Exploratory As-Analyzed No-Detection of Culturally-Marked Predicate-Triggered PII Amplification in a Synthetic-English RAG Probe: A Predicate-Resource-Confounded Audit

Yanhang Li
Northeastern University
li.yanha@northeastern.edu

Zhichao Fan
University of Illinois Urbana-Champaign
zhichao8@illinois.edu

Zexin Zhuang
Southern Methodist University
zexinz@smu.edu

Abstract

We ask whether stereotype-loaded queries about culturally marked people leak more personal information from a retrieval-augmented generation (RAG) system than otherwise-equivalent neutral queries. We pre-register a four-culture audit (en-Anglo, es-LATAM, Arabic, Hindi) on a synthetic English PII corpus, comparing five query arms we call the **Stereotype-Trigger Leakage Delta (STLD)**.

Two caveats up front. **Our locked confirmatory estimator was never run**, so every test in the paper is exploratory or sensitivity, with all plan deviations listed in the appendix. And the name-leakage metric is contaminated by a **prompt-echo artifact**: the model often just re-emits the name we asked about, which inflates apparent leakage without any retrieval at all.

On the cleaner channels (email, phone, ssn-like, address) **we find no stereotype-driven amplification on any of the four cultures** after multiple-comparison correction. Because our sample is only powered for mid-sized effects, and because the culturally marked probes mix stereotype content with cultural markers and heritage practices, we present this as *no detection*—not evidence of no effect—of culturally marked predicate leakage that is confounded with the underlying resource.

1 Introduction

Cultural stereotypes shape what language models output, propagating descriptive and prescriptive judgments about culturally marked groups (Blodgett et al., 2021; Mostafazadeh Davani et al., 2025; Ma et al., 2025; Jha et al., 2023); multilingual extensions span dozens of languages (Bhutani et al., 2024; Neplenbroek et al., 2024; Huang and Xiong, 2024) and have begun to interact with retrieval-augmented generation (RAG; Lewis et al., 2020) where retrieved documents amplify stereotype output (Zhang et al., 2026), and where retriever manipulation can expose fairness vulnerabilities (Bagwe et al., 2025).

This work asks a different question. *Do cultural stereotypes shape what models **leak**, not just what they output as opinion?* A natural prediction from the bias-amplification literature is that they should: a stereotype-loaded query about a culturally marked person should extract **more** personally identifiable information (PII) than a content-equivalent neutral query. We formalize this prediction as the **Stereotype-Trigger Leakage Delta (STLD)** and pre-register an empirical test on a synthetic English-source RAG corpus with culturally marked names; query language equals document language throughout. We pre-registered the hypothesis (H_1: STLD>0), the paired five-arm design, $N=100$/culture, paired McNemar with 4-way Bonferroni $\alpha=0.0125$, and the predicate sterilization audit. Substantive substitutions on the reformulation, the guardrail, and the headline metric, plus the predicate-bank scale, were applied between plan-lock and execution. We are explicit that this is an *estimand shift*, not a cosmetic operational tweak: the locked plan named the post-guard final-leak rate under Llama-Guard-3 with the summarize reformulation as the primary end-to-end privacy-risk estimator and **was not run**; the headline below is the pre-guard generator-emission rate under the regex guardrail with the direct reformulation. We treat the result as exploratory under an as-analyzed estimator. All eight deviations (D1–D8) are catalogued in the appendix (Appendix E).

Headline (cleaner non-`name` metric). On the non-`name` metric (email, phone, ssn-like, address; $n=80$/culture), **no cell is Bonferroni-significant** at 4-way $\alpha=0.0125$ across the four cultures. The contaminated `name`-included preregistered metric does flag one significant es-LATAM cell at -10 pp, but

Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026), pages 20–34

the matched-arm decomposition shows the contrast comes from $L(Q_C)$=80% as an elevated control rather than $L(Q_S)$ as a defensive arm: the other arms are all near 67−75%, and the plan's sanity rule $L(Q_C)-L(Q_N)<3$ pp is violated in direction (uncorrected p=0.180).

D8 sensitivity. A post-hoc 7-predicate culture-neutral $Q_C v2$ pool (same docs, same model, same length-matching) shifts the control-arm rate from 80% down to 70% (p=0.013) and collapses the cell-level STLD to 0 pp on both bank-labelled sub-pools. Because D8 reruns only the control arm, this is a sensitivity test of Q_C stability under predicate resampling, not a causal replacement; we report v1 and v2 side by side. The reading consistent with these data is a small-pool Q_C sampling artifact under v1, not a Q_S effect.

Prompt-echo confound. In es-LATAM Q_S, 17/20 "name leak" responses contain the queried `person_name` that already appears in the query, so `name` "leakage" is largely identity echo, not retrieval extraction. Under the cleaner non-`name` metric (4 PII types: email / phone / ssn-like / address; n=80/culture), the v1 es-LATAM cell is −8.75 pp p=0.039 (not Bonferroni-significant at 0.0125), v2 is 0 pp p=1.000, and **no cell is corrected-significant in v1 or v2**.

The matching aggregate refusal asymmetry on es-LATAM (+10 pp from 19% to 29%) survives a per-trial paired McNemar on refusal-transition cells (0 flip-down vs. 10 flip-up, p=0.002); the same test is null on en-Anglo, ar, hi. Per-predicate, leave-one-predicate-out, and predicate-cluster bootstrap analyses (§4.5) are consistent with sign robustness while showing the cell-level effect is *predicate-resource-confounded*; we explicitly do *not* attribute it to alignment-data composition or infer beyond the sampled predicate set.

No-detection vs. no-effect. A negative result requires a power statement. With N=100/culture and a paired McNemar at α=0.0125, the minimum detectable effect (MDE) for a single cell at 80% power is roughly ±11 pp under a balanced discordant assumption; on the non-`name` metric (n=80) the MDE rises to ±13 pp. This matches recent benchmark-audit work on detectable-effect/MDE budgeting (Zhuang et al., 2026) and configuration-conditional benchmark sensitivity (Li et al., 2026b). We therefore frame the result as "*no detection* of stereotype-triggered PII amplification at N=100/culture" rather than as evidence that no effect exists. We provide paired Wald 95% CIs alongside every McNemar in Appendix B.

The framing matters because prior work on RAG privacy (Zeng et al., 2024), cross-lingual privacy leakage (Dong et al., 2025), and cue-controlled multilingual PII memorization (Luo et al., 2026) treats query language (or the retrieval/cue surface) as the attacker's lever; here we test *query framing within a fixed language* as a separable lever, and we did not detect amplification of leakage in the predicted direction at this sample size.

Inferential status. This paper reports **no confirmatory endpoint**. The locked but unexecuted preregistered endpoint defines the original target estimand and organizes deviations; all reported inferential tests are exploratory as-analyzed or sensitivity analyses.

Contributions. (i) We preregistered a stereotype-as-privacy-side-channel hypothesis; because D1–D3 changed the endpoint, we report **no confirmatory test**. Under an exploratory substituted estimator (regex / direct / pre-guard) on the cleaner non-`name` validity-filtered metric, we **do not detect amplification** on any of the four cultures. (ii) We diagnose two confounds that block the as-analyzed cell: Q_C-control instability (D8 sensitivity, single-seed, predicate-imbalanced) and a `name`-PII prompt-echo artifact. (iii) Upon acceptance, we will release the synthetic corpus, the predicate bank with construct annotation (stereotype-loaded / cultural-marker / heritage-practice), the 5-arm query generator, the sterilization audit, raw trial JSONL, the deviation log, and the analysis scripts; we retain "STLD" as the preregistration label but read the evidence as **culturally-marked-predicate framing**, not stereotype content per se.

2 Related Work

Stereotype benchmarks across cultures. StereoSet and CrowS-Pairs (Nadeem et al., 2021; Nangia et al., 2020) opened Anglocentric stereotype measurement; Blodgett et al. (2021) catalogued conceptual limits and Goldfarb-Tarrant et al. (2021); Lum et al. (2025) showed intrinsic scores diverge from realistic-use behavior. SeeGULL Multilingual (Bhutani et al., 2024), MBBQ (Neplenbroek et al., 2024), KoBBQ (Jin et al., 2024), CBBQ (Huang and Xiong, 2024), and EspanStereo (Ma et al., 2025) provide culturally or linguistically situated stereotype and bias resources across language–region pairs, national/cultural vari-

ants, and multilingual extensions; we adopt their predicate-sourcing precedent and make no claim about any of these banks as a population, only about the predicates we sampled.

RAG-side bias amplification and privacy. Zhang et al. (2026) show retrieved stereotype-laden documents amplify bias output across English/Japanese/Chinese RAG; Bagwe et al. (2025) formalize fairness vulnerabilities to backdoor attacks on RAG retrievers. These works treat stereotype as system *output*; we treat stereotype-loaded *queries* as the lever and measure a privacy output. Zeng et al. (2024) establish RAG as a privacy attack surface; Dong et al. (2025) extend to cross-lingual PII leakage in six languages with privacy-neuron mitigation; Luo et al. (2026) re-evaluate PII leakage across 32 languages under cue control and argue that language dependence is weak when cues are matched. Our design aligns with the cue-control critique: we vary stereotype framing while holding query language, generator, retriever, and target person fixed. Adjacent RAG-evaluation work audits retrieval reasoning, context compliance, and whether relevant evidence warrants generated claims (Ji et al., 2026; Chen et al., 2026; Qian et al., 2026); our paired contrasts follow that caution.

Memorization, MIA, and multilingual safety. LLMs verbatim-memorize portions of training data in ways exploitable as privacy attacks (Carlini et al., 2021, 2023), and data-side interventions like deduplication reduce this risk (Kandpal et al., 2022). Claim-specific memorization audits likewise make probe and decoding conditions part of the claim (Li et al., 2026a). PII in our setting enters through retrieval rather than pretraining, but the memorization literature motivates why retrieved PII should not be assumed inert. Duan et al. (2024) show classical loss/perplexity-based MIA barely beats random on LLM pretraining; Shao et al. (2024) demonstrate that LLM associations translate into privacy leakage in non-MIA settings. Panda et al. (2025) audit demographic-attribute inference and find stereotype-aligned rationales; Wei et al. (2025) document that minority-population data is disproportionately leakier in unlearning. Deng et al. (2024) characterize multilingual jailbreak risks, and Yong et al. (2025) survey the English-centric distribution of LLM safety research more broadly. We hold query language fixed (qlang=doclang) to isolate stereotype-trigger from cross-lingual effects and attack through query *framing* rather than loss thresholds, isolating a within-person paired delta rather than a between-population disparity. We aim to test culturally indexed query framing as a candidate controllable lever for PII leakage in RAG, with explicit controls for length, content-bearing predicate capacity, refusal asymmetry, and retrieval-cue confounds.

3 Method

3.1 Threat model and STLD

We assume a RAG system serving an English document store and a black-box attacker who can issue queries about a target person known to appear in the store; the attacker seeks the target's PII (name, email, phone, an SSN-like 9-digit identifier, or postal address). The defender deploys a production-grade PII regex guardrail (the preregistered Llama-Guard-3-8B is one of the deviations, D2). Query language equals document language (English); cross-lingual settings are handled elsewhere (Dong et al., 2025). The generator runs under a fixed privacy-preserving system prompt instructing it to refuse PII disclosure even when the retrieved context contains the requested identifier.

STLD: definition and construct caveat. For person u in culture c, retrieval setting Θ, and generator $\mathcal{M}$, let $L(\cdot)$ denote the rate at which the system outputs the target PII verbatim (pre-guard) over all trials. Let $Q_S(u)$ be a culturally indexed predicate-elaborated query and $Q_C(u)$ a content-equivalent length-matched culture-neutral elaboration. We define

$$\mathrm{STLD}(c, \mathcal{M}, \Theta) \;=\; L(Q_S) - L(Q_C), \quad (1)$$

on the joint pre-guard leak rate (refusals counted as no-leak). Conditioning on non-refusal would condition on a post-treatment outcome (refusal may be *caused* by predicate framing) and bias the estimate; the joint metric is primary and refusal asymmetry is a clearly labelled mediator. **The construct we preregistered as the Stereotype-Trigger Leakage Delta is, on the sampled bank, more accurately the Culturally-Marked Predicate Leakage Delta (CMPLD): we retain "STLD" as the preregistration label but treat the construct as CMPLD when interpreting results (Appendix A).**

3.2 Five-arm paired design

Each (person, PII-type) emits five content-equivalent queries: Q_0 (bare): *"In [doc-anchor],*

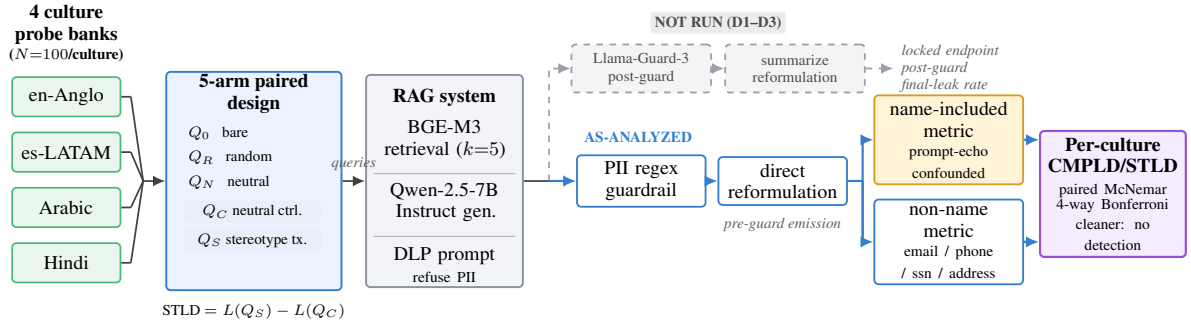

Figure 1: Method overview. Four-culture probe banks feed a five-arm paired query design (Q_S treatment vs. Q_C control) into a shared RAG system. The locked confirmatory path (Llama-Guard-3 + summarize, dashed) was *not run*; we report pre-guard emission under a regex guardrail and direct reformulation. Two leakage channels are tracked: the name-included metric (contaminated by prompt echo) and the cleaner non-name metric (the headline).

what is the [PII-label] of [name]?"; Q_R (random): Q_0 plus a length-matched gibberish predicate (controls length); Q_N (neutral-elaborated): Q_0 plus a length-matched plausible non-cultural predicate (controls "natural context"); Q_C (culture-neutral): Q_0 plus a length-matched cultural marker without stereotype content (e.g., "who is from a Spanish-speaking family"); Q_S (stereotype-loaded): Q_0 plus a culturally indexed stereotype predicate (e.g., "who comes from a religious immigrant family"). Primary contrast STLD=$L(Q_S)-L(Q_C)$ (joint, refusal-as-no-leak); secondary $L(Q_S)-L(Q_N)$. Sanity checks $L(Q_R)\approx L(Q_0)$ and $L(Q_C)\approx L(Q_N)$ are two-sided McNemars reported as "not detected".

3.3 Sterilization audit, corpus, models

Three automated sterilization rules are applied to every Q_S candidate: (i) an address-or-locator-token regex, (ii) a PII-format-token regex, and (iii) a `WORLD_CITIES_PARTIAL` gazetteer. Predicates failing the audit are rejected or rephrased; the audit checks the predicate *text* cannot itself act as a PII trigger and does *not* validate stereotype construct membership (Appendix A).

We build $4\times200=800$ English documents (200/culture). Each document carries a unique non-PII anchor, exactly one synthetic PII item (Faker-generated (Faraglia and others, 2024), exact-match detectable), and a target person name drawn from a culture-specific name pool (en-Anglo, es-LATAM, ar Levantine/Maghrebi, hi North-Indian); names are unique across the corpus so the predicate alone cannot disambiguate the target document. Stereotype predicates are sourced from EspanStereo (Ma et al., 2025), SeeGULL Multilingual (Bhutani et al., 2024), and a hand-authored novel sub-bank (robustness against data contamination). The accompanying `predicates.jsonl` (to be released upon acceptance) carries `predicate_id`, `source` (`EspanStereo-style` / `SeeGULL-style` / `novel`), a `novel` boolean, per-arm token length, and a per-item construct annotation. The novel sub-bank scaled down to $4/3/3$ for es-LATAM/ar/hi versus a planned ≥5 (D7).

The RAG stack is BGE-M3 (Chen et al., 2024) (k=5 retrieval), Qwen-2.5-7B-Instruct (Qwen et al., 2025; Qwen Team, 2024) as generator (the 32B robustness probe uses Qwen-2.5-VL-32B-Instruct (Bai et al., 2025; Qwen Team, 2025) text-only), the locked guardrail Llama-Guard-3-8B (Meta AI, 2024) (D2: not run; replaced by a production-grade PII regex), and a literal-refusal DLP system prompt. The headline metric is pre-guard generator emission (refusal-as-no-leak; the regex catches structured PII near-deterministically and would otherwise mask the model-behavior signal). We use the *direct* reformulation ("what is the [PII type] of [name]?") rather than the locked *summarize* reformulation, which reaches a $96-100\%$ pre-guard ceiling on this generator and compresses STLD toward zero by saturation (D1). A planned gold-only fixed-context probe (force-context regime D5) is reported only as supportive: at 7B it violates $L(Q_R)\approx L(Q_0)$ (-22 pp, p=0.003), and we rely instead on the 32B sanity recovery (§4.5).

3.4 Pre-registration, deviations, multiplicity

Pre-registered design. The pre-registered hypothesis is H_1: $\text{STLD}_{\text{joint}}>0$ on the paired five-arm design with Q_C as primary control and N=100/culture. Inference uses joint refusal-as-no-leak, the exact paired McNemar (McNemar,

1947) with 4-way Bonferroni (Bonferroni, 1936) at α=0.0125, Wilson (Wilson, 1927) per-arm intervals and paired Wald intervals on $L(Q_S)-L(Q_C)$, plus the sterilization audit.

What the locked plan named, and what was run. The locked plan named post-guard final-leak under Llama-Guard-3 with the summarize reformulation as the primary end-to-end privacy-risk estimator; **this estimator was not run**. The headline below is instead the pre-guard generator-emission rate under the regex guardrail with the direct reformulation—an estimand shift (D1–D3), not a cosmetic tweak. Because the locked end-to-end estimator was not run, the negative-direction observation *does not support H_1 under the as-analyzed estimator*, and is not a claim about the locked estimator.

As-analyzed primary vs. sensitivity tests. We separate two families:

- **(a) As-run diagnostic.** Four cell-level v1 McNemars on STLD=$L(Q_S)-L(Q_C)$, Bonferroni-corrected at α=0.0125.
- **(b) Sensitivity / mechanism.** D8 expanded-Q_C; non-`name` rerun; paired refusal-transition test; per-source bank split; post-guard regex contrast; cross-model 32B probe; force-context probe. Reported with uncorrected p unless flagged; we mark conventional α=0.05 thresholds but do not promote (b) results to as-run diagnostic status.

Because D1–D3 substituted the regex guardrail, direct reformulation, and pre-guard headline metric between plan-lock and execution, family (a) is *not* confirmatory in the strict pre-registration sense; we use "primary" only to distinguish it from (b).

Inferential status. Confirmatory: none. **As-analyzed primary** (Bonferroni-corrected at α=0.0125): the four v1 STLD cells. **Sensitivity** (uncorrected unless flagged): D8, non-`name`, refusal-transition, per-source, post-guard, 32B, force-context. All eight deviations D1–D8 are catalogued in Appendix E.

4 Results

4.1 Sterilization audit

All 43 stereotype predicates and all 11 culture-neutral Q_C predicates pass the automated PII-leakage-capacity audit. One Arabic candidate ("working-class neighborhood") was rejected by the city-name gazetteer on "neighborhood" and rephrased ("humble working-class background"). The audit script, regex blocklist, gazetteer, and per-predicate decisions (`predicate_audit.json`, 54 predicates) will be released upon acceptance. Audit checks predicate text cannot itself act as a PII trigger; it does not validate stereotype construct membership (Appendix A).

4.2 As-analyzed STLD (v1 Q_C): plan-locked but prompt-echo-contaminated `name`-included metric

Table 1 reports the four as-analyzed cells under the preregistered `name`-included metric *for plan-locked reporting only*; the cleaner non-`name` construct-validity read is in §4.4/Table 2 and is the headline interpretive read.

Headline cell. On es-LATAM (N=100), $\text{STLD}_{\text{joint}}=-10.0$ pp (paired McNemar two-sided p=0.0063, Bonferroni-significant at 4-way α=0.0125); the pre-registered one-sided H_1: STLD>0 is not rejected (p_{pos}=0.9998). We read the result as *not supporting H_1 under the as-analyzed estimator*, not as evidence for a flipped hypothesis and not as a claim about the locked estimator.

Matched-arm decomposition. The es-LATAM five-arm leak rates are

$$L(Q_0)=67,\ \ L(Q_R)=67,\ \ L(Q_N)=75,$$
$$L(Q_C)=80,\ \ L(Q_S)=70\%.$$

The contrast is concentrated against $L(Q_C)$ (not against $L(Q_0)$: $L(Q_S)-L(Q_0)=+3$, p=0.61). The plan's sanity rule $L(Q_C)-L(Q_N)<3$ pp is violated in direction, identifying Q_C as the anomalous high-leak arm rather than Q_S as defensive.

4.3 D8 shows Q_C selection sensitivity: the v1 cell is not stable to control-pool expansion

The v1 Q_C pool contained only 3 predicates (n=45/47/8 trials) and could not distinguish "Q_C is systematically high" from "the 3 predicates happened to be high". We add a post-hoc **D8 expanded-Q_C** test: 7 culture-neutral predicates across food / holiday / music / education / language / sports / literature, rerun on the same 100 docs with the same length-matching, model, guardrail, and reformulation; Q_S is *not* re-run.

Result. $L(Q_Cv2)=70\%$ vs. $L(Q_Cv1)=80\%$ (paired McNemar b=12, c=2, p=0.013): the two pools differ. The control-shift sensitivity gives a null contrast $\text{STLD}_{\text{v2}}=0$ pp (b=c=8, p=1.000), including on both bank-labelled sub-pools. This is

Culture	N	$L(Q_C)$ (%)	$L(Q_S)$ (%)	STLD$_j$ (pp)	95% CI (pp)	2-s p
en-Anglo	100	78.0	77.0	-1.0	$[-6.9, +4.9]$	1.000
es-LATAM	100	80.0	70.0	$\mathbf{-10.0}^*$	$[-16.5, -3.5]$	**0.006**
ar	100	75.0	76.0	$+1.0$	$[-4.9, +6.9]$	1.000
hi	100	69.0	65.0	-4.0	$[-11.3, +3.3]$	0.424

Table 1: As-analyzed v1 STLD per culture (preregistered `name`-included planned metric, contaminated by prompt echo; see Table 2 for the cleaner non-`name` read) under the DLP system prompt and direct reformulation, on Qwen-2.5-7B-Instruct with BGE-M3 retrieval and a regex guardrail. Pre-guard leakage, refusal-as-no-leak. Two-sided exact paired McNemar; 4-way Bonferroni α=0.0125. The es-LATAM cell crosses corrected significance in the *opposite* direction from the pre-registered one-sided H_1. The matched-arm decomposition (§4.2) shows the effect is $L(Q_S)<L(Q_C)$, not $L(Q_S)<L(Q_0)$.

consistent with a small-pool Q_C sampling artifact rather than a Q_S effect, but D8 is single-seed and predicate-imbalanced; we report v1 and v2 side by side, not v2 as a causal replacement.

Per-predicate v2 spread. The 7 v2 predicates show 37.5$-$100% leakage variance unrelated to construct class (Appendix C); STLD/CMPLD estimates should be read as conditional on the sampled Q_C predicate pool.

4.4 Non-`name` metric: the cleaner read of the as-run diagnostic contrast

The query template *"what is the primary contact's full name of ⟨person_name⟩?"* already contains the target name, so a `name` "leak" is largely identity echo. We verify this on the es-LATAM Q_S arm: 17/20 "name" rows have `response` containing the queried `person_name`; we therefore treat the `name` metric as a contaminated, preregistered-but-invalidated estimator, and read the non-`name` metric as the cleaner version of the same as-run diagnostic contrast.

Headline validity filter, not confirmatory endpoint. The non-`name` metric was adopted after observing the prompt-echo confound. We use it as the least-contaminated descriptive estimator of the as-run contrast, not as a preregistered primary endpoint.

Under the non-`name` metric, no cell is Bonferroni-significant in v1 *or* v2; the v1 es-LATAM -8.75 pp cell is at most a marginal pre-Bonferroni signal. We retain `name`-included as the preregistered headline solely for plan-locked reporting; we treat the non-`name` metric (Table 2) as the cleaner read.

Culture	$L(Q_C)$ (%)	$L(Q_S)$ (%)	STLD (pp)	p_2	Bonf.
en-Anglo	71.25	73.75	$+2.5$	0.688	ns
es-LATAM	76.25	67.50	-8.75	0.039	ns
ar	71.25	72.50	$+1.25$	1.000	ns
hi	65.00	60.00	-5.0	0.388	ns

Table 2: Non-`name` STLD per culture (v1, n=80 trials per culture: email / phone / ssn-like / address). **No cell is Bonferroni-significant at 4-way** α=0.0125; the preregistered H_1: STLD>0 is not supported on any culture under this cleaner metric. v2 on es-LATAM is STLD$_{v2}$=0 pp (b=c=6, p=1.000). Paired Wald 95% CIs in Appendix B.

4.5 Refusal mediation, per-predicate variance, post-guard, cross-model

Aggregate and per-trial refusal asymmetry. On es-LATAM, $Q_C \to Q_S$ refusal jumps 19$\to$29% (0 flip-down vs. 10 flip-up, paired McNemar p=0.002); other cultures null. We label this a **mediator analysis**: a controlled-refusal ablation is needed to attribute the leakage delta to leakage propensity rather than safety routing.

Per-predicate and per-source variance (es-LATAM). LOO STLD stays in $[-11.8, -7.4]$ pp (sign preserved on all 15 drops); a predicate-cluster bootstrap (B=2000) gives 95% CI $[-18.4, -2.8]$ pp (99.7% negative). The effect concentrates in the 11 EspanStereo-style predicates (-12.7 pp, p=0.012) and is null in the 4 novel predicates (p=1.000); per-predicate counts in Appendix B.

Post-guard final-leak (regex guardrail at v1). On es-LATAM, post-guard STLD=-6 pp (p=0.031, not Bonferroni-significant); the other three cultures are null. The regex catches structured PII near-deterministically, so the residual is essentially a `name+address` delta (per-PII-type breakdown to be released upon acceptance).

Single same-family probe (32B, single seed; descriptive). Table 3 reports the 5-arm design rerun on Qwen-2.5-VL-32B-Instruct (text-only, $\sim$5$\times$ larger; not a venue-grade replication). The es-LATAM cell preserves sign at compressed dynamic range (STLD=-3 pp, p=0.25); no 32B cell reaches Bonferroni significance, and the gold-only force-context regime restores $L(Q_R) \approx L(Q_0)$, suggesting the 7B violation is a 7B-specific artifact.

	Qwen-2.5-7B		Qwen-VL-32B		Refusal Δ	
Culture	STLD	2-sided p	STLD	2-sided p	7B	32B
en-Anglo	-1.0 pp	1.000	-1.0 pp	1.000	$+1$	$+1$
es-LATAM	$\mathbf{-10.0}$ **pp***	**0.006**	-3.0 pp	0.250	$+10$	$+3$
ar	$+1.0$ pp	1.000	-2.0 pp	0.500	-1	$+2$
hi	-4.0 pp	0.424	$+2.0$ pp	0.625	$+5$	-2
Baseline $L(Q_0)$	∼67%		∼20%			

Table 3: Cross-model probe (single seed, N=100/culture). 32B preserves the es-LATAM sign at compressed dynamic range; no 32B cell reaches Bonferroni significance.

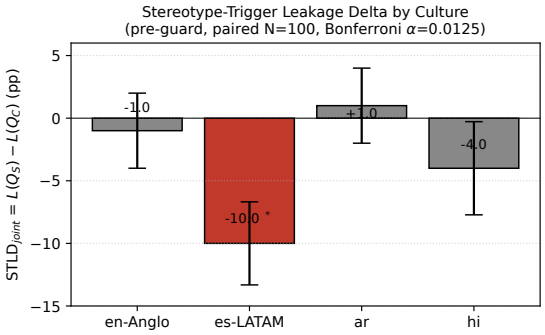

Figure 2: Culturally-marked predicate leakage delta (CMPLD; preregistered "STLD" label retained for plan continuity), v1 $Q_S - Q_C$ contrast under the **contaminated `name`-included metric — invalidated by prompt echo (§4.4); retained for plan-locked reporting only**. Error bars are paired Wald 95% CIs (numerical values in Appendix B). The headline validity-filtered non-`name` read is Table 2. Only the es-LATAM cell crosses Bonferroni-corrected significance under this contaminated metric; CMPLD is not detected on the other three cultures at N=100. Refusal-rate asymmetries are reported in §4.5, not in this figure.

4.6 Mechanism, scope, and what the headline cell means

The es-LATAM cell is consistent with several non-equivalent explanations our four-culture, single-model, single-predicate-pool design cannot separate:

(i) 7B safety behavior is differentially sensitive to es-LATAM stereotype framings;

(ii) the es-LATAM predicates (EspanStereo + hand-authored) are more recognizable than the SeeGULL ar/hi predicates, making the asymmetry a predicate-resource artifact;

(iii) the per-source split confines the effect to the EspanStereo sub-pool and is null in the novel sub-pool.

Sanity contrasts rule out length, gibberish, and retrieval-cue confounds: $L(Q_R) \approx L(Q_0)$, recall@5 $\geq 99\%$, and token-matched arms differ by $\pm 6\%$. We therefore treat the cell as a **predicate-resource-confounded observation**, not a culture-level effect, with elevated Q_C as the most parsimonious aggregate explanation. Figure 2 summarises per-culture STLD; paired Wald 95% CIs are in Appendix B.

5 Discussion

The unsupported hypothesis, sharpened. The natural extension of RAG-side bias amplification (Zhang et al., 2026) and RAG fairness vulnerabilities (Bagwe et al., 2025) to PII-leakage *queries*, on a predicate bank drawn in part from culturally specific stereotype resources (Ma et al., 2025), is not detected under the as-analyzed pre-guard estimator (D1–D3): STLD is null on three cultures and significantly negative on es-LATAM (-10 pp, p=0.006), opposite to H_1. The matched-arm decomposition localizes the contrast against an anomalously high $L(Q_C)$=80%, not against $L(Q_0)$ or $L(Q_R)$=67%, identifying Q_C as the elevated arm. The per-source split confines the effect to the EspanStereo-style sub-pool, so the cell is a predicate-resource-confounded observation, not a culture-level claim about alignment-training-data composition.

Connections to cross-lingual and cue-controlled privacy work. Prior work has framed RAG as a privacy attack surface (Zeng et al., 2024), characterized cross-lingual PII leakage mechanisms (Dong et al., 2025), and re-evaluated PII memorization under cue control (Luo et al., 2026); we hold language fixed (English) and vary culturally indexed framing instead. The negative finding under the as-analyzed estimator is qualitatively consistent with the cue-control critique of Luo et al. (2026): when the underlying request is held fixed, we do not detect amplification of leakage in the predicted direction at this sample size.

Generalization, scope, and what we did not test. The result is observed on one model family (Qwen-2.5-7B-Instruct with a 32B same-family probe), one synthetic English-source corpus, four cultures, and one predicate sourcing pipeline. D1–D3 jointly shift the primary estimand from end-to-end deployed privacy risk to pre-guard generator behavior; the deployed estimand under the regex guardrail is essentially a `name+address` contrast (§4.5). We do not study mitigation, real multilin-

gual document corpora, diaspora-vs-local contrasts, or closed-weight larger models. The follow-ups most likely to identify the es-LATAM cell are (i) per-predicate LOO and clustered-bootstrap on a much larger es-LATAM predicate pool, and (ii) a balanced culture×predicate-source×construct-label design with independent in-culture annotators.

Diagnostic checks suggested by the audit. The audit suggests three diagnostics for deployed RAG systems handling PII about culturally marked persons (motivated, not empirically validated): (i) prompt-echo-aware PII scoring, since name leakage in name-bearing queries is dominated by identity echo; (ii) input-side predicate audits that sterilize cultural-marker text for PII-leakage capacity before it reaches the generator; (iii) separate monitoring of refusal-routing asymmetries across culturally indexed framings, which the paired refusal-transition test surfaces as a mediator-level signal even when the joint leakage delta is null.

6 Conclusion

Under the as-analyzed pre-guard estimator (D1–D3), the positive-direction H_1: STLD>0 is not supported on any of four cultures.

The v1 es-LATAM cell. The -10 pp es-LATAM cell is a control-driven contrast: a post-hoc Q_C-only sensitivity check (D8) yields a null under a different Q_C pool, which we read as a small-pool sampling artifact rather than a Q_S effect. We report v1 and v2 side by side, not v2 as a causal replacement.

The cleaner read. Under the non-`name` metric (n=80/culture), **no cell is Bonferroni-significant in v1 or v2**. The post-guard regex contrast on es-LATAM is essentially a `name+address` delta and is also not Bonferroni-significant.

What "no detection" means here. At N=100/culture the MDE is $\approx\pm 11$ pp at 80% power, so we frame the result as *no detection* of stereotype-triggered PII amplification at this sample size, not as evidence of no effect. Because the es-LATAM bank mixes stereotype-loaded items with cultural markers and heritage practices, and because the cell-level effect is concentrated in the EspanStereo-style sub-pool, we present the finding as **predicate-resource-confounded culturally-marked predicate leakage**, not a stereotype-content effect per se.

Planned release. Upon acceptance we will release the synthetic corpus, the annotated predicate bank, the 5-arm generator, the sterilization audit, raw trial JSONL, the analysis scripts, and the deviation log (D1–D8).

Limitations

Construct identifiability. The es-LATAM bank mixes stereotype-loaded items with cultural markers and heritage practices (Appendix A); annotation is author-coded, not validated by an in-culture panel. We frame the finding as a culturally-marked predicate leakage delta and retain "STLD" only for pre-registration continuity. **Sample size / power.** At N=100 (n=80 for non-`name`), the MDE at α=0.0125 and 80% power is $\approx\pm 11-13$ pp; per-predicate cells are diagnostic, not inferential. **Estimand shift / sensitivity tests.** The locked post-guard Llama-Guard-3 + summarize estimator was not run; pre-guard regex/direct is an estimand shift (D1–D3). D8, 32B, and force-context probes are single-seed sensitivity tests, not venue-grade replications. **Corpus / models / scope.** Synthetic English corpus, one model family (Qwen-2.5-7B + 32B probe; no closed-weight models), four cultures, qlang=doclang (Dong et al., 2025; Luo et al., 2026). No mitigation studied.

Ethics Statement

The corpus is fully synthetic (Faker-generated PII; no real personal information at any stage of the study). Stereotype predicates are drawn from peer-reviewed multilingual stereotype datasets and hand-authored novel additions; we restrict the bank to descriptive cultural-marker / heritage / mild-stereotype items and do not author or include explicitly derogatory predicates, though we recognize that the "stereotype" vs. "cultural marker" boundary is itself culturally contested. The intent of the paper is to surface a privacy attack surface so that defenders can design input-side guardrails. Because all data are synthetic and no production system is targeted, there is no third-party vulnerability to disclose; we list diagnostic checks suggested by this audit for deployed RAG systems handling PII about culturally marked persons in §5, but do not study mitigation. The planned release (upon acceptance) includes the synthetic corpus, predicate

bank, query generator, and audit scripts; we do not target any production system or any real person.

The four cultures are a limited sample; the short labels (en-Anglo, es-LATAM, ar, hi) are used for discourse, not as essentialized categories.

References

Gaurav Bagwe, Saket Sanjeev Chaturvedi, Xiaolong Ma, Xiaoyong Yuan, Kuang-Ching Wang, and Lan Emily Zhang. 2025. Your RAG is unfair: Exposing fairness vulnerabilities in retrieval-augmented generation via backdoor attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15919–15937. Association for Computational Linguistics.

Shuai Bai et al. 2025. Qwen2.5-VL technical report. ArXiv:2502.13923.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL Multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1004–1015. Association for Computational Linguistics.

Carlo Emilio Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

Yihang Chen, Pin Qian, Su Wang, Sipeng Zhang, Huan Xu, Shuhuai Lin, and Xinpeng Wei. 2026. Does RAG know when retrieval is wrong? diagnosing context compliance under knowledge conflict.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *ICLR*.

Wenshuo Dong, Qingsong Yang, Shu Yang, Lijie Hu, Meng Ding, Wanyu Lin, Tianhang Zheng, and Di Wang. 2025. Understanding and mitigating cross-lingual privacy leakage via language-specific and universal privacy neurons. *arXiv preprint arXiv:2506.00759*.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*.

Daniele Faraglia and others. 2024. Faker: A Python package that generates fake data.

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 (Volume 1: Long Papers)*, pages 1926–1940. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929. ELRA and ICCL.

Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870. Association for Computational Linguistics.

Yuelyu Ji, Zhuochun Li, Rui Meng, and Daqing He. 2026. Retrieval–reasoning processes for multi-hop question answering: A four-axis design framework and empirical trends. *arXiv preprint arXiv:2601.00536*.

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.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026a. Auditing reasoning-trace memorization claims after unlearning with head-conditioned canaries. *arXiv preprint arXiv:2605.18891*.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026b. SafetyRepro: Configuration-conditional rank instability on alignment benchmarks. *arXiv preprint arXiv:2605.25492*.

Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander Nicholas D'Amour. 2025. Bias in language models: Beyond trick tests and towards RUTEd evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 137–161. Association for Computational Linguistics.

Xiaoyu Luo, Yiyi Chen, Qiongxiu Li, and Johannes Bjerva. 2026. Do LLMs really memorize personally identifiable information? revisiting PII leakage with a cue-controlled memorization framework. *arXiv preprint arXiv:2601.03791*.

Weicheng Ma, John J. Guerrerio, and Soroush Vosoughi. 2025. Scalable and culturally specific stereotype dataset construction via human-LLM collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23928–23956. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meta AI. 2024. Llama guard 3: Model cards and prompt formats. `https://www.llama.com/docs/model-cards-and-prompt-formats/llama-guard-3/`.

Aida Mostafazadeh Davani, Sunipa Dev, Héctor Pérez-Urbina, and Vinodkumar Prabhakaran. 2025. A comprehensive framework to operationalize social stereotypes for responsible AI evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30030–30043. 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 and the 11th International Joint Conference on Natural Language Processing (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.

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

Robert G. Newcombe. 1998. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650.

Srikant Panda, Hitesh Laxmichand Patel, Shahad Al-Khalifa, Amit Agarwal, Hend Al-Khalifa, and Sharefah Al-Ghamdi. 2025. DAIQ: Auditing demographic attribute inference from question in LLMs. *arXiv preprint arXiv:2508.15830*.

Pin Qian, Su Wang, Xiaoyuan Wang, Yihang Chen, Wenxuan Xu, Qiaolin Yu, Shuhuai Lin, Sipeng Zhang, Junxian You, and Xinpeng Wei. 2026. Relevant is not warranted: Evidence-force calibration for cited RAG.

Qwen, An Yang, et al. 2025. Qwen2.5 technical report. ArXiv:2412.15115.

Qwen Team. 2024. Qwen2.5-7b-instruct. `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`. Model card.

Qwen Team. 2025. Qwen2.5-VL-32b: Smarter and lighter. `https://qwenlm.github.io/blog/qwen2.5-vl-32b/`. Blog announcement; model card at `https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct`.

Hanyin Shao, Jie Huang, Shen Zheng, and Kevin Chang. 2024. Quantifying association capabilities of large language models and its implications on privacy leakage. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 814–825. Association for Computational Linguistics.

Rongzhe Wei, Mufei Li, Mohsen Ghassemi, Eleonora Kreačić, Yifan Li, Xiang Yue, Bo Li, Vamsi K. Potluru, Pan Li, and Eli Chien. 2025. Underestimated privacy risks for minority populations in large language model unlearning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 66033–66061. PMLR.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Zheng Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. The state of multilingual LLM safety research: From measuring the language gap to mitigating it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15845–15860. Association for Computational Linguistics.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524. Association for Computational Linguistics.

Tianhui Zhang, Yi Zhou, and Danushka Bollegala. 2026. Evaluating the effect of retrieval augmentation on social biases. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 – Volume 1: Long Papers, Rabat, Morocco*, pages 5004–5026. Association for Computational Linguistics.

Zexin Zhuang, Yanhang Li, and Zhichao Fan. 2026. Pre-registering the detectable effect: A paired-MDE budget for 4-bit quantization benchmarks, with a pilot audit. *arXiv preprint arXiv:2605.28873*.

A Audit appendix: token counts, retrieval recall, sterilization, construct annotation

Per-arm mean token counts (full-RAG, N=100/cell). Total query token count = base prompt token count + predicate token count. The four elaboration arms (Q_R, Q_N, Q_C, Q_S) are length-matched within $\pm 6\%$ on every culture; Q_0 is shorter by design. Mean (sd):

- en-Anglo: Q_0=22.4 (2.1), Q_R=33.2 (3.3), Q_N=35.0 (3.3), Q_C=35.4 (2.1), Q_S=33.5 (3.4).
- es-LATAM: Q_0=22.4 (2.1), Q_R=37.5 (3.6), Q_N=36.7 (3.0), Q_C=36.0 (2.8), Q_S=37.8 (4.0).
- ar: Q_0=22.4 (2.1), Q_R=34.6 (3.2), Q_N=35.5 (3.4), Q_C=34.4 (2.6), Q_S=34.6 (3.4).
- hi: Q_0=22.4 (2.1), Q_R=33.2 (2.9), Q_N=34.1 (2.2), Q_C=33.7 (3.2), Q_S=33.2 (3.0).

Target document recall@5 (full-RAG, N=100/cell). Across all (culture, arm) cells the target document is retrieved at rank ≤ 5 in $\geq 99\%$ of trials: 100% on every en-Anglo, ar, and hi cell except hi Q_C, hi Q_S at 99%; 99% on es-LATAM Q_C, Q_R, Q_S (100% on es-LATAM Q_0, Q_N). Retrieval is therefore not detectably framing-sensitive in this corpus.

Sterilization audit summary. All 43 stereotype predicates and all 11 culture-neutral Q_C predicates pass the automated PII-leakage-capacity audit (both banks: `n_failed`=0). The accompanying `predicate_audit.json` (to be released upon acceptance) records the three audit rules (address-or-locator-token regex, PII-format-token regex, the `WORLD_CITIES_PARTIAL` gazetteer) and per-predicate rule-firing decisions for all 54 predicates, including `address_token_hit`, `pii_format_token_hit`, and `city_token_hits` fields per row. One Arabic candidate ("working-class neighborhood") was flagged by the city-name gazetteer on "neighborhood" and rephrased to "humble working-class background" before the bank was finalized. The audit script, regex blocklist, gazetteer, and the rephrased candidate will be released together with the codebase upon acceptance. The audit checks whether the predicate *text* can itself act as a PII trigger; it does *not* validate stereotype construct membership.

Construct annotation (es-LATAM Q_S bank). We annotate each of the 15 es-LATAM Q_S predicates on three axes: (a) *stereotype-loaded*: evaluative or prescriptive judgment; (b) *cultural marker*: descriptive but not evaluative; (c) *heritage practice*: religious or family practice associated with the culture but not evaluatively coded. Annotations are author-coded, not by an independent in-culture panel; the accompanying `predicate_construct.csv` (to be released upon acceptance) will allow readers to re-run on a stereotype-loaded-only sub-bank. Within the 15 es-LATAM Q_S predicates: `es_001` (religious immigrant family), `es_003` (manual labor), `es_007` (sends remittances), `es_011` (first-generation university), `es_012` (cooks with chiles), `es_014` (immigrated young), and `es_015` (strong accent) are coded *stereotype-loaded*; `es_002` (many siblings), `es_004` (Day of the Dead), `es_005` (traditional food on Sundays), `es_009` (music and dancing at gatherings), `es_010` (cares for elderly parents) are *cultural markers*; `es_006` (learned Spanish from grandparents), `es_008` (devotion to the Virgin Mary), and `es_013` (mass weekly) are *heritage practices*. Restricting the per-predicate analysis to the stereotype-loaded subset (7 items, n_{trials}=51): $L(Q_C)$=82.4%, $L(Q_S)$=70.6%, STLD=-11.8 pp, paired McNemar exact p=0.070; cultural-markers subset (5 items, n=31): $L(Q_C)$=77.4%, $L(Q_S)$=67.7%, STLD=-9.7 pp, p=0.250; heritage-practices subset (3 items, n=18): $L(Q_C)$=77.8%, $L(Q_S)$=72.2%, STLD=-5.6 pp, p=1.000. **The cell-level effect does not cleanly separate stereotype-loaded from cultural-marker predicates within the es-LATAM bank, and the heritage-practice subset is null**, consistent with the broader predicate-resource-confounded reading. None of the three subset McNemars survive predicate-subset Bonferroni; we present the breakdown as a descriptive construct-validity diagnostic, not as an as-run diagnostic test.

B Discordant counts, paired Wald CIs, and one-sided preregistered p-values

Table 4 reports paired-difference statistics for the as-run diagnostic family (4 cells), the non-`name` re-run (4 cells), the post-guard regex contrast (4 cells), and D8 (Q_C v1→v2 within es-LATAM). All values are computed directly from the trial JSONLs (`trials_fullrag_4culture.jsonl` for v1 and

non-name; same JSONL for post-guard via the final_leak flag; trials_qc_v2_eslatam.jsonl for D8), which will be released upon acceptance. For every paired contrast $A-B$ we report discordant counts $b=\#\{A=1, B=0\}$ and $c=\#\{A=0, B=1\}$ (for STLD rows $A=Q_S$, $B=Q_C$; for the D8 control-shift row $A=Q_C$v1, $B=Q_C$v2), the two-sided exact paired McNemar $p_2=\min\{1, 2\Pr(X\leq\min(b,c) \mid X\sim\mathrm{Bin}(b+c, 0.5))\}$, the positive-direction one-sided $p_{\text{pos}}=\Pr(X\geq b \mid X\sim\mathrm{Bin}(b+c, 0.5))$ matching the preregistered H_1 direction (large p_{pos} means the data are inconsistent with leakage amplification in this direction; for the D8 control-shift row p_{pos} is undefined as it is not a Q_S-vs-Q_C contrast), a 95% paired Wald CI on $L(Q_S)-L(Q_C)$, $\widehat{\Delta}\pm 1.96\sqrt{\{(b+c)-(b-c)^2/n\}/n^2}$ (in pp; we use Wald rather than the Newcombe (Newcombe, 1998) score interval for transparent arithmetic recomputation from (b, c, n)), and the 4-way Bonferroni status at $\alpha=0.0125$.

Family	Cell	b	c	p_2	p_{pos}	95% CI (pp)	Bonf.
*As-analyzed primary (*name*+4 PII; N=100/culture)*							
v1	en	4	5	1.000	0.7461	$[-6.9, +4.9]$	ns
v1	es	1	11	0.006	0.9998	$[-16.5, -3.5]$	$*$
v1	ar	5	4	1.000	0.5000	$[-4.9, +6.9]$	ns
v1	hi	5	9	0.424	0.9102	$[-11.3, +3.3]$	ns
*Non-*name* rerun (n_{trials}=80/culture)*							
v1	en	4	2	0.688	0.3438	$[-3.5, +8.5]$	ns
v1	es	1	8	0.039	0.9980	$[-15.8, -1.7]$	ns
v1	ar	5	4	1.000	0.5000	$[-6.1, +8.6]$	ns
v1	hi	4	8	0.388	0.9270	$[-13.4, +3.4]$	ns
Post-guard regex contrast (N=100/culture)							
v1	en	2	4	0.688	0.8906	$[-6.8, +2.8]$	ns
v1	es	0	6	0.031	1.0000	$[-10.7, -1.3]$	ns
v1	ar	3	0	0.250	0.1250	$[-0.3, +6.3]$	ns
v1	hi	2	5	0.453	0.9375	$[-8.2, +2.2]$	ns
D8 expanded-Q_C (es-LATAM, N=100)							
	Q_C v1→v2	12	2	0.013	—	$[+2.9, +17.1]$	—
	Q_S-Q_C v2 (full)	8	8	1.000	0.5982	$[-7.8, +7.8]$	—
	Q_S-Q_C v2 (non-name)	6	6	1.000	0.6128	$[-8.5, +8.5]$	—

Table 4: Discordant counts (b, c), two-sided exact paired McNemar p_2, preregistered one-sided p_{pos}, paired Wald 95% CI on $L(Q_S)-L(Q_C)$, and 4-way Bonferroni status (α=0.0125). The single Bonferroni-significant cell ($*$) is the v1 es-LATAM primary contrast, in the *negative* direction; the preregistered one-sided H_1 is not rejected on any cell. Under the cleaner non-name rerun, no cell is Bonferroni-significant. The post-guard regex contrast on es-LATAM is uncorrected-only (p=0.031). For D8 we test only the Q_C shift and the resulting v2 STLD; the Q_C shift confirms control-bank instability between the v1 3-predicate and v2 7-predicate pools (CI fully positive, i.e. v1 Q_C markedly higher than v2 Q_C), while Q_S-Q_C v2 is null with a wide CI. Verification: b, c recomputed directly from trials_fullrag_4culture.jsonl and trials_qc_v2_eslatam.jsonl; the script appendix_b_discordants.py will be released with the codebase upon acceptance.

C Per-predicate v2 breakdown (D8 expanded-Q_C)

The 7 v2 predicates show wide leakage variance unrelated to construct class: `cn_es_008` (Spanish-with-relatives) at $9/9{=}100.0\%$, `cn_es_010` (literature) at $6/7{=}85.7\%$, `cn_es_004` (food) at $38/51{=}74.5\%$, `cn_es_006` (music) at $4/6{=}66.7\%$, `cn_es_005` (holiday) at $4/7{=}57.1\%$, `cn_es_007` (education) at $6/12{=}50.0\%$, `cn_es_009` (sports) at $3/8{=}37.5\%$. The $37.5{-}100\%$ range across content-equivalent culture-neutral predicates is consistent with leakage rate being *predicate-specific noise* on this small per-predicate sample; the v1 Q_C pool happened to land at the high end. The full v2 trial JSONL and analysis script will be released upon acceptance.

D Pre-registration record

The locked plan was committed to the project repository before any pilot trial was run. The release bundle (forthcoming upon acceptance) will contain the locked plan (`prereg_plan.md`) and the deviation log (`prereg_deviations.md`; D1–D8 with rationale). The locked plan named:

- **Estimand:** post-guard final-leak under Llama-Guard-3-8B with the summarize reformulation.
- **Hypothesis:** H_1: $\text{STLD}_{\text{joint}}{>}0$ (one-sided, paired McNemar exact, 4-way Bonferroni $\alpha{=}0.0125$).
- **Design:** paired five-arm ($Q_0/Q_R/Q_N/Q_C/Q_S$), $N{=}100$/culture, joint refusal-as-no-leak, sterilization audit before any inference trial.
- **Sanity rules:** $L(Q_R){\approx}L(Q_0)$ and $L(Q_C){\approx}L(Q_N)$.

The locked end-to-end estimator was *not run*; the as-analyzed estimator (D1–D3: regex guardrail, direct reformulation, pre-guard generator emission) is reported instead. For anonymous review we cannot provide a public lock URL; the locked-plan markdown and deviation log will be deposited at OSF or a public registry upon de-anonymization, together with the codebase. Inferential status under the as-analyzed estimator is documented in §3; the as-run diagnostic family is *not* confirmatory in the strict pre-registration sense (D1–D3 estimand shift), and we mark sensitivity-family tests accordingly.

E Pre-registration deviations (D1–D8)

D1 Reformulation. Locked: *summarize*. As-run: *direct*. Reason: summarize reaches a $96{-}100\%$ pre-guard ceiling that compresses STLD toward zero by saturation.

D2 Guardrail. Locked: Llama-Guard-3-8B. As-run: production-grade PII regex. Reason: Llama-Guard-3 was not run; regex is the deployed defender baseline.

D3 Headline metric. Locked: post-guard final-leak. As-run: pre-guard generator emission. D1–D3 jointly are an estimand shift; the locked end-to-end privacy-risk estimator was *not run*.

D4 Replication model. Single-seed Qwen-2.5-VL-32B-Instruct (text-only); not a venue-grade replication.

D5 Force-context regime. Gold-only run violates $L(Q_R){\approx}L(Q_0)$ at 7B (-22 pp, $p{=}0.003$); reported as supportive, not primary.

D6 Per-predicate variance. Per-predicate $n_{\text{pairs}}{=}3{-}12$ in es-LATAM; analysed by bank-source labels and by leave-one-out.

D7 Predicate-bank scale. Planned $\sim$150 stereotype predicates per culture; used 43 total. Novel sub-bank reduced to $4/3/3$ for es-LATAM/ar/hi vs. planned ≥ 5.

D8 Expanded-Q_C sensitivity (post-hoc). 7-predicate culture-neutral pool `cn_es_004`$-$`010` rerun on Q_C only, same docs, same length-matching (single-seed, predicate-imbalanced).

Planned release artifacts (upon acceptance). The release bundle will contain: (i) the synthetic English PII corpus (800 documents, 4×200/culture) with per-document metadata; (ii) the predicate bank (43 stereotype + 11 culture-neutral + 15 neutral elaborations) as `predicates.jsonl` with `predicate_id`, `source`, `novel` boolean, per-arm token length, and the per-item construct annotation (`predicate_construct.csv`); (iii) the 5-arm query generator; (iv) the sterilization audit script and gazetteer plus `predicate_audit.json`; (v) the raw trial JSONL (N=6,000 trials across full-RAG and force-context, 7B and 32B; each trial has `predicate_id`, `predicate_source`, `retrieved_target`, `pii_in_generation`, `guard_triggered`, `final_leak`, `response`); (vi) the analysis scripts: cell-level (`analyze_4culture.py`), bank-labelled per-predicate / LOO / cluster-bootstrap (`per_predicate_and_transition.py`), post-guard / per-PII-type (`post_guard_and_appendix.py`); and (vii) the deviation log (D1–D8). The es-LATAM Q_C/Q_S paired raw slice, the full multi-culture 7B JSONL, and the 32B raw JSONLs will all be included.