language and translation technology team

ContraGAND

Auditing and Repairing Gender Ambiguity Failures in LLMs with Neurosymbolic Contrastive Data Augmentation

Thomas Moerman[1], Eleni Gkovedarou[1,2], Janiça Hackenbuchner[1]

[1] LT3, Language and Translation Technology Team, Ghent University [2] CLiPS, University of Antwerp

We automate gender ambiguity detection. LLMs tend to read gender cues correctly only for a given word. If not given, they default to "ambiguous". We trained on contrastive pairs and fixed this.

1 The task

Can a model accurately detect a referent as ambiguous? Or masculine/feminine, where context provides a cue?

"…you to become a top influencer in your industry or niche."

vs.

"…you to become a top influencer in your industry or niche, sir."

Closed task: referent given

given the sentence and the word **influencer**

expected **masculine**, because of **sir**

Open task: referent not given

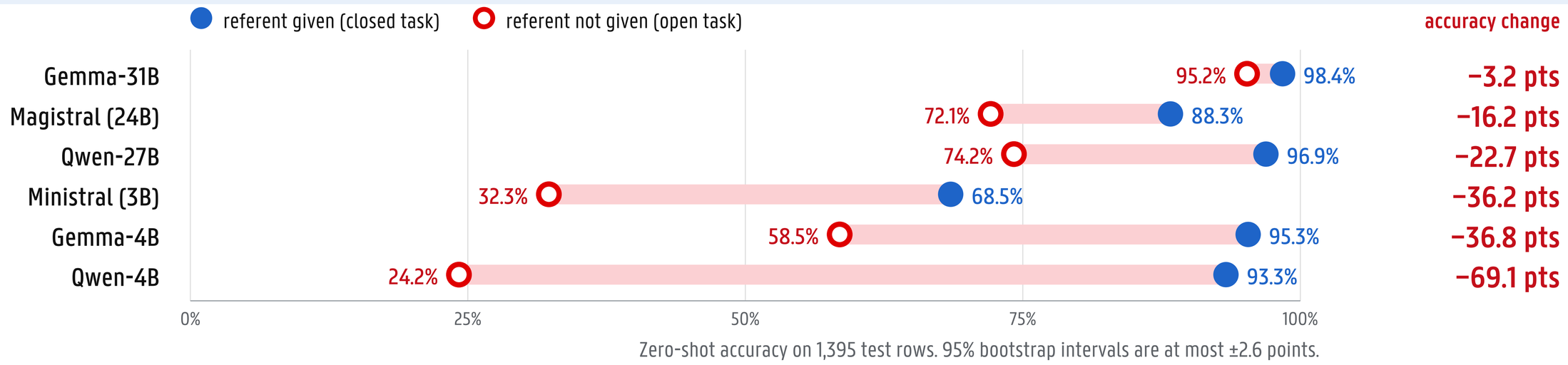
given the sentence only

Qwen-27B influencer: **ambiguous**, sir: masculine

Qwen-27B tags the cue **sir** but does not link it to the **influencer**.

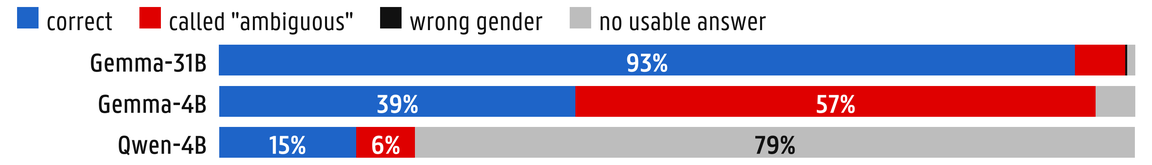
2 The gap

When the model must first find the referent, accuracy drops by 3 to 69 points.



Zero-shot accuracy on 1,395 test rows. 95% bootstrap intervals are at most ±2.6 points.

Why? Models are uncommitted, not misdirected: wrong answers are "ambiguous" or unusable, almost never the other gender.

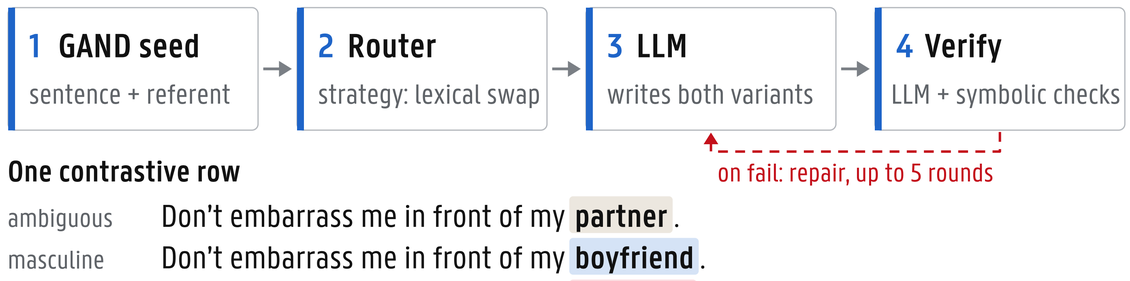


Open task, zero-shot: share of each model's 930 masculine and feminine test rows. Wrong gender: 0 to 2 rows.

3 The repair

Fine-tuning on contrastive pairs lifts both 4B models above 95%, level with the 31B model.

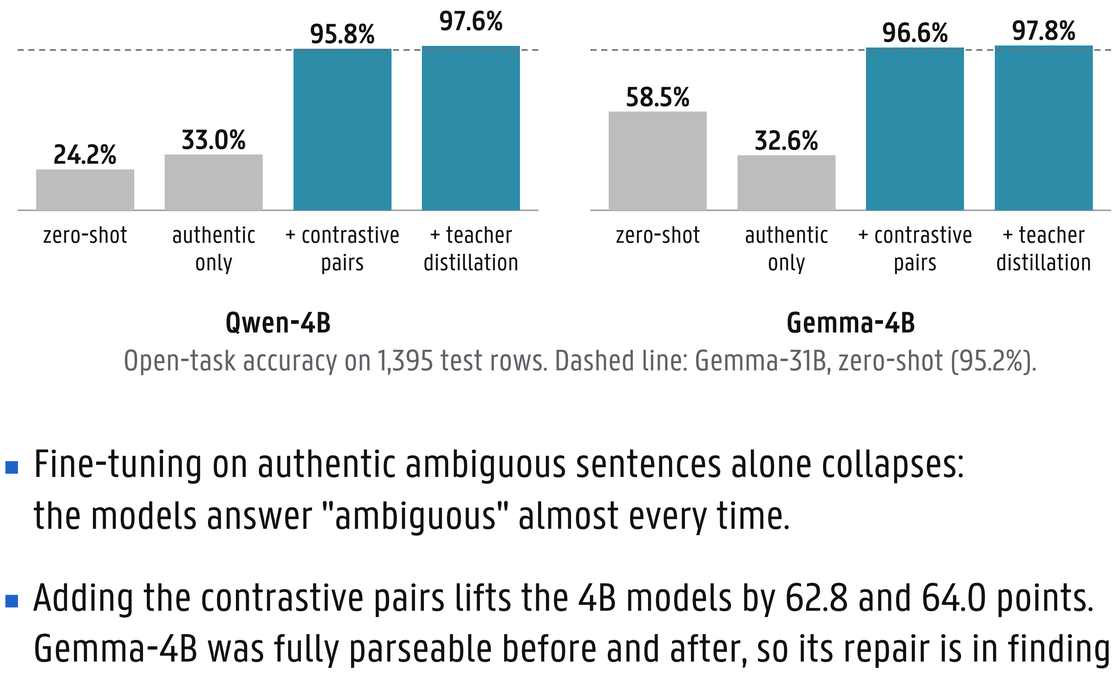
- **ContraGAND:** nine rewrite strategies add a gender cue to authentic ambiguous sentences, six next to the referent and three in another clause.
- An LLM adds both m/f variants. An LLM evaluator and symbolic validators can each veto, with up to five repair rounds. The test set is human-validated.



One contrastive row

ambiguous	Don't embarrass me in front of my **partner**.
masculine	Don't embarrass me in front of my **boyfriend**.
feminine	Don't embarrass me in front of my **girlfriend**.

4,852 sentences × 3 variants = 14,556 rows. Human audit kept 465 of 505 test sentences (92.1%).



Open-task accuracy on 1,395 test rows. Dashed line: Gemma-31B, zero-shot (95.2%).

- Fine-tuning on authentic ambiguous sentences alone collapses: the models answer "ambiguous" almost every time.
- Adding the contrastive pairs lifts the 4B models by 62.8 and 64.0 points. Gemma-4B was fully parseable before and after, so its repair is in finding referents and binding cues, not formatting.

1 Closed-task scores overstate ability.
Test the open task, where the model must find the referent itself.

2 The failure is under-commitment.
Models answer "ambiguous" or give no usable answer. It is not stereotype-driven.

3 Contrastive pairs are the training signal.
With them, a 4B model is competitive with a 31B model on this task.

Limits. English only, three labels, one fine-tuning seed. We measure a prerequisite skill, not downstream translation harm.

GHENT UNIVERSITY

Code, data, models, demo

hf.co/spaces/TomMoeras/contragand

github.com/TomMoeras/ContraGAND hf.co/collections/TomMoeras/contragand

firstname.lastname@ugent.be

Compute: VSC, funded by FWO and the Flemish Government. J.H. is funded by FWO project 1SH5V24N.

fwo TransTech EMNLP2026 BUDAPEST