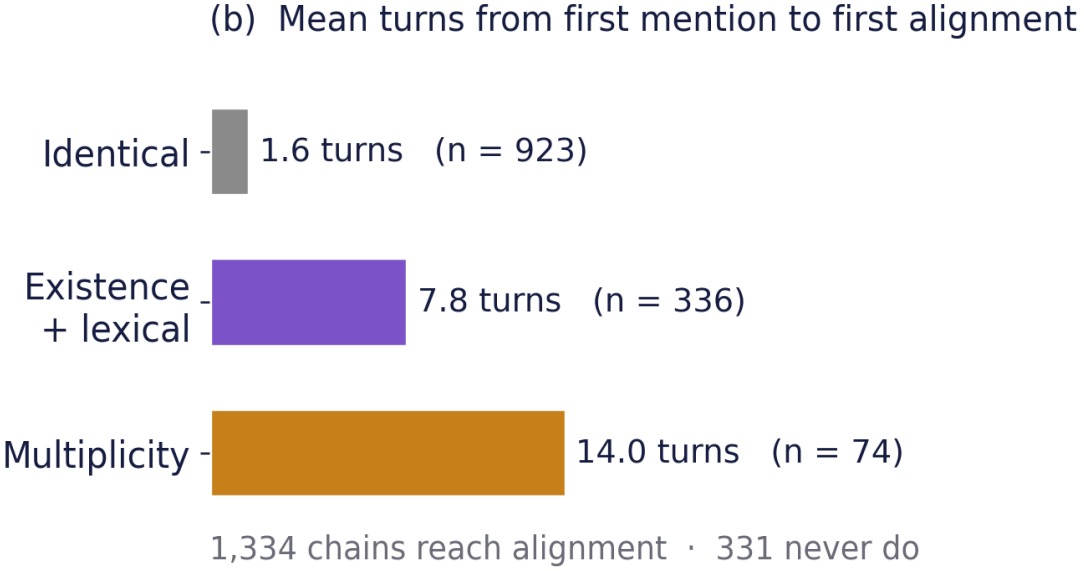
```
[giver utt:131] i've got a meadow ref.112 to the left of that due south pee porashin
[follower utt:132] a vast meadow ref.114
[giver utt:133] a vast meadow ref.115
[follower utt:134] i've got a vast meadow ref.116 up just below the blacksmiths and carpenters
[giver utt:135] well is that huge big and there's nothing at all in that space
[follower utt:136] uh-huh except
[giver utt:137] about halfway up the page
[follower utt:138] vast meadow ref.119 about halfway up the page it's empty on mine
[giver utt:139] well you want to well anyway follow crane bay round the curve
```

… (utterances 140–160: participants navigate along crane bay and attractive cliffs) …

```
[giver utt:161] just below i've got a vast meadow ref.127 there
[follower utt:162] right where will i stop that line across the top
```

Negotiation has a length





1,665

chains covering 11,473 of the 13,077 REs, mean 6.89 each

331

chains never reach alignment: participants abandon the negotiation

PART 2 · SIGDIAL 2026

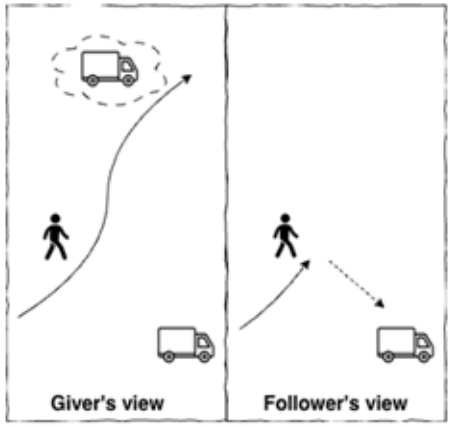
Seeing is not sharing

Some Vision-Language Models Overestimate Common Ground in Asymmetric Dialogue

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

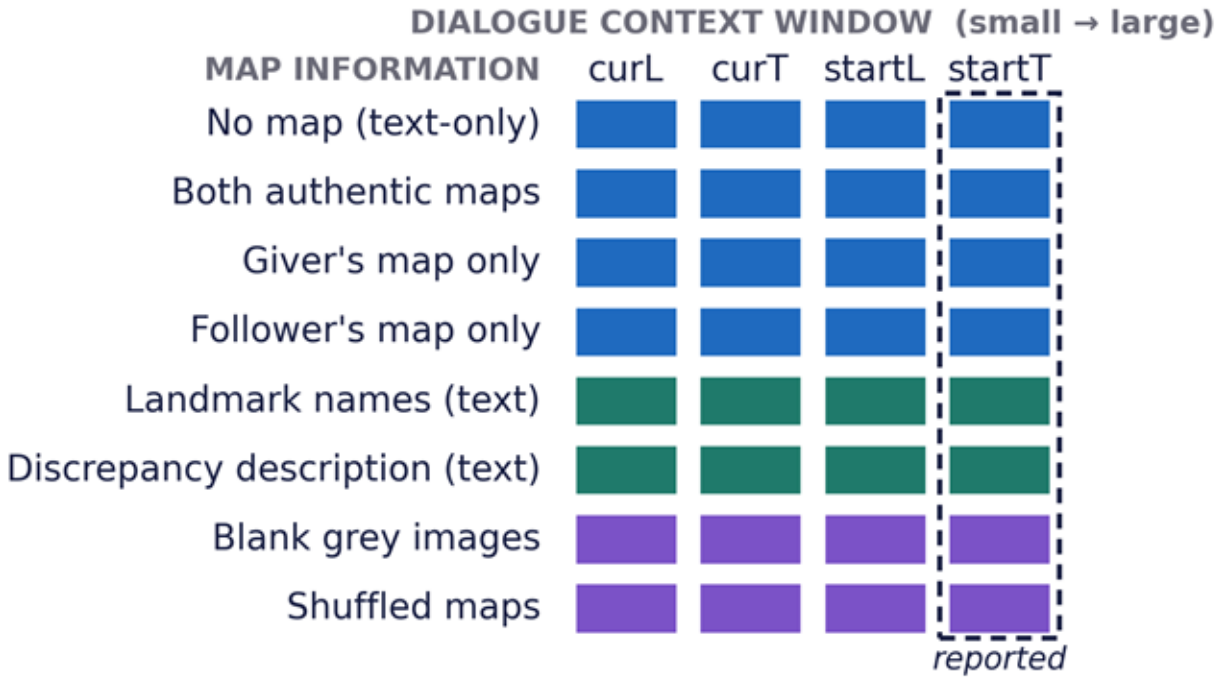
Four windows, from the current transaction up to the target line (curL) to the whole dialogue through the end of that transaction (startT).

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.

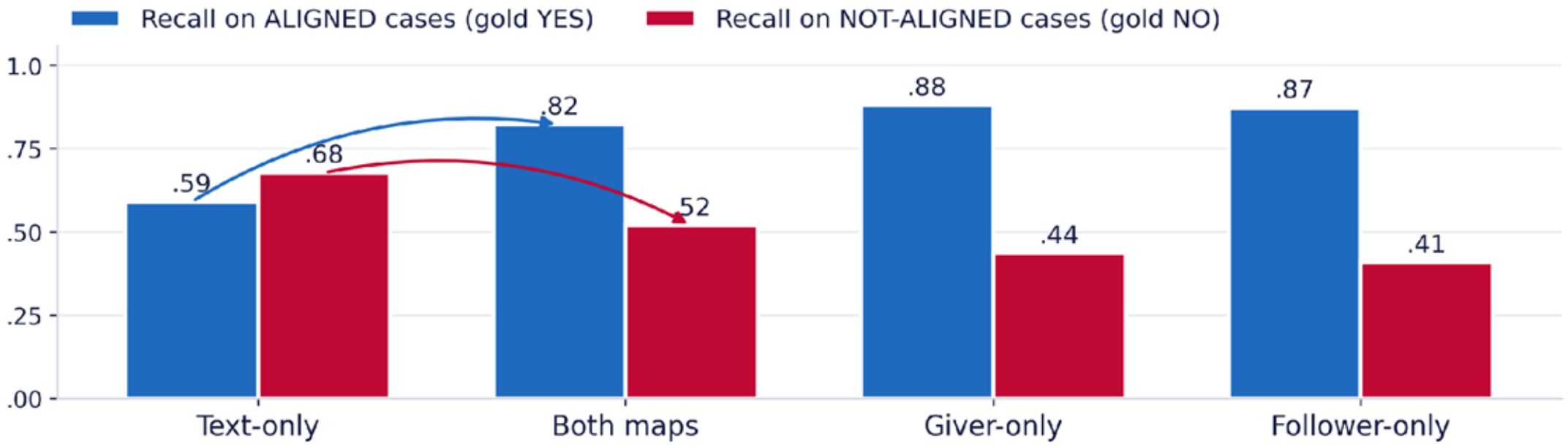
Qwen3-VL-8B runs the full grid; the other four run the four primary map conditions, also across all four windows.

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

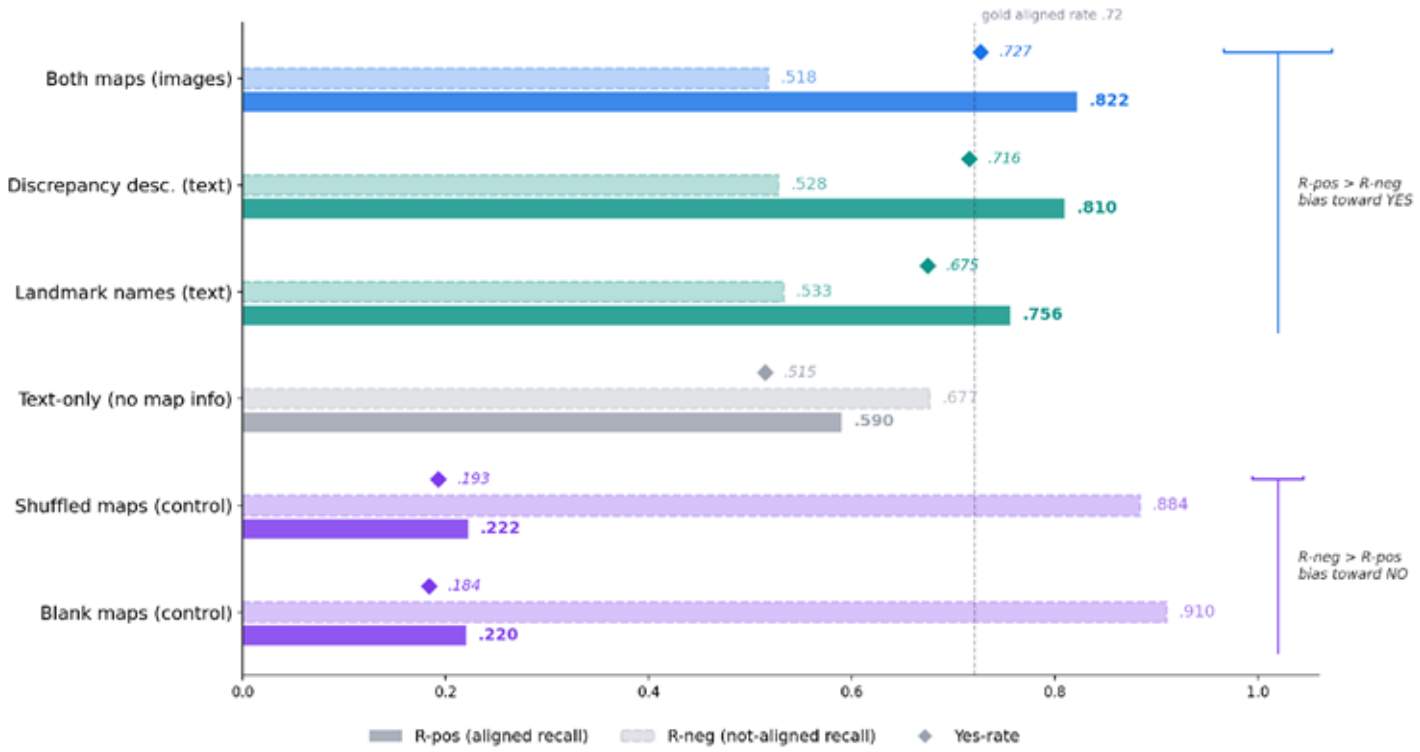


Qwen3-VL-8B at startT. The gold aligned rate is .721 — the yes-rate alone is not the test; the direction of the recall trade-off is.

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.

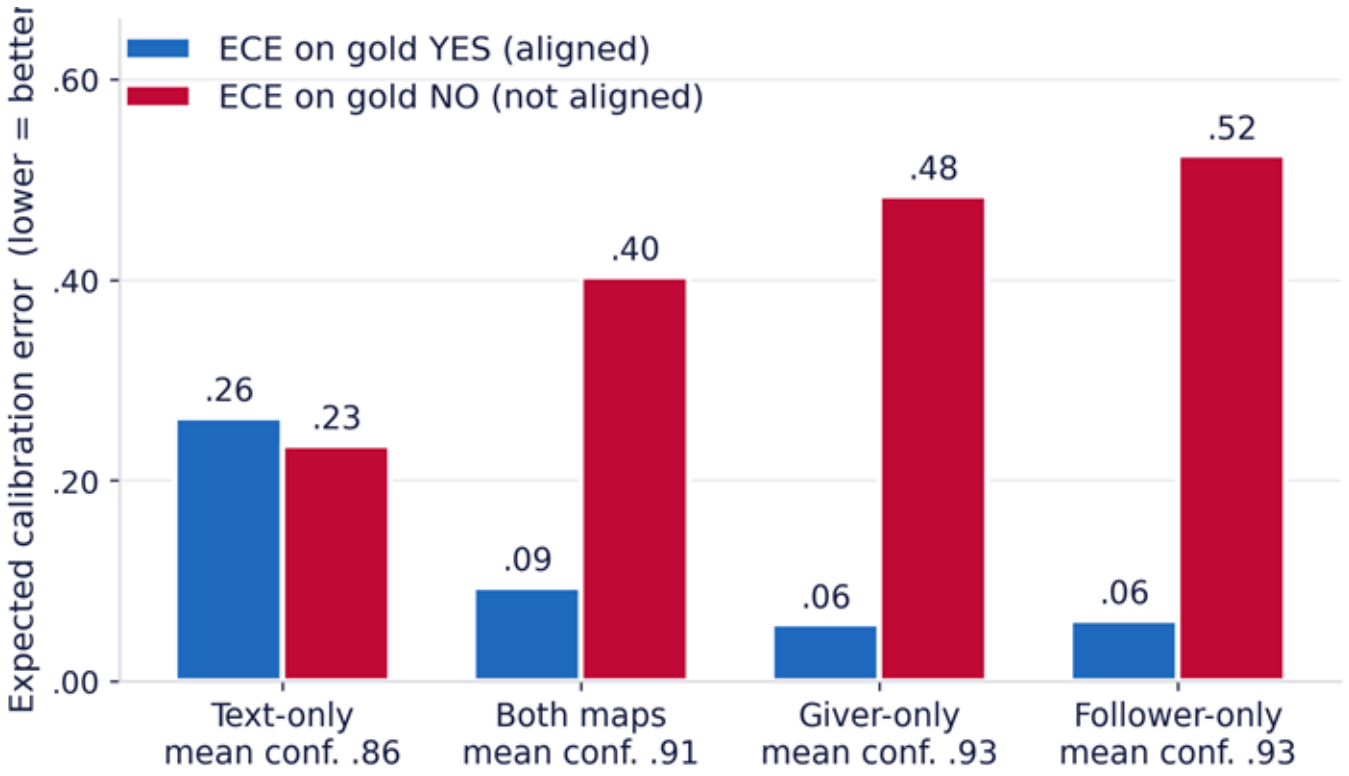
SO THE CUE IS MAP CONTENT

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.

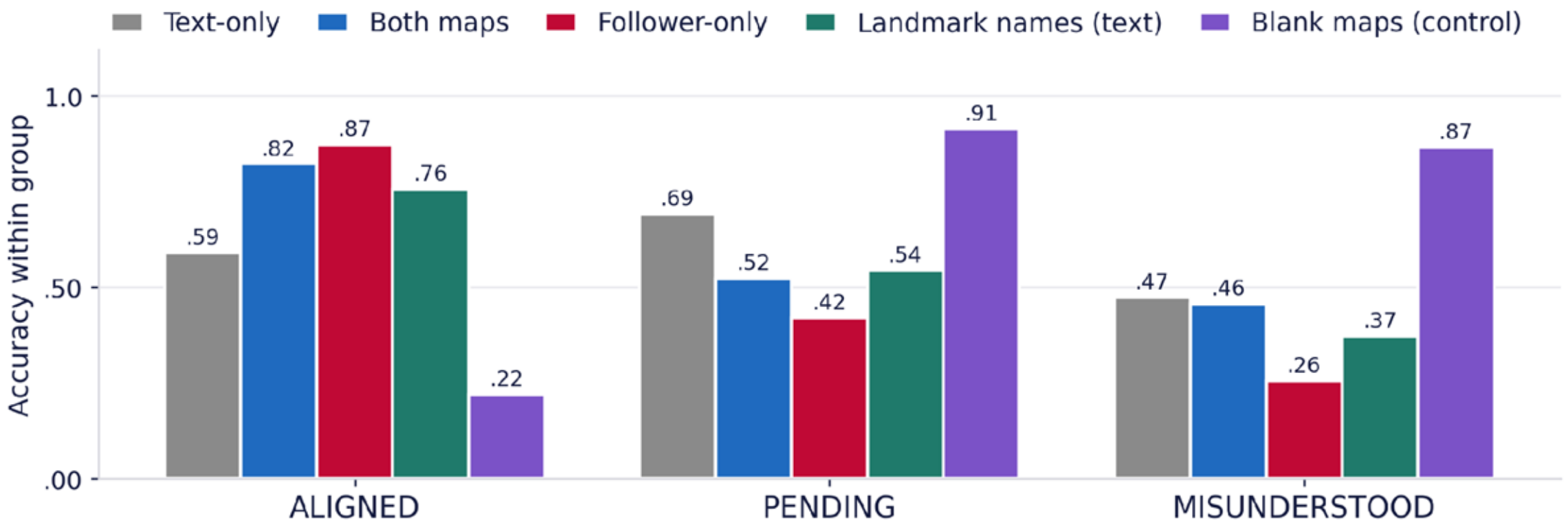
THE ASYMMETRY IS THE EVIDENCE

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

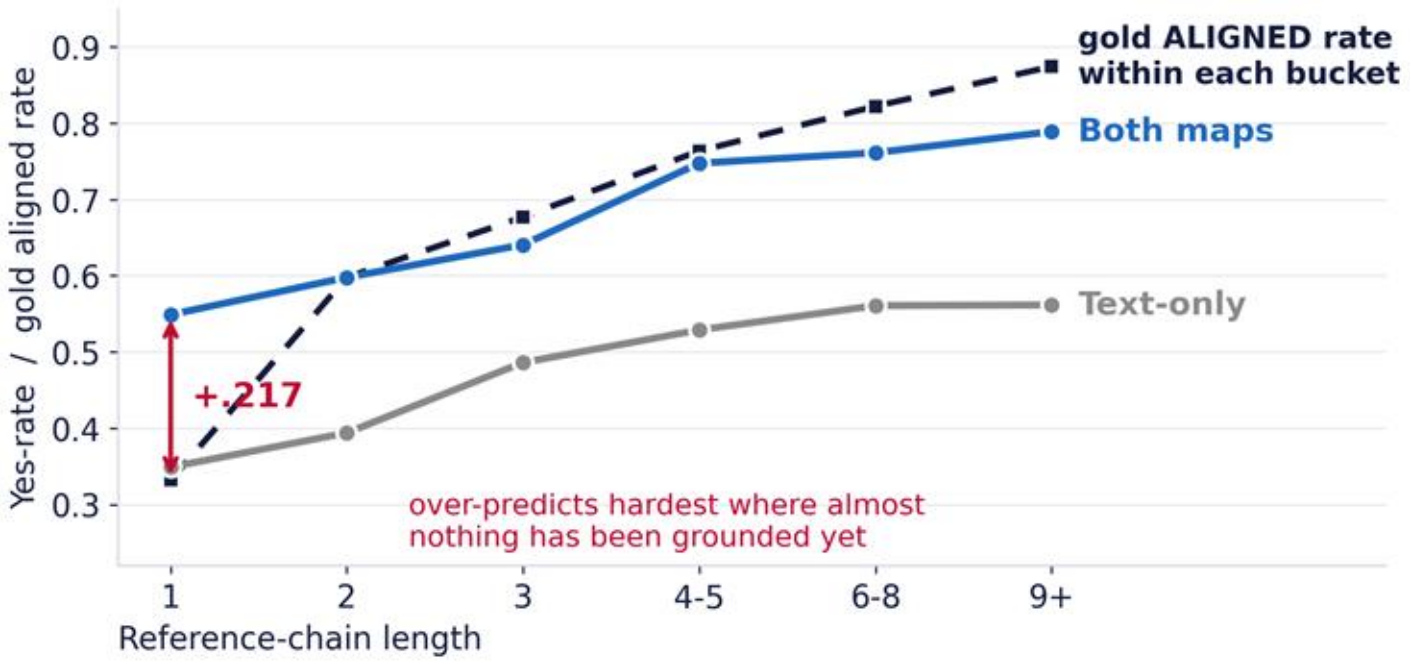
n = 3,403
−16.8 pp $p < 10^{-6}$

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.

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.

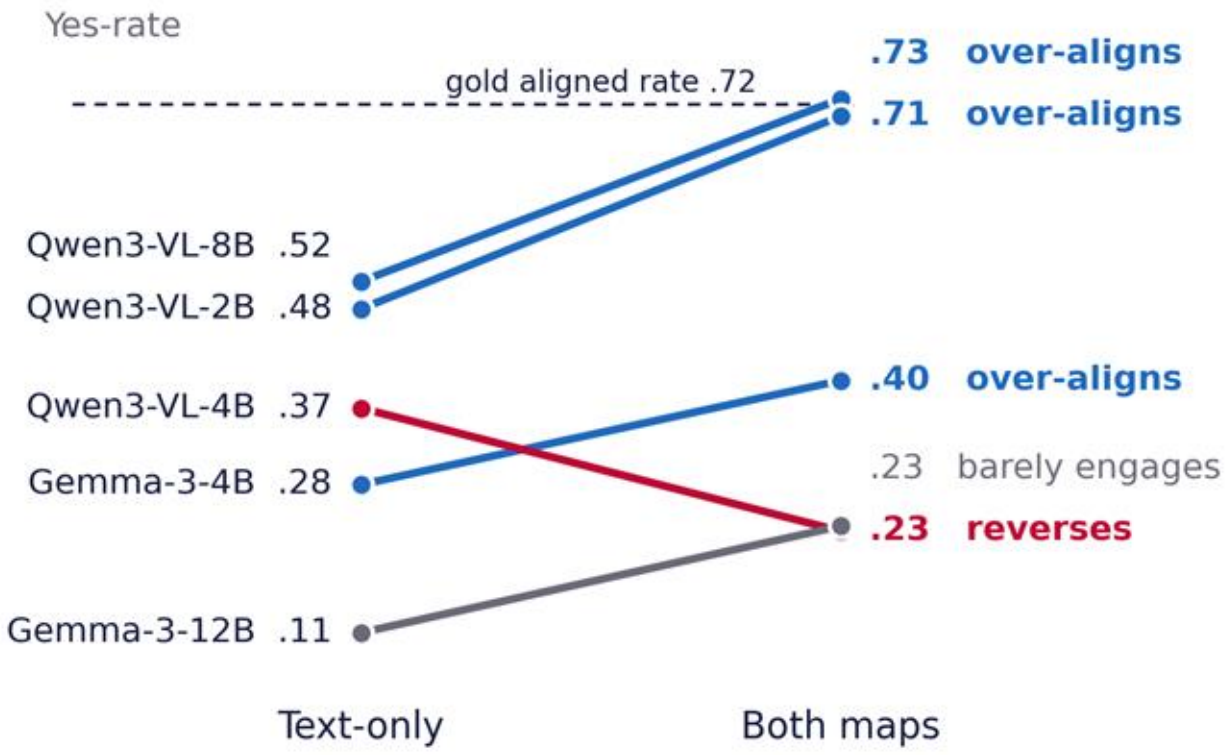
REPEATED MENTION

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.

PART 3 · HVL AT THE DISCOURSE LEVEL

Ground truth that is perspective-conditioned

CONCLUSION

Four take-aways

1 Methodological

Perspectivist, reference-level annotation makes silent misalignment measurable. Two interpretations instead of one gold label.

2 Empirical

Human misunderstanding is rare (1.8%) but systematic: multiplicity discrepancies run 6× the base rate.

3 About models

The VLMs we tested conflate potential with established common ground — and are confidently wrong.

4 For this workshop

Discourse ground truth can be perspective-conditioned.

Grounding is a process, and it has a perspective view.

Utrecht University

Thank you

— and, in the spirit of this workshop, thank you for your disagreement.

Nan Li · n.li@uu.nl · Utrecht University

LREC 2026
paper
lrec.elra.info

GMMT
dataset
HF

SIGDIAL 2026
paper
ACL Anthology

SINS
dataset + code
HF

Funded by NWO, as part of the AiNed project “Dealing with Meaning Variation in NLP” (AiNed Fellowship NGF.1607.22.002).

BACKUP

Full baseline grid · Qwen3-VL-8B at all four windows

Map access	Text	Acc.	F1 macro	R pos	R neg	Yes-rate
Text-only	curL	.442	.439	.261	.910	.213
	curT	.639	.604	.647	.616	.574
	startL	.533	.532	.409	.855	.336
	startT	.614	.591	.590	.677	.515
Both maps	curL	.627	.593	.633	.609	.566
	curT	.654	.618	.665	.625	.584
	startL	.694	.630	.769	.501	.694
	startT	.737	.671	.822	.518	.727
Giver-only	curL	.626	.586	.650	.565	.590
	curT	.688	.632	.747	.536	.668
	startL	.723	.644	.827	.454	.749
	startT	.756	.669	.879	.436	.791
Follower-only	curL	.619	.569	.663	.504	.617
	curT	.653	.594	.716	.490	.658
	startL	.718	.632	.832	.422	.761
	startT	.743	.650	.872	.408	.794

n = 13,077. Best macro-F1 per map-access level is startT in every map condition; for text-only, curT very slightly beats startT.

SIGDIAL 2026, Table 1.