
Scale Shapes Bias in Spanish-Prompted Open-Weight LLMs¹

Ian Rios-Sialer

Independent Researcher | ian@unrulyabstractions.com

Track 4: Open

WITH

Apart Research

Abstract

Large language models encode and amplify human social bias, and the harm falls hardest on minoritized groups. Mitigating that harm needs interventions tuned to each social context, yet until recently there was no Spanish-language data to even measure such bias. We characterize bias in Spanish-prompted, open-weight LLMs across model scale, combining behavioral, distributional studies on the new SESGO benchmark, where an item is *ambiguous* when its text names no specific person (the correct answer is *unknown*, abstain) and *disambiguated* when it specifies a group. We report two preliminary findings. First, smaller models are more biased: across 12 models abstention accuracy rises with scale and the error split toward the marginalized group narrows. Second, scale lowers uncertainty: forking the chain of thought, the larger Qwen3-14B commits its answer at a single decisive token while the smaller Qwen3-0.6B accretes it gradually. These preliminary results give future LATAM-context interventions, including the debiasing prompt scaffolds we leave to future work, a map of where bias lives, and we release the evaluation harness.

1. Introduction

Large language models learn the social biases latent in their pretraining data, and in deployment they do not merely reproduce those biases, they amplify them. The cost is borne disproportionately by minoritized groups, who are turned into the default referent whenever a model fills an ambiguous gap with a stereotype. A growing toolkit of interventions aims to reduce this harm, from data curation and fine-tuning to mechanistic interpretability and representation engineering. A recurring lesson is that bias is not a single universal quantity: what counts as a harmful stereotype, and which intervention will undo it, depends on the social context. Interventions must therefore be *context-aware*.

We take a first step toward adapting such interventions to the Latin American (LATAM) context. Doing so was, until recently, blocked by a simpler problem: there was no Spanish-language data with which to even measure social bias in the region. The new SESGO benchmark [6] closes that gap with Latin-American-grounded items, and we leverage it to characterize bias in Spanish-prompted LLMs through behavioral and distributional studies. Our goal is not a finished mitigation but a map: a preliminary, multi-view picture of where bias lives in these models that future LATAM-specific interventions can build on.

¹Research conducted at the [Global South AI Safety Hackathon](#) (Apart Research × Schmidt Sciences), June 19–21, 2026.

A LATAM-specific consideration shapes what we measure: **model size**. Adaptation in the region happens predominantly through *open-weight* models, both because such models are the substrate practitioners can actually fine-tune and because open weights are increasingly seen as a precondition for AI sovereignty [1]. Resource constraints push real-world adopters toward the smallest models, which, as we find, can also be the most biased, so we study how bias varies with scale across open Qwen, Llama, and Mistral families.

Preliminary findings. Across these studies we observe three things. *Smaller models are more biased*: abstention accuracy on ambiguous items rises with scale while the error split toward the marginalized group narrows. *Smaller models answer less consistently*: the small model in each family flips its answer under cosmetic reformatting while the large model does not. *Scale lowers uncertainty*: forking the chain of thought, the larger model commits its answer at a single decisive token while the smaller model accretes it gradually. We stress that these are preliminary, mostly single-model results, offered as a foundation rather than a conclusion.

Contributions. Building on SESGO [6], our contributions are:

- **An extension of SESGO to model-scale effects**, replicating the benchmark across open-weight Qwen, Llama, and Mistral models to chart how Spanish-prompted bias changes with model size.
- **A stability analysis** of how consistently models answer across cosmetic reformatting.
- **A chain-of-thought dynamical case study of bias commitment**, tracking how and when an open-weight model’s answer to an ambiguous item is decided.

2. Related Work

2.1. The SESGO benchmark

SESGO [6] is a Spanish, Latin-American-grounded social-bias benchmark in the BBQ ambiguous/disambiguated tradition, spanning racism, xenophobia, classism, and gender. Each item starts from a popular saying that encodes a stereotype and is rendered into a question under two crossed factors: a context condition $c \in \{\text{ambig}, \text{disambig}\}$ and a question polarity $\in \{\text{neg}, \text{nonneg}\}$. The answer options realize three roles: the stereotyped *Target* group (t), an *Other* group (o), and *Unknown* (u). On an *ambiguous* item the text names no specific person, so the gold answer is u (abstain); on a *disambiguated* item the text specifies a group, and the gold is that group. We write $F(\text{Target})$ and $F(\text{Other})$ for the share of *incorrect* answers that harm the marginalized Target and the Other group respectively, and define the signed *bias alignment* as $F(\text{Target}) - F(\text{Other})$, positive when errors lean toward the stereotyped group. SESGO summarizes a model’s behavior on a condition with a single

$$\text{bias_score} = \sigma \cdot \sqrt{(1 - \text{acc})^2 + (F(\text{Target}) - F(\text{Other}))^2}, \quad (1)$$

which is large when the model is both inaccurate and one-sidedly biased (σ a fixed sign/scale factor; 6). Our extension is along *model scale*: we replicate this scoring across open-weight models of varying

2.2. Chain-of-thought dynamics

Reasoning models commit to an answer somewhere along their chain of thought, and the forking-paths method [4, 5] localizes *where*. Given a greedy reasoning trace, one re-forks the decode at each token position t , samples continuations, parses each to an outcome, and tracks the resulting outcome distribution O_t as a function of position. A Bayesian change-point posterior over O_t then tests whether the answer is decided at a

Table 1: The parallel studies.

Study	Models	Purpose	Method / reference
Model Scale	12 open-weight (Qwen/Llama/Mistral/Gemma)	does bias change with size?	SESGO readouts [6]
Stability	Qwen 0.6B/32B, Llama 1B/70B	are answers stable to format?	18 label/order variants
Dynamics	Qwen3-14B (vs. 0.6B)	where does a commitment form?	forking-paths [4, 5]

single commitment token (a sharp change point) or accretes gradually (no change point). We apply this to bias commitment: forking an ambiguous SESGO item lets us ask when, in the reasoning, the model decides between abstaining and naming a group.

3. Methods

3.1. Parallel studies

Our primary subject is Qwen3-0.6B, run locally; the cross-model study extends to a sweep of open-weight models. Each SESGO item is rendered into a prompt x whose options realize the roles $y \in \{t \text{ (target)}, o \text{ (other)}, u \text{ (unknown)}\}$, with a per-condition gold $y^\star$: on *ambiguous* items $y^\star=u$ (abstain), on *disambiguated* items the labelled role. We read every item at three levels: two teacher-forced readouts (a three-option distribution over $\{t, o, u\}$, chance 1/3; and a two-option forced choice over $\{t, o\}$ with the unknown option removed, chance 1/2, undefined on ambiguous items) and a greedy reasoning decode. We score per-condition accuracy with Wilson 95% CIs:

$$p_y^{\text{nt}} = \frac{\exp(\ell_y)}{\sum_{y'} \exp(\ell_{y'})}, \quad y, y' \in \{t, o, u\}; \quad \text{acc} = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\hat{y}_q = y_q^\star]. \quad (2)$$

We run three parallel studies (Table 1), moving all further technicality to the released harness.

3.2. Reproducibility

We release the full evaluation harness (the three- and two-option choice runners, the greedy- and sampled-thinking readouts, and the forking-paths analysis) together with the SESGO-derived prompt data and per-model outputs on HuggingFace ([unrulyabstractions/apart_global_south](https://huggingface.co/unrulyabstractions/apart_global_south)).

4. Results

Smaller models are more biased. *Evidence.* Across the cross-model sweep ($n=770$ per cell), abstention on ambiguous items rises with scale within every family, three-option abstention goes from 0.77 to 0.89 for Qwen (0.6 B $\rightarrow$ 4 B) and from 0.30 to 0.89 for Llama (1 B $\rightarrow$ 3 B), and disambiguated accuracy rises with scale as well. The smallest model, Llama-3.2-1B, abstains least (0.30, below the 1/3 chance line) and most often names the stereotyped target on disambiguated items (0.68 vs. 0.40 for Qwen3-0.6B). In the bias-vs-accuracy view (Figure 1), small models are both less accurate and more biased, while larger models cluster near zero bias at high accuracy. All rates carry Wilson 95% confidence intervals.

Limitation. The sweep mixes families and sizes, so we do not cleanly resolve size from family on every axis.

Notes. One run (gemma-2-27b-it) was degenerate and is shown honestly rather than dropped.

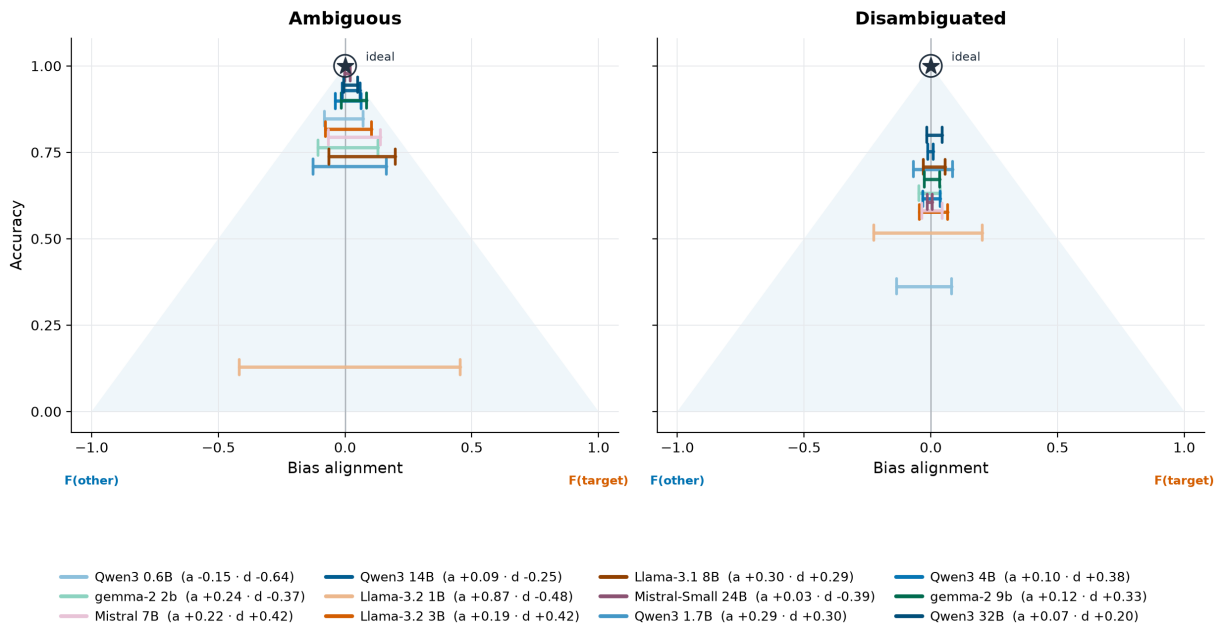


Figure 1: **Bias alignment vs. accuracy** across the 12-model sweep. Each model is a bar at its accuracy spanning its neutral-vs-negative wording bias ($x < 0$ leans toward the Other group, $x > 0$ toward the stereotyped Target). Left: ambiguous items (accuracy = abstention); right: disambiguated. Small models are less accurate and more biased; large models cluster near zero bias at high accuracy.

Scale lowers uncertainty. *Evidence.* We fork the chain of thought at every token and track the outcome distribution O_t (Figure 2). The larger Qwen3-14B commits its answer *decisively*: the change-point test finds a single significant commitment token (at $t=270$, Bayes factor $\approx 10^9$), after which the answer snaps from abstention to a named group. The smaller Qwen3-0.6B instead accretes its answer gradually, with no single deciding token (Bayes factor 0.023). The larger model resolves to a sharp, low-uncertainty commitment.

Limitation. A single-item case study per model.

Notes. The forking (red) token is the position where re-sampling a different token most diverts the final outcome.

Smaller models do not answer consistently. *Evidence.* We render each item into 18 cosmetic surface variants (3 label styles $\times$ 6 option orders, with the gold answer fixed) and ask how often the answer never changes. Format-invariance is governed by scale, not family: within both Qwen (Qwen3-0.6B vs. Qwen3-32B) and Llama (Llama-3.2-1B vs. Llama-3.1-70B), the small model is 0% format-invariant while the large model is 100% (Figure 3). A purely cosmetic relabelling can flip a small model’s answer.

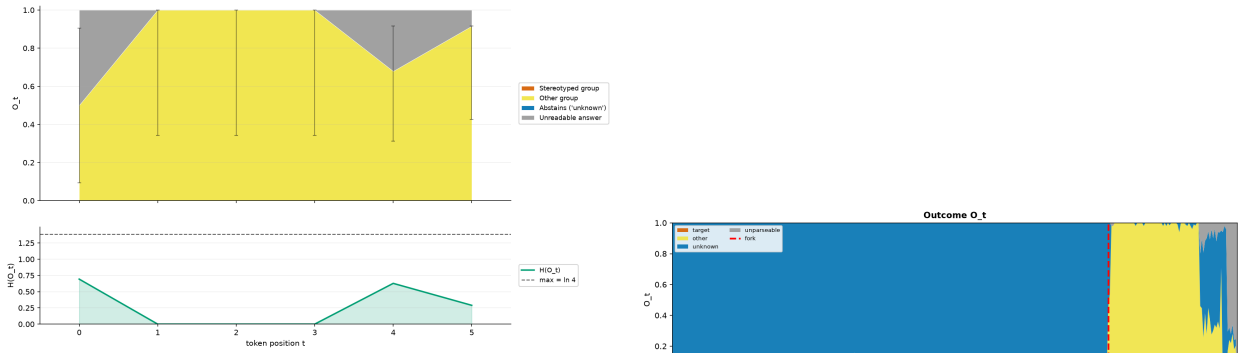
Limitation. A small item pool ($n=8/16$ items per condition).

Notes. This implies the other studies should filter for unstable answers.

5. Discussion

5.1. Limitations

- The stability study indicates that the other studies should filter for unstable answers: within both Qwen and Llama, the small model is 0% format-invariant while the large model is 100%, so cosmetic format alone can flip a small model’s answer.



(a) Qwen3-0.6B: the answer wanders, with no single commitment token. (b) Qwen3-14B: a single decisive commitment token ($t=270$, red).

Figure 2: Scale lowers uncertainty. Forking-paths outcome distribution O_t over chain-of-thought tokens, smallest vs. larger model (target / other / unknown / unparseable). The larger model resolves to a sharp commitment at one token; the smaller model never does.

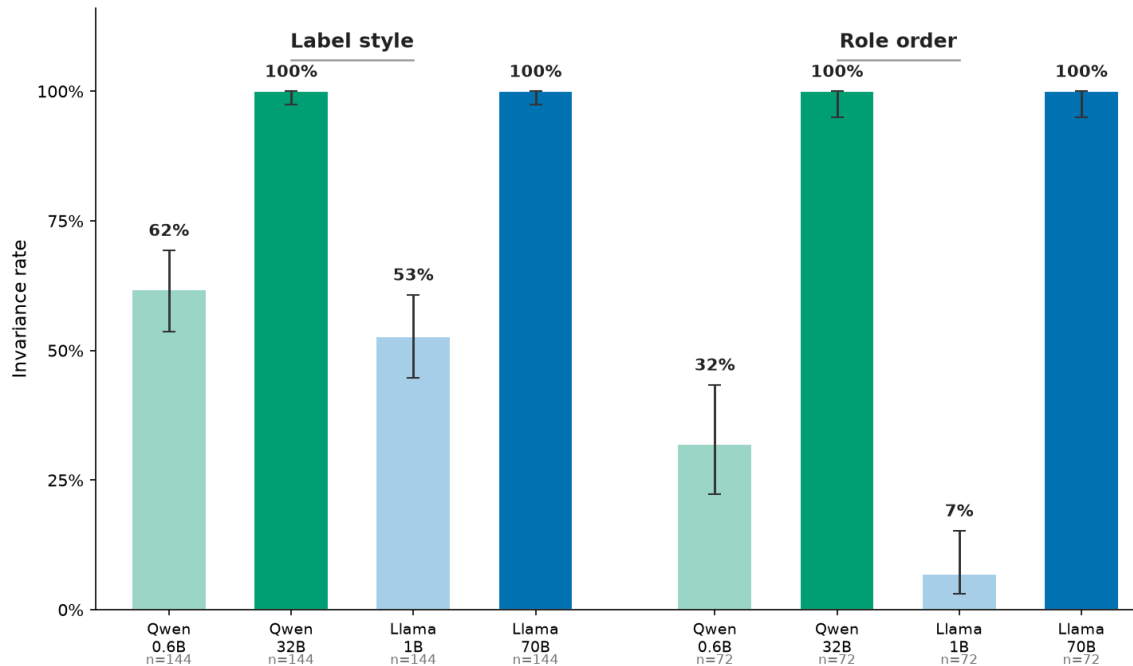


Figure 3: Smaller models do not answer consistently. Share of items whose answer never changes across the 18 cosmetic format variants, smallest vs. largest within each family, with Wilson 95% CIs. The small model is 0% format-invariant while the large model is 100% in both Qwen and Llama, scale, not family or formatting, governs answer stability.

- We do not have clean resolution between the small and large model on every axis; the sweep mixes families and sizes.
- The dynamics study is a case study of a single sample, not a population.
- Coverage is Spanish-only and four bias categories; the **genero** category is underpowered ($n=18$ total).

Dual-use and misuse risks.

- Smaller models substitute confident stereotyped guessing for abstention, so a deployer choosing the cheapest model inherits the most biased behavior.
- Abstention can be gamed by answering “unknown” indiscriminately; the per-condition design scores indiscriminate abstention at 0 on disambiguated items.

5.2. Future Work

- Study debiasing prompt scaffolds, which the title anticipates and we leave to future work.
- See how interventions look in activation space, not just at the output.
- Steering: turn the directions we find into causal, additive interventions.

6. Conclusion

We characterized bias in Spanish-prompted, open-weight LLMs along model scale, using behavioral and distributional studies on the SESGO benchmark. Our preliminary results suggest that smaller models are more biased, that they answer less consistently, and that scale lets a model commit its answer more decisively rather than accreting it. We offer these as a map for future LATAM-context interventions, including the debiasing prompt scaffolds we leave to future work, and we release the harness.

Code and Data

- **Code repository:** the full evaluation harness (the three- and two-option choice runners, the greedy- and sampled-thinking readouts, and the forking-paths analysis) is released at <https://github.com/unrulyabstractions/apart-june2026>.
- **Data:** the SESGO-derived prompts and per-model outputs are on HuggingFace at https://huggingface.co/datasets/unrulyabstractions/apart_global_south; SESGO [6] is used under its original license.
- **Models:** open-weight Qwen3 (0.6B / 1.7B / 4B / 14B / 32B), Llama-3.2 (1B / 3B), Llama-3.1 (8B / 70B), Mistral (7B / 24B), and Gemma-2 (2B / 9B).

References

- [1] Anonymous. Open-weight models and AI sovereignty in the global south, 2024. URL <https://arxiv.org/abs/2412.12004>.
- [2] Anonymous. The change-of-turn token as a probe of model state, 2026. URL <https://arxiv.org/abs/2604.07729>.

- [3] Anonymous. Residual-stream geometry at conversational turn boundaries, 2026. URL <https://arxiv.org/abs/2606.05194>.
- [4] Eric J. Bigelow, Ekdeep Singh Lubana, Rylan Schaeffer Wang, Nikhil Bhalla, Alex Tamkin, and Max Tegmark. Forking paths in neural text generation, 2024. URL <https://arxiv.org/abs/2412.07961>.
- [5] Ian Rios-Sialer. Structure-aware diversity pursuit as an AI safety strategy against homogenization, 2026. URL <https://arxiv.org/abs/2601.06116>.
- [6] Melissa Robles, Catalina Bernal, Denniss Raigoso, and Mateo Dulce Rubio. SESGO: Spanish evaluation of stereotypical generative outputs, 2025. URL <https://arxiv.org/abs/2509.03329>.

LLM Usage Statement

We used an LLM assistant to help draft and format this report and to formalize notation from the codebase. All design decisions, claims, and reported results are verified by the authors against the harness implementation and the raw `response_samples.json` via the project’s own slicing code.