

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

Thomas Moerman¹, Eleni Gkovedarou^{1,2} and Jani Hackenbuchner¹

¹LT³, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

²CLiPS, University of Antwerp, Lange Winkelstraat 40, Belgium
{firstname.lastname}@ugent.be

Abstract

Recent work on gender bias in the automatic translation of gender-ambiguous scenarios has highlighted the need for gender ambiguity detection. We address this gap with **CONTRAGAND**, a contrastive corpus of 4,852 natural gender-ambiguous English sentences, each minimally paired with a masculine and a feminine variant generated by a neurosymbolic pipeline. We use CONTRAGAND to perform and evaluate gender ambiguity detection in two settings: (i) a *closed task*, where the referent is given, and (ii) an *open task*, where it is not. For both tasks, we evaluate six open-weight large language models by prompting, and test selected models with fine-tuning and knowledge distillation. We show that (i) the closed task is saturated in current models, whereas (ii) the open task produces a gap that varies considerably by model. We find that two failure modes dominate: a *binding* failure (the model surfaces the referent but defaults to ‘ambiguous’ when the gender cue is cross-clause) and a *format* failure that, on local cues, is primarily under-commitment rather than a binding deficit (long unparseable reasoning chains on small models). Further analysis reveals that both are forms of over-prediction of the ‘ambiguous’ class rather than stereotype-driven misclassification. While explicit prompts and decoding interventions do not close the gap, both fine-tuning on the augmented contrastive variants and distillation from a 31 B teacher are effective at teaching a 4 B student to successfully perform gender ambiguity detection.

 github.com/TomMoeras/ContraGAND

 Dataset · closed- / open-task models · Demo

1 Introduction

Years of research show that Large Language Models (LLMs) and Machine Translation (MT) systems exhibit and exacerbate gender biases (Sun et al., 2019; Stanczak and Augenstein, 2021; Kotek et al., 2023; Vanmassenhove, 2024; Savoldi et al., 2025).

EXAMPLE

“Don’t embarrass me in front of my *partner*.”

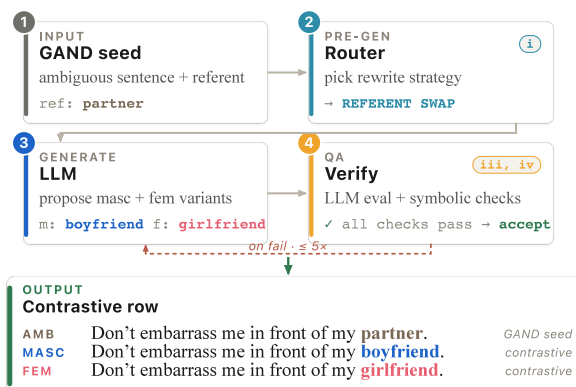


Figure 1: Overview of the CONTRAGAND generation pipeline. The full pipeline is explained in Section 3.1.

State-of-the-art models rather reliably resolve intra- and inter-sentential coreference for (gendered) pronouns (Lee et al., 2017; Otmazgin et al., 2023; Le and Ritter, 2023), as in “The *doctor* finished *her* shift and went home.”, where *her* disambiguates *doctor* as feminine. Without the gendered pronoun, however, the referent’s gender can no longer be inferred. This source-text neutrality proves problematic for translation: when the target language has grammatical gender, the system must commit to one and typically falls back on stereotypes or a masculine default (Hackenbuchner et al., 2025; Mastromichalakis et al., 2025). Arbitrarily marking entities with a gender, even if not implied in the source (Rarrick et al., 2023), perpetuates occupational stereotypes and representational harm (Blodgett et al., 2020).

Prior gender-coreference benchmarks (Zhao et al., 2018; Rudinger et al., 2018; Webster et al., 2018; Levy et al., 2021) measure referent-pronoun attribution, as in “The *physician* hired the *secretary* because *he* was overwhelmed with clients” (Zhao et al., 2018). Two gaps remain: these benchmarks have not been tested on purely ambiguous data (without a pronoun to resolve), and they

pre-define the referent-pronoun pair rather than asking the model to identify referents bottom-up. We address both with a multifaceted pipeline for gender *ambiguity* detection: deciding, for a referent in a sentence, whether context provides a cue that marks it as unambiguously masculine or feminine, or whether the referent’s gender remains genuinely ambiguous (e.g., in “My *client* has not been charged with a thing!”, the gender of *client* is truly ambiguous).

To address the first gap, we present **CONTRAGAND**: a contrastive corpus of 4,852 natural gender-ambiguous English sentences, each minimally paired with a masculine and a feminine variant generated by a neurosymbolic pipeline. The ambiguous sentences are authentic data seeded from GAND (Hackenbuchner et al., 2026)¹. The minimal contrastive pairs have been generated using nine disambiguation strategies, and the test split has been human-audited. An overview of how these contrastive pairs are produced is shown in Figure 1. We leverage CONTRAGAND to detect gender ambiguity, primarily to identify genuinely ambiguous cases in the authentic data. The model should abstain when no cue is present and commit to a contrastive variant (masculine or feminine) when one is. The masc./fem. accuracy scores serve as a sanity check, confirming the model abstains only when warranted rather than defaulting to “ambiguous”. If a referent is provided in the detection process, this is dubbed a **closed task**.

To address the second gap, we further evaluate an **open task** in which the model must first enumerate the referent(s) in the sentence and then assign each one a gender label, exercising self-enumeration and cross-clause cue binding. We identify and treat the closed-to-open gap as a **fairness diagnostic**: a proxy for three skills a downstream system (e.g., translation into a gendered language) would need on these sentences: producing well-formed outputs, finding the referents itself, and linking cross-clause gender cues back to them. The gap is closed by fine-tuning on CONTRAGAND’s contrastive variants, optionally with multi-referent teacher distillation.

To summarise, our contributions are: (i) CONTRAGAND, a 4,852-source $\times$ 3-variant contrastive corpus for English referent-gender ambiguity, used to benchmark six open-weight LLMs; (ii) the closed-to-open gap as a fairness diagnostic, de-

composed into binding- and format-failure modes; and (iii) fine-tuning on its augmented contrastive variants, optionally with teacher distillation, which closes the open-task gap on a 4 B student.

2 Related work

Gender bias is systematic prejudice or discrimination based on gender (Friedman and Nissenbaum, 1996). It has been documented in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Zhao et al., 2019) and downstream tasks including machine translation (Savoldi et al., 2021), coreference (Rudinger et al., 2018; Webster et al., 2018; Zhao et al., 2018), classification (Dixon et al., 2018), generation (Sheng et al., 2021), abusive-language detection (Park et al., 2018), speech recognition (Tatman, 2017), and parsing (Garimella et al., 2019), with downstream propagation risking reinforcement of harmful stereotypes (Sun et al., 2019).

A related challenge is **gender ambiguity**, where the text does not clearly specify a referent’s gender and models should preserve ambiguity rather than infer gender from occupational or social stereotypes. Existing benchmarks (WinoBias (Zhao et al., 2018), Winogender (Rudinger et al., 2018), GAP (Webster et al., 2018), BUG (Levy et al., 2021)) study this problem using ambiguous pronoun relations and counter-stereotypical examples. More recent benchmarks assess downstream model bias, mostly in translation, but do not provide an automated gender ambiguity detection task (Rarrick et al., 2023; Kotek et al., 2023; Garg et al., 2024; Hackenbuchner et al., 2025; Mastromichalakis et al., 2025). We address this gap by constructing a contrastive corpus on top of two existing technical foundations: counterfactual data augmentation (CDA) and LLM-based data generation.

CDA constructs minimal pairs by swapping protected-attribute terms in place. Lu et al. (2020) introduced gender-swap CDA (*he* $\leftrightarrow$ *she*) for bias mitigation, and Kaushik et al. (2020) generalised it to annotator-written minimal-pair edits that improve downstream robustness. Yet CDA usually applies a single substitution rule (typically a pronoun swap), thereby underrepresenting other devices that mark referential gender in English, such as titles, vocatives, kinship terms, and adjectival modifiers. A parallel line uses **LLMs as data generators**. Self-Instruct (Wang et al., 2023) and Unnatural Instructions (Honovich et al., 2023) boot-

¹<https://huggingface.co/datasets/jhacken/GAND>

strap large training corpora from a small human-written seed, approaching human-written data quality after light filtering, while Constitutional AI (Bai et al., 2022) and Self-Refine (Madaan et al., 2023) add a critique-and-revise loop that improves outputs against a written rubric. However, Zheng et al. (2023) show that LLM-as-a-judge, though broadly aligned with human preference, inherits systematic biases on length, position, and self-preference, so the judge alone is not a sufficient quality signal. Garcez and Lamb (2023) argue that hybrid neurosymbolic systems, where neural generation is constrained and audited by symbolic rules, can deliver both flexibility and verifiable quality. Building on these two complementary lines, we combine an extended CDA methodology with LLM-based generation to construct CONTRAGAND.

3 Methodology

3.1 Creating CONTRAGAND

The GAND seed. The Gender-Ambiguous Natural Data resource² (GAND; Hackenbuchner et al., 2026) is a corpus of 5,047 authentic English sentences drawn from natural text resources. Each sentence is gender-ambiguous with respect to a certain *referent*. Role and occupation referents are the noun classes most consistently associated with stereotype-aligned bias in NLP (Sun et al., 2019; De-Arteaga et al., 2019). To ensure ambiguity, GAND sentences are automatically and manually filtered to exclude gender markers (e.g., pronouns, kinship relations)³.

Contrastive variant generation. GAND is single-class: every sentence is gender-ambiguous by construction. We extend it into a contrastive benchmark by generating masculine and feminine minimal-pair variants for each ambiguous source sentence, building on the LLM-as-data-generator paradigm (Bai et al., 2022; Wang et al., 2023; Honovich et al., 2023) with two adaptations. Extending an ambiguous dataset with masc./fem. variants might seem counterproductive in gender-ambiguity detection. However, the unambiguous variants are there precisely so we can detect a model that would simply label everything ambiguous: without them, a model that always answers “ambiguous” would score perfectly on the ambiguous data, and its failure would go unnoticed.

²<https://huggingface.co/datasets/jhacken/GAND>

³The full list of excluded gender markers is available in the GAND repository: <https://github.com/jhacken/GAND>.

(i) A strategy taxonomy for wider linguistic coverage. English marks referent gender through several devices: pronouns, titles, vocatives, kinship terms, and adjectival modifiers (Sun et al., 2019; Stanczak and Augenstein, 2021). We define a nine-strategy taxonomy, split into two binding regimes. Six *local cues* place the gender signal adjacent to the referent: LEXICAL SWAP (replace the referent with its gendered lexical pair, drawn from Bartl and Leavy, 2022 and Papadopoulos et al., 2025), PRO-NOUN REWRITE, POSSESSIVE TAG, DISTAL NOUN CUE, GENDER ADJECTIVE, and DASHED APPOSITIVE. Three *cross-clause cues* place the gender signal across a clause boundary, requiring the model to bind the cue back to the referent: KIN APPOSITIVE, TITLE APPOSITIVE, and VOCATIVE ADDRESS, e.g., “Oh, right, *sir*, you’re a *cook*.”, where *sir* must be bound back to *cook*. Each row is generated under whichever strategy the sentence most naturally supports, widening the coverage that single-rule counterfactual augmentation cannot reach (Section 2). We exploit the local-vs-cross-clause split as a diagnostic in Section 4.2. Appendix A contains full definitions and worked examples.

(ii) A neurosymbolic verification layer. Given that the LLM-as-a-judge paradigm can itself exhibit systematic biases (Zheng et al., 2023), we treat the LLM evaluator as merely one signal among several. Four QA mechanisms can veto an LLM-accepted output. (i) A *symbolic pre-generation router* forces a strategy when sentence structure makes one canonical. (ii) *Cross-row rotators* prevent lexical repetition across a batch. (iii) An *LLM evaluator* scores five rubric dimensions with hard-fail flags. (iv) *Symbolic post-validators* check edit-distance, formatting parity, variant-distinctness, referent preservation, and changes in parts-of-speech (POS). Failed rows re-enter an evaluate-and-repair loop for up to five rounds, with feedback combining the LLM critique and the symbolic checker output. This follows the iterative-refinement pattern of Madaan et al. (2023) and Bai et al. (2022), and instantiates the third-wave neurosymbolic programme (Garcez and Lamb, 2023).⁴ Generation uses Claude Opus 4.6 (Anthropic, 2026). Appendix B explains the full

⁴We use the term ‘neurosymbolic’ in the loosely coupled sense of Kautz (2022): the symbolic layer derives new symbols from a curated lexicon and routes each item through a priority-ordered decision procedure, and the coupling is bidirectional because deterministic validators can veto the LLM evaluator and their violation messages feed the next generation attempt.

Split	Source sentences	Total rows
Train	3,902	11,706
Validation	485	1,455
Test (audited)	465	1,395
Total	4,852	14,556

Table 1: CONTRAGAND splits. Each source sentence yields three rows: the ambiguous original plus a masculine and a feminine minimal-pair variant. The test split is manually cleaned.

Alias	Checkpoint	Size
<i>Medium (24–31 B)</i>		
Gemma-31B	Gemma-4-31B-it (Google DeepMind, 2026)	31 B
Qwen-27B	Qwen3.6-27B (Qwen Team, 2026b)	27 B
Magistral	Magistral-Small-2506 (Mistral AI, 2025)	24 B
<i>Small (3–4 B)</i>		
Gemma-4B	Gemma-4-E4B-it (Google DeepMind, 2026)	4 B
Qwen-4B	Qwen3.5-4B (Qwen Team, 2026a)	4 B
Ministral	Ministral-3.3B-2512 (Mistral AI, 2026)	3 B

Table 2: Three model families evaluated with a medium and a small variant each. We refer to each model by its short alias throughout the paper. Full HuggingFace checkpoint URLs are in the footnote of Section 3.2.

pipeline in detail.

CONTRAGAND comprises three splits (Table 1). Every row carries the source sentence, both variants, a strategy label (Appendix A), and an annotated target referent, so the same row supports both closed and open evaluation (Section 3.2). Two annotators with backgrounds in linguistics and translation studies independently audited disjoint halves of the 505 generated test rows (253 and 252), following the same accept/reject rubric as the LLM verifier: the gender-marking edit must be correct, and the resulting sentence must be grammatical and natural. They retained 92.1 % (465 of 505) of the audited test set, compared with the verifier’s 97.0 % acceptance on the same rows. Raw agreement between each annotator and the verifier is 0.874 and 0.940 (mean 0.907). We release both annotators’ per-row labels with the corpus. For distillation (Section 4.3), we generate Gemma-31B’s open-task multi-referent enumeration with reasoning on the train and validation splits and use it as silver supervision.

3.2 Experimental setup

Two task framings. We evaluate every model on the same test sentences under two framings (Figure 2): the **closed task** supplies a referent, and the model returns one label from {masculine, feminine, ambiguous}. The **open task** supplies only the sentence, and the model must enumerate ev-

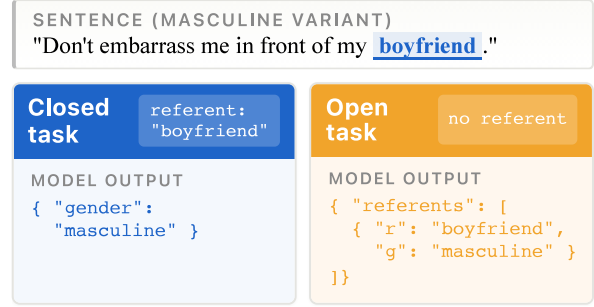


Figure 2: Closed and open task framings on the same input, applied to the masculine variant of a contrastive row produced by the pipeline in Figure 1.

ery possible referent and label each, exercising self-enumeration and cue binding. The primary metric is ambiguous-class F1. The masculine and feminine metrics guard against a model that labels everything *ambiguous*. Verbatim prompts are in Appendix C.

Models. We evaluate six open-weight LLMs from three model families (Table 2), with one model per family in each of two brackets: *medium* (24–31 B) and *small* (3–4 B). Qwen-4B is also the student we fine-tune in Section 4.3. Appendix D contains inference details and HuggingFace checkpoint URLs.

Main-text results are reported in a **zero-shot** setting: only the system prompt and the test query are passed to the model. Evaluation uses the 1,395 (465×3) long-form rows of the audited test split. Accuracies are computed on the full 1,395-row denominator throughout. Parse errors and missing responses count as incorrect. Few-shot results are in Appendix E.

To close the gap exposed in Section 4.2, we fine-tune Qwen-4B and Gemma-4B with QLoRA (Hu et al., 2022) under the conditions in Table 3. The closed-task series compares **FT-GAND** (authentic GAND only, single-class) to **FT-CONTRA** (FT-GAND plus our masculine and feminine variants) to isolate the contribution of synthetic data. The open-task series mirrors this and adds **FT-MULTI**, a teacher-student condition in which the student learns Gemma-31B’s successful multi-referent enumeration. Full hyperparameters are in Appendix D.

4 Results

4.1 Closed-task results

The closed task is saturated (Table 4). Four of the six evaluated models achieve a macro-F1 of at least 0.93 in zero-shot, including the 4 B mod-

Condition	Task	Supervision
FT-GAND	closed	ambiguous sources only
FT-CONTRA	closed	FT-GAND + masc./fem. variants
FT-GAND	open	ambiguous sources only
FT-CONTRA	open	FT-GAND + masc./fem. variants
FT-MULTI	open	FT-CONTRA rows with teacher multi-referent enumeration

Table 3: Fine-tuning conditions for the Qwen-4B and Gemma-4B students. FT-MULTI distills Gemma-31B as teacher. The others use canned per-class reasoning.

Closed Task			
Model	Acc	mF1	ambF1
Gemma-31B	0.984	0.984	0.979
Qwen-27B	0.969	0.969	0.958
Magistral	0.883	0.887	0.854
Gemma-4B	0.953	0.954	0.936
Qwen-4B	0.933	0.958	0.968
Ministral	0.685	0.718	0.757

Table 4: Closed-task zero-shot results (accuracy, macro-F1, ambiguous-class F1) across all six models, on the final human-reviewed test split (465 source rows, 1,395 long-form rows). 95 % bootstrap confidence intervals (CIs) are tabulated in Appendix F. Few-shot results are in Appendix E.

els Qwen-4B at 0.933 accuracy and Gemma-4B at 0.953. While Gemma-31B reaches macro-F1 0.984 and Qwen-27B 0.969, Magistral (0.887) and Ministral (0.718) fall below 0.90. The ordering does not cleanly follow parameter count. Magistral (24 B) lands below the two 4 B models in accuracy. Most of the current open-weight LLMs tested solve the closed framing on this dataset. The saturated closed task therefore serves as a validity control, namely it shows the labels are learnable.

Our primary metric, ambiguous-class F1 (ambF1), exceeds 0.93 zero-shot for the top four models and broadly mirrors the macro-F1 ordering. Magistral and Ministral show more pronounced drops, further discussed in Section 5.2.

The per-class bottleneck is *masculine recall*: zero-shot masculine F1 falls below ambiguous F1 for two of the small models (Qwen-4B, Ministral; per-class metrics in Table 27, Appendix K). Specifically, for masculine rows, Ministral predicts “ambiguous” rather than the correct “masculine” class, while Qwen-4B’s masculine misses are mostly unparseable responses rather than wrong labels. Per-strategy breakdowns, few-shot effects, and the ambiguous-only ablation are in Appendices G, E, and H.

Open Task			
Model	Acc	mF1	ambF1
Gemma-31B	0.952	0.956	0.942
Qwen-27B	0.742	0.757	0.726
Magistral	0.721	0.741	0.725
Gemma-4B	0.585	0.580	0.628
Qwen-4B	0.242	0.356	0.548
Ministral	0.323	0.390	0.550

Table 5: Open-task zero-shot results (accuracy, macro-F1, ambiguous-class F1). The model must first identify all referents in the sentence, then classify each. A row is correct only if every gold-labelled referent is found and classified correctly. 95 % bootstrap CIs are tabulated in Appendix F. Few-shot results are in Appendix E.

4.2 Open-task gap

We rerun all six models on the same 1,395 rows (verified test set) under the open framing. Open-task accuracy ranges from 0.952 for Gemma-31B to 0.242 for Qwen-4B (Table 5). Throughout this section, we distinguish two failure modes. A *binding failure* is a response that parses cleanly and lists the gold referent but labels it ambiguous because the cue in another clause is not linked back to it. A *format failure* is a response that never yields a parseable answer. Table 6 separates the two for every model.

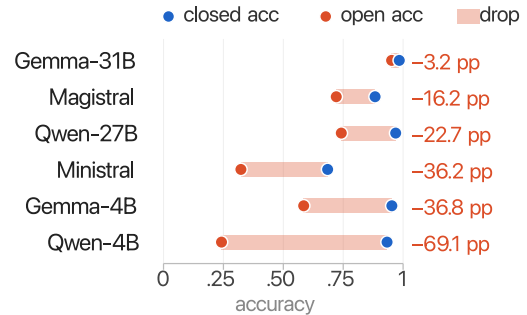


Figure 3: Closed-to-open accuracy drop per model. Each row pairs the closed-task accuracy (blue) with the open-task accuracy (red), labelled in percentage points.

Figure 3 isolates the closed-to-open accuracy drop, which ranges from -3.2 percentage points (pp) (Gemma-31B) to -69.1 pp (Qwen-4B). Concretely, a Qwen-4B response on this test population would carry an unjustified or under-committed gender outcome on roughly three of every four sentences, against about one in twenty for Gemma-31B. Two observations can be made. First, closed-task accuracy overstates the performance of any model that must surface the referent itself. Gemma-4B illustrates the pattern most cleanly because it has no format-failure confound. It scores 0.95 on the closed task and emits a clean response on ev-

Model	parse %	found %	Acc	w-match
Gemma-31B	99.6	99.2	0.952	0.960
Qwen-27B	100.0	97.5	0.742	0.761
Magistral	98.8	95.2	0.721	0.758
Gemma-4B	100.0	96.6	0.585	0.606
Qwen-4B	30.7	28.5	0.242	0.849
Ministral	67.2	52.1	0.323	0.619

Table 6: Open-task funnel decomposition (zero-shot). *parse %* is the share of rows producing a parseable response. *found %* is the share where the gold referent was in the model’s list. *Acc* is the headline accuracy (Table 5). *w-match* is the gender-correct rate conditional on the gold referent being matched. Qwen-4B’s 0.849 within-matched accuracy is the key signal: when it parses and finds the referent, it classifies as well as the other small models on the closed task.

ery test row, but drops to 0.585 on the open task. The entire 36.8 pp loss comes from failing to link the gender cue to the referent. Second, Gemma-31B’s 3.2 pp loss shows that capable models can enumerate and bind referents without catastrophic degradation. Between these extremes, the gap takes two shapes (Table 6). On the medium models, it comes from responses that find the gold referent but label it ambiguous. On the small models, it comes from unparseable output. In these cases, binding is a secondary contributor, visible only when the model does parse.

However, open-task accuracy in isolation obscures which *kind* of failure dominates. To get a row right, the model must do three things in order. First, produce a clean response. Second, list the gold referent. Third, label the referent’s gender correctly. Table 6 reports the success rate of each model at each step, and three distinct patterns emerge.

Gemma-31B loses almost nothing along the way: 95 % of its predictions clear all three steps (Table 6). Qwen-27B, Magistral, and Gemma-4B parse cleanly and surface the gold referent on roughly 95–97 % of rows, but mislabel the gender on 24–39 % of those matched rows (100 minus the *w-match* column). In these cases, the model finds the right referent, but cannot bind a cross-clause cue to it.

Qwen-4B fails differently. Only 30 % of its outputs are even parseable (as further illustrated in Figure 4), but among those that do parse, the gender-correct rate is 0.849, second only to Gemma-31B’s 0.960. Among the rows it does parse, the model is 95.7 % accurate when the gender cue sits next to the referent and only 65.7 % when the cue sits

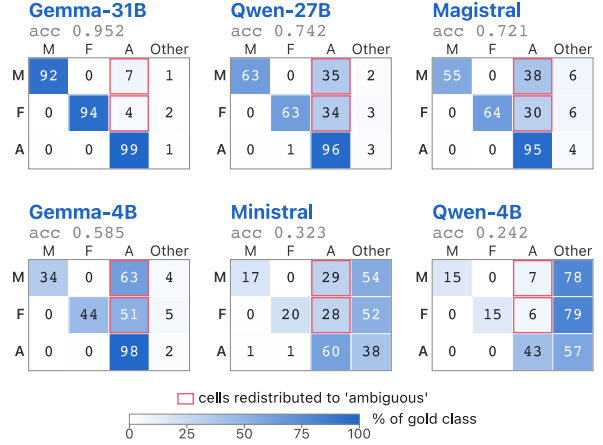


Figure 4: Open-task confusion matrices (zero-shot), row-normalised. *Other* combines parse-error and not-found predictions.

in a different clause (Appendix I, Table 25). Parse rates themselves also differ by cue locality (36.5 % local against 23.8 % cross-clause, $\chi^2_1 = 25.8$, $p < 10^{-6}$). A parser-rescue pass recovers 0 of 967 unparseable rows as valid embedded JSON, and the mean response length on those rows is 14,429 characters, ruling out token-budget truncation. Ministral fails on both axes, with 33 % parse errors while a further 20 % of all rows match the referent but carry a wrong gender label.

Local vs. cross-clause cues. Grouping the nine strategies into *local cues* (gender signal adjacent to the referent) and *cross-clause cues* (binding requires coreference across a clause boundary), every model is more accurate on local than cross-clause cues. On parseable-output models, the row-weighted local-to-cross-clause drop widens with the headline gap, from 3.6 pp on Gemma-31B to 32.6 pp on Gemma-4B (Table 20). The penalty concentrates on VOCATIVE ADDRESS, where the cue (*sir/ma’am*) sits at the sentence boundary and must be bound back to the role referent: Gemma-31B holds 0.935 while Qwen-27B falls to 0.381. A representative VOCATIVE ADDRESS row is “Oh, right, *sir*, you’re a *cook*.”, where the cue *sir* refers across a clause boundary to *cook*. Weaker models flag *sir* as masculine in isolation but still leave *cook* labelled ambiguous. TITLE APPOSITIVE is the easiest cross-clause cue because the appositive typically sits within the same noun phrase as the referent (“the doctor, Mr. Chen”). Per-strategy detail is in Appendix G.

Figure 4 shows row-normalised open-task confusion matrices. Gemma-31B is near-diagonal (≥ 0.92 on every true class). Qwen-27B and

Gemma-4B redistribute masculine and feminine mass into the ambiguous column (≈ 0.35 and ≈ 0.57 respectively). Ministral shows the same redistribution with an additional parse-error tail. An asymmetric m/f residual in these matrices is discussed in Section 5.4.

The gap is not closed by prompt or decoding interventions. Neither explicit binding prompts nor structural decoding interventions close the open-task gap on the three small models (Table 7). These fixes lift accuracy by at most 21 pp without altering the classification posterior. Schema-constrained decoding eliminates parse errors (parse-err 0.4 %) but caps accuracy at 0.417. The model now answers cleanly, but it answers ambiguous on almost every row (ambiguous recall 0.92). The residual gap is a capability ceiling that motivates the fine-tuning recipe in Section 4.3. The full nine-fix sweep is in Appendix I.

Intervention	parse-err %	Acc	amb recall
Baseline (zero-shot)	69.3	0.242	0.426
Repetition penalty 1.3	34.1	0.450	0.634
GUIDED_JSON (schema)	0.4	0.417	0.923

Table 7: Decoding interventions on Qwen-4B (open task, zero-shot). Full nine-fix sweep in Appendix I.

4.3 Closing the open-task gap

Condition	Acc	mF1	mR	fR	aR
Qwen-4B (ICL zs)	0.242	0.356	0.153	0.146	0.426
FT-GAND	0.330	0.191	0.009	0.030	0.953
FT-CONTRA	0.958	0.973	0.970	0.970	0.935
FT-MULTI	0.976	0.978	0.970	0.976	0.983
Gemma-4B (ICL zs)	0.585	0.580	0.335	0.443	0.976
FT-GAND	0.326	0.179	0.000	0.019	0.959
FT-CONTRA	0.966	0.980	0.974	0.974	0.948
FT-MULTI	0.978	0.982	0.966	0.978	0.989
Gemma-31B (ICL zs)	0.952	0.956	0.920	0.948	0.987

Table 8: Open-task LoRA fine-tuning conditions for both students, with recall by class. mR/fR/aR = masculine/feminine/ambiguous recall. ICL zs = the in-context zero-shot baseline. 95 % bootstrap CIs for Qwen-4B in Appendix F.

We fine-tune two of the three small models with QLoRA and serve them via vLLM (Kwon et al., 2023) at inference time: Qwen-4B, because it demonstrated the biggest drop from the closed task and performed worst on the open task, and Gemma-4B, because it produces no format failures, so its gains isolate enumeration and cue binding from

output formatting (Section 4.2). Full hyperparameters are in Appendix D. We focus on the open task because the size-dependent gap appears there. The closed task is largely saturated on this benchmark, and the same training recipe matches the 31 B teacher’s closed-task accuracy at one-eighth the parameter count for both students. Closed-task fine-tuning details and table are in Appendix J.

We design three open-task data conditions. **FT-GAND** trains on authentic GAND source rows only (every gold label is ambiguous). **FT-CONTRA** adds our masculine and feminine contrastive variants (3,902 sources $\times$ 3 variants = 11,706 examples), each supervised on the single annotated referent. **FT-MULTI** layers teacher distillation on top, training the student on complete multi-referent enumerations produced by Gemma-31B. All conditions are trained for both students.

Table 8 and Figure 5 show the results. FT-GAND collapses to predicting ambiguous on both students: accuracy 0.330 for Qwen-4B and 0.326 for Gemma-4B, with masculine and feminine recall *below* the zero-shot baseline. Naive single-class fine-tuning on GAND, therefore, is unsuccessful. Instead, contrastive variants are needed as a training signal. FT-CONTRA jumps to 0.958 on Qwen-4B and 0.966 on Gemma-4B. For Gemma-4B this is a 0.381 rise from 0.585 zero-shot that exceeds even its closed-task zero-shot accuracy of 0.953 (Table 4). Since the model emits parseable output on every test row both before and after fine-tuning, the gain reflects enumeration and cue binding alone. At baseline (ICL zs) it answers cleanly but labels masculine and feminine referents ambiguous (recalls 0.335 and 0.443). After fine-tuning, both recalls reach 0.974 with no formatting to fix. The contrastive variants therefore repair the binding deficit itself, not merely output formatting. FT-MULTI reaches accuracies of 0.976 and 0.978 respectively, surpassing the teacher’s 0.952 open zero-shot baseline (for Qwen-4B: non-overlapping bootstrap CIs, significant McNemar)⁵. As FT-MULTI is trained on the teacher’s labels, it naturally also inherits the teacher’s mistakes (see Section 6). We read this as a competitive student at roughly one-eighth the parameter count, not as evidence of capability beyond the teacher.⁶

In sum, contrastive variants provide the dominant signal (+62.8 pp on Qwen-4B and +64.0 pp

⁵Bootstrap CIs and McNemar statistics are in Appendix F.

⁶The comparison against the teacher’s few-shot accuracy is in Appendix E.

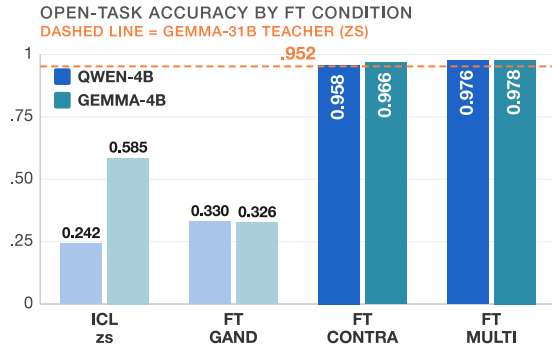


Figure 5: Open-task accuracy under the LoRA fine-tuning conditions for both students. The dashed line marks the Gemma-31B teacher’s zero-shot reference (0.952).

on Gemma-4B from FT-GAND to FT-CONTRA). The additional +1.8 pp for Qwen-4B from FT-MULTI is attributable to the model’s ability to find all the referents. Not-found errors drop from 42 to 6 (counted over the released per-row predictions), accounting for most of the gain. Masculine and feminine recall barely change (0.970 $\rightarrow$ 0.970, 0.970 $\rightarrow$ 0.976). Full McNemar statistics are in Appendix F.

Beyond the open-task gap itself, two residual gender biases are visible across the evaluated models. Per-model breakdowns are in Figure 6 and Appendix L. On rows carrying a gender cue, accuracy is slightly higher on feminine than on masculine referents in 10 of 12 model-task cells, with magnitudes up to +10.8 pp on Gemma-4B open and +16.1 pp on Ministral closed (the counter-cells are Qwen-27B closed at -1.1 pp and Qwen-4B open at -0.7 pp). On truly ambiguous rows, the few definite-gender predictions that do occur cluster on occupational and caregiving stereotypes (*champion*, *soldier*, *scientist* on the masculine side, *housekeeper*, *nanny*, *homemaker* on the feminine side). The most stereotype-prone model commits 9 such predictions across 465 ambiguous rows, and the open framing reduces this further (Gemma-4B 9 $\rightarrow$ 3, Qwen-4B 5 $\rightarrow$ 0, Ministral unchanged at 7 $\rightarrow$ 7).

5 Discussion

5.1 The closed-to-open gap as a fairness diagnostic

In the closed task, the model receives the referent and decides its gender from in-context cues, or absence thereof. Most current open-weight LLMs handle this reliably (Section 4.1). Ideally, however, a model should detect referents on its own (e.g.,

at the document level, in a translation pipeline) and infer their correct gender. The open task adds those deployment-relevant steps (Section 4.2). We read the closed-to-open accuracy delta as a fairness diagnostic: a proxy for the errors a model might make on the same sentences when it must identify referents itself. It tells us about the skills a downstream system would need (identifying referents and linking in-context cues to them, with well-formed output as a measurement precondition) rather than directly measuring the harm in deployment. The funnel decomposition in Table 6 and the decoding sweep in Table 7 spell out which of these skills each model lacks.

The following example shows the binding failure concretely. On “...you to become a top *influencer* in your industry or niche, *sir*.” (gold referent *influencer*, gold label masculine), Gemma-31B finds both candidate role nouns, links the address term *sir* back to *influencer*, and labels *influencer* masculine. Qwen-27B flags *sir* as masculine in isolation but does not link it back, and so leaves *influencer* labelled ambiguous (Appendix H, Table 21). The following subsections examine the errors driving the closed-to-open separation.

5.2 Uncommitted, not misdirected

The dominant open-task error (see Section 5.1) is the failure to link a present gender cue to the referent. The cue is detected (e.g., *sir* correctly flagged as masculine), but the model does not link it to the referent, and so the referent is labelled ambiguous. The pattern recurs across all models, visible in the confusion matrices of Figure 4.

A model that never abstains can still appear accurate when it commits to a gender, but it will rarely surface the ambiguous label, even when that label is the right answer. On CONTRAGAND, the open task shows that the models say ambiguous nearly as often on masculine-labelled sentences as on feminine-labelled ones (the residual asymmetry is quantified in Section 5.4). The systems do not prefer either binary gender. Most models simply do not reliably detect when ambiguous is right. The fine-tuning recipe in Section 4.3 restores balanced precision and recall on the ambiguous class without re-introducing a masculine-feminine asymmetry, supporting the same reading.

5.3 Under-commitment on local cues, residual binding deficit on cross-clause cues

We label the open-task error pattern in Section 4.2 *under-commitment*: the model’s tendency to refuse a clean answer or to default to ambiguous rather than commit to a gender. The phenomenon is a behavioural one rather than a lack of knowledge. It echoes the tendency to abstain that has been documented in instruction-tuned LLMs on sensitive-attribute prompts (Bai et al., 2022; Röttger et al., 2024; Bianchi et al., 2024), and runs opposite to confident stereotype-based commitment (Kotek et al., 2023; Wan et al., 2023). The small models err by abstaining rather than by following a stereotype.

The results in Section 4.2 reveal two different kinds of failure. When the gender cue appears next to the referent, the model links the two correctly whenever it answers. The drop occurs because the model refuses to answer at all. When the cue appears in a different clause, the model misses the link even when it answers. The two failures compound. The two error modes, therefore, have different causes. Under-commitment responds to simple decoding tweaks such as temperature or a repetition penalty, while cross-clause failure does not.

Qwen-4B’s parse failures fit the same picture. The model is reasoning at length about the gender choice without ever producing a clean answer (Appendix I). The other small models default to ambiguous, while Qwen-4B defaults to endless reasoning. Both are forms of over-predicting *ambiguous* under uncertainty. That a single training step (fine-tuning on the contrastive variants, Section 4.3) fixes both is the strongest evidence that they share a cause.

5.4 Residual gender bias

The two residual biases (see Section 4.3) appear in different settings, depending on whether the sentence carries a gender cue. If it does, the models are slightly more accurate on feminine referents. This aligns with findings by Manna et al. (2025), who showed that models pay more attention to feminine pronouns. This finding opposes the masculine default reported in earlier embedding work (Bolukbasi et al., 2016; Caliskan et al., 2017) and suggests that instruction tuning has overcorrected the residual rather than removed it. When no cue is present, the few guesses that do commit to a specific gender revert to older occupational and caregiving role-

gender stereotypes (Garg et al., 2018; Sun et al., 2019). The footprint is small enough that we caution against over-reading it, but the direction is consistent across models.

6 Conclusion

We presented a multifaceted pipeline for detecting gender ambiguity. Using CONTRAGAND’s naturally ambiguous sentences and the synthetic minimal masculine/feminine contrastive pairs added in this work, we evaluated six models on a closed and an open task. On the closed task, where a model is required to detect whether a given referent is gender-ambiguous within a sentence, all models perform well in the baseline setup. Therefore, we consider this task saturated. For the open task, a model must first detect a referent and then determine whether it is gender-ambiguous. Here, the gap varies considerably across models, is not stereotype-driven, and is dominated by under-commitment on local cues and a residual binding deficit on cross-clause cues, where models over-predict “ambiguous” rather than binding the available signal. Fine-tuning on contrastive variants, optionally augmented with teacher distillation, closes the gap on a student at one-eighth the teacher’s parameter count. In sum, this work presents an effective pipeline for both referent and gender (ambiguity) detection for small and medium-sized models.

Limitations

CONTRAGAND is English-only and built from C4/OpenSubtitles. Generalisation to other genres (e.g., biomedical, legal) and to grammatical-gender languages is open. In the latter, gender marking is morphologically pervasive rather than confined to a small set of cue devices, so the evaluation framing does not port directly. The label set is three-way (masculine, feminine, ambiguous) and does not cover singular they or other gender-neutral pronouns. Extending CONTRAGAND with additional gender categories is left to future work (Cao and Daumé III, 2020; Brandl et al., 2022).

Two of the three small models were fine-tuned end-to-end: Qwen-4B because it is the weakest open-task model, and Gemma-4B because its zero parse errors isolate binding from formatting (Section 4.3). No fine-tuning run was repeated with a different random seed, so the bootstrap CIs in Appendix F reflect sample-level rather than training-

run variance.

The open-task evaluator (Section 3.2) counts a row as a model miss if the model fails to match any referent through any of the three matching stages. This is conservative against the model. FT-MULTI is trained on Gemma-31B’s silver enumerations and can inherit wrong bindings on multi-referent edge cases: e.g., on “While the Japanese *fighter* flew off, Bill and a major from his barrack . . .”, the student labels *fighter* masculine, although the gold label is *ambiguous* (*his* refers to *Bill*, not *fighter*). The Section 4.3 parity result should therefore be read as “competitive at one-eighth the parameters” rather than as distilled capability beyond the teacher. Whether the recipe is teacher-agnostic (e.g., distilling from Qwen-27B or Magistral) remains an open question.

Finally, this paper measures the prerequisite capability of referent enumeration and cue binding rather than downstream harm: the intended two-stage use, in which the released detector labels each referent and a downstream system such as translation commits only where a cue exists, is the subject of an ongoing study.

Ethical considerations

This paper studies gender ambiguity detection in LLMs and proposes both a contrastive dataset (CONTRAGAND) and a fine-tuning recipe that improves model behaviour on ambiguous referents. We acknowledge several ethical considerations.

Label-set scope. Our three-class label set {masculine, feminine, ambiguous} is restricted to binary gender categories with an ambiguous abstention class. It does not represent non-binary gender identities, singular-they referents, or culture-specific gender systems. Extending the corpus and evaluation to include additional gender categories is an important direction for future work.

Stereotype residuals. Section 5.4 reports that even instruction-tuned models still occasionally guess gender based on occupational or relational stereotypes on truly ambiguous source rows. The fine-tuned student (FT-CONTRA, FT-MULTI) reduces this behaviour but does not eliminate it. Practitioners should not interpret our headline accuracies as evidence that the underlying model is stereotype-free outside the evaluated distribution.

Data sourcing. CONTRAGAND extends the publicly available GAND corpus (Hackenbuchner et al., 2026), which is drawn from OpenSubtitles

and C4. No new private text was collected. We manually audit the test split and disclose acceptance rates (Section 3.1).

Compute and environmental cost. The generation pipeline uses Claude Opus 4.6 (Anthropic, 2026). The fine-tuning runs use one node of 4× A100-80 GB for the wall-clock times listed in Table 15 (Appendix D), totalling at most five GPU-hours per condition. We did not run repeated seeds (Limitations). Doing so would multiply the compute cost linearly.

Use of AI. AI assistance was used in three ways in this work. First, Claude Opus 4.6 (Anthropic, 2026) served as a data generator within the contrastive-variant pipeline, as described in Section 3.1. Second, AI was used as a writing assistant for rewriting sentences. Third, AI was used to document the scripts and repository accompanying this paper. In all cases, the final wording and scripts are our own.

Reproducibility.⁷ Hyperparameters, prompts, and the demonstration-construction recipe are in Appendices D and C. CONTRAGAND is released at <https://huggingface.co/datasets/TomMoeras/ContraGAND> under the ODC-By licence, matching the GAND seed corpus.

Acknowledgments

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government. Janiça Hackenbuchner was funded by The Research Foundation – Flanders (FWO), research project 1SH5V24N (from 01.11.2023 until 31.10.2027).

References

- Anthropic. 2026. [Claude Opus 4.6 system card](#). System card, Anthropic.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. [Constitutional AI: Harmlessness from AI feedback](#). *arXiv preprint arXiv:2212.08073*.

⁷Code, per-row predictions, and the evaluation harness: <https://github.com/TomMoeras/ContraGAND>. Fine-tuned detectors: <https://huggingface.co/TomMoeras/contragand-qwen3.5-4b-closed> and <https://huggingface.co/TomMoeras/contragand-qwen3.5-4b-open>. Interactive demo: <https://huggingface.co/spaces/TomMoeras/contragand-demo>.

- Marion Bartl and Susan Leavy. 2022. [Inferring gender: A scalable methodology for gender detection with on-line lexical databases](#). In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 47–58, Dublin, Ireland. Association for Computational Linguistics.
- Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. [Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. [Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. [How conservative are language models? Adapting to the introduction of gender-neutral pronouns](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630. Association for Computational Linguistics.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. [Semantics derived automatically from language corpora contain human-like biases](#). *Science*, 356(6334):183–186.
- Yang Trista Cao and Hal Daumé III. 2020. [Toward gender-inclusive coreference resolution](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595. Association for Computational Linguistics.
- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnamurthy Kenthapadi, and Adam Tauman Kalai. 2019. [Bias in Bios: A case study of semantic representation bias in a high-stakes setting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. [Measuring and mitigating unintended bias in text classification](#). In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, pages 67–73, New York, NY, USA. Association for Computing Machinery.
- Batya Friedman and Helen Nissenbaum. 1996. [Bias in computer systems](#). *ACM Trans. Inf. Syst.*, 14(3):330–347.
- Artur d’Avila Garcez and Luís C. Lamb. 2023. [Neurosymbolic AI: the 3rd wave](#). *Artificial Intelligence Review*, 56(11):12387–12406.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. [Word embeddings quantify 100 years of gender and ethnic stereotypes](#). *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- Sarthak Garg, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao, and Matthias Paulik. 2024. [Generating gender alternatives in machine translation](#). *Preprint*, arXiv:2407.20438.
- Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. [Women’s syntactic resilience and men’s grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.
- Google DeepMind. 2026. [Gemma 4](#). Model collection.
- Janiča Hackenbuchner, Jasper Degrauwe, Arda Tezcan, and Joke Daems. 2026. [Explaining GAND: A resource on gender-ambiguous natural data & contrastive attribution](#). In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 245–267, Tilburg, The Netherlands. European Association for Machine Translation.
- Janiča Hackenbuchner, Eleni Gkovedarou, and Joke Daems. 2025. [GENDEROUS: Machine translation and cross-linguistic evaluation of a gender-ambiguous dataset](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 302–319, Vienna, Austria. Association for Computational Linguistics.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. [Unnatural instructions: Tuning language models with \(almost\) no human labor](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. [Learning the difference that makes a difference with counterfactually-augmented data](#). In *International Conference on Learning Representations*.
- Henry A. Kautz. 2022. [The third AI summer: AAAI Robert S. Engelmore Memorial Lecture](#). *AI Magazine*, 43(1):105–125.
- Hadas Kotek, Rikker Dockum, and David Sun. 2023. [Gender bias and stereotypes in large language models](#). In *Proceedings of The ACM Collective Intelligence Conference*, CI ’23, pages 12–24. ACM.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the*

- 29th Symposium on Operating Systems Principles (SOSP), pages 611–626. Association for Computing Machinery. vLLM.
- Nghia T. Le and Alan Ritter. 2023. [Are large language models robust zero-shot coreference resolvers?](#) *Preprint*, arXiv:2305.14489.
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. [End-to-end neural coreference resolution](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.
- Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. [Collecting a large-scale gender bias dataset for coreference resolution and machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. [Gender bias in neural natural language processing](#). In *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202. Springer.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with self-feedback](#). *Advances in Neural Information Processing Systems*, 36.
- Chiara Manna, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2025. [Are we paying attention to her? investigating gender disambiguation and attention in machine translation](#). In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 1–16, Geneva, Switzerland. European Association for Machine Translation.
- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. [Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32233–32249, Suzhou, China. Association for Computational Linguistics.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics.
- Mistral AI. 2025. [Magistral](#). *Preprint*, arXiv:2506.10910.
- Mistral AI. 2026. [Ministral 3](#). *Preprint*, arXiv:2601.08584.
- Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. [LingMess: Linguistically informed multi expert scorers for coreference resolution](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ben Papadopoulos, Cory Bedin, Sebastian Clendenning-Jiménez, and Z. K. Robinson. 2025. [Gender In Language Project: English](#). Accessed 2026-05-18.
- Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. [Reducing gender bias in abusive language detection](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.
- Qwen Team. 2026a. [Qwen3.5: Towards native multi-modal agents](#). Qwen Team blog.
- Qwen Team. 2026b. [Qwen3.6-27B: Flagship-level coding in a 27B dense model](#). Qwen Team blog.
- Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. [GATE: A challenge set for gender-ambiguous translation examples](#). In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES '23*, pages 845–854, New York, NY, USA. Association for Computing Machinery.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. [XSTest: A test suite for identifying exaggerated safety behaviours in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. [A decade of gender bias in machine translation](#). *Patterns*, 6(6):101257.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. [Gender bias in machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. [Societal biases in language generation: Progress and challenges](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

- Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.
- Karolina Stanczak and Isabelle Augenstein. 2021. [A survey on gender bias in natural language processing](#). *arXiv preprint arXiv:2112.14168*.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. [Mitigating gender bias in natural language processing: Literature review](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.
- Rachael Tatman. 2017. [Gender and dialect bias in YouTube’s automatic captions](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.
- Eva Vanmassenhove. 2024. [Gender bias in machine translation and the era of large language models](#). In *Gendered Technology in Translation and Interpreting*, pages 225–252. Routledge.
- Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. [“kelly is a warm person, joseph is a role model”: Gender biases in LLM-generated reference letters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. [Mind the GAP: A balanced corpus of gendered ambiguous pronouns](#). *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. [Gender bias in contextualized word embeddings](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and Chatbot Arena](#). *Advances in Neural Information Processing Systems*, 36.

A Strategy taxonomy

The nine disambiguation strategies used by the generation pipeline, plus the COMBINATION meta-strategy, are listed in Table 9.

Hard rules encoded in the system prompt. Disambiguate the referent’s gender, not any other word. Do not add a new sentence after a period. Both variants must differ from each other and from the source. Preserve formatting quirks (non-standard contractions). Pronouns must refer to the referent. When the referent is the speaker, apply a symmetric first-person swap (*I/me/my* → *he/him/his* or *she/her/her*). Prefer embedded edits over appended dashes. When the referent surface form is *you*, use *sir/ma’am* rather than rewriting *you* to *he/she*.

Worked examples per strategy. Table 10 shows one contrastive row per strategy, drawn from the audited test split. Apostrophe formatting follows the OpenSubtitles source.

B Synthetic Data Generation

Four independent quality-assurance mechanisms run on every row, and any of them can veto an output that the LLM has accepted. A *symbolic pre-generation router* emits a deterministic strategy hint when one of four signals fires (Table 11). *Cross-row rotators* distribute surname and relational-pair choices uniformly across a batch. An *LLM evaluator* scores five rubric dimensions with hard-fail flags. *Symbolic post-validators* apply deterministic checks and override LLM acceptance on mechanical failures. An autopilot loop runs evaluate → retry for up to five rounds, with feedback for each retry assembled from the previous round’s failures.

B.1 Symbolic pre-generation routing

A grounding module checks four signals in priority order before sending each row to the model, emitting at most one hint per row. Rows matching no signal receive no hint, and the LLM chooses freely.

Priority 1 is gated by a forbidden-pair list: (hero, heroine), (master, mistress),

Strategy	Description
LEXICAL SWAP	Replace the referent with its gendered pair (e.g., <i>partner</i> → <i>boyfriend/girlfriend</i>).
PRONOUN REWRITE	Swap an existing neutral pronoun (<i>they</i> → <i>he/she</i>).
POSSESSIVE TAG	Insert a bound possessive (<i>his / her</i>).
DISTAL NOUN CUE	Swap an upstream neutral noun (<i>person</i> → <i>man/woman</i>), with the referent left unchanged.
KIN APPOSITIVE	Insert a kin appositive (e.g., <i>the janitor, his brother</i>).
TITLE APPOSITIVE	Add <i>Mr./Mrs.</i> (Surname) as an appositive.
VOCATIVE ADDRESS	Vocative (<i>sir / ma'am</i>), only when the referent is the addressee.
GENDER ADJECTIVE	Prepend <i>male / female</i> .
DASHED APPOSITIVE	Dashed appositive containing <i>a man / a woman</i> .
COMBINATION	Two strategies applied together.

Table 9: The nine disambiguation strategies plus COMBINATION.

Strategy	Referent	Ambiguous / Masculine / Feminine
LEXICAL SWAP	<i>partner</i>	<i>Don't embarrass me in front of my partner.</i> <i>Don't embarrass me in front of my boyfriend.</i> <i>Don't embarrass me in front of my girlfriend.</i>
PRONOUN REWRITE	<i>lover</i>	<i>I'm a lover, not a fighter.</i> <i>He's a lover, not a fighter.</i> <i>She's a lover, not a fighter.</i>
POSSESSIVE TAG	<i>driver</i>	<i>I was just getting a pep talk from my driver.</i> <i>I was just getting a pep talk from him, my driver.</i> <i>I was just getting a pep talk from her, my driver.</i>
DISTAL NOUN CUE	<i>student</i>	<i>That's it, well done my student!</i> <i>That's it, well done my student, good lad!</i> <i>That's it, well done my student, good lass!</i>
KIN APPOSITIVE	<i>client</i>	<i>My client wants the public to know the truth.</i> <i>My client, my nephew, wants the public to know the truth.</i> <i>My client, my niece, wants the public to know the truth.</i>
TITLE APPOSITIVE	<i>patient</i>	<i>A patient of the hospital wanted to see it.</i> <i>A patient of the hospital, Mr. Chen, wanted to see it.</i> <i>A patient of the hospital, Mrs. Chen, wanted to see it.</i>
VOCATIVE ADDRESS	<i>cook</i>	<i>Oh, right, you're a cook.</i> <i>Oh, right, sir, you're a cook.</i> <i>Oh, right, ma'am, you're a cook.</i>
GENDER ADJECTIVE	<i>explorer</i>	<i>I am a doctor, not a space explorer.</i> <i>I am a doctor, not a male space explorer.</i> <i>I am a doctor, not a female space explorer.</i>
DASHED APPOSITIVE	<i>member</i>	<i>So you got yourself a member of the crazy folks tribe?</i> <i>So you got yourself a member — a man — of the crazy folks tribe?</i> <i>So you got yourself a member — a woman — of the crazy folks tribe?</i>

Table 10: One example contrastive row per atomic strategy, drawn from rows of the audited test split. The differing tokens are bolded.

Pri.	Signal	Hint
1	Referent has a masculine/feminine pair in the gendered-noun lexicon	LEXICAL SWAP with explicit pair
2	Addressee regex (vocative position)	VOCATIVE ADDRESS
3	Neutral pronoun (<i>they/their/them/themselves</i>) within ± 10 tokens of the referent	PRONOUN REWRITE
4	First-person markers (<i>I'm, as a...</i>) within 50 characters of the referent	PRONOUN REWRITE (symmetric <i>I/me/my</i> $\rightarrow$ <i>he/him/his</i> or <i>she/her/her</i>)

Table 11: Priority-ordered symbolic routing signals. Only the highest-priority matching signal fires per row.

(confidant, confidante). Human review surfaced these as systematically problematic because the gendered forms carry asymmetric connotations. When a forbidden pair is detected, the router falls through to the next signal, and the system prompt independently instructs the model not to apply LEXICAL SWAP to these referents (Section 3.1).

B.2 Cross-row rotators

To increase lexical variety, the pipeline rotates through a fixed list of twelve invented surnames for TITLE APPOSITIVE and five kinship pairs (brother/sister, uncle/aunt, father/mother, son/daughter, nephew/niece) for KIN APPOSITIVE. The assignment picks the least-used option so far, which is deterministic for a fixed batch order. The selected surname or relation is only used if the model chooses the corresponding generation strategy. Without this rotation step, the model tends to repeatedly generate the same few surnames or kinship pairs across many examples.

B.3 LLM evaluator

A separate LLM evaluator scores each (source, masculine, feminine) triple on a 1–5 scale across five dimensions: whether each variant disambiguates the referent’s gender, whether the inserted material is semantically correct, naturalness, whether no extra sentence was appended, and whether the variants are minimally different from the source and from each other. A row is rejected if the referent-disambiguation or extra-sentence dimension falls below 3, if the referent’s part-of-speech appears to have changed, or if the mean rubric score falls below the threshold. A free-text suggestion field captures the evaluator’s recommendation for retries, including an alternative strategy and sample

candidates.

B.4 Symbolic post-generation validators

For every row the LLM evaluator accepts, a deterministic check suite runs.

- **Edit distance.** Levenshtein on whitespace tokens. Soft-fail above 5 edits, hard-fail above 8. Catches over-editing.
- **Formatting parity.** Non-standard contractions from the source (e.g., I’ m, don’ t) must remain unchanged. Models often silently “correct” them.
- **No extra sentence.** A sentence-boundary count (with title-abbreviation awareness) must not increase.
- **Referent preservation.** For non-LEXICAL SWAP strategies, the referent word must appear unchanged in both variants.
- **POS-change guard.** A regex matching the referent followed by *man, woman, boy, girl, lady,* or *gentleman* catches cases where the referent was turned into an attributive adjective (e.g., *doctor man*).

Any failure sets `accepted=False` and exposes the failure message to the retry mechanism.

B.5 Autopilot retry with directed feedback

The autopilot runs `evaluate` $\rightarrow$ `retry` until all rows have been accepted or each row has exhausted its per-row retry cap of five. Round 0 evaluates every generated row. Later rounds evaluate only the rows newly generated as retries, so cost scales with the remaining failures rather than the full batch.

Each retry’s user message carries a feedback block assembled from four sources.

1. **Hard-rule violations.** Symbolic post-validator messages: identical variants, unchanged variants, POS change, extra sentence.
2. **Score-based complaints.** Low-scoring rubric dimensions (correctness or naturalness below 4, referent disambiguation or extra-sentence at 3 or below) are listed explicitly.
3. **Evaluator suggestion.** The free-text suggestion field, typically naming a preferred alternative strategy together with sample candidates.
4. **Symbolic checker failure list.** The full symbolic failure messages, reproduced verbatim, so the model can see exactly which constraint was violated.

C System prompts and few-shot demonstrations

Closed-task system prompt. The closed-task prompt is summarised in Table 12.

Section	Content
Task	Classify the gender of a named REFERENT in an English sentence.
Labels	masculine: the referent is unambiguously male. feminine: the referent is unambiguously female. ambiguous: the sentence contains no signal that reveals the referent’s gender.
Output	Respond with ONLY a JSON object with keys gender (one of the three labels), confidence (integer 1–5, where 5 = certain), and reasoning (one short sentence citing the textual evidence: pronoun, title, gendered noun, etc.).

Table 12: Closed-task system prompt.

Open-task system prompt. The open-task prompt is summarised in Table 13.

Section	Content
Task	Read an English sentence, identify every role, occupation, or relational REFERENT in it (e.g. <i>doctor</i> , <i>teacher</i> , <i>brother</i> , <i>neighbor</i> , <i>captain</i>), and classify each one’s gender.
Labels	masculine: unambiguously male. feminine: unambiguously female. ambiguous: no signal in the sentence reveals this referent’s gender.
Surface form	Use the surface form as it appears in the sentence (e.g. <i>doctor</i> , not <i>the doctor</i>). List referents in the order they appear.
Output	Respond with ONLY a JSON object with a referents list. Each list element is an object with keys referent, gender (one of the three labels), confidence (integer 1–5), and reasoning (one short sentence per referent).

Table 13: Open-task system prompt.

Few-shot demonstration construction. The per-row-sampled few-shot condition draws six demonstrations from the training split per test row, with the canonical 2/2/2 class mix: two masculine variants (sentence + masculine referent, gold masculine), two feminine variants, and two ambiguous source rows. The six demonstrations are shuffled and prepended to the test query as alternating user/assistant chat turns. Sampling is deterministic per test-row index with seed 42. The canned reasoning strings used in the assistant turns are listed in Table 14.

Gold label	Reasoning string
masculine	“the sentence contains a male-marking signal (title, pronoun, or gendered noun) bound to the referent”
feminine	“the sentence contains a female-marking signal (title, pronoun, or gendered noun) bound to the referent”
ambiguous	“no gendered title, pronoun, or noun in the sentence reveals the referent’s gender”

Table 14: Canned per-class reasoning strings used in few-shot assistant turns.

For the ambiguous-only ablation (Appendix H), all six demonstrations are sampled from the ambiguous source pool. Only the class mix changes. The pool size and seed are identical.

Open-task referent resolver. The open-task evaluator scores each model-emitted referent against the gold referent through three sequential matching stages. Pseudocode:

```
def resolve(predicted, gold, sentence):
    # Stage 1: exact-string match
    (case-insensitive,
    # whitespace-normalised)
    if normalise(predicted) ==
    normalise(gold):
        return MATCH
    # Stage 2: lemma match (singular
    form, ignoring
    # definite/indefinite articles)
    if lemma(predicted) == lemma(gold):
        return MATCH
    # Stage 3: containment in
    sentence-anchored
    # span around gold
    if predicted within +/-3 tokens of
    gold
        in sentence:
            return MATCH
    return NOT_FOUND
```

A row is scored NOT_FOUND only if no model-emitted referent passes any of the three stages. This is conservative with respect to the model: a misspelt but semantically correct referent counts as a miss. Rows that satisfy stage 1 or 2 contribute to the *matched* count in Table 6. The *w-match* column is the gender-correct rate conditional on matching.

D Fine-tuning hyperparameters

Model checkpoints. The six evaluated models in Table 2 correspond to the following HuggingFace checkpoint URLs: Gemma-31B (<https://huggingface.co/google/gemma-4-31B-it>), Qwen-27B (<https://huggingface.co/Qwen/Qwen3.6-27B>), Magistral (<https://huggingface.co/mistralai/Magistral-Small-2506>),

Gemma-4B (<https://huggingface.co/google/gemma-4-E4B-it>), Qwen-4B (<https://huggingface.co/Qwen/Qwen3.5-4B>), Ministral (<https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512>).

Fine-tuning configuration. All Qwen-4B runs share a QLoRA configuration with `lora_r=64`, `lora_alpha=128`, and `lora_dropout=0.05` on all linear projections (q, k, v, o, gate, up, down). The optimiser uses a learning rate of 2×10^{-5} with cosine schedule, 100 warmup steps, weight decay 0.01, and `micro_batch_size=8` with gradient accumulation 2 (effective batch 16). Training runs five epochs with early-stopping patience 3 on `eval_loss` at `eval_steps=100`, `sequence_len=1024`, and bf16 precision with flash attention. Complete YAML configs, one per student $\times$ condition, are released in the companion repository⁸.

Condition	Train rows	Val rows	Wall-time
FT-GAND (closed)	3,902	485	41 min
FT-CONTRA (closed)	11,706	1,455	1 h 01
FT-GAND	3,902	485	~30 min
FT-CONTRA	11,706	1,455	1 h 06
FT-MULTI	11,684	1,451	1 h 15

Table 15: Per-condition train sizes and wall-times on 4 $\times$ A100-80 GB.

E Few-shot results

The main text reports zero-shot accuracies (Tables 4 and 5). This appendix reports the matched few-shot condition: setup, headline numbers, and the divergent effect of demonstrations across closed and open framings.

Few-shot setup. For each model–task pair, few-shot (fs) prepends six demonstrations as user/assistant chat turns, sampled per test-row index (seed 42) from the training split. Each set contains two masculine-variant, two feminine-variant, and two ambiguous source-row demonstrations (the canonical 2/2/2 mix). A hygiene check guarantees that no demonstration sentence appears in the test set. All conditions are evaluated on the same 1,395 long-form rows used in the main text (full-denominator policy). Demonstration construction and the canned per-class reasoning strings are in Appendix C. The ambiguous-only class-mix ablation is in Appendix H.

⁸Code and configs: <https://github.com/TomMoeras/ContraGAND>.

Closed task. Table 16 pairs the zero-shot accuracies (Table 4) with the matched few-shot accuracies on the closed task. Demonstrations have divergent effects. Gains are modest at the top end (Gemma-31B +0.4 pp, Qwen-27B +0.8 pp accuracy) and Ministral picks up +4.1 pp. For three of the models, few-shot *hurts*: Magistral drops 6.7 pp, Gemma-4B drops 7.5 pp, and Qwen-4B drops 4.3 pp. We hypothesise that the six demonstrations introduce label-distribution and surface-format cues that compete with the system-prompt class definitions, consistent with the demonstration-locked behaviours reported by Min et al. (2022) and Lu et al. (2022).

Open task. Table 17 pairs the zero-shot accuracies (Table 5) with the matched few-shot accuracies on the open task. In contrast to the closed-task pattern where they hurt strong models, demonstrations lift *every* model under the open framing: with six demos, Gemma-31B improves to 0.971 and Qwen-27B recovers to 0.907. Demonstrations in the open format serve a dual purpose: they are structural examples of the required output, not only classification exemplars.

Distilled student vs. few-shot teacher. Against the Gemma-31B teacher’s few-shot accuracy (0.971), the Qwen-4B student trained under FT-MULTI (0.976) is within statistical noise: paired McNemar over the 1,395 long-form rows gives $\chi^2_1 = 0.98$, $p = 0.32$ (Appendix F). We do not claim a lift over the few-shot teacher. The student matches it at one-eighth the parameter count.

F Statistical tests

Bootstrap confidence intervals. We compute 95 % bootstrap CIs on accuracy from per-row predictions on the 1,395-row long-form test split, resampling rows with replacement for 10,000 iterations under a fixed seed. The full-denominator policy is applied throughout: parse errors and missing responses count as wrong. Per-condition CIs appear in Table 18.

Fine-tuned student vs. teacher. The Qwen-4B student under FT-MULTI reaches accuracy 0.976 (95 % CI [0.969, 0.984]), exceeding the Gemma-31B teacher’s open zero-shot baseline (0.952, [0.941, 0.963]) with non-overlapping CIs. We additionally run a paired McNemar test on the 1,395 long-form test rows. Of the rows where the two systems disagree, the student is correct and the teacher wrong on 51 rows. The reverse holds on 17 rows.

Model	Zero-shot			Few-shot (2/2/2)		
	Acc	mF1	ambF1	Acc	mF1	ambF1
Gemma-31B	0.984	0.984	0.979	0.988	0.988	0.983
Qwen-27B	0.969	0.969	0.958	0.977	0.977	0.970
Magistral	0.883	0.887	0.854	0.816	0.818	0.782
Gemma-4B	0.953	0.954	0.936	0.878	0.879	0.838
Qwen-4B	0.933	0.958	0.968	0.890	0.909	0.930
Ministral	0.685	0.718	0.757	0.726	0.730	0.713

Table 16: Closed-task results (accuracy, macro-F1, ambiguous-class F1) under zero-shot and few-shot (2/2/2 contrastive) conditions across all six models. 95 % bootstrap CIs are tabulated in Appendix F.

Model	Zero-shot			Few-shot (2/2/2)		
	Acc	mF1	ambF1	Acc	mF1	ambF1
Gemma-31B	0.952	0.956	0.942	0.971	0.978	0.978
Qwen-27B	0.742	0.757	0.726	0.907	0.922	0.898
Magistral	0.721	0.741	0.725	0.744	0.776	0.755
Gemma-4B	0.585	0.580	0.628	0.670	0.751	0.759
Qwen-4B	0.242	0.356	0.548	0.345	0.501	0.637
Ministral	0.323	0.390	0.550	0.507	0.592	0.638

Table 17: Open-task results (accuracy, macro-F1, ambiguous-class F1) under zero-shot and few-shot (2/2/2 contrastive) conditions. A row is correct only if every referent is found and classified correctly. 95 % bootstrap CIs are tabulated in Appendix F.

Model	Closed task		Open task	
	Acc (zs) [95% CI]	Acc (fs) [95% CI]	Acc (zs) [95% CI]	Acc (fs) [95% CI]
Gemma-31B	0.984 [.976, .990]	0.988 [.982, .993]	0.952 [.941, .963]	0.971 [.961, .979]
Qwen-27B	0.969 [.959, .978]	0.977 [.969, .985]	0.742 [.719, .765]	0.907 [.891, .922]
Magistral	0.883 [.866, .900]	0.816 [.796, .837]	0.721 [.698, .744]	0.744 [.721, .766]
Gemma-4B	0.953 [.942, .964]	0.878 [.860, .895]	0.585 [.559, .611]	0.670 [.644, .694]
Qwen-4B	0.933 [.919, .946]	0.890 [.873, .906]	0.242 [.220, .265]	0.345 [.320, .370]
Ministral	0.685 [.660, .709]	0.726 [.703, .749]	0.323 [.298, .347]	0.507 [.480, .533]

Table 18: 95 % bootstrap CIs on accuracy for all six paper models under both task framings and both prompting conditions. 10,000 resamples, seed 42. Full-denominator policy (Section 3.2).

With continuity correction, $\chi_1^2 = 16.0$, two-sided $p < 10^{-4}$. Against the teacher’s few-shot accuracy (0.971), the discordant counts are 29 and 21 in the student’s favour ($\chi_1^2 = 0.98$, $p = 0.32$). The few-shot comparison is therefore inside the bootstrap noise floor and we do not claim a lift. The student matches the few-shot teacher within statistical noise at one-eighth the parameter count.

FT-CONTRA vs. FT-MULTI. The marginal +1.8 pp lift from FT-CONTRA (0.958, [0.948, 0.969]) to FT-MULTI (0.976, [0.969, 0.984]) sits at the boundary of the marginal CIs. Paired McNemar gives $b_{01} = 43$, $b_{10} = 18$ (FT-MULTI correct, FT-CONTRA wrong vs. the reverse), $\chi_1^2 = 9.4$, $p = 2.1 \times 10^{-3}$, so the gain is statistically supported despite the touching marginal CIs.

G Per-strategy breakdowns

Closed-task accuracy by strategy. Table 19 reports per-strategy closed-task accuracy for the three medium models, computed over the masculine and feminine variants of the audited test split with the same strategy labels as Table 20. VOCATIVE ADDRESS is the hardest strategy for Qwen-27B (0.922) and Magistral (0.552): the cue is a remote vocative that must be bound across clauses to the role being discussed. For Gemma-31B, LEXICAL SWAP is hardest (0.881): when the referent itself is replaced (*waiter* $\rightarrow$ *waitress*), the model must recognise the swap as the sole gender signal. Few-shot demonstrations have uneven effects within a single model: Magistral’s VOCATIVE ADDRESS accuracy rises by 22 pp under few-shot, but its KIN APPOSITIVE accuracy drops by 54 pp (computed from the released per-row predictions), a sign of demonstration-induced miscalibration on strategies it previously handled correctly.

Open-task accuracy by strategy. Table 20 reports the same per-strategy accuracy on the open task across all six models. The cross-clause penalty is concentrated on VOCATIVE ADDRESS and KIN APPOSITIVE for every model except Gemma-31B. On VOCATIVE ADDRESS, the accuracy spread is largest: Gemma-31B holds 0.935 while Qwen-27B falls to 0.381, Magistral to 0.541, and Gemma-4B to 0.364, a 57 pp range driven by a single strategy. On KIN APPOSITIVE, Magistral is the worst of the medium models (0.458), and Gemma-4B drops further to 0.375. TITLE APPOSITIVE is the

easiest cross-clause cue because the appositive sits within the same noun phrase as the referent (*the doctor, Mr. Chen*) rather than at sentence boundaries. Even there, Gemma-4B loses to 0.446 while Gemma-31B keeps 0.955.

By contrast, GENDER ADJECTIVE holds above 0.92 across all four parseable-output models (0.921 to 0.979), and Gemma-31B keeps every local-cue strategy at 0.90 or higher, confirming that the open-task degradation concentrates on cue binding rather than being across-the-board. The format-failing small models (Qwen-4B, Ministral) sit far below all others on every strategy: their open-task accuracy is dominated by parse failures and not_found errors rather than by strategy-specific binding deficits.

The full per-strategy closed-to-open delta substantiates the grouped local vs. cross-clause cue probe in Section 4.2. Strategies whose cue is adjacent to the referent (GENDER ADJECTIVE, LEXICAL SWAP) survive the open-task transition within 4 pp on all four parseable-output models (for Gemma-4B, computed from the released per-row predictions). Cue-distal strategies lose up to 54.1 pp (VOCATIVE ADDRESS on Qwen-27B), and Gemma-31B is the only model that keeps every such drop below 19 pp. Magistral’s small VOCATIVE ADDRESS delta (1.1 pp) reflects its already-low closed accuracy (0.552) rather than robust binding.

H Few-shot class-mix ablation and a verbatim failure trace

A verbatim coreference-blind output. The compact narration of the binding failure in Section 5.1 and Section 5.2 draws on test row 8 (source “*you to become a top influencer in your industry or niche, sir*”, annotated referent *influencer*, gold label *masculine*). Qwen-27B’s verbatim open-task response (predicted *ambiguous*, incorrect) is in Table 21.

Class-mix ablation. To isolate whether the contrastive (masculine + feminine) class mix carries the few-shot signal, we replace the canonical 2/2/2 mix with six demonstrations all sampled from the ambiguous source pool, keeping the same demonstration count, seed, and sentence pool. Table 22 reports the result on the three medium models.

Magistral collapses by 23 pp on the closed task and 27 pp on the open task when given ambiguous-only demonstrations. Its masculine recall drops from 0.641 to 0.269 (72.5 % of masculine test rows

Strategy	n (masc+fem)	Qwen-27B	Gemma-31B	Magistral
GENDER ADJECTIVE	126	0.984	0.968	0.929
DASHED APPOSITIVE	22	0.955	1.000	0.955
DISTAL NOUN CUE	20	1.000	1.000	0.950
POSSESSIVE TAG	116	0.983	0.991	0.940
PRONOUN REWRITE	178	0.972	0.983	0.815
LEXICAL SWAP	42	0.929	0.881	0.810
KIN APPOSITIVE	48	0.979	1.000	0.771
VOCATIVE ADDRESS	154	0.922	0.994	0.552
TITLE APPOSITIVE	224	0.951	0.978	0.902

Table 19: Per-strategy closed-task accuracy (averaged over masculine and feminine variants), zero-shot. Strategies that bind the gender cue to a remote vocative (VOCATIVE ADDRESS) or that require recognising a lexical swap (LEXICAL SWAP) are systematically harder than strategies that annotate the referent directly. Accuracies are computed over the 930 masculine and feminine rows of the audited test split, using the same strategy labels as Table 20. Note that DISTAL NOUN CUE ($n=20$) and DASHED APPOSITIVE ($n=22$) accuracies are point estimates over small denominators and should not be over-interpreted.

Strategy	n (rows)	Gemma-31B	Qwen-27B	Magistral	Gemma-4B	Qwen-4B	Ministral
<i>Local cues</i>							
GENDER ADJECTIVE	189	0.952	0.979	0.926	0.921	0.275	0.497
POSSESSIVE TAG	174	0.989	0.977	0.902	0.770	0.471	0.540
PRONOUN REWRITE	267	0.981	0.798	0.659	0.558	0.255	0.236
LEXICAL SWAP	63	0.921	0.937	0.778	0.873	0.413	0.460
DISTAL NOUN CUE	30	0.900	0.633	0.567	0.600	0.333	0.267
DASHED APPOSITIVE	33	1.000	0.758	0.727	0.758	0.152	0.303
<i>Cross-clause cues</i>							
TITLE APPOSITIVE	336	0.955	0.688	0.744	0.446	0.179	0.280
VOCATIVE ADDRESS	231	0.935	0.381	0.541	0.364	0.117	0.182
KIN APPOSITIVE	72	0.819	0.625	0.458	0.375	0.097	0.222

Table 20: Per-strategy open-task accuracy (zero-shot, full-denominator policy), all six models. Strategies are grouped by cue locality (Section 4.2). Row-weighted local-cue accuracy exceeds cross-clause accuracy on every parseable-output model (by 3.6 to 32.6 pp), and VOCATIVE ADDRESS alone spans 57 pp between Gemma-31B and Gemma-4B. Small denominators (DISTAL NOUN CUE, DASHED APPOSITIVE) should not be over-interpreted.

Referent	Gender	Reasoning
influencer	ambig.	“The term ‘influencer’ is gender-neutral and the sentence provides no contextual clues to indicate a specific gender.”
sir	masc.	“‘Sir’ is a traditional honorific explicitly used to address males.”

Table 21: Qwen-27B open-task response on test row 8 (gold referent *influencer*, gold label *masculine*). The vocative *sir* is correctly tagged but the model does not bind it to *influencer*.

tagged *ambiguous*), consistent with a reasoning model locking onto the demonstration class distribution. Qwen-27B drops 15 pp on the open task but barely moves on the closed task. Gemma-31B shows only a 3 pp drop on either task, consistent with a strong pretraining baseline that has already absorbed the classification format. The contrastive class mix, therefore, carries a signal mainly for reasoning models that lock onto demonstration statistics. For non-reasoning frontier models the con-

trastive signal is largely redundant with the system-prompt class definitions.

I Prompt and decoding interventions on the open task

We tested whether the open-task gap on small models could be closed without fine-tuning. The two experiments below run on the headline test split (1,395 long-form rows over 465 source sentences). All numbers are accuracy on the same metric used in Table 5.

Selection of interventions. The nine decoding fixes evaluated for Table 24 (eight tabulated, with ASSISTANT_PREFILL omitted per the caption) are not exhaustive but cover the canonical cheap interventions: token-budget changes (MAXTOK8K), temperature (LOWTEMP), in-band parsing (PARSER_RESCUE), in-context conditioning (PER_ROW_FEWSHOT), thinking-mode toggles (NOTHINK, STRICTJSON_NOTHINK), schema-

Model	Closed task			Open task		
	fs (2/2/2)	fs-ambonly	Δ Acc	fs (2/2/2)	fs-ambonly	Δ Acc
Gemma-31B	0.988	0.984	-0.004	0.971	0.939	-0.032
Qwen-27B	0.977	0.968	-0.009	0.907	0.761	-0.146
Magistral	0.816	0.591	-0.225	0.744	0.470	-0.274

Table 22: Few-shot contrastive (2/2/2) vs. ambiguous-only demonstrations on closed and open tasks. Replacing contrastive demos with ambiguous-only demos collapses Magistral by 23 pp on the closed task and 27 pp on the open task. The other two models are largely unaffected on the closed task.

constrained decoding (GUIDED_JSON), and a sampling tweak (REPETITION_PENALTY). We did not run vendor-specific structured-output APIs (function-calling, JSON-mode), which are not uniformly available on open-weight checkpoints and would conflate vendor instruction-tuning with the underlying capability claim.

Prompt counterfactual across three small models. Table 23 reports the open-task accuracy of each small model under four system-prompt variants: the default open-task prompt (Appendix C), a binding-explicit prompt (the model is instructed to enumerate referents, find the closest gender cue, and bind it before labelling), a chain-of-thought variant, and a minimal-instruction control. Only Gemma-4B gains substantially from the binding prompt (+8.5 pp), and even then it lands 28.2 pp below Gemma-31B’s open-task zero-shot accuracy. Qwen-4B moves by only +1.1 pp. Ministral collapses entirely under the binding prompt: parse_err jumps from 32.8 % to 100 %, suggesting the longer system prompt exhausts its instruction-following budget.

Model	default	binding	CoT	minimal
Qwen-4B	0.242	0.253	0.186	0.131
Gemma-4B	0.585	0.670	0.525	0.389
Ministral	0.323	0.000	—	0.000

Table 23: Open-task zero-shot accuracy across four system-prompt variants on the three small models. The default prompt is the system prompt in Appendix C. The binding prompt instructs the model to bind a gender cue to each enumerated referent. CoT prepends a chain-of-thought scaffold. Minimal strips instructions to a single sentence. Ministral’s binding and minimal columns are 100 % parse-error.

Qwen-4B parse-error diagnosis. Of the 967 parse-error rows in the Qwen-4B baseline, 86.6 % emit prose-wrapped JSON, 13.1 % emit fenced markdown, and 0.3 % are truncated mid-JSON. The mean response length is 14,429 characters: the model is talking around the JSON, not

running out of token budget. We tested nine fixes (Table 24). Schema-constrained decoding (GUIDED_JSON) eliminates parse failures (0.4 %) but caps accuracy at 0.417 because the structural constraint also overrides the model’s trained-in ambiguous default (ambiguous recall jumps from 0.43 to 0.92). The best non-destructive fix is a repetition penalty of 1.3 (REP_PENALTY), which lifts accuracy from 0.242 to 0.450 by raising the JSON commit rate without altering the model’s classification posterior. The PARSER_RESCUE run shows that 0 of the 967 parse-error rows contain a schema-valid embedded JSON object: format and capability fail together on these rows, not independently.

Fix	parse-err %	not-found %	Acc
baseline	69.3	2.2	0.242
MAXTOK8K	67.6	1.6	0.252
LOWTEMP (T=0.3)	70.4	1.6	0.232
PARSER_RESCUE	69.3	2.1	0.242
PER_ROW_FEWSHOT	35.9	29.0	0.340
NOTHINK	19.1	20.0	0.280
STRICTJSON_NOTHINK	0.6	17.1	0.394
GUIDED_JSON	0.4	13.4	0.417
REP_PENALTY (1.3)	34.1	12.6	0.450

Table 24: Qwen-4B open-task zero-shot accuracy under nine prompt and decoding interventions. ASSISTANT_PREFILL (omitted from the table, accuracy 0) broke output formatting completely. The best fix lifts accuracy by +21 pp but stays well below Gemma-4B’s 0.585.

The composite picture: prompts and decoding hyperparameters can move Qwen-4B’s accuracy from 0.242 to roughly 0.45, but the remaining ~ 50 pp gap to a healthy open-task baseline (Gemma-31B at 0.952) is a capability ceiling. Fine-tuning is required to close it (Section 4.3).

Within-parseable breakdown for Qwen-4B.

The headline “0.849 within-parseable accuracy” (Section 5.3, Table 6) is computed over the ~ 400 rows on which the model both parses and surfaces the gold referent. To test whether this subset is representative of the test population, Table 25 stratifies it by disambiguation strategy and by cue local-

ity. The LOCAL group contains gender adjective, pronoun rewrite, possessive tag, distal noun cue, lexical swap, and dashed appositive. The CROSS-CLAUSE group contains kin appositive, title appositive, and vocative address.

Strategy	<i>n</i>	pars.%	matc.%	acc/m
<i>Local cues</i>				
PRONOUN REWRITE	267	30.0	26.2	97.1
GENDER ADJECTIVE	189	30.7	27.5	100.0
POSSESSIVE TAG	174	47.7	47.7	98.8
LEXICAL SWAP	63	44.4	44.4	92.9
DASHED APPOSITIVE	33	27.3	18.2	83.3
DISTAL NOUN CUE	30	60.0	50.0	66.7
LOCAL (all)	756	36.5	33.6	95.7
<i>Cross-clause cues</i>				
TITLE APPOSITIVE	336	23.2	22.9	77.9
VOCATIVE ADDRESS	231	25.1	22.1	52.9
KIN APPOSITIVE	72	22.2	20.8	46.7
CROSS-CLAUSE (all)	639	23.8	22.4	65.7

Table 25: Qwen-4B open-task zero-shot, stratified by disambiguation strategy and grouped by cue locality. *pars.%* is the share of strategy-*n* rows that produce a parseable response. *matc.%* is the share where the gold referent is matched. *acc/m* is the gender-correct rate conditional on matching. Two patterns emerge. (i) Parse rates are higher on local strategies ($\chi^2_1 = 25.8$, $p < 10^{-6}$), so the parseable subset is not a representative slice. (ii) Within-parseable accuracy varies by 53 pp across strategies, from 100 % (GENDER ADJECTIVE) to 46.7 % (KIN APPOSITIVE). Strategy-level small denominators (DISTAL NOUN CUE, DASHED APPOSITIVE, KIN APPOSITIVE) should not be over-interpreted.

The two-row summary makes three implications precise. (i) The parseable subset is not representative of the full test population. Parse rates differ by cue locality ($\chi^2_1 = 25.8$ with continuity correction, $p = 3.9 \times 10^{-7}$ over the 1,395-row test split). Local cues parse at 36.5 % against 23.8 % for cross-clause cues. The within-parseable accuracy of 0.849 reported in Section 4.2 is therefore aggregated over an easier subset. (ii) Within that subset, local-cue accuracy is near-ceiling on the four largest strategies (92.9–100 %) while cross-clause-cue accuracy drops to 46.7 % on KIN APPOSITIVE and 52.9 % on VOCATIVE ADDRESS, comparable to the other small models in Table 6. (iii) Two small-denominator local strategies (DISTAL NOUN CUE $n_{\text{matched}} = 15$ and DASHED APPOSITIVE $n_{\text{matched}} = 6$) sit below the local average, but neither denominator supports a strong claim. The under-commitment story in Section 5.3 is therefore most accurate for local cues. Cross-clause cues exhibit a residual binding deficit even when the format failure is removed.

Condition	Acc	mF1	mR	fR	aR
Qwen-4B (ICL zs)	0.933	0.958	0.890	0.944	0.963
FT-GAND	0.376	0.252	0.043	0.084	1.000
FT-CONTRA	0.992	0.992	0.991	0.996	0.989
Gemma-4B (ICL zs)	0.953	0.954	0.927	0.955	0.978
FT-GAND	0.386	0.268	0.019	0.140	1.000
FT-CONTRA	0.993	0.993	0.991	0.991	0.996
Gemma-31B (ICL zs)	0.984	0.984	0.972	0.987	0.991

Table 26: Closed-task LoRA fine-tuning conditions for both students. Authentic-only training (FT-GAND) collapses to predicting *ambiguous* on both students. Adding synthetic contrastive variants (FT-CONTRA) reaches the 31 B teacher’s accuracy for both students.

Representative parse-failure excerpt. To make the “reasoning without committing” pathology concrete, we reproduce one representative Qwen-4B parse-failure response below, on test row 39 (masculine variant “You’ re truly the friendliest *clerk* I ever met, *sir*.”, gold referent *clerk*, gold label *masculine*). The response is 12,723 characters long, never emits a JSON object, and runs out of token budget mid-loop. We show the first and last ~ 100 tokens. The elision marks the omitted middle. The full per-row prediction parquet contains the unabridged outputs.

```

2. Analyse the Sentence:
  * Sentence:  "You' re truly the
               friendliest
               clerk I ever met, sir."
  * Breakdown:
    * "You' re": Pronoun (2nd person).
                  Gender unknown/ambiguous.
    * "clerk": Occupation/Role.
    [ . . . ]
  * *Wait, is "sir" a referent?* Yes.
  * *Wait, is "clerk" a referent?* Yes.
  * *Wait, is "I" a referent?* No.
  * *Wait, is "You" a referent?* No.
  * *Wait, is "sir" a referent?* Yes.
  * *Wait, is "clerk" a referent?* Yes.
  * *Wait, is
[response ends mid-loop, no JSON emitted]

```

J Closed-task fine-tuning

We complement the open-task fine-tuning in Section 4.3 with the same two data conditions applied to the closed task. **FT-GAND** trains on authentic GAND source rows only (every gold label *ambiguous*). **FT-CONTRA** adds our masculine and feminine contrastive variants (3,902 sources $\times$ 3 variants = 11,706 examples), each with a canned per-class rationale.

The pattern mirrors the open task (Table 26). FT-GAND collapses to predicting *ambiguous* (overall accuracy 0.376, masculine recall 0.043, fem-

inine recall 0.084, ambiguous recall 1.000). FT-CONTRA reaches 0.992, macro-F1 0.992, with per-class recalls of 0.991/0.996/0.989. The +61.6 pp lift from FT-GAND is attributable entirely to the 7,804 contrastive variants. FT-CONTRA slightly exceeds the 31 B teacher’s closed-task zero-shot accuracy (Gemma-31B 0.984, Table 4) at one-eighth the parameter count. The same pattern holds for Gemma-4B: FT-GAND collapses to 0.386 (ambiguous recall 1.000) and FT-CONTRA reaches 0.993 from its 0.953 zero-shot baseline.

K Full per-class tables

Table 27 reports per-class precision, recall, and F1 for all six models, both prompting conditions (zero-shot and few-shot), and both tasks (closed and open). The headline tables in Section 4.1 and Section 4.2 are the macro-summary of these. Figure 6 visualises the masculine vs. feminine recall asymmetry discussed in Section 5.4.

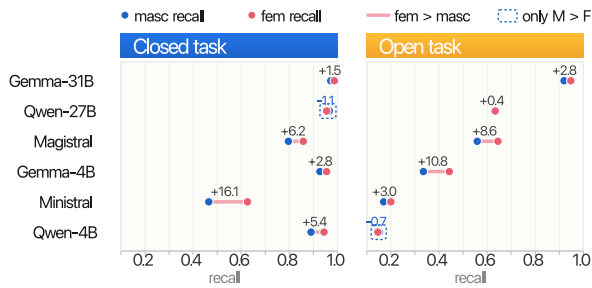


Figure 6: Masculine vs. feminine recall across all six models and both task settings. Open-task recalls of the small models are depressed by parse errors, which count as misses under the full-denominator policy (Appendix I).

Row-aligned closed-to-open transition counts.

Table 28 reports the per-model decomposition of the closed-to-open transition, joined row-by-row over the 1,395 long-form test rows. *both correct* is the count of rows the model gets right on both tasks. *lost amb* is the count where the model is correct on closed and emits ambiguous on open. *lost parse* is the count where the model is correct on closed and emits an unparseable response on open.

L Residual gender biases: supporting numbers

The two-direction summary in Section 5.4 (a small feminine-favouring direction when a cue is present, and stereotype-following when no cue is present) draws on the following supporting numbers. Figure 6 (Appendix K) plots the per-model masculine

vs. feminine recall asymmetry across both task settings. The feminine-favouring direction is most pronounced on Gemma-4B open (+10.8 pp) and Ministral closed (+16.1 pp), and the two counter-direction cells are Qwen-27B closed (−1.1 pp) and Qwen-4B open (−0.7 pp, with both recalls depressed by parse errors counted as misses). The stereotype-following direction rests on small counts of confident definite-gender predictions (confidence ≥ 4) per model. The most stereotype-prone model commits 9 such predictions across 465 ambiguous test rows, and the open framing reduces this further (Gemma-4B $9 \rightarrow 3$, Qwen-4B $5 \rightarrow 0$, Ministral unchanged at $7 \rightarrow 7$). Specific lexical clusters are *champion*, *soldier*, *scientist* in the masculine direction and *housekeeper*, *nanny*, *homemaker* in the feminine direction, mirroring the occupational and caregiving role-gender associations of Bolukbasi et al. (2016); Caliskan et al. (2017); Garg et al. (2018); Sun et al. (2019); Zhao et al. (2018).

	Model	Acc	Masculine			Feminine			Ambiguous		
			P	R	F1	P	R	F1	P	R	F1
Closed task, zero-shot.	Gemma-31B	0.984	0.989	0.972	0.980	0.998	0.987	0.992	0.966	0.991	0.979
	Qwen-27B	0.969	0.989	0.966	0.977	0.991	0.955	0.973	0.931	0.987	0.958
	Magistral	0.883	0.995	0.796	0.884	1.000	0.858	0.924	0.748	0.996	0.854
	Gemma-4B	0.953	0.980	0.927	0.952	0.991	0.955	0.973	0.897	0.978	0.936
	Qwen-4B	0.933	0.990	0.890	0.938	0.993	0.944	0.968	0.972	0.963	0.968
	Ministral	0.685	0.982	0.465	0.631	0.986	0.626	0.766	0.623	0.963	0.757
	Model	Acc	Masculine			Feminine			Ambiguous		
			P	R	F1	P	R	F1	P	R	F1
Closed task, few-shot.	Gemma-31B	0.988	0.989	0.983	0.986	1.000	0.989	0.995	0.975	0.991	0.983
	Qwen-27B	0.977	0.983	0.976	0.980	0.991	0.974	0.983	0.960	0.981	0.970
	Magistral	0.816	0.946	0.641	0.764	0.964	0.858	0.908	0.664	0.951	0.782
	Gemma-4B	0.878	0.919	0.834	0.875	0.934	0.916	0.925	0.797	0.884	0.838
	Qwen-4B	0.890	0.926	0.856	0.889	0.942	0.877	0.909	0.922	0.938	0.930
	Ministral	0.726	0.991	0.501	0.666	0.979	0.692	0.811	0.559	0.985	0.713
	Model	Acc	Masculine			Feminine			Ambiguous		
			P	R	F1	P	R	F1	P	R	F1
Open task, zero-shot.	Gemma-31B	0.952	0.993	0.920	0.955	0.995	0.948	0.971	0.900	0.987	0.942
	Qwen-27B	0.742	0.993	0.630	0.771	0.990	0.634	0.773	0.583	0.961	0.726
	Magistral	0.721	0.992	0.559	0.715	1.000	0.645	0.784	0.582	0.959	0.725
	Gemma-4B	0.585	0.994	0.335	0.502	0.990	0.443	0.612	0.462	0.976	0.628
	Qwen-4B	0.242	1.000	0.153	0.265	1.000	0.146	0.255	0.767	0.426	0.548
	Ministral	0.323	0.919	0.170	0.287	0.969	0.200	0.332	0.510	0.598	0.550
	Model	Acc	Masculine			Feminine			Ambiguous		
			P	R	F1	P	R	F1	P	R	F1
Open task, few-shot.	Gemma-31B	0.971	0.991	0.948	0.969	1.000	0.976	0.988	0.968	0.987	0.978
	Qwen-27B	0.907	0.993	0.862	0.923	0.995	0.897	0.943	0.843	0.961	0.898
	Magistral	0.744	0.970	0.559	0.709	0.953	0.791	0.865	0.660	0.882	0.755
	Gemma-4B	0.670	0.975	0.576	0.724	0.967	0.639	0.769	0.728	0.794	0.759
	Qwen-4B	0.345	0.992	0.262	0.415	0.986	0.292	0.451	0.949	0.480	0.637
	Ministral	0.507	0.970	0.344	0.508	0.965	0.469	0.631	0.580	0.708	0.638

Table 27: Full per-class precision, recall, and F1 for all six models under all four (task, condition) combinations. Accuracy is derived from per-class true positives and the total number of examples.

Model	both correct	lost amb	lost parse
Gemma-31B	1,321	42	4
Qwen-27B	1,029	289	0
Magistral	962	217	16
Gemma-4B	803	482	0
Qwen-4B	334	54	884
Ministral	406	114	295

Table 28: Per-model closed-to-open transition counts over the 1,395-row test split. *both correct* = correct on both tasks. *lost amb* = correct closed, predicted ambiguous on open. *lost parse* = correct closed, un-parseable on open. Polar masculine↔feminine swaps total 0 or 1 on every paper model and are omitted.