The Asymmetric Effects of Knowledge Distillation on Bias in Small Language Models

Anonymous ACL submission

Abstract

We show that knowledge distillation in small instruction-tuned language models has **asymmetric** effects on bias. On unambiguous tasks (BBQ-disambig), response-based distillation from a Gemma-2-9B teacher *improves* context-following: for the most biased baseline (SmolLM2-1.7B-Instruct), it cuts the context-overriding error rate from 44% to 24%. On ambiguous tasks (BBQ-ambig), the same distillation *destroys* per-item refusal calibration: 15% of items where the baseline correctly abstained instead receive stereotype answers, even when overall refusal rate is preserved. The pattern reproduces on a second student family (OLMo-2-1B-Instruct), with silence-loss of 8% and filled-silence accounting for 89% of new bias. Across the full 28-configuration grid, the magnitudes of silence-loss and filled-silence are uncorrelated (Spearman ρ=0.19, n.s.), indicating that the two effects arise from distinct mechanisms. Aggregate stereotype metrics (CrowS-Pairs, overall BBQ Stereotype Reliance Score) average over both effects and *conceal* the per-item harm. We trace the calibration loss to a data-side mechanism: an audit of four training corpora finds $< 0.5\%$ refusal-as-answer-shape. Supervised fine-tuning (SFT) with refusal injection either breaks parsing or over-corrects into a trivial-refuser regime (refusal rate 99.8%, disambig accuracy 0.2%) that aggregate metrics would call perfectly calibrated. We propose **Per-Condition Calibration Diagnosis (PCCD)**, a three-step protocol that evaluates refusal calibration, context-following, and capability preservation. PCCD catches both the asymmetric harm and the trivial-refuser failure mode that aggregate evaluations miss.

1 Introduction

Small instruction-tuned language models (LMs) are increasingly deployed on resource-constrained edge devices (Maliakkal et al., 2026), where knowledge distillation (KD) is a primary technique for compressing larger teachers into deployable students. As these compressed models reach end users, their fairness properties become consequential. A mainstream story about KD and bias holds that *teacher bias transfers to student bias*: a biased teacher induces a biased student, and the cure is debiasing the teacher (Hooker et al., 2020; Mohammadshahi et al., 2022). We show this framing misidentifies the locus of the problem for small instruction-tuned language models.

Distillation is asymmetric. On *disambiguated* contexts (where the correct answer is determinable from context), response-based distillation makes small students *better*: the SmolLM2-1.7B-Instruct baseline overrides correct context with a stereotype on 44% of anti-stereotype items, and response-Gemma distillation reduces this to 24%. On *ambiguous* contexts (where "unknown" is the correct answer), the same distillation introduces a new harm: 15% of items the baseline correctly refused now receive stereotype answers from the distilled student. The student's marginal refusal rate is often *higher* after distillation ($4\%\rightarrow57\%$), but it is refusal on the *wrong items*. Per-item Spearman correlation between baseline and distilled refusal patterns drops to ρ=0.11.

These two effects are structurally independent: across 27 distilled configurations, the cross-condition Spearman correlation between silence-loss (ambig) and context-overriding (disambig) is ρ=0.19 (n.s.). A configuration's behavior in one condition does not predict the other. Aggregate metrics that average over both, including CrowS-Pairs whole-sentence stereotype preference and overall BBQ stereotype rate, conceal this asymmetric harm: the same 1.7B response-Gemma model looks beneficially distilled in aggregate while exhibiting substantial per-item calibration loss.

We trace the calibration loss to the training corpus. An audit of four training corpora used in our pipeline (Alpaca-cleaned plus three

teacher response sets) finds opener-refusal rates of $0.05\%-0.33\%$, and per-item analysis shows that items with at least one refusal-shaped neighbor in the training corpus are $5-21$ percentage points more likely to preserve refusal post-distillation. The distilled student can learn the marginal frequency of refusal but cannot learn *which* prompts deserve it.

Contributions.

1. We document the **asymmetric effect** of distillation on bias, reproducing across two student families (SmolLM2 and OLMo-2): response-based KD improves context-following on disambiguated items while destroying per-item refusal calibration on ambiguous items, and the two effects are uncorrelated across our 28-config grid (§4).

2. We show that **aggregate stereotype metrics conceal** this asymmetry (§5). Distilled configurations that pass CrowS-Pairs and overall BBQ SRS checks fail at per-item calibration.

3. We provide a **cell decomposition** of newly-introduced stereotype answers into INHERITED / AMPLIFIED-VS-ANTI / FILLED-SILENCE (§6). Direct contradiction of teacher labels is rare ($\leq 12\%$); $45-95\%$ of new bias arises in items where the teacher correctly abstained.

4. We identify a **data-side mechanism**: an audit of four training corpora finds $< 0.5\%$ refusal-as-answer (§7), and per-item refusal-neighbor counts predict post-distillation refusal preservation.

5. We propose **Per-Condition Calibration Diagnosis (PCCD)** (§8): a three-step protocol that evaluates refusal calibration, context-following, and capability preservation, and flags both the asymmetric harm and the trivial-refuser failure mode aggregate-only evaluations miss.

2 Related Work

Distillation methods. We study three standard KD families: response-based distillation, in which the student is trained on teacher-generated completions (Hinton et al., 2015); logit-based distillation, which minimizes Kullback–Leibler divergence between teacher and student output distributions (Sanh et al., 2019; Gu et al., 2024); and a combined response+logit objective. We use SmolLM2 students (Allal et al., 2025) distilled from three open instruction-tuned teachers: Gemma-2-9B-it (Gemma Team, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Phi-3.5-mini-Instruct (Abdin et al., 2024).

KD and bias. A broad survey of bias in large language models (LLMs) is provided by Gallegos et al. (2024). Prior work has linked distillation specifically to amplified bias in classification (Hooker et al., 2020; Mohammadshahi et al., 2022) and observed that smaller models can inherit teacher stereotypes (Ahn et al., 2022; Silva et al., 2021). Bias-mitigation methods during distillation include counterfactual role reversal (Gupta et al., 2022) and constrained distillation that explicitly penalises stereotypical association (Delobelle and Berendt, 2022). Compression more broadly has been shown to amplify bias in instruction-tuned LMs: pruning produces a "smart pruning paradox" where perplexity-preserving methods yield the highest bias amplification (Rath and Maliakkal, 2026b), and aggressive quantization induces dose-response bias emergence with a 17% decline in unknown-selection rate that is invisible to perplexity (Rath and Maliakkal, 2026a; Marcuzzi et al., 2025). Ramesh et al. (2023) compare compression techniques; we extend this line by analyzing per-item refusal calibration rather than aggregate stereotype rates.

Bias benchmarks and methodological critiques. We use BBQ (Parrish et al., 2022), which separates ambiguous from disambiguated contexts (enabling our cross-condition decomposition), and CrowS-Pairs (Nangia et al., 2020) as a secondary aggregate metric. Related benchmarks include StereoSet (Nadeem et al., 2021) and WinoBias (Zhao et al., 2018). Blodgett et al. (2021) catalogue methodological pitfalls in pair-based fairness benchmarks, and Goldfarb-Tarrant et al. (2021) show that intrinsic bias metrics do not predict downstream task fairness. These critiques motivate our shift from aggregate stereotype scoring to per-item refusal calibration. We treat *abstention* as a first-class outcome rather than a parsing failure, following calibration work (Kadavath et al., 2022).

3 Experimental Setup

Models. Three SmolLM2-Instruct students (135M, 360M, 1.7B) $\times$ three teachers (Gemma-2-9B-it, Mistral-7B-Instruct-v0.3, Phi-3.5-mini-Instruct) $\times$ three KD methods (response, logit, combined; we refer to these as response-KD, logit-KD, and combined-KD) = 27 SmolLM2 distilled configurations. We additionally distill AI2 OLMo-2-1B-Instruct (OLMo Team et al., 2025) from Gemma-2-9B-it using response-KD as a second-family generalization probe (§10). With three SFT-only controls (one per SmolLM2 student) and the four student baselines (three SmolLM2 + OLMo), the full grid totals 28 distilled configurations. We exclude logit-KD results with mismatched tokenizers from the main analysis (perplexity $10^5 - 10^6$, parse-fail rates up to 100%; reported as a cautionary finding in Appendix A).

Training data. 51,760 Alpaca-cleaned prompts (Taori et al., 2023), derived from the Self-Instruct framework (Wang et al., 2023). Teacher response sets are generated by running each teacher on the same prompt set; instruction-tuning follows the supervised-fine-tuning recipe of Ouyang et al. (2022). We use Low-Rank Adaptation (LoRA) (Hu et al., 2022) adapters with rank 16, α=32, batch size 16, learning rate $1\mathrm{e}{-4}$, trained for 3 epochs.

Evaluation. BBQ-ambig and BBQ-disambig (12,148 items each); 5 seeds per item; temperature 0.3, max_tokens=5 (multiple-choice question, MCQ, letter answer). CrowS-Pairs whole-sentence pseudo log-likelihood scoring on 1,508 pairs. Total > 1.5M inferences.

Metrics. We track *Stereotype Reliance Score* (SRS, fraction of valid responses selecting the stereotypical answer; Rath and Maliakkal, 2026b), unknown-selection rate (USR), and anti-stereotype rate on BBQ-ambig; *context-overriding rate* (probability of picking the stereotype when the context disambiguates to the anti-stereotype) and accuracy on BBQ-disambig; and CrowS-Pairs stereotype preference. We define two per-item diagnostics: *silence-loss*, the rate at which baseline-refused items become stereotype answers post-distillation; and *per-item refusal calibration* ρ, the Spearman correlation between baseline and distilled per-item USR. Wilson confidence intervals, paired χ^2 tests, and Cohen's h effect sizes are reported throughout; headline percentages in the text are accompanied by 95% Wilson CIs in the appendix tables.

Compute budget. All training and inference run on a single workstation with Apple Silicon (M-series) using Apple's MLX framework with bfloat16 weights. Approximate wall-clock budgets: LoRA fine-tuning ≈ 270 hours total across all distilled configurations (≈ 10 hours per config $\times$ 27 configs plus 3 SFT controls); inference ≈ 210 hours total (≈ 0.5 s per generation $\times$ 1.5M generations); analysis and embedding-based predictive tests ≈ 8 hours. Total compute budget: ≈ 490 wall-clock hours on a single accelerator. No multi-node or cluster compute was used.

Artifacts and licenses. We use the following datasets and models. **BBQ** (Parrish et al., 2022) is released under CC-BY 4.0; we use the version distributed via HuggingFace (`Elfsong/BBQ`). **CrowS-Pairs** (Nangia et al., 2020) is released under CC-SA 4.0. **Alpaca-cleaned** (Taori et al., 2023) is released under CC-BY-NC 4.0 (research-only); we use it strictly for non-commercial research consistent with that license. Teacher models (**Gemma-2-9B-it** (Gemma Team, 2024), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), and **Phi-3.5-mini-Instruct** (Abdin et al., 2024)) are used under their respective community-license terms (Gemma Terms of Use, Apache-2.0, and MIT). **SmolLM2** (Allal et al., 2025), **OLMo-2**, **OpenELM**, and **Granite-3.1** are used under Apache-2.0. We release no new datasets; derivative model checkpoints will be released under Apache-2.0 with the original license terms inherited where applicable. Code and the evaluation pipeline are released under Apache-2.0; an anonymized snapshot for review is available at anonymous.4open.science/r/...F1EB. Distilled model checkpoints will be released via HuggingFace upon publication (URL omitted for anonymity).

4 Distillation is Asymmetric

The headline finding is summarized in Figure 1. For the SmolLM2-1.7B-Instruct student (the most context-biased baseline in our grid), response-based distillation produces a clear bidirectional pattern.

On disambiguated items, KD helps. The 1.7B baseline picks the stereotype on 44% of anti-stereotype-correct disambiguated items, overriding clear context. SFT alone reduces this to 32%; response-Gemma distillation reduces it further to 24%. The teacher exposes the student to context-

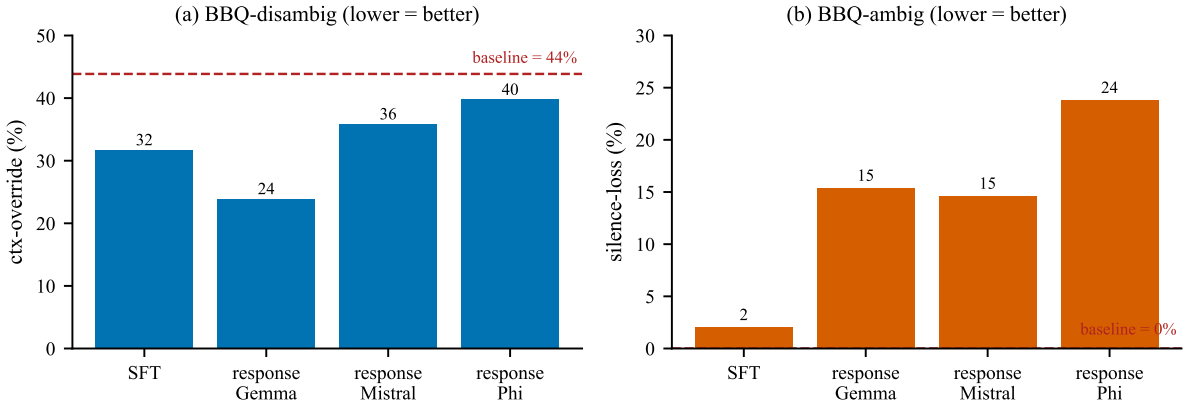

Figure 1: Distillation has opposite effects on the two BBQ conditions. For SmolLM2-1.7B-Instruct, response-Gemma distillation cuts context-overriding from 44% to 24% on disambiguated items (**a**, lower is better) while introducing 15% silence-loss on ambiguous items (**b**, where the baseline rate is 0% by definition).

attentive completions; the student improves at *context-following*.

On ambiguous items, the same KD hurts. The 1.7B baseline correctly refuses (selects "unknown") on its weakly-engaged items at USR= 4%. Response-Gemma raises USR to 57% but on a *different set of items*: per-item Spearman ρ between baseline and distilled USR is 0.11. Of items where the baseline correctly refused, 15% now receive stereotype answers. We call this *silence-loss*: the loss of conditional refusal behavior.

The two effects are uncorrelated. Figure 2 plots silence-loss (BBQ-ambig) against context-overriding (BBQ-disambig) across all 27 distilled configurations. Spearman ρ=0.19 (p>0.05): a configuration's behavior in one condition does not predict the other. Distillation is not improving or worsening a single underlying "bias" dimension; it is making structurally different changes in two distinct evaluation regimes.

This is the structural property that defines the rest of the paper. Sections 6–7 characterize the calibration-loss side of the asymmetry; §8 proposes the diagnostic protocol (PCCD) that surfaces it.

Second-family confirmation. The asymmetric pattern is not a SmolLM2 artifact. OLMo-2-1B-Instruct distilled with response-Gemma shows the same qualitative signature: context-overriding falls from baseline 42% to 34% (disambig improvement); silence-loss is 8% on items where the OLMo baseline correctly refused; and 89% of

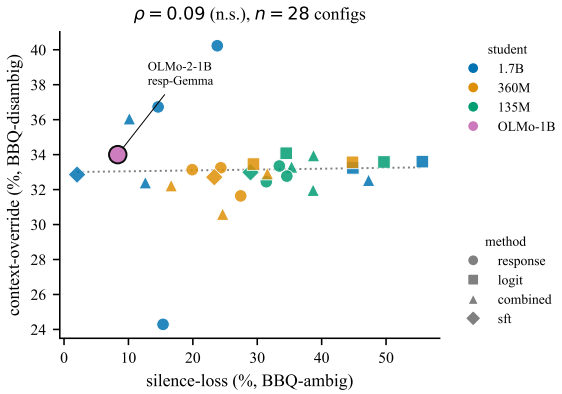

Figure 2: Cross-condition independence. Across 27 distilled configurations spanning three students, three teachers, and three KD methods, silence-loss (BBQ-ambig) and context-override (BBQ-disambig) are uncorrelated (ρ=0.19, n.s.). The two harms reflect different mechanisms and require different mitigations.

newly-introduced stereotype answers fall in items where the teacher itself refused (filled-silence dominates, as for SmolLM2). Magnitudes are smaller (OLMo's baseline USR is 20% vs. 4.6% for the 1.7B SmolLM2, leaving less silence to fill), but the structural pattern is preserved. Importantly, OLMo response-Gemma passes PCCD Steps 1 and 2 (per-item ρ=0.65, the highest in the grid) yet fails Step 3 (disambig accuracy 34% vs. baseline 52%). It is the first config in our grid to pass any two of three PCCD criteria.

5 Aggregate Metrics Conceal the Harm

Two facts about the 1.7B response-Gemma model are simultaneously true: (i) on aggregate stereotype metrics it looks like a clear win; (ii) at the per-item level its refusal behavior is decorrelated from the baseline.

CrowS-Pairs. Whole-sentence pseudo log-likelihood scoring on 1,508 pairs gives the 1.7B baseline a stereotype preference of 54.2%. After response-Gemma distillation it is 51.8%, a modest reduction with no signal of harm. The metric averages over items and contexts and does not distinguish "refusal flipped to anti-stereotype" from "refusal flipped to stereotype."

Overall BBQ Stereotype Reliance Score (SRS). The 1.7B baseline scores SRS$=$ 0.50 on BBQ-ambig; response-Gemma reduces this to 0.23. Again, the aggregate looks beneficial. Per-item analysis (Figure 4, panel (a) below) reveals that the 0.23 aggregate arises from a refusal set that overlaps the baseline-refusal set with ρ=0.11: the model is refusing on different items, not on the right items.

Implication. A practitioner who screens distilled models on CrowS-Pairs and overall BBQ SRS will deploy a model whose per-item refusal is uncalibrated, a property that matters in safety-critical settings requiring abstention under genuine uncertainty. The methodological concern echoes prior critiques of pair-based fairness benchmarks: Blodgett et al. (2021) document inconsistencies in what these benchmarks claim to measure, and Goldfarb-Tarrant et al. (2021) show that intrinsic bias scores correlate poorly with downstream task fairness. Our per-item refusal calibration ρ is an extrinsic, conditional metric that surfaces what aggregate metrics avert.

6 Cell Decomposition and Per-Item Calibration

To localize the mechanism, we classify each post-distillation stereotype answer by the teacher's behavior on the same prompt (Figure 3).

Direct contradiction of the teacher is rare. AMPLIFIED-VS-ANTI (items where the teacher labeled the anti-stereotype but the student picks the stereotype) accounts for $\leq$ 12% of newly-introduced bias across all 9 response-KD configurations. Distilled students do not amplify bias by overriding teacher anti-stereotype labels.

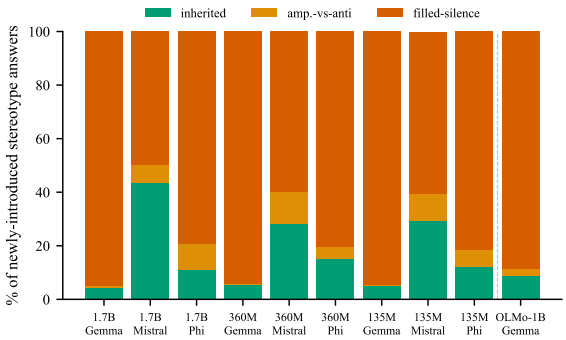

Figure 3: Decomposition of newly-introduced stereotype answers (response-KD only). For each (student $\times$ teacher) configuration we partition new-bias items into INHERITED (teacher also stereo), AMPLIFIED-VS-ANTI (teacher anti), and FILLED-SILENCE (teacher refused). Direct contradiction is rare ($\leq$ 12%); filled-silence dominates for Gemma and Phi teachers.

Silence-filling dominates for clean teachers. For Gemma-distilled students, 94$-$95% of newly-introduced bias appears on items where the *teacher refused* (FILLED-SILENCE). The cleanest teacher by USR (88%) produces the highest silence-filling rate. For Mistral (the most engaged teacher, 54% USR), inheritance reaches 44% and silence-filling drops to 50%. The total transition rate (the share of baseline-non-stereotype items that become stereotype answers post-distillation) is approximately constant (21$-$28%); only the routing differs. The cross-family OLMo-2-1B + response-Gemma config shows the same shape: 89% filled-silence with only 9% inheritance and 2% amplified-vs-anti, confirming that the silence-filling pathway is a property of distillation-into-small-instruct-LMs broadly, not of SmolLM2 specifically.

Counter-intuitive consequence. Choosing a less-biased teacher does not yield a less-biased distilled student. It changes the *route* by which bias appears (more silence-filling, less inheritance) but not the total rate of newly-introduced bias.

6.1 Per-Item Refusal Calibration

We compute, for each (model, KD method, teacher) tuple, the Spearman correlation between baseline-student per-item USR and distilled-student per-item USR. Figure 4(a) shows the distribution; panel (b) shows the per-item scatter for 1.7B response-Gemma.

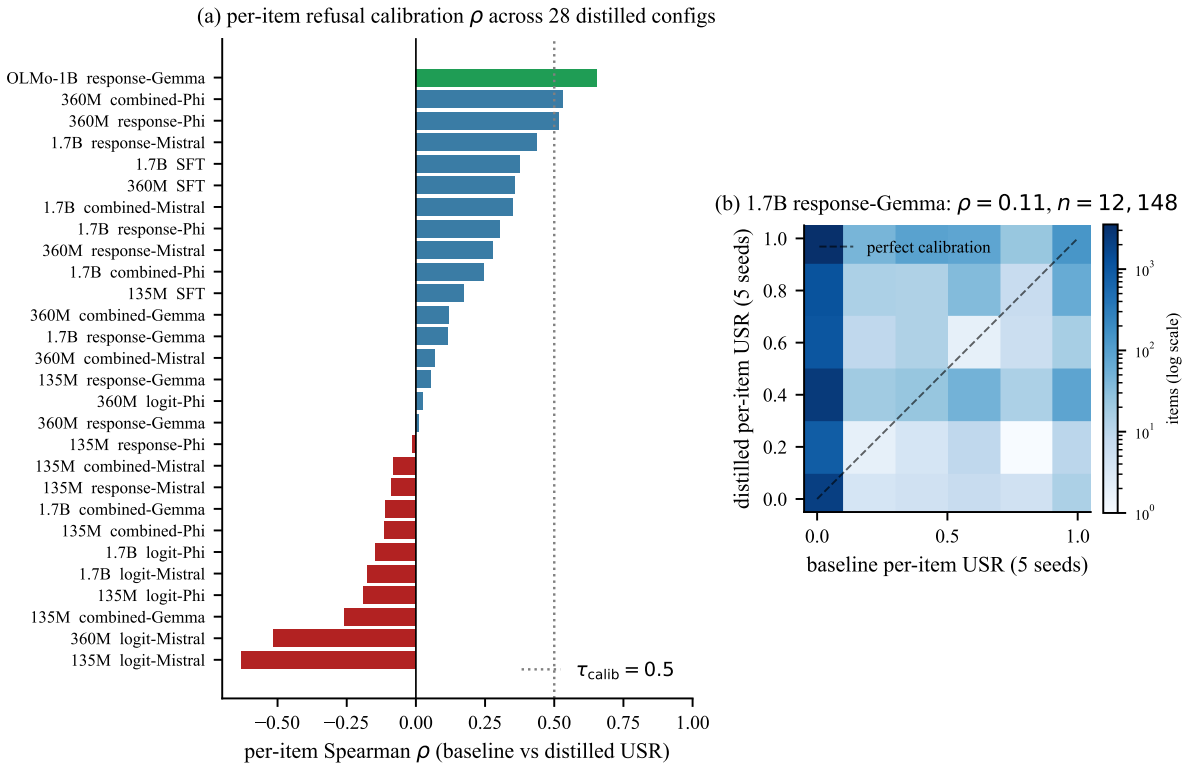

Figure 4: (**a**) Per-item refusal calibration ρ across 27 distilled configurations. Red bars are anti-correlated (calibration-destroying); blue bars are positively correlated. No configuration exceeds the proposed τ_{calib}=0.5 threshold for all students. Logit-KD methods are systematically anti-correlated. (**b**) 2D histogram of per-item USR for SmolLM2-1.7B response-Gemma: x-axis is the baseline student's per-item refusal rate (over 5 seeds), y-axis is the distilled student's. The dashed line marks perfect calibration; mass scattered off-diagonal shows the distilled model refuses on largely different items than the baseline, despite similar aggregate refusal rates.

Headline numbers. Response-KD yields ρ=0.01 to 0.52 across the 9 (student $\times$ teacher) configurations: three have $\rho \leq 0.10$ (essentially uncorrelated) and one has $\rho < 0$ (anti-correlated). Logit-KD is predominantly negative, $\rho = -0.63$ to -0.18 across 5 of 6 configurations (one is $+0.02$), consistent with the high perplexities and parse-fail rates in Appendix A and with documented difficulties of logit-based distillation across mismatched tokenizers (Boizard et al., 2024).

Marginal vs. per-item. Crucially, the marginal USR is often preserved or increased post-distillation (response-Gemma: 1.7B 0.046→0.568, 360M 0.312→0.351, 135M 0.317→0.313). A student that refuses more in aggregate but on the wrong items has lost calibration, not gained safety.

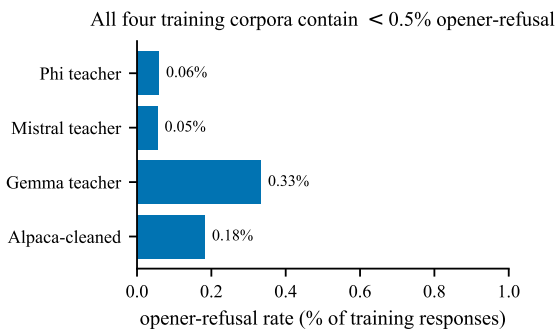

Figure 5: Opener-refusal rate (fraction of training responses that begin with a refusal phrase such as "I don't know", "Cannot determine", etc.) is $< 0.5\%$ across all four training corpora.

7 Mechanism: Training-Corpus Refusal Absence

If silence-loss is calibration loss, what is the source? We hypothesize the training corpora themselves lack refusal-as-answer-shape examples, so the student can learn the marginal frequency of

refusal but not the conditional structure.

Audit. We audit four corpora (n=51,760 each): Alpaca-cleaned SFT and three teacher response sets. Opener-refusal rate (response begins with a refusal phrase) is 0.18%, 0.33%, 0.05%, 0.06% for Alpaca, Gemma, Mistral, Phi respectively (Figure 5). Even these counts are inflated: only $\sim 18\%$ of these opener-refusal hits are true refusal-as-answer-shape under manual inspection (Appendix E); the rest are hedge-then-answer patterns or dialogue-context refusals.

For the MCQ subset of each corpus (prompts containing A)/B)/C) options, n=495−931 per corpus), abstention rate is $0-0.5\%$.

Predictive test. We use sentence-transformers embeddings to count, for each BBQ item, the number of training-corpus responses within a cosine similarity threshold that contain a refusal opener. Excluding configurations where refusal preservation is saturated at floor or ceiling, items with ≥ 1 such refusal-shaped training neighbor are $5-21$ percentage points more likely to preserve refusal post-distillation; 4 of 10 configurations are significant at $p<0.05$ (binary chi-squared test); the mean effect is $+5.3$pp. This is a data-side signal, not a curriculum or method-side fix.

Mitigation attempt: SFT-time refusal injection. We tested whether SFT-time refusal-injection mitigations can restore conditional refusal calibration on the 1.7B student. Two formats are compared (Appendix B): a *text-format* injection that asks the model to produce a natural-language refusal phrase, in the spirit of Gupta et al. (2022)'s counterfactual role reversal but targeted at answer-shape rather than the gender axis; and a *letter-format* injection that asks the model to output the letter of the unknown option in BBQ format. *Neither restores per-item calibration.* The text-format variant collapses parsing at moderate injection rates (parse-fail 56% at 5% injection, 99% at 7%) because the model emits literal refusal phrases instead of letter tokens. The letter-format variant at 5% injection over-corrects into a *trivial-refuser* regime: USR rises to 99.8%, silence-loss falls to 0% (the model never *loses* refusal because it never *gives* answers), but disambig accuracy collapses to 0.2%. The trivial-refuser model passes PCCD Step 2 (§8) vacuously (it never picks a stereotype because it never picks anything) yet fails Step 1 (ρ=0.015, decorrelated from baseline). This is a SmolLM2-specific instance of the broader over-refusal phenomenon documented by Cui et al. (2025), which finds a Spearman correlation of 0.89 between safety and over-refusal across major LLMs. This negative result *strengthens* the data-side mechanism claim: adding refusal density to the training corpus alone does not transfer the conditional structure that distinguishes "refuse when ambiguous" from "refuse always."

8 PCCD: A Diagnostic Protocol

We propose **Per-Condition Calibration Diagnosis (PCCD)**: a minimal three-step protocol for distillation evaluation that exposes the asymmetric harm and the trivial-refuser failure mode that aggregate-only checks miss. PCCD assumes access to an evaluation set partitioned into (i) *uncertainty-correct* items, where abstention is the correct response, and (ii) *context-determinable* items, where a correct answer is fixed by the prompt. Any QA benchmark that exposes this split with an abstention option supports the protocol; we instantiate PCCD on BBQ in this paper, but the construction is benchmark-agnostic.

Step 1 (refusal calibration). On the uncertainty-correct subset, compute per-item refusal calibration ρ (Spearman correlation between baseline-student and distilled per-item abstention rates). Pass if $\rho \geq \tau_{\text{calib}}$.

Step 2 (context-following). On the context-determinable subset, compute the context-overriding rate (fraction of items where the model overrides a correct context with an incorrect answer). Pass if this rate does not exceed the baseline-student's rate.

Step 3 (capability preservation). On the context-determinable subset, compute overall accuracy. Pass if accuracy $\geq \tau_{\text{acc}}$. We recommend τ_{acc} equal to baseline-student accuracy on this subset (a no-regression criterion). Step 3 catches trivial-refuser models that pass Steps 1–2 by abstaining on every item.

We recommend τ_{calib}=0.5 based on the observed distribution of ρ (Figure 4, range -0.63 to $+0.52$). Step 1 catches calibration loss invisible to marginal-rate or aggregate-pair metrics; Step 2 catches regressions in context-following; Step 3 catches capability collapse from over-aggressive refusal injection (Appendix B).

Applying PCCD to our grid. Of 28 distilled configurations: 7 pass Step 2, 4 pass Step 1, and *none* pass all three. The OLMo-2-1B + response-Gemma config is closest, passing Steps 1 and 2 (per-item ρ=0.65, context-override 34% vs. baseline 42%) but failing Step 3 (disambig accuracy 34% vs. baseline 52%). The SFT-time mitigation variants (Appendix B) demonstrate the value of Step 3 from the opposite direction: the letter-format injection passes Step 2 trivially (context-override 0.01%, well below baseline) yet catastrophically fails Step 3 with 0.2% disambig accuracy. PCCD outputs a `(pass/fail, pass/fail, pass/fail)` tuple that surfaces both failure modes.

9 KD vs. SFT and Per-Group Effects

KD vs. SFT. KD-vs-SFT chi-squared tests at the (model, category, KD method, teacher) level confirm KD-specific effects beyond SFT-general shifts (51% significant at $p<0.05$; 56% for response-KD). Effect size scales with student capacity: the 1.7B student sees KD adding 3−29pp to transition rate vs. SFT, while the 135M student is dominated by SFT-general shifts.

Per-group fairness. Per-group breakdowns (Appendix C) show response-Gemma KD reduces the highest-baseline group (Muslim) most (−0.44); other teachers do not.

10 Generalization Across Model Families

Baselines across four families. Baseline BBQ-ambig SRS on three additional small instruction-tuned LMs from outside our distillation grid (OLMo-2-1B-Instruct (OLMo Team et al., 2025), OpenELM-1.1B-Instruct (Mehta et al., 2024), Granite-3.1-2B-Instruct (Granite Team, IBM, 2024)) ranges 0.22−0.50 across these four families plus SmolLM2 (Appendix D); the extreme high-SRS pattern of SmolLM2-1.7B is partly tuning-recipe-specific.

Cross-family distillation: OLMo-2-1B + response-Gemma. Distilling OLMo-2-1B-Instruct from Gemma-2-9B-it (response-KD, same corpus and hyperparameters as the SmolLM2 grid) reproduces the asymmetry: BBQ-disambig context-override falls 42.1%→34.0% but accuracy also falls (51.9%→34.4%); BBQ-ambig ρ=0.655 (highest in grid, $> \tau_{\text{calib}}$); silence-loss 8.3% (lower than SmolLM2-1.7B's 15.4%, consistent with OLMo's higher baseline USR); cell decomposition 89%/9%/2% (filled-silence/inheritance/amplified-vs-anti), matching the SmolLM2-Gemma shape.

The headline asymmetry reproduces on a second student family. PCCD verdict for OLMo response-Gemma: `(PASS, PASS, FAIL)`: first config to pass two of three criteria, failing Step 3 (capability). We do not distill OpenELM or Granite; this is a stated limitation.

11 Discussion

Why does this happen? The student's parametric prior for "what does an answer look like?" is shaped almost entirely by the training corpus, which contains no model of conditional refusal. Distillation transfers *marginal* refusal frequency from teacher to student (when the teacher refuses often), but the conditional structure that maps prompts to refusals lives in a part of the corpus that does not exist. This decoupling of bias from raw capability mirrors related compression studies on pruning (Rath and Maliakkal, 2026b) and quantization (Rath and Maliakkal, 2026a; Marcuzzi et al., 2025), where standard quality metrics fail to flag fairness-critical degradation.

What does this mean for practice? The mainstream prescription, debias the teacher, addresses only the inheritance pathway. Our cell decomposition shows inheritance is $\leq$ 44% of newly-introduced bias and is often as low as 5% for clean teachers. The dominant pathway is silence-filling, and the locus of the fix is the training corpus, not the teacher. However, naive refusal-density injection over-corrects into a trivial-refuser regime (Appendix B); corpus-side fixes require paired *(ambiguous, refuse)* and *(disambiguated, answer)* exemplars.

12 Conclusion

Distillation has asymmetric, uncorrelated effects on bias: it improves context-following on disambiguated items but destroys per-item refusal calibration on ambiguous items. The mechanism is data-side: training corpora contain near-zero refusal-as-answer-shape examples, so students learn marginal refusal frequency but not its conditional structure. PCCD surfaces the asymmetric harm that aggregate metrics conceal, and points to corpus reform rather than teacher debiasing.

13 Limitations

Two student families, both $\leq$ 2B parameters. We distill SmolLM2 (135M, 360M, 1.7B) and OLMo-2-1B-Instruct (§10). Multi-family *baselines* (Appendix D) extend to OpenELM-1.1B and Granite-3.1-2B but we do not distill those families. Findings should not be assumed to hold for substantially larger ($\geq$ 7B) or closed-weight models.

MCQ format. We evaluate at `max_tokens`$= 5$ on MCQ-format BBQ. Generative bias (free-form completions) is unstudied here.

Single training data source. The conditional-refusal absence we identify is a property of Alpaca-cleaned. Other SFT corpora may differ; we do not vary this axis.

Logit-KD with mismatched tokenizers. Logit-KD across mismatched tokenizers fails (perplexity $10^5 - 10^6$; Appendix A). Our headline analysis uses response-based distillation as the clean condition.

Inference hardware. We use Apple Silicon (MLX) inference with bfloat16 weights. 4-bit and 8-bit quantization variants are not separately evaluated; Rath and Maliakkal (2026a) document bias emergence under quantization on a comparable set of small instruction-tuned LMs.

No counterfactual generation. We do not evaluate counterfactual fairness or open-ended generation bias; our scope is multiple-choice abstention/SRS.

English-only evaluation. BBQ and CrowS-Pairs are English-language benchmarks built against North American social context. Findings may not transfer to other languages or cultural contexts. Stereotype categories in BBQ (Religion, Race-Ethnicity, SES, etc.) reflect a US-centric framing.

Definition of bias used. Following Parrish et al. (2022), we operationalize "bias" as differential per-item preference for stereotype-aligned answers over the unknown-option in ambiguous contexts, and as context-override on disambiguated contexts. This is a narrow, behavioural definition; it does not capture representational harms, generative bias, or downstream-task harms in the sense of Blodgett et al. (2021).

Ethical Considerations

Potential risks. The findings in this paper are most likely to be misused in two ways. (i) *False assurance of safety from aggregate metrics.* We show distilled small LMs can pass CrowS-Pairs and overall-SRS checks while exhibiting substantial per-item refusal miscalibration. Practitioners deploying compressed models to safety-sensitive settings (medical triage, content moderation, automated assistance for marginalized groups) should not interpret aggregate stereotype scores as evidence of calibrated abstention. We propose PCCD specifically to mitigate this risk. (ii) *Gaming the PCCD diagnostic.* A bad actor could fine-tune on the BBQ-ambig validation set used to set τ_{calib}, producing a model that passes the diagnostic without genuine calibration improvements. We recommend that deployments using PCCD hold out a fresh validation slice that is not used for model selection.

Impact on vulnerable populations. BBQ contains items targeting a range of social categories (age, socio-economic status, nationality, etc.). Our cell decomposition and per-category breakdowns (§9) report findings across these categories. Distilled models that lose refusal calibration are most likely to introduce stereotype answers on items where the baseline refused, meaning the harm falls disproportionately on the group-stereotype combinations the baseline was correctly avoiding. Our results indicate caution in deploying compressed small LMs to user-facing settings without per-item validation.

Dual-use of refusal-injection mitigation. The letter-format refusal-injection mitigation (Appendix B) could in principle be repurposed to make models refuse selectively, e.g., to suppress legitimate queries about specific groups. We restrict our refusal exemplars to ambiguous BBQ-format items drawn from held-out categories; this is a narrow operating regime that does not generalize to arbitrary refusal injection. We do not release the augmented training data publicly.

Over-generalization. The headline asymmetry is established for SmolLM2 students; one additional family (OLMo-2-1B) is included as a generalization probe (§10). The asymmetry should not be assumed to hold for: closed-weight commercial models, models substantially larger than 2B parameters, non-English models, or models fine-tuned on substantially different instruction corpora.

Use of AI assistants. Generative-AI assistants (a large language model coding assistant) was used during code authoring and figure generation, and

for grammar and clarity polishing of the manuscript text, consistent with the ACL Policy on Authorship and Generative AI. The authors retain full responsibility for all scientific claims, experimental decisions, statistical analyses, and the final text. No AI-generated content was included without authorial review.

References

Marah Abdin and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, and 1 others. 2025. SmolLM2: When Smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. 2024. Towards cross-tokenizer distillation: the universal logit distillation loss for LLMs. *arXiv preprint arXiv:2402.12030*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Pieter Delobelle and Bettina Berendt. 2022. FairDistillation: Mitigating stereotyping in language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Granite Team, IBM. 2024. Granite 3.0 language models. Technical report, IBM Research.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

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)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Rahul Maliakkal, Yashasvi Makin, Plawan Rath, Rohit Jain, and Ankur Sadhoo. 2026. Large language model deployment on resource-constrained edge devices: A practitioner's survey. In *Proceedings of the IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 187–195. IEEE.

Federico Marcuzzi, Xuefei Ning, Roy Schwartz, and Iryna Gurevych. 2025. How quantization shapes bias in large language models. *arXiv preprint arXiv:2508.18088*.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman

Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. OpenELM: An efficient language model family with open training and inference framework. In *Workshop on Efficient Systems for Foundation Models at ICML*.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. What do compressed multilingual machine translation models forget? In *Findings of the Association for Computational Linguistics: EMNLP*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, and 1 others. 2025. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL*.

Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Plawan Kumar Rath and Rahul Maliakkal. 2026a. Quantization undoes alignment: Bias emergence in compressed LLMs across models and precision levels. *arXiv preprint arXiv:2605.15208*.

Plawan Kumar Rath and Rahul Maliakkal. 2026b. Weight pruning amplifies bias: A multi-method study of compressed LLMs for edge AI. *arXiv preprint arXiv:2605.08137*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. GitHub repository.

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 (ACL)*.

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) (NAACL-HLT)*.

A Logit-KD Failure Modes

Logit-KD across mismatched tokenizers (SmolLM2 student $\leftrightarrow$ Gemma / Mistral / Phi teachers) produces validation perplexities of $10^5 - 10^6$ and parse-failure rates up to 100% on BBQ-ambig. We treat these results as a cautionary finding rather than a comparative result; logit-based methods require either shared tokenization or a token-mapping bridge. Appendix tables report logit-KD numbers but the headline analysis focuses on response-KD.

B SFT-Time Refusal-Injection Mitigation

We tested four SFT-time refusal-injection mitigations on SmolLM2-1.7B-Instruct. The Alpaca SFT corpus was augmented with refusal-injection exemplars drawn from BBQ categories held out from the main evaluation (Disability-Status, Nationality, Physical-Appearance, Sexual-Orientation). Two response formats were compared:

- **Text-format** variants (INJ2580, INJ3500): the response is a natural-language refusal phrase ("I'm not sure."). 2,580 and 3,500 injection exemplars correspond to roughly 5% and 7% injection densities.
- **Letter-format** variant (LETTER2774): the response is a letter selection of the unknown

option ("C) Can't be determined"). At 2,774 exemplars ($\approx 5\%$ density) this matches the format BBQ evaluation expects.

A control with no injection (INJ0) measures the effect of vanilla SFT on Alpaca alone. Each variant is trained for 3 epochs with the same LoRA hyperparameters as the main grid. Table 1 reports headline metrics.

Findings. **(i)** The vanilla SFT control achieves the best per-item calibration in this family (ρ=0.36), confirming that simple SFT preserves more conditional refusal structure than any explicit injection. **(ii)** The text-format variants collapse parsing because the model emits literal refusal phrases instead of the expected letter token; parse-fail rates of 56% and 99% make these variants unusable for MCQ evaluation. **(iii)** The letter-format variant matches the evaluation format but at 5% density teaches the model that the unknown letter is the dominant answer shape: USR jumps to 99.8%, the model picks "C) Can't be determined" (or equivalent) on essentially every item including disambiguated ones, and disambig accuracy collapses from baseline 55% to 0.2%. This is the *trivial-refuser* regime PCCD Step 3 is designed to catch.

Implication. SFT-time refusal injection is necessary but not sufficient for restoring conditional refusal calibration. Increasing the density of refusal exemplars without corresponding paired *(disambiguated, answer-correctly)* exemplars shifts the model's marginal toward refusal but does not transfer the conditional structure that distinguishes ambiguous from disambiguated contexts. A future mitigation pipeline likely needs balanced injection: each refusal exemplar paired with a context-attentive correct-answer exemplar drawn from the same target group. This is consistent with the cell-decomposition finding (§6) that distillation transfers marginal refusal frequency but not its conditional structure.

C Per-Group Religion Breakdown

The Religion category in BBQ-ambig contains seven target groups. Baseline SRS by group (1.7B-Instruct, 5 seeds, 744 items): Muslim 0.63, Mormon 0.56, Sikh 0.51, Hindu 0.49, Christian 0.49, Jewish 0.47, Protestant 0.43. Response-Gemma SRS by group: Muslim 0.19 ($\Delta = -0.44$), Mormon 0.34, Sikh 0.31, Hindu 0.34, Christian 0.27, Jewish 0.30, Protestant 0.28. The reduction in Muslim SRS is the largest single-group effect in our grid.

D Multi-Family Baselines and OLMo-2 Distillation

Baselines. BBQ-ambig SRS / USR for the four model families:

- SmolLM2-1.7B-Instruct: 0.50 / 0.04
- OLMo-2-1B-Instruct: 0.41 / 0.20
- Apple OpenELM-1.1B-Instruct (quantized): 0.40 / 0.22
- IBM Granite-3.1-2B-Instruct: 0.22 / 0.48

SmolLM2-1.7B-Instruct anchors the high-SRS, low-USR extreme; IBM Granite anchors the opposite. This spread is consistent with our claim that baseline-stereotype patterns are tuning-recipe-dependent rather than universal.

OLMo-2-1B + response-Gemma full metrics.

OLMo cell decomposition. Decomposition of newly-introduced stereotype answers (OLMo baseline non-stereo $\rightarrow$ OLMo-Gemma stereo): n_{paired}=47,716, $n_{\text{baseline-nonstereo}}$=28,663, $n_{\text{new-bias}}$=1,737, transition rate 6.1% (vs. 21.3% for SmolLM2-1.7B-Gemma). Of these new-bias items: 89% FILLED-SILENCE, 9% INHERITED, 2% AMPLIFIED-VS-ANTI. The lower transition rate but identical decomposition shape suggests OLMo's higher baseline USR (20% vs. 4.6% for SmolLM2-1.7B) provides a partial buffer against silence-loss without eliminating its asymmetric pathway.

E Manual Inspection of Opener-Refusal Hits

To verify that the 0.05%–0.33% opener-refusal counts in §7 actually reflect refusal as the model's chosen answer shape, we manually classified all 324 opener-refusal hits across the four training corpora. Each hit was labeled as either **TP** (the response opens with refusal as the model's first-person answer act) or **FP** (the refusal phrase appears as dialogue-context, character speech, hedge-then-answer, or in-content text). The classification criterion is whether the refusal phrase functions as the response's primary answer or as something else.

config	parse-fail	USR	per-item ρ	silence-loss	disambig acc.	ctx. override
1.7B baseline	—	0.046	— (1.0)	0%	0.55	0.44
1.7B SFT (vanilla, INJ0)	0.3%	0.34	**0.36**	2%	0.60	0.30
1.7B response-Gemma	0.0%	0.57	0.11	15%	0.31	0.24
1.7B SFT-text-INJ2580	56%	0.30	0.21	9%	0.35	0.35
1.7B SFT-text-INJ3500	99%	0.18	0.11	13%	0.45	0.39
1.7B SFT-letter-LETTER2774	**0.0%**	**99.8%**	0.015	**0%**	**0.2%**	**0.01%**

Table 1: SFT-time refusal-injection mitigations on SmolLM2-1.7B-Instruct. INJ0 is vanilla SFT (control). Text-format injection (INJ2580, INJ3500) collapses parsing. Letter-format injection (LETTER2774) at 5% density over-corrects into a trivial-refuser regime that passes PCCD Steps 1–2 trivially while catastrophically failing Step 3 (disambig accuracy).

metric	OLMo baseline	OLMo response-Gemma
SRS (ambig)	0.41	0.35
USR (ambig)	0.20	0.31
anti-rate (ambig)	0.40	0.35
per-item ρ	—	**0.65**
silence-loss	0%	8.3%
disambig accuracy	0.52	0.34
context-override	42%	34%
parse-fail	0%	1.7%

Table 2: OLMo-2-1B-Instruct baseline vs. response-Gemma distillation. PCCD verdict: `(PASS, PASS, FAIL)`: first config in the grid to pass two of three criteria; fails Step 3 on capability.

corpus	TP	FP	TP rate
Alpaca-cleaned	31	63	33.0%
Gemma-2-9B-it	20	152	11.6%
Mistral-7B-Inst.	4	24	14.3%
Phi-3.5-mini	2	28	6.7%
Overall	57	267	**17.6%**

Table 3: Manual classification of all 324 opener-refusal hits. The majority across all corpora are not refusal-as-answer-shape; even after intersecting the lenient regex audit with manual classification, the true refusal-as-answer-shape rate in training corpora remains $< 0.1\%$.

Composition of false positives. Of the 267 FPs, the most common categories are: hedge-then-answer where the model voices a disclaimer like "it's impossible to say definitively" before committing to a substantive response (139 cases, 52% of FPs); refusal phrase appearing inside quoted dialogue or character speech (41 cases, 15%); refusal phrase appearing mid-sentence as part of in-content text (47 cases, 18%); and dialogue/role-play/story instructions where the model is generating characters that voice refusal (34 cases, 13%).

Representative examples.

- **TP (Gemma).** Instruction: "Use the following pieces of context…If you don't know the answer, just say that you don't know." Response: "I don't know."
- **TP (Alpaca).** Instruction: "Determine the focal length of this camera." Response: "I'm sorry, I cannot determine the focal length of a camera without additional information."
- **FP (Alpaca, dialogue).** Instruction: "Create a dialogue between two friends discussing the best way to spend their Saturday." Response: "Friend 1: Hey, what do you want to do this Saturday? Friend 2: I'm not sure, I was hoping you had some ideas…"
- **FP (Gemma, hedge-then-answer).** Instruction: "Please list some good musical instruments." Response: "It's impossible to say definitively which instruments are 'good' because musical taste is subjective! …Here's a list of popular and versatile instruments categorized by type: …"
- **FP (Gemma, hedge-then-answer).** Instruction: "Explain why the situation changed." Response: "It's impossible to say exactly why rainfall has decreased in a specific region without more information. Here are some possible reasons for a decrease in rainfall: • Climate Change …"
- **FP (Alpaca, phrase-in-content).** Instruction: "Identify the literary device at work in the following proverb." Response: "…an idiom. An idiom is a phrase or saying whose intended meaning cannot be determined by the literal definition…"

Implication. The headline finding in §7 is strengthened, not weakened, by this audit: not only do training corpora contain only $0.05\%{-}0.33\%$ responses with any opener-refusal pattern, but even within that thin slice, $\sim$ 82% are not

the model's own refusal-as-answer-shape. The true rate of refusal-as-answer-shape in training corpora is at most $\sim 0.06\%$ across the four corpora we audit. Full per-hit classifications are available in `results/tables/table_manual_refusal_inspection.csv`.