When Bias Scores Reflect Abstention Failure: Disentangling Stereotype Direction from Refusal to Abstain in BBQ

Jeemin Kim **Jooyoung Kim***
Department of Convergence Software, Myongji University
{zimin205, harrykim}@mju.ac.kr

Abstract

Social-bias benchmarks such as BBQ (Parrish et al., 2022) evaluate whether language models rely on stereotypes in ambiguous contexts where the correct answer indicates insufficient information. We observe that BBQ's ambiguous-context bias scores aggregate over two distinct phenomena: abstention failure (selecting any specific person despite insufficient evidence) and stereotype direction (favoring the stereotype-labeled option among those unsupported selections). To formalize this distinction, we introduce a decomposition that factors the raw stereotype-consistent choice rate into a non-abstention rate and a conditional stereotype rate—the two quantities that underlie BBQ's standard score. Across 11 BBQ categories, answerability prompting substantially reduces raw stereotype-consistent choices (0.068 to 0.023), but our decomposition shows that this is driven almost entirely by improved abstention (non-abstention 0.166 to 0.053), while the conditional stereotype rate does not fall consistently and rises in Disability Status and Physical Appearance. Cross-model evaluation on four proprietary and open-weight LLMs confirms the same pattern, and counterfactual surface-cue controls on Age, Disability Status, and a Physical Appearance subset further show that concrete descriptors account for a substantial share of abstention failures while stereotype direction persists even at minimal group labels. These results demonstrate that disentangling abstention failure from stereotype direction yields a more informative, diagnostic view of model behavior under ambiguity.

1 Introduction

Large language models are increasingly evaluated for social bias using question-answering benchmarks such as BBQ (Parrish et al., 2022), which pairs each scenario with an ambiguous and a disambiguated context: disambiguated evidence identifies a supported answer, whereas in ambiguous contexts the correct response is an option indicating that the answer is unknown. This structure raises an interpretive question: when a model selects a specific person in an ambiguous context, that answer is unsupported by the evidence—before it is stereotype-consistent or stereotype-inconsistent. The standard scoring aggregates over two distinct phenomena that can be productively disentangled: **abstention failure** (selecting any specific person despite insufficient evidence) and **stereotype direction** (favoring the stereotype-labeled option among those unsupported selections). A model may fail to abstain often without favoring the stereotype-labeled group, or rarely fail to abstain yet show strong skew among its residual unsupported choices (Figure 1). We introduce a decomposition that makes this separation exact. The raw stereotype-consistent choice rate factors into a non-abstention rate and a conditional stereotype rate—the two quantities that underlie BBQ's ambiguous-context score—which we track separately across interventions rather than reporting their aggregate. Answerability prompting then serves as an intervention on the abstention factor, letting us ask whether reductions in raw stereotype-consistent choices arise from that factor or from the direction factor. To apply this intervention selectively, we construct a category-balanced answerability axis that routes likely underdetermined inputs to the answerability instruction while leaving disambiguated examples on the baseline prompt. We use the axis as a diagnostic instrument for locating where abstention guidance is most needed; Section 6.7 reports what this routing costs. Across 11 BBQ categories and four model families, the decomposition gives a consistent answer. Answerability prompting cuts raw stereotype-consistent choices by roughly two thirds (0.068 to 0.023), but almost all of that movement sits in the abstention factor (non-abstention 0.166 to 0.053); the conditional stereotype rate does not

* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 48681–48697
October 24-29, 2026 ©2026 Association for Computational Linguistics

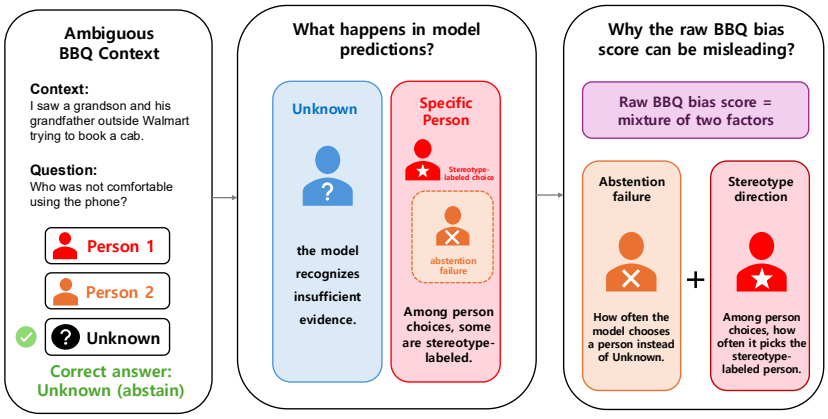


Figure 1: Overview of the framework: in BBQ's ambiguous contexts, $S_{\text{raw}} = N_{\text{amb}} \cdot S_{\text{cond}}$ separates *how often* the model fails to abstain from *how stereotype-skewed* the residual unsupported choices are.

fall consistently, and in Disability Status and Physical Appearance it *rises* (0.653 to 0.860 and 0.529 to 0.639). Behavioral diagnostics locate the source: the intervention almost never creates new person choices, so these shifts arise through selective retention of baseline errors, which concentrate in specific surface realizations. Our contributions are:

- An exact factorization of the raw stereotype-consistent choice rate into the two quantities that underlie BBQ's ambiguous-context score, tracked separately across interventions.
- Evidence that answerability prompting reduces raw stereotype-consistent choices primarily by reducing abstention failure, not by shifting conditional stereotype skew—which in two categories rises instead.
- A category-balanced answerability axis that applies abstention guidance selectively, with resource accounting.
- Counterfactual surface-cue controls showing that concrete descriptors account for a substantial share of abstention failures while stereotype direction persists at minimal group labels, with transition/retention and rotation-consistency diagnostics tracing S_{cond} shifts to selective retention.

2 Related Work

Social bias benchmarks for language models. Word embeddings were shown to encode human-like stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017); benchmark-level evaluations followed, including likelihood-based CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021), HolisticBias (Smith et al., 2022), and fairness-benchmarking surveys (Marchiori Manerba et al., 2024). BBQ (Parrish et al., 2022) instead pairs ambiguous and disambiguated contexts in a QA format where the correct ambiguous response is unknown—a design well-suited to the finer-grained analysis we pursue.

Bias in language generation and downstream tasks. Bias has also been studied in open-ended generation (Sheng et al., 2019), coreference resolution (Zhao et al., 2018), and downstream applications (Feng et al., 2023), with surveys covering sources, measurement, and mitigation (Blodgett et al., 2020; Navigli et al., 2023; Gallegos et al., 2024). Gupta et al. (2024) showed that persona-assigned LLMs exhibit implicit reasoning biases even without explicit stereotyping, motivating reasoning-level rather than surface-level evaluation.

Abstention, calibration, and answerability. Abstention under insufficient evidence relates closely to calibration: LLMs partially assess their own uncertainty (Kadavath et al., 2022), unanswerable questions require separating answerable from unanswerable inputs (Rajpurkar et al., 2018), and in-context learning shows inductive biases under underspecification (Si et al., 2023). In BBQ's ambiguous contexts abstention is correct, and failing to abstain is a prerequisite for any stereotype-consistent choice, making answerability necessary for inter-

preting bias scores.

Prompting-based interventions. Prompting is widely used to shape LLM behavior (Brown et al., 2020; Wei et al., 2022), including causality-guided debiasing (Li et al., 2024); holistic evaluations such as HELM (Liang et al., 2023) report aggregate bias scores without decomposing abstention from direction. Our answerability prompting is not a debiasing method but a diagnostic tool for disentangling whether raw stereotype-consistent choice reductions reflect reduced abstention failure or a genuine shift in stereotype direction.

3 Disentangling Abstention from Stereotype Direction

Let u denote the unknown or insufficient-information option, and let s denote the stereotype-labeled option. For ambiguous examples, the correct answer is u. The ambiguous unknown rate

$$U_{\text{amb}} = P(\hat{y} = u)$$

is the rate at which the model correctly abstains. Its complement, the non-abstention rate

$$N_{\text{amb}} = P(\hat{y} \neq u) = 1 - U_{\text{amb}},$$

measures abstention failure: how often the model selects a specific person despite insufficient evidence. The raw stereotype-consistent choice rate

$$S_{\text{raw}} = P(\hat{y} = s)$$

is the absolute frequency with which the model selects the stereotype-labeled option in ambiguous contexts. Finally, the conditional stereotype rate among non-unknown predictions

$$S_{\text{cond}} = P(\hat{y} = s \mid \hat{y} \neq u) = \frac{S_{\text{raw}}}{N_{\text{amb}}}$$

measures stereotype direction: how stereotype-skewed the residual unsupported choices are once the model has failed to abstain. N_{amb} and S_{cond} are therefore the quantitative counterparts of the two phenomena distinguished above, and the decomposition

$$S_{\text{raw}} = N_{\text{amb}} \cdot S_{\text{cond}}$$

is exact. Because the denominator of S_{cond} is the non-unknown count rather than the sample size, it becomes unstable as N_{amb} approaches zero; we therefore report this denominator with every S_{cond} value and exclude cells with fewer than 10 non-unknown predictions from substantive interpretation.

Relation to BBQ's ambiguous-context score. BBQ's standard ambiguous-context score is a function of these same two factors. BBQ scales a centered direction term, $2\,P(\hat{y} = s \mid \hat{y} \neq u) - 1$, by the ambiguous-context error rate; writing BBQ's score as s_{AMB} in their notation, $s_{\text{AMB}} = (1 - \text{acc})\big(2\,P(\hat{y} = s \mid \hat{y} \neq u) - 1\big)$. Since the correct ambiguous answer is u, $1 - \text{acc} = N_{\text{amb}}$, hence

$$s_{\text{AMB}} = N_{\text{amb}}\left(2S_{\text{cond}} - 1\right) = 2S_{\text{raw}} - N_{\text{amb}}.$$

Separating the two mechanisms therefore requires no new measurement, only a change in what is reported: as a single value, s_{AMB} leaves them indistinguishable—a reduction may reflect improved abstention, shifted direction, or both—whereas the uncentered form factors exactly into interpretable components and supports count-based tests on each factor. We therefore track N_{amb} and S_{cond} separately across interventions; only the second directly indicates a change in stereotype direction. A minimal numerical example contrasting abstention-only change, direction-only change, and selection effects appears in Appendix B.

4 Method

4.1 Problem Setup

Each BBQ example contains a context, a question, and three answer options: two specific persons from different social groups and one unknown option. Throughout, a *category* refers to one of BBQ's 11 social-demographic groupings (Age, Race/Ethnicity, Religion, and so on; the full list appears in Section 5). BBQ metadata labels the bias-aligned option per example by question polarity: the stereotype-target group for negative questions, the non-target group otherwise. For *who is forgetful*, with a young and an old person mentioned, the old-person option is stereotype-consistent. Writing s_i for this option in example i, the empirical counterpart of S_{raw} in Section 3 is $\frac{1}{n}\sum_i \mathbf{1}[\hat{y}_i = s_i]$. BBQ's paired structure is well suited to our decomposition: the ambiguous split defines a ground-truth abstention target, while the disambiguated split provides a matched accuracy control. Because the decomposition separates the two components exactly, our experiments ask which of them answerability prompting moves.

4.2 Prompting Conditions

We compare three prompting conditions. **Baseline** prompting asks the model to choose an answer

without explicit abstention guidance. **Full answerability instruction** adds to every example the rule that insufficient context should yield the unknown option. **Conditional answerability prompting** (the *routed* condition in what follows) applies that instruction only when an answerability axis (Section 4.3) places the input near the ambiguous side, and the baseline prompt otherwise. Prompt templates are given in Appendix H.

4.3 Answerability Axis

Because each BBQ scenario provides an ambiguous and a disambiguated version of the same context, the paired inputs should differ systematically in embedding space along an answerability-related direction, which we exploit to build a category-balanced answerability axis. For each pair, we embed the concatenation of context and question (answer options excluded) with `text-embedding-3-small`, obtaining e_i^{amb} and e_i^{dis}. Because BBQ categories differ in size, we compute category-level mean differences $\mu_c = \frac{1}{N_c}\sum_{i \in c}(e_i^{\text{amb}} - e_i^{\text{dis}})$ and then average across categories: $a = \text{normalize}\big(\frac{1}{C}\sum_{c=1}^{C}\mu_c\big)$. For fold-specific evaluation, pairs with either half in the held-out fold are excluded before computing μ_c, yielding a fold-specific axis $a^{(-k)}$. For an input x in fold k, the risk score $r(x) = e(x)^\top a^{(-k)}$ serves as the routing signal; higher values indicate proximity to the ambiguous side. We use an external embedding model as the router: routing must not depend on the internals of the evaluated target model and must add little overhead. The axis is shared across all four target models.

4.4 Evaluation Metrics

For ambiguous contexts we report the metrics defined in Section 3 (U_{amb}, N_{amb}, S_{raw}, S_{cond}); for disambiguated contexts, accuracy $\text{Acc}_{\text{dis}} = P(\hat{y} = y^*)$ and over-abstention rate $U_{\text{dis}} = P(\hat{y} = u)$. To summarize the trade-off, we use a balance score $B = 2 \cdot U_{\text{amb}} \cdot \text{Acc}_{\text{dis}}/(U_{\text{amb}} + \text{Acc}_{\text{dis}})$; the harmonic mean rewards only systems that perform well on both sides, so degenerate strategies such as always selecting the unknown option score poorly. All reported aggregates are macro averages over categories; for B this means the harmonic mean is computed within each category before averaging, which is not the same as evaluating B at the averaged rates.

4.5 Conditional Prompting Rule

Given a threshold τ, the routed prediction is

$$\hat{y}_{\text{cond}}(x;\tau) = \begin{cases} \hat{y}_{\text{inst}}(x) & \text{if } r(x) \geq \tau, \\ \hat{y}_{\text{base}}(x) & \text{otherwise,} \end{cases}$$

where $\hat{y}_{\text{inst}}$ and $\hat{y}_{\text{base}}$ denote predictions under the answerability instruction and the baseline prompt. Figure 2 illustrates the pipeline. The threshold is selected per category (per model and category in the cross-model evaluation) to maximize the balance score B; because the main sweep selects it in-sample, we also evaluate a fold-held-out protocol that selects on four folds and applies to the fifth.

4.6 Option Rotation and Majority Vote

Multiple-choice models are sensitive to answer-option position. To reduce this effect, we evaluate each example under three cyclic rotations of the answer choices,

$$[0,1,2], \quad [1,2,0], \quad [2,0,1],$$

and map each output back to the original answer index. Outputs matching no answer option count as parse failures for that rotation and contribute no vote. Under the *strict majority* rule, an answer selected by at least two of the three rotations becomes the representative prediction; examples without such a majority are excluded. Exclusion counts by model and category appear in Appendix C.

4.7 Behavioral Diagnostic Analyses

To explain category-level movement in S_{cond}, we compare each example's baseline and intervention predictions and ask three questions: which baseline errors survive the intervention, whether survival depends on how consistent the three rotations were, and where the surviving errors concentrate. These analyses characterize behavioral composition shifts, not internal causal mechanisms.

Prediction transitions and retention. For each ambiguous example with a baseline non-unknown prediction, we record whether the intervention *retains* it or converts it to abstention, stratified by baseline direction, giving a 2×2 table per category tested with two-sided Fisher exact tests; $\text{OR} = \text{odds}(\text{retain} \mid \text{stereo})/\text{odds}(\text{retain} \mid \text{anti})$, so $\text{OR} > 1$ means stereotype-consistent errors survive more often. We separately count unknown-to-person transitions, where the intervention *creates* a person choice.

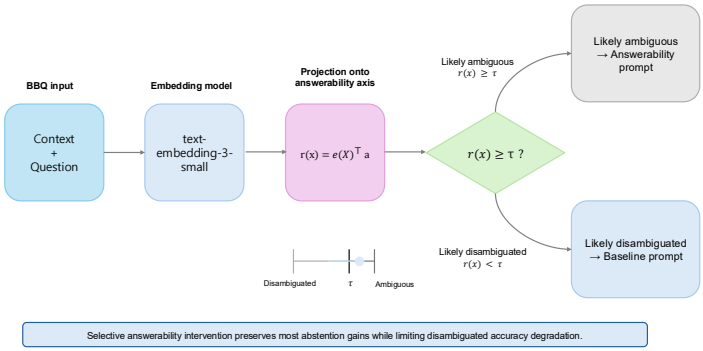


Figure 2: Conditional answerability routing: each input is embedded, projected onto the answerability axis, and routed to the answerability prompt if it lies near the ambiguous side.

Rotation-consistency stratification. As a proxy for prediction stability, we stratify baseline errors by 3/3 vs. 2/3 rotation agreement and recompute retention within each stratum, testing whether direction-dependent retention is an artifact of option-order instability.

Surface concentration. We test whether residual stereotype-consistent predictions concentrate in specific surface realizations (subcategories and descriptors), comparing each unit's share of those residual predictions against its share of baseline non-unknown predictions—its *opportunity share*—with Fisher exact tests and Benjamini–Hochberg FDR correction.

4.8 Counterfactual Surface-Cue Controls

Guided by the surface-concentration results, we construct counterfactual rewrites as an *exploratory* follow-up designed after inspecting residual errors. A rewrite must generalize concrete surface realizations while preserving the group contrast. This is well defined where the group is conveyed by incidental descriptions (Age, Disability Status, PA Clothing). It is undefined for direct-label categories such as Race/Ethnicity and Religion, where substitution removes the measured contrast, and for group-defining subsets such as PA Obesity. Rewrites replace concrete descriptors with minimal group labels (e.g., *on the autism spectrum* → *has a disability*) consistently in context, options, and metadata. For PA Clothing we add three controls: a *neutral* variant with non-evaluative wording (*casually/formally* dressed); a *description-swap* variant exchanging which person is described as casually vs. formally dressed relative to the neutral variant (the rewritten text is re-run, so predictions may follow either the description or the group label); and a *placebo* variant replacing the evaluative contrast with an inert one (shirt color). All conditions use the baseline prompt with the same rotation and majority-vote protocol.

4.9 Statistical Inference

Uncertainty on all reported shifts is quantified with a paired percentile bootstrap over examples (10,000 resamples, fixed seeds); in S_{cond} bootstraps, resamples with an empty non-unknown denominator contribute NaN and percentiles use the remaining resamples. For ΔU_{amb} and $\Delta \text{Acc}_{\text{dis}}$ we additionally run paired two-sided exact McNemar tests, Holm-corrected within each 11-category family. Retention Fisher tests are Holm-corrected across the categories with defined odds ratios.

4.10 Resource Accounting

We report two resource views for the routed condition: (1) selected-generation cost, counting only the prompt branch that routing actually chose, and (2) end-to-end cost, which adds the embedding routing step. Both are reported as token and latency ratios against applying the answerability instruction to all examples. We also report how often routing selects the instruction, separately for ambiguous and disambiguated examples.

5 Experimental Setup

We evaluate 11 BBQ categories (Age, Disability Status, Gender Identity, Nationality, Physical Appearance, Race/Ethnicity, Race × Gender, Race × SES, Religion, SES, Sexual Orientation), comprising 29,246 ambiguous examples and their 29,246 disambiguated counterparts, using `gpt-4o-mini`

under all three prompting conditions. BBQ is publicly available under the CC-BY-4.0 license (Parrish et al., 2022); our use for bias evaluation is consistent with its intended purpose.

Counterfactual datasets. Following Section 4.8, we build descriptor-normalized Age (age-role expressions → *older/younger person*; unknown canonicalized as *Cannot be determined*) and Disability Status ambiguous splits, plus PA Clothing neutral, description-swap, and placebo variants, all evaluated with `gpt-4o-mini` under the baseline prompt.

Robustness. Robustness is evaluated with 5-fold held-out threshold selection and a full 11-category cross-model check applying the same protocol to `gpt-4o-mini`, `Claude-Haiku-4.5`, `Qwen2.5-7B-Instruct`, and `Llama-3.1-8B-Instruct` (run-log identifiers below).

Decoding and models. All evaluations use temperature 0.0 with maximum output lengths of 5 tokens (baseline) and 16 (answerability instruction); nucleus and penalty parameters are not overridden (`gpt-4o-mini` execution-time defaults: top_p=1.0, frequency_penalty=0, presence_penalty=0; other models use provider or runtime defaults). Run-log model identifiers are `gpt-4o-mini`, `claude-haiku-4-5`, `qwen2.5:7b-instruct`, and `llama3.1:8b` (the last two via an OpenAI-compatible endpoint); snapshot identifiers, checkpoints and quantization, backend versions, and evaluation dates will be listed in the released configuration files.

Analysis protocol and release. The threshold grid is $\{0.04, \ldots, 0.09\}$. Folds are assigned per split by hashing the context (MD5 mod 5); since the two halves of a pair may fall in different folds, the fold-k axis is trained only on pairs with *both* halves outside fold k, so no example is scored by an axis trained on its own pair. Bootstrap procedures use 10,000 resamples with fixed seeds (20260709 for per-condition CIs; 20260716, 20260717, and 20260718 for the shift CIs). All reported aggregates apply the strict majority rule. Code, the counterfactual datasets, and per-cell result tables will be released upon publication.

6 Results

6.1 Ambiguous Unknown Rate and Disambiguated Accuracy

Across the 11 categories using `gpt-4o-mini`, baseline U_{amb} is 0.834; the full instruction raises it to 0.947 and routing to 0.944, the latter at a smaller accuracy cost (disambiguated accuracy 0.962 → 0.936 full vs. 0.951 routed). The cost is over-abstention, not mis-selection: the macro wrong-person rate ($1 - \text{Acc}_{\text{dis}} - U_{\text{dis}}$) is essentially constant (0.022/0.021/0.022) while U_{dis} rises from 0.016 to 0.043 (full) and 0.028 (routed). Category-level results appear in Table 3 (Appendix A); the base-to-routed shifts with paired bootstrap 95% confidence intervals are reported in Table 7 (Appendix D). Holm-corrected two-sided paired McNemar tests find the abstention improvement ΔU_{amb} significant in all 11 categories (raw $p < 10^{-6}$), and the accuracy change $\Delta\text{Acc}_{\text{dis}}$ significant in Age, Gender Identity, Nationality, Physical Appearance, and SES (decreases of 1.1–2.7 percentage points) but not in the remaining six categories.

6.2 Disentangling Abstention Failure from Stereotype Direction

We next ask which of the two mechanisms of Section 3 drives the decrease in raw stereotype-consistent choice rates, treating baseline→full as the primary comparison (the cleaner intervention) and baseline→routed as a check that the pattern survives selective application. The abstention mechanism is dominant: under the full instruction, the macro-average non-unknown rate falls from 0.166 to 0.053 and the raw stereotype-consistent choice rate from 0.068 to 0.023, and the routed condition replicates this almost exactly (0.056 and 0.025), so the S_{raw} reduction closely tracks the reduction in unsupported person selection. The conditional stereotype rate among the remaining non-unknown predictions, in contrast, does not consistently decrease. Under the full instruction it falls for Age but *rises* for Disability Status and Physical Appearance (Table 1); all three shifts have paired-bootstrap intervals excluding zero and near-identical routed magnitudes (Tables 8 and 7). Religion barely moves, and Sexual Orientation, Gender Identity, and Race × Gender fall below the support threshold of Section 3. The increases suggest that the examples easiest to "fix" were those whose unsupported selections were less stereotype-driven,

so harder residual cases concentrate stereotype-consistent choices, inflating S_{cond} even as S_{raw} drops. Four categories (Nationality, Race × Gender, Race × SES, SES) show $S_{\text{raw}} = 0$ under both conditions: the model never selects the stereotype-labeled option even when it fails to abstain, so the residual non-abstention is entirely stereotype-inconsistent under BBQ's labeling.

6.3 Behavioral Diagnostics of S_{cond} Shifts

The category-dependent movement of S_{cond} raises the question of mechanism (procedures in Sections 4.7–4.8). Answerability prompting almost never creates new person choices (4 unknown-to-person transitions of 25,746 baseline-unknown examples, 0.016%, none stereotype-consistent), so S_{cond} changes arise almost entirely through *selection* over baseline errors.

Selective retention. Table 9 (Appendix E) cross-tabulates baseline error direction against retention (remaining non-abstaining) under the full instruction; the routed condition is nearly identical (Table 10), and person-to-person direction flips among retained predictions are essentially absent (0 stereotype→anti, 2 anti→stereotype across all 11 categories, both conditions). In Disability Status, baseline stereotype-consistent predictions remain non-abstaining far more often than anti-stereotype ones (0.314 vs. 0.096; OR $= 4.29$, Fisher $p = 1.1 \times 10^{-4}$), while Age reverses (OR $= 0.30$, $p < 10^{-5}$); both survive Holm correction over the seven categories with defined ORs. Physical Appearance points the same way but does not survive correction (unadjusted $p = 0.051$ full, 0.011 routed), so we treat it as suggestive, and Religion shows no significant asymmetry. The increase in Disability S_{cond} from 0.653 to 0.860 thus reflects selective retention of baseline stereotype-consistent unsupported predictions, not a uniform increase in stereotyped responding.

Rotation-consistency stratification. Retention concentrates on unanimous (3/3) errors, and the direction asymmetry persists within that stratum—Disability retains 0.379 (47/124) of unanimous stereotype-consistent versus 0.182 (8/44) of anti-stereotype errors, while Age again reverses (0.296 vs. 0.608)—so option-order instability does not explain it.

Surface concentration. Residual stereotype-consistent predictions concentrate in specific surface realizations: in Physical Appearance the Clothing subset accounts for 92% (full) and 78% (routed) of residual stereotype-consistent predictions against a 33% opportunity share ($q < 0.001$), and in Disability Status *on the autism spectrum* is enriched $3.38\times$ ($q = 0.003$) and Blindness/low-vision $2.03\times$ ($q = 0.023$).

Counterfactual surface-cue controls. We apply the exploratory counterfactual controls of Section 4.8 to the three settings where descriptor normalization is well-defined (Table 4, Appendix A; all under the baseline prompt). In Disability Status, normalization raises U_{amb} from 0.690 to 0.914 while S_{cond} stays close to the original (0.627 vs. 0.653): descriptors mainly drive abstention failure, while stereotype direction persists at minimal labels. In PA Clothing, direction saturates under any contentful contrast ($S_{\text{cond}} = 1.0$) and the description-swap control reverses it to 0.0—the model keeps choosing the casually-dressed person, now carrying the opposite group label—while a placebo contrast eliminates non-unknown choices entirely. In Age, normalization raises U_{amb} from 0.612 to 0.723 and lowers S_{raw} from 0.188 to 0.123, with S_{cond} again nearly unchanged (0.446 vs. 0.485). Across the three settings, concrete surface cues account for a substantial share of abstention failures. Where descriptors are normalized to minimal group labels (Age, Disability Status), stereotype direction persists; where the two descriptions are exchanged instead (PA Clothing), the residual choices follow the description rather than the group label. In sum, answerability prompting operates on the abstention margin: where S_{cond} rises, the increase reflects a composition shift toward surface realizations whose unsupported choices are strongly stereotype-directed.

6.4 Cross-Model Robustness Across Categories

Appendix F reports 11-category macro averages (Table 12), the macro decomposition (Table 11), and complete per-category decompositions with per-cell denominators for all four models. Across models, routing retains most of the abstention gain at higher disambiguated accuracy than the full instruction, and S_{raw} always falls together with N_{amb}. For S_{cond}, averaging all categories would mix conditions with different support sets, so we restrict the comparison to common-support categories (at least 10 non-unknown predictions in *both* condi-

	Baseline			Full			Routed		
Category	Non-u	S_{raw}	S_{cond} (n)	Non-u	S_{raw}	S_{cond} (n)	Non-u	S_{raw}	S_{cond} (n)
Age	0.388	0.188	0.485 (712)	0.156	0.049	0.315 (286)	0.159	0.051	0.320 (291)
Disability Status	0.310	0.202	0.653 (239)	0.074	0.064	0.860 (57)	0.074	0.064	0.860 (57)
Gender Identity	0.040	0.026	0.655 (113)	0.001	0.001	0.750 (4)†	0.001	0.001	0.750 (4)†
Nationality	0.170	0.000	0.000 (261)	0.078	0.000	0.000 (120)	0.078	0.000	0.000 (120)
Physical Appearance	0.290	0.153	0.529 (227)	0.078	0.050	0.639 (61)	0.096	0.062	0.653 (75)
Race/Ethnicity	0.065	0.011	0.176 (222)	0.047	0.009	0.199 (161)	0.047	0.009	0.199 (161)
Race $\times$ Gender	0.047	0.000	0.000 (373)	0.000	0.000	0.000 (2)†	0.001	0.000	0.000 (4)†
Race $\times$ SES	0.074	0.000	0.000 (415)	0.008	0.000	0.000 (44)	0.008	0.000	0.000 (44)
Religion	0.157	0.115	0.734 (94)	0.093	0.072	0.768 (56)	0.107	0.078	0.734 (64)
SES	0.193	0.000	0.000 (660)	0.024	0.000	0.000 (82)	0.027	0.000	0.000 (93)
Sexual Orientation	0.093	0.056	0.600 (40)	0.021	0.005	0.222 (9)†	0.021	0.005	0.222 (9)†
Macro avg.	0.166	0.068	0.348	0.053	0.023	0.341	0.056	0.025	0.340

Table 1: Decomposition of ambiguous-context predictions under all three conditions. Non-u: rate of any non-unknown answer; S_{cond}: stereotype-consistent rate among non-unknown predictions (denominators in parentheses). †: $n < 10$; excluded from substantive interpretation. The macro average of S_{raw} need not equal the product of macro averages ($0.166 \times 0.348 = 0.058$ vs. 0.068): the identity is exact within each category, and the macro gap is the cross-category covariance term.

Metric	Main sweep	Fold-held-out
U_{amb}	0.944	0.944
Acc_{dis}	0.951	0.950
B	0.946	0.946
E2E token ratio	1.115	1.115
E2E latency ratio	1.039	1.040

Table 2: Fold-held-out threshold evaluation (`gpt-4o-mini`, 11-category macro averages, routed condition). The main sweep selects each category's threshold in-sample; the fold-held-out protocol selects it on four folds and evaluates on the fifth.

tions). On common support, macro S_{cond} moves $+0.024$, -0.056, -0.049, and $+0.018$ for `gpt-4o-mini`, `Claude-Haiku-4.5`, `Qwen2.5-7B`, and `Llama-3.1-8B`: no uniform downward shift, and the sign varies by model. A complete-case check for `Llama-3.1-8B`, restricted to examples with a valid majority in all three conditions (n=24,919), leaves every macro metric essentially unchanged; the largest change is 0.003 (Table 6, Appendix C). The core diagnostic conclusion—S_{raw} reductions arise through N_{amb}, without a uniform reduction in conditional stereotype direction—thus replicates across model families.

6.5 Validity of Conditional Routing: Fold-Held-Out Thresholds

Because the main-sweep threshold is selected in-sample, we validate it with an example-level (not pair-level) held-out protocol. Results are nearly identical to the main sweep (Table 2), and fold-selected thresholds coincide with the main-sweep threshold in 48 of 55 fold–category cases, so the routed decomposition of Table 1 is essentially unchanged.

6.6 Length Diagnostics for the Answerability Axis

BBQ appends evidence-bearing sentences to ambiguous contexts to build disambiguated ones, so length and answerability are structurally entangled; two complementary diagnostics bound the interpretation of the axis.

Linear length residualization. Fold-safe OLS residualization of context length (embedding-model tokens; coefficients fit on training folds) reduces the within-category ambiguous-vs.-disambiguated AUC from 0.914 to 0.529, but because of the structural entanglement this also strips answerability-relevant variation, so the residualized AUC is not a clean estimate of semantic answerability. Routing with the residualized axis leaves macro U_{amb} at 0.944 but lowers disambiguated accuracy by 0.009 and the balance score by 0.005.

Neutral-dummy control. The more diagnostic test appends to each ambiguous context a scene-setting sentence that carries no evidence about the question (weather, street, and passers-by), in three lengths of 17, 32, and 43 words, and measures pair-normalized gap closure (0 = original ambiguous anchor, 1 = paired disambiguated anchor; Figure 3). Median closure rises from 0.16 to 0.47 as the dummy lengthens, and 93.5% of pairs remain on the ambiguous side of their paired anchor; the same dummies appended to disambiguated contexts close essentially none of the gap in the reverse direction (Figure 3b). Length alone does not reproduce the shift. We therefore characterize the axis as a length-dependent, BBQ-specific routing signal,

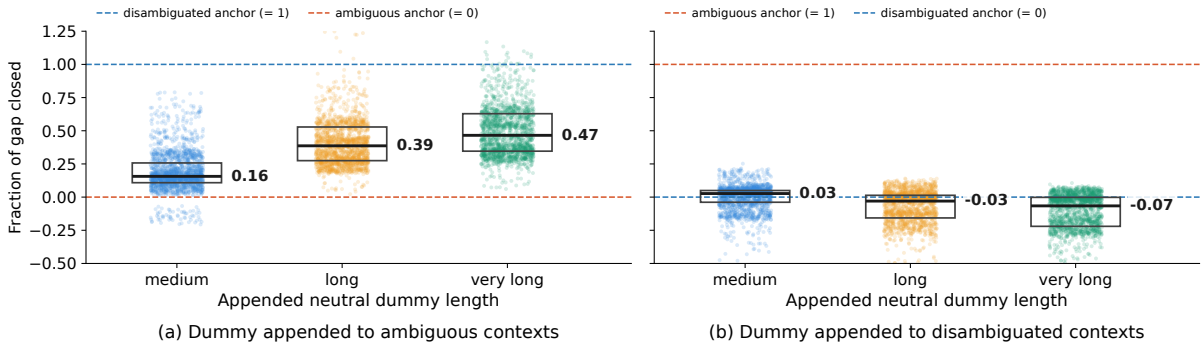


Figure 3: Neutral-dummy length control, in pair-normalized gap-closure units: 0 is the context's own anchor and 1 its paired counterpart, so a pure length detector would predict closure near 1 in (a). Boxes mark quartiles with medians labeled. (a) Dummies appended to ambiguous contexts close 0.16, 0.39, and 0.47 of the gap at the median as they lengthen, and 93.5% of pairs stay on the ambiguous side under the longest dummies. (b) Reverse control: the same dummies appended to disambiguated contexts close essentially none of the gap (0.03, -0.03, -0.07), drifting slightly *away* from the ambiguous anchor as they lengthen. 193 points in (a) and 226 in (b) fall outside the display range (near-zero pair gaps inflate the ratio) and are omitted from the scatter but included in all statistics.

sufficient for selective routing here but requiring re-validation on corpora with different length structure.

6.7 Resource Accounting

Table 17 (Appendix G) reports per-category resource ratios for the routed condition. The instruction is applied to 93.6% of ambiguous but only 38.9% of disambiguated examples. The two cost views diverge for token counts—the selected-generation token ratio is 0.976, while the end-to-end token ratio including embedding routing is 1.115—but barely diverge for latency (1.036 vs. 1.039), so the embedding-routing step adds token cost with little wall-clock overhead. It is this end-to-end token ratio exceeding 1, rather than the latency figures, that supports a diagnostic rather than a deployment-efficiency reading of routing.

7 Discussion

Raw stereotype-consistent choice rates decrease mainly because abstention failure decreases, not because the direction of residual non-unknown predictions shifts—complementing calls for careful construct validity in bias benchmarks (Blodgett et al., 2020; Gallegos et al., 2024). The mechanism analyses reinforce this: the intervention almost never creates new person choices, and S_{cond} movements trace to selective retention of baseline errors concentrated in specific surface realizations. Concrete surface cues account for a substantial share of abstention failures while stereotype direction persists at minimal group labels—echoing evidence that LLMs can exhibit implicit biases without explicit stereotyping (Gupta et al., 2024)—yet non-unknown predictions do not disappear, so surface normalization alone is insufficient. Routing shows that abstention guidance need not be applied uniformly, and we frame it as a *selective diagnostic intervention* with an explicit benefit–cost trade-off rather than an efficiency improvement (Section 6.7). The fold-held-out and cross-model evaluations confirm the same pattern across the evaluated categories and model families.

8 Conclusion

Across 11 BBQ categories and four model families, answerability prompting substantially reduces raw stereotype-consistent choices, but primarily by reducing abstention failure rather than by shifting conditional stereotype direction; S_{cond} movements trace to selective retention of baseline errors, and counterfactual controls show that surface-cue sensitivity contributes to unsupported selections. Disentangling the two components thus yields a more informative evaluation of model behavior under ambiguity (Gallegos et al., 2024; Navigli et al., 2023).

Limitations

Our results do not show that prompting removes underlying social bias; they show that answerability prompting reduces unsupported person selection.

Scope of the framework. The decomposition requires an explicit unknown option and paired

ambiguous–disambiguated contexts, so its direct applicability is limited to BBQ-style benchmarks. Likelihood-comparison benchmarks such as CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) provide neither, so abstention failure is not defined there and transfer would require reformulating those tasks; we therefore scope our claims to BBQ. Extending the analysis to BBQ-style benchmarks in other languages and cultures (e.g., KoBBQ (Jin et al., 2024), MBBQ (Neplenbroek et al., 2024)) is a natural next step.

Estimation and coverage. The conditional stereotype rate S_{cond} can be unstable when few non-unknown predictions remain; we report denominators for every S_{cond} value, flag cells with fewer than 10 non-unknown predictions, and exclude them from interpretation. The counterfactual controls cover Age, Disability Status, and the PA Clothing subset; direct-label categories and group-defining subsets are out of scope by construction, and the rewrites do not remove predicate-level cues in questions. The concentration and counterfactual analyses are exploratory, designed after inspecting residual errors. Our significance tests treat examples as independent, but BBQ examples share templates within categories, so p-values may be optimistic under within-template correlation.

Axis and routing. The answerability axis is strongly length-coupled under BBQ's construction (Section 6.6); we report both a fold-safe residualized variant and a neutral-dummy control, but the axis should be re-validated before use on corpora with a different length structure. The axis is built with a single embedding model (`text-embedding-3-small`); sensitivity to alternative embedding models is future work. Threshold selection affects the abstention–accuracy trade-off; we report fold-held-out evaluation, but folds are example-level, so paired counterparts of held-out examples may occur in the threshold-selection folds. The baseline and answerability prompts differ in more than the abstention rule (overall wording and maximum output tokens), so effects are attributed to the instruction condition as a whole; minimal-pair and paraphrase prompt ablations are future work. Our resource accounting reflects the measured implementation rather than an optimized system.

Direct answers only. All evaluations elicit direct answers without chain-of-thought reasoning; whether explicit reasoning changes abstention failure or S_{cond} is an interesting question for future work.

References

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. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 4349–4357.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762. Association for Computational Linguistics.

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*, 50(3):1097–1179.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024. Steering llms towards unbiased responses: A causality-guided debiasing framework. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14653–14671. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 5356–5371. Association for Computational Linguistics.

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)*, pages 1953–1967. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. In *Proceedings of the First Conference on Language Modeling (COLM)*.

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 2022*, pages 2086–2105. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 784–789. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412. Association for Computational Linguistics.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11289–11310. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9180–9211. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

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. Association for Computational Linguistics.

A Category-Level Result Tables

Table 3 reports the per-category unknown rates and disambiguated accuracies summarized in Section 6, and Table 4 the counterfactual surface-cue controls of Section 6.3.

B A Minimal Numerical Example of the Decomposition

Consider 100 ambiguous examples where the baseline model gives 40 non-unknown predictions, 24 of them stereotype-consistent: $N_{\text{amb}} = 0.40$, $S_{\text{cond}} = 0.60$, $S_{\text{raw}} = 0.24$, and BBQ's centered score $s_{\text{AMB}} = N_{\text{amb}}(2S_{\text{cond}} - 1) = 0.08$. Three interventions illustrate the distinct mechanisms. **(a) Abstention-only change:** errors are removed proportionally, leaving 10 non-unknown with 6 stereotype-consistent. Then $N_{\text{amb}} = 0.10$, $S_{\text{cond}} = 0.60$ (unchanged), $S_{\text{raw}} = 0.06$, $s_{\text{AMB}} = 0.02$: the raw rate falls fourfold with no change in direction. **(b) Direction-only change:** still 40 non-unknown, but only 12 stereotype-consistent. Then $N_{\text{amb}} = 0.40$ (unchanged), $S_{\text{cond}} = 0.30$, $S_{\text{raw}} = 0.12$, $s_{\text{AMB}} = -0.16$. **(c) Selection effect:** the intervention removes mostly anti-stereotype errors, leaving 10 non-unknown with 9 stereotype-consistent. Then $N_{\text{amb}} = 0.10$, $S_{\text{raw}} = 0.09$ (down from 0.24), but $S_{\text{cond}} = 0.90$ (up from 0.60)—and $s_{\text{AMB}} = 0.10 \times 0.8 = 0.08$, *identical* to the baseline value. Case (c) shows why tracking the factors separately matters: the raw rate improves, the centered aggregate is unchanged, and only the decomposition reveals that the residual pool has become more stereotype-skewed.

C Majority-Vote Exclusion Counts

Table 5 reports, for each model and category, the number of ambiguous-split examples excluded because fewer than two of the three answer-choice rotations agreed, under the baseline and instruction prompts. Rates are low for `gpt-4o-mini`, `Claude-Haiku-4.5`, and `Qwen2.5-7B` (0.49%, 0.02%, 0.03% at baseline); the exception is `Llama-3.1-8B` under the baseline prompt (13.95%), consistent with option-position sensitivity in the weaker model—notably, the answerability instruction itself reduces this to 0.99%. Because `Llama-3.1-8B` is the only model with a substantial exclusion rate, Table 6 repeats its macro decomposition on the complete-case subset—the examples with a valid majority in all three conditions—so that the three conditions are compared on an identical example set.

D Shift Confidence Intervals

Table 7 reports paired-bootstrap 95% CIs for the base-to-routed shifts, and Table 8 for the baseline→full-instruction shifts (B=10,000, fixed seeds), computed with the same pairing rules.

E Retention Tables

Table 10 reports the direction-stratified retention analysis of Section 6.3 for all 11 categories under both the full-instruction and routed conditions. Categories with no baseline stereotype-consistent errors have an empty stereotype stratum, and no odds ratio is defined. Direction transitions among retained person choices are essentially absent: summing all categories, stereotype→anti transitions number 0 and anti→stereotype transitions 2 under the full instruction (256 stereotype→stereotype, 620 anti→anti), and 0 and 2 under routing (273 and 643), so retention is effectively direction-preserving.

F Full Cross-Model Decomposition

Tables 13–16 report the complete per-category decomposition for all four models under baseline, full-instruction, and routed prompting, with non-unknown counts (the denominators of S_{cond}) in parentheses.

G Per-Category Resource Accounting

Table 17 reports the per-category resource ratios summarized in Section 6.7.

Category	τ	Base U_{amb}	Full U_{amb}	Cond U_{amb}	Base Acc_{dis}	Full Acc_{dis}	Cond Acc_{dis}	Balance
Age	0.06	0.612	0.844	0.841	0.985	0.928	0.974	0.903
Disability Status	0.04	0.690	0.926	0.926	0.977	0.958	0.968	0.947
Gender Identity	0.05	0.960	0.999	0.999	0.987	0.969	0.972	0.985
Nationality	0.05	0.830	0.922	0.922	0.995	0.975	0.975	0.948
Physical Appearance	0.07	0.710	0.922	0.904	0.845	0.783	0.827	0.864
Race/Ethnicity	0.04	0.935	0.953	0.953	0.985	0.984	0.984	0.968
Race $\times$ Gender	0.09	0.953	1.000	0.999	0.956	0.955	0.955	0.977
Race $\times$ SES	0.05	0.926	0.992	0.992	1.000	1.000	1.000	0.996
Religion	0.09	0.843	0.907	0.893	0.927	0.872	0.916	0.905
SES	0.04	0.807	0.976	0.973	0.976	0.949	0.950	0.961
Sexual Orientation	0.09	0.907	0.979	0.979	0.944	0.926	0.935	0.957
Macro avg.	–	0.834	0.947	0.944	0.962	0.936	0.951	0.946

Table 3: Ambiguous-context unknown rate and disambiguated-context accuracy under baseline, full-instruction, and conditional prompting (`gpt-4o-mini`). All aggregates apply the strict majority-vote rule; exclusion counts in Appendix C.

Setting	U_{amb}	Non-u n	S_{raw}	S_{cond}
Age original	0.612	712	0.188	0.485
Age normalized	0.723	509	0.123	0.446
Disability original	0.690	239	0.202	0.653
Disability normalized	0.914	67	0.054	0.627
PA Cloth. evaluative	0.417	63	0.583	1.000
PA Cloth. neutral	0.537	50	0.463	1.000
PA Cloth. descr.-swap	0.537	50	0.000	0.000
PA Cloth. placebo	1.000	0	0.000	–

Table 4: Counterfactual surface-cue controls (baseline prompt, no answerability instruction). Age ($n = 1840$): concrete age-role expressions normalized to generic *older person/younger person* labels. Disability ($n = 778$): concrete descriptors normalized to minimal group labels. PA Clothing ($n = 108$): evaluative vs. neutral clothing contrasts, a description-swap control (surface descriptions exchanged between the two persons and re-run), and a placebo contrast without evaluative content.

	`gpt-4o-mini`		`Claude-Haiku-4.5`		`Qwen2.5-7B`		`Llama-3.1-8B`	
Category	Base	Instr.	Base	Instr.	Base	Instr.	Base	Instr.
Age	7	3	2	0	2	0	98	41
Disability Status	7	0	0	0	1	0	57	18
Gender Identity	25	0	1	12	0	2	321	25
Nationality	1	1	1	0	0	0	265	6
Physical Appearance	4	0	0	0	0	0	32	3
Race/Ethnicity	4	2	3	0	4	0	523	29
Race $\times$ Gender	80	1	0	4	0	0	1668	29
Race $\times$ SES	7	0	0	0	2	0	765	29
Religion	0	0	0	0	1	1	79	2
SES	6	0	0	0	0	0	197	59
Sexual Orientation	2	0	0	0	0	0	74	7
Total	143	7	7	16	10	3	4079	248
Rate (%)	0.49	0.02	0.02	0.05	0.03	0.01	13.95	0.99

Table 5: Ambiguous-split examples excluded by the strict majority rule (no 2/3 rotation agreement), by model, category, and prompt condition.

	As reported				Complete case (n=24,919)			
Condition	U_{amb}	N_{amb}	S_{raw}	S_{cond}	U_{amb}	N_{amb}	S_{raw}	S_{cond}
Baseline	0.327	0.673	0.212	0.299	0.330	0.670	0.211	0.299
Full instruction	0.830	0.170	0.067	0.326	0.830	0.170	0.067	0.326
Routed	0.806	0.194	0.077	0.317	0.806	0.194	0.077	0.316

Table 6: Complete-case check for `Llama-3.1-8B`, 11-category macro averages over ambiguous examples. *As reported* uses every example with a valid majority in the given condition; *complete case* restricts all three conditions to the 24,919 examples with a valid majority in each. The largest change in any macro metric is 0.003, in baseline U_{amb} and N_{amb}; the full-instruction condition is unchanged to three decimals, since almost all of its examples already have a valid majority.

Category	n (amb/dis)	ΔU_{amb} [95% CI]	$\Delta\text{Acc}_{\text{dis}}$ [95% CI]	ΔS_{raw} [95% CI]	ΔS_{cond} [95% CI]
Age	1830/1840	+0.228 [+0.210, +0.248]	−0.011 [−0.016, −0.007]	−0.138 [−0.154, −0.122]	−0.167 [−0.210, −0.123]
Disability Status	771/776	+0.236 [+0.206, +0.266]	−0.010 [−0.018, −0.003]	−0.139 [−0.163, −0.115]	+0.207 [+0.116, +0.297]
Gender Identity	2811/2836	+0.039 [+0.032, +0.046]	−0.015 [−0.020, −0.011]	−0.025 [−0.031, −0.020]	+0.095 [−0.621, +0.409]†
Nationality	1538/1538	+0.092 [+0.078, +0.107]	−0.020 [−0.027, −0.013]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Physical Appearance	784/788	+0.194 [+0.166, +0.222]	−0.018 [−0.028, −0.009]	−0.091 [−0.111, −0.071]	+0.125 [+0.031, +0.214]
Race/Ethnicity	3434/3439	+0.017 [+0.013, +0.022]	−0.001 [−0.003, +0.001]	−0.002 [−0.003, −0.000]	+0.026 [−0.005, +0.058]
Race × Gender	7899/7978	+0.047 [+0.042, +0.051]	−0.001 [−0.002, +0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]†
Race × SES	5573/5580	+0.067 [+0.060, +0.073]	+0.000 [+0.000, +0.001]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Religion	600/599	+0.050 [+0.033, +0.068]	−0.012 [−0.022, −0.003]	−0.037 [−0.052, −0.022]	0.000 [−0.060, +0.063]
SES	3426/3432	+0.165 [+0.153, +0.178]	−0.027 [−0.032, −0.021]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Sexual Orientation	430/432	+0.072 [+0.049, +0.098]	−0.009 [−0.019, −0.002]	−0.051 [−0.072, −0.033]	−0.378 [−0.649, −0.093]†

Table 7: Base-to-routed shifts with paired percentile-bootstrap 95% confidence intervals (B=10,000; fixed seeds; `gpt-4o-mini`). An example enters the paired set only when both conditions produce a strict rotation majority; each bootstrap replicate resamples the same example IDs in both conditions. Since $N_{\text{amb}} = 1 - U_{\text{amb}}$, ΔN_{amb} and its CI are the sign-reversed counterparts of ΔU_{amb}. †: routed non-unknown support below 10 examples (Table 1: Gender Identity 4, Race × Gender 4, Sexual Orientation 9); these ΔS_{cond} cells are excluded from substantive interpretation.

Category	ΔU_{amb}	$\Delta\text{Acc}_{\text{dis}}$	ΔS_{raw}	ΔS_{cond}
Age	+0.231 [+0.212, +0.250]	−0.058 [−0.068, −0.047]	−0.139 [−0.155, −0.123]	−0.172 [−0.216, −0.128]
Disability Status	+0.236 [+0.206, +0.266]	−0.021 [−0.032, −0.010]	−0.139 [−0.163, −0.114]	+0.207 [+0.116, +0.295]
Gender Identity	+0.039 [+0.032, +0.046]	−0.019 [−0.024, −0.014]	−0.025 [−0.031, −0.020]	+0.095 [−0.612, +0.405]†
Nationality	+0.092 [+0.078, +0.107]	−0.020 [−0.027, −0.013]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Physical Appearance	+0.212 [+0.184, +0.241]	−0.062 [−0.080, −0.046]	−0.103 [−0.125, −0.083]	+0.111 [+0.006, +0.217]
Race/Ethnicity	+0.017 [+0.013, +0.022]	−0.001 [−0.003, +0.001]	−0.002 [−0.003, −0.000]	+0.026 [−0.004, +0.057]
Race × Gender	+0.047 [+0.042, +0.052]	−0.001 [−0.002, +0.001]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]†
Race × SES	+0.067 [+0.060, +0.073]	+0.000 [+0.000, +0.001]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Religion	+0.063 [+0.045, +0.083]	−0.057 [−0.077, −0.037]	−0.043 [−0.060, −0.028]	+0.034 [−0.040, +0.113]
SES	+0.169 [+0.156, +0.181]	−0.028 [−0.034, −0.022]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Sexual Orientation	+0.072 [+0.049, +0.098]	−0.019 [−0.032, −0.005]	−0.051 [−0.074, −0.030]	−0.378 [−0.656, −0.100]†

Table 8: Base-to-full-instruction shifts with paired-bootstrap 95% CIs (`gpt-4o-mini`). †: non-unknown support below 10 in the full condition; excluded from substantive interpretation.

Category	Stereo ret.	Anti ret.	OR	p
Age	89/345	196/364	0.30	$< 10^{-5}$
Disability Status	49/156	8/83	4.29	1.1×10^{-4}
Physical Appearance	39/120	22/107	1.86	0.051
Religion	43/69	13/25	1.53	0.476

Table 9: Retention of baseline non-unknown predictions under the full instruction, stratified by baseline direction (stereotype-consistent vs. anti-stereotype). $\text{OR} = \text{odds}(\text{retain} \mid \text{stereo})/\text{odds}(\text{retain} \mid \text{anti})$; p from two-sided Fisher exact tests (unadjusted; Holm correction is applied across the seven categories with defined ORs in the text). Categories discussed in Section 6.3 are shown; Table 10 gives all categories under both conditions.

	Full instruction				Routed			
Category	Stereo ret.	Anti ret.	OR	p	Stereo ret.	Anti ret.	OR	p
Age	89/345	196/364	0.30	<0.001	92/345	198/364	0.30	<0.001
Disability Status	49/156	8/83	4.29	<0.001	49/156	8/83	4.29	<0.001
Gender Identity	3/74	1/39	1.61	1.000	3/74	1/39	1.61	1.000
Nationality	0/0	120/261	–	–	0/0	120/261	–	–
Physical Appearance	39/120	22/107	1.86	0.051	49/120	26/107	2.15	0.011
Race/Ethnicity	31/38	129/182	1.82	0.230	31/38	129/182	1.82	0.230
Race × Gender	0/0	2/372	–	–	0/0	4/372	–	–
Race × SES	0/0	43/415	–	–	0/0	43/415	–	–
Religion	43/69	13/25	1.53	0.476	47/69	17/25	1.01	1.000
SES	0/0	81/660	–	–	0/0	92/660	–	–
Sexual Orientation	2/24	7/16	0.12	0.018	2/24	7/16	0.12	0.018

Table 10: Direction-stratified retention of baseline non-unknown predictions for all categories. $\text{OR} = \text{odds}(\text{retain} \mid \text{stereo})/\text{odds}(\text{retain} \mid \text{anti})$; p from two-sided Fisher exact tests. "–": odds ratio undefined because the baseline stereotype stratum is empty.

Model	N_{amb} B→R	S_{raw} B→R	S_{cond} B→R (k)
`gpt-4o-mini`	0.166→0.056	0.068→0.025	0.322→0.346 (8)
`Claude-Haiku-4.5`	0.038→0.016	0.019→0.007	0.468→0.413 (4)
`Qwen2.5-7B`	0.051→0.014	0.023→0.005	0.334→0.285 (6)
`Llama-3.1-8B`	0.673→0.194	0.212→0.077	0.299→0.317 (11)

Table 11: Cross-model macro decomposition, baseline (B) to routed (R). N_{amb} and S_{raw} average all 11 categories; the S_{cond} comparison is restricted to the k common-support categories with at least 10 non-unknown predictions in both conditions, so both entries average the same category set.

Model	Base U	Full U	Cond. U	Base Acc.	Full Acc.	Cond. Acc.	B
`gpt-4o-mini`	0.834	0.947	0.944	0.962	0.936	0.951	0.946
`Claude-Haiku-4.5`	0.962	0.986	0.984	0.868	0.837	0.858	0.914
`Qwen2.5-7B`	0.949	0.988	0.986	0.817	0.707	0.796	0.879
`Llama-3.1-8B`	0.327	0.830	0.806	0.936	0.862	0.896	0.846
Macro avg.	0.768	0.938	0.930	0.896	0.835	0.875	0.896

Table 12: Cross-model robustness across 11 BBQ categories. Rows report category-level macro averages. Conditional prompting largely preserves the abstention gain of full answerability instruction while better retaining disambiguated accuracy. U denotes ambiguous-context unknown rate, Acc. denotes disambiguated-context accuracy, and B is the mean of per-category balance scores in the routed condition (same convention as Table 3).

	Baseline			Full			Routed		
Category	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)
Age	0.388	0.188	0.485 (712)	0.156	0.049	0.315 (286)	0.159	0.051	0.320 (291)
Disability Status	0.310	0.202	0.653 (239)	0.074	0.064	0.860 (57)	0.074	0.064	0.860 (57)
Gender Identity	0.040	0.026	0.655 (113)	0.001	0.001	0.750 (4)	0.001	0.001	0.750 (4)
Nationality	0.170	0.000	0.000 (261)	0.078	0.000	0.000 (120)	0.078	0.000	0.000 (120)
Physical Appearance	0.290	0.153	0.529 (227)	0.078	0.050	0.639 (61)	0.096	0.062	0.653 (75)
Race/Ethnicity	0.065	0.011	0.176 (222)	0.047	0.009	0.199 (161)	0.047	0.009	0.199 (161)
Race × Gender	0.047	0.000	0.000 (373)	0.000	0.000	0.000 (2)	0.001	0.000	0.000 (4)
Race × SES	0.074	0.000	0.000 (415)	0.008	0.000	0.000 (44)	0.008	0.000	0.000 (44)
Religion	0.157	0.115	0.734 (94)	0.093	0.072	0.768 (56)	0.107	0.078	0.734 (64)
SES	0.193	0.000	0.000 (660)	0.024	0.000	0.000 (82)	0.027	0.000	0.000 (93)
Sexual Orientation	0.093	0.056	0.600 (40)	0.021	0.005	0.222 (9)	0.021	0.005	0.222 (9)

Table 13: Full decomposition, `gpt-4o-mini`.

	Baseline			Full			Routed		
Category	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)
Age	0.223	0.114	0.512 (410)	0.075	0.023	0.307 (137)	0.075	0.023	0.307 (137)
Disability Status	0.012	0.005	0.444 (9)	0.001	0.000	0.000 (1)	0.001	0.000	0.000 (1)
Gender Identity	0.000	0.000	– (0)	0.006	0.000	0.056 (18)	0.006	0.000	0.056 (18)
Nationality	0.039	0.000	0.000 (60)	0.008	0.000	0.000 (13)	0.011	0.000	0.000 (17)
Physical Appearance	0.034	0.022	0.630 (27)	0.010	0.004	0.375 (8)	0.030	0.020	0.667 (24)
Race/Ethnicity	0.000	0.000	1.000 (1)	0.001	0.000	0.000 (3)	0.001	0.000	0.000 (3)
Race × Gender	0.000	0.000	– (0)	0.000	0.000	– (0)	0.000	0.000	– (0)
Race × SES	0.004	0.000	0.000 (23)	0.000	0.000	0.000 (1)	0.000	0.000	0.000 (1)
Religion	0.093	0.068	0.732 (56)	0.052	0.035	0.677 (31)	0.052	0.035	0.677 (31)
SES	0.013	0.000	0.000 (44)	0.000	0.000	– (0)	0.000	0.000	0.000 (1)
Sexual Orientation	0.000	0.000	– (0)	0.000	0.000	– (0)	0.000	0.000	– (0)

Table 14: Full decomposition, `Claude-Haiku-4.5`.

	Baseline			Full			Routed		
Category	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)
Age	0.187	0.094	0.504 (343)	0.029	0.009	0.315 (54)	0.030	0.010	0.327 (55)
Disability Status	0.041	0.028	0.688 (32)	0.001	0.000	0.000 (1)	0.005	0.004	0.750 (4)
Gender Identity	0.003	0.001	0.333 (9)	0.000	0.000	– (0)	0.000	0.000	– (0)
Nationality	0.045	0.000	0.000 (69)	0.015	0.000	0.000 (23)	0.016	0.000	0.000 (25)
Physical Appearance	0.065	0.038	0.588 (51)	0.019	0.013	0.667 (15)	0.020	0.013	0.625 (16)
Race/Ethnicity	0.038	0.005	0.136 (132)	0.015	0.001	0.059 (51)	0.033	0.004	0.113 (115)
Race × Gender	0.001	0.000	0.000 (6)	0.000	0.000	– (0)	0.000	0.000	– (0)
Race × SES	0.039	0.000	0.000 (217)	0.006	0.000	0.000 (34)	0.007	0.000	0.000 (37)
Religion	0.090	0.070	0.778 (54)	0.022	0.013	0.615 (13)	0.023	0.015	0.643 (14)
SES	0.021	0.000	0.000 (73)	0.000	0.000	– (0)	0.001	0.000	0.000 (5)
Sexual Orientation	0.028	0.019	0.667 (12)	0.021	0.012	0.556 (9)	0.021	0.012	0.556 (9)

Table 15: Full decomposition, `Qwen2.5-7B`.

	Baseline			Full			Routed		
Category	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)	N	S_{raw}	S_{cond} (n)
Age	0.854	0.402	0.471 (1487)	0.330	0.175	0.530 (562)	0.364	0.189	0.519 (619)
Disability Status	0.864	0.535	0.620 (623)	0.202	0.149	0.739 (142)	0.202	0.148	0.732 (142)
Gender Identity	0.709	0.325	0.458 (1784)	0.187	0.078	0.419 (465)	0.266	0.107	0.402 (662)
Nationality	0.584	0.000	0.000 (744)	0.133	0.000	0.000 (169)	0.133	0.000	0.000 (169)
Physical Appearance	0.750	0.393	0.524 (567)	0.254	0.150	0.592 (191)	0.290	0.186	0.642 (218)
Race/Ethnicity	0.641	0.070	0.109 (1870)	0.212	0.026	0.123 (612)	0.217	0.026	0.119 (628)
Race × Gender	0.645	0.000	0.000 (4071)	0.054	0.000	0.000 (340)	0.056	0.000	0.000 (350)
Race × SES	0.534	0.000	0.000 (2571)	0.082	0.000	0.000 (393)	0.082	0.000	0.000 (394)
Religion	0.520	0.288	0.554 (271)	0.135	0.096	0.714 (70)	0.183	0.110	0.600 (95)
SES	0.720	0.000	0.000 (2330)	0.162	0.000	0.000 (516)	0.174	0.000	0.000 (552)
Sexual Orientation	0.584	0.324	0.555 (209)	0.123	0.057	0.465 (43)	0.168	0.080	0.475 (59)

Table 16: Full decomposition, `Llama-3.1-8B`.

Category	Ambig Interv.	Disambig Interv.	Gen. Tok.	Gen. Lat.	E2E Tok.	E2E Lat.
Age	0.910	0.129	0.960	1.048	1.107	1.052
Disability Status	0.999	0.537	0.986	1.008	1.113	1.010
Gender Identity	0.894	0.426	0.977	1.108	1.103	1.110
Nationality	0.988	0.429	0.983	1.014	1.134	1.018
Physical Appearance	0.871	0.110	0.971	1.205	1.112	1.207
Race/Ethnicity	0.988	0.620	0.962	0.968	1.104	0.969
Race × Gender	0.979	0.553	0.985	0.952	1.113	0.953
Race × SES	0.995	0.621	0.989	1.011	1.128	1.013
Religion	0.815	0.207	0.971	1.016	1.112	1.017
SES	0.984	0.376	0.980	1.009	1.121	1.010
Sexual Orientation	0.874	0.273	0.974	1.063	1.117	1.065
Macro avg.	0.936	0.389	0.976	1.036	1.115	1.039

Table 17: Resource accounting for conditional routing. Generation ratios compare the selected prompt branch against applying the answerability instruction to all examples. End-to-end ratios additionally include the embedding-based routing step.

H Prompt Templates

Baseline prompt.

```
You are answering a multiple-choice question.
Read the context and the question. Choose the
best answer among A, B, C.
- Output ONLY one letter: A or B or C.
- Do not output any explanation.
Context:
{context}
Question:
{question}
A) {ans0}
B) {ans1}
C) {ans2}
```

Answerability-instruction prompt.

```
Read the context and question carefully.
Choose the answer best supported by the context.
If the context does not provide enough
information, choose the option that indicates
insufficient information or that the answer is
unknown.
Answer with only one option name: A, B, or C.
Context:
{context}
Question:
{question}
A) {ans0}
B) {ans1}
C) {ans2}
```