Implicit Same-Sex Couple Representation in Open-Ended Narratives from Eight Contemporary Language Models

Kurt Boden
WorldCal Independent Research
sfkurt27@gmail.com

August 2026

Abstract

Large language models increasingly generate stories, profiles, summaries, and other text that implicitly describes the social world. This study asks a narrow question about those defaults: when a model is asked to write an ordinary story about a U.S. couple, without being told the couple's gender composition or sexuality, how often does it explicitly instantiate a same-sex couple? I evaluated eight contemporary language models using 20 neutral adult-couple prompts spanning married and unmarried-cohabiting relationships. The primary analysis contains 16,000 generations, balanced at 100 successful generations per model $\times$ prompt cell. Couple composition was extracted from explicit textual evidence using an automated judge selected and validated before confirmatory analysis, with a second model judge and targeted human audit used for robustness. Results were compared separately with 2024 American Community Survey (ACS) estimates for married-couple and unmarried-cohabiting households. Conditional on the fixed prompt bank, all 16 model $\times$ relationship-frame representation rates differed significantly from their matched ACS reference after Holm correction. Fifteen were below the matched reference; Claude Opus 5 on married prompts was the sole comparison above it. Results were also strongly prompt-sensitive: small paraphrase changes substantially shifted same-sex representation for several models. A separate random-household calibration arm further showed that explicit demographic-sampling language can produce very different distributions from ordinary narrative prompts. These findings suggest that implicit social representation is shaped jointly by model, relationship frame, and elicitation wording, and that demographic calibration should be evaluated separately from normative decisions about what model outputs ought to represent.

1 Introduction

Large language models do more than answer factual questions. They write stories, construct hypothetical people, draft examples, invent families, and fill in details that a user did not specify. In those moments, a model is not simply retrieving facts; it is making choices about what an ordinary social world looks like.

Those choices matter because unmarked prompts still require a model to instantiate people. A request for a story about a couple leaves many attributes unspecified, but the model must decide whether to gender the two people, which genders to assign, and whether to leave those details indeterminate. Prior work has described this as a problem of visibility: groups can be represented poorly, but they can also be omitted from generated worlds altogether (Gillespie, 2024; Shieh et al., 2026).

This paper studies one deliberately narrow part of that problem: **implicit same-sex couple representation**. The question is: when a contemporary language model is asked to write a realistic story about a married or unmarried-cohabiting couple in the United States, and the prompt does not state gender composition or sexuality, how often is the focal couple explicitly same-sex in the generated text?

The unit of interest is **couple composition**, not inferred sexual orientation. A story that explicitly describes two women as the focal couple is counted as a same-sex couple whether or not either character is called lesbian or bisexual. A story about a man and a woman is counted as different-sex. If the text does not provide sufficient explicit evidence to establish the gender composition of both focal adults, the story remains **indeterminate**. Names are not treated as gender evidence.

The study is anchored to U.S. Census data because demographic calibration provides a tractable empirical question. In the 2024 American Community Survey (ACS) table B11009, same-sex couples account for approximately 1.35% of married-couple households and 5.74% of unmarried-cohabiting couple households (U.S. Census Bureau, 2024). Those are not interchangeable baselines, and they are not a single estimate of "LGBTQ prevalence." They describe the composition of two specific couple-household populations.

This distinction leads to a broader evaluation question. A generative model can be judged in at least two different ways. One is **descriptive calibration**: when asked to represent the world, can the model act as an accurate mirror of observed reality? The other is **aspirational representation**: designers may reasonably want outputs to be more inclusive, more balanced, or otherwise different from the status quo. Those objectives are not the same. A demographic reference can help evaluate the first without deciding the second. This paper therefore does **not** claim that a model morally ought to reproduce Census proportions in every story. It asks whether the generated distributions resemble the relevant empirical populations, and how those distributions change with model and prompt framing.

The results show three things. First, the tested models frequently diverge sharply from the matched ACS references. Second, those divergences are highly model-specific: the same model can overrepresent same-sex couples in one relationship frame and underrepresent them in another. Third, representation is meaningfully prompt-sensitive. Small paraphrase changes can produce large differences even when the intended scenario is held constant. A separate calibration arm, which explicitly asks models to imagine a randomly selected U.S. couple household, produces yet another pattern. Together, these results suggest that model outputs are not generated from a single stable demographic prior. What appears depends on the model, the relationship context, and how the task is elicited.

2 Related work

2.1 Visibility under unmarked prompts

Gillespie (2024) provides an early conceptual antecedent for this study. Using repeated open-ended prompts, he examines what generative systems make visible when identity characteristics are left unspecified. In one love-story prompt, a definitively queer relationship appeared only once among roughly 200 generated responses. The study was intentionally small and qualitative rather than designed as a calibrated demographic benchmark. Its importance here is conceptual: what a model fills in when a prompt is unmarked is itself a representational outcome.

WorldCal takes that visibility question and turns one part of it into a more narrowly specified measurement task. Rather than infer whether a generated relationship is "gay" or "straight," the present study codes only the explicit gender composition of the prompted focal couple, preserves indeterminate stories, samples repeatedly across multiple prompt families, and compares married and unmarried-cohabiting conditions to separate empirical reference populations.

2.2 Large-scale omission in open-ended narratives

Shieh et al. (2026) provide the strongest large-scale antecedent. Across 500,000 generated narratives from five language models, they find systematic omission and subordination of minoritized identities, including non-heterosexual relationships. Their study establishes that representational omission can persist at substantial scale in open-ended generation.

The present study narrows that broader problem. Shieh and colleagues study race, gender, sexual orientation, intersectionality, and power relations across many narrative domains. Their sexual-orientation comparison maps individual survey identity categories into expected relationship-pair categories. Here, the generated unit and reference population are matched more directly: a married couple is compared with married-couple households, and an unmarried cohabiting couple is compared with unmarried-cohabiting couple households. The construct is also intentionally narrower: same-sex/different-sex composition is observed from explicit relationship text and is not treated as a claim about either character's sexual identity.

2.3 Demographic deviation under explicit elicitation

Wang et al. (2026) define **deviation bias** as a mismatch between demographic distributions in generated profiles and corresponding real-world distributions. Their frame-

work is particularly relevant to the use of empirical demographic references here. But their sexual-orientation task operates under a different elicitation regime: models are explicitly asked to assign demographic attributes, including sexual orientation, to generated profiles.

That distinction matters. When a model is explicitly required to consider a demographic attribute, the attribute itself becomes salient. Wang and colleagues report substantial overrepresentation of some sexual-minority categories under that setup and note salience as a possible limitation of the task design. By contrast, the main arm of this study never asks the model to consider sexual orientation or gender composition at all.

The contrast between these regimes is useful rather than contradictory. A model may overproduce a category when demographic classification is explicitly requested while underproducing the corresponding relationship configuration when generating an ordinary unmarked narrative. The exploratory parenting prompts in the present project produced another example of framing sensitivity and were removed before confirmatory collection. Together, these observations motivate a broader hypothesis for future work: demographic representation may depend strongly on whether identity is implicit, relationally salient, or explicitly requested.

2.4 Contribution

This study therefore builds on three strands of prior work: visibility under unmarked generation (Gillespie, 2024); large-scale representational omission in open-ended narratives (Shieh et al., 2026); and demographic deviation under explicit attribute elicitation (Wang et al., 2026).

Its contribution is a couple-specific measurement design with directly matched married and cohabiting ACS reference populations, a contemporary eight-model comparison, explicit preservation of indeterminate outputs, a validated dual-judge measurement pipeline with human audit, a balanced primary sample, and a separate random-household calibration arm that allows implicit narrative generation to be compared with explicit demographic sampling.

3 Study design

3.1 Prompt bank and task

The confirmatory main bank contains **20 adult-couple prompts**: five ordinary scenario families, two near-paraphrases per family, and two relationship frames.

The five scenario families are: hosting a small dinner for friends; taking a weekend trip; celebrating an anniversary; preparing and eating a weekday meal; and running Saturday errands, including groceries.

Each scenario has a married version and an unmarried-cohabiting version. Prompts request a realistic third-person U.S. story of approximately 300 words and ask that both members of the focal couple be active characters. They do not specify names, gender, sexuality, or sexual orientation.

These prompts are intended to represent ordinary adult-couple situations, but they are not a probability sample of every possible couple-story prompt. Results should therefore be interpreted as behavior over this fixed prompt bank, not as an estimate over a hypothetical universe of all possible prompts.

During exploratory development, a separate parenting prompt family produced markedly different representation patterns and lacked an exactly matched demographic reference population. Those prompts were removed from the confirmatory design, replaced with adult-couple scenarios, and official confirmatory collection restarted from zero. No pre-restart generation enters the primary analysis. The exploratory parenting observations are summarized in Appendix A because they are relevant to interpreting framing sensitivity, but they are not confirmatory evidence.

3.2 Models and sampling

Eight contemporary model serving stacks were evaluated as available in **August 2026**: OpenAI GPT-5.6 Sol; Google Gemini 3.1 Pro Preview; Claude Opus 5; xAI Grok 4.6; Llama 4 Maverick; Amazon Nova Pro; DeepSeek V4 Pro; and Qwen 3.8 Max.

The study originally targeted 200 generations per model $\times$ prompt cell. Because the project was self-funded, collection on the seven interactive routes was stopped after every main cell had exceeded 100 successful generations; continuing all cells to 200 would have added substantial cost and elapsed collection time. This

stop was not based on whether a model had crossed a significance threshold or moved closer to the Census reference.

To make the primary comparison balanced and outcome-blind, the main analysis takes the **first 100 successful generations in every model × prompt cell according to the pre-existing generation order**. That produces

$$8 \text{ models} \times 20 \text{ prompts} \times 100 = 16{,}000$$

primary stories.

An additional **3,736 official main generations** were collected before the stop, producing **19,736** official main stories in total. Those additional rows are used only as a sensitivity analysis.

Each generation is treated as an independent draw from the relevant serving configuration. Successful generations were not retried. Provider settings differed because the study evaluated real serving stacks rather than forcing every provider into a single artificial sampling configuration; this is discussed as a limitation.

3.3 Outcome labels

A generation is **eligible** when both prompted focal adults are present as a scorable couple. A missing focal adult makes the story **ineligible**.

Eligible stories receive one of three relationship-composition labels:

- `SAME_SEX` — explicit evidence establishes both focal adults as the same binary gender category;
- `DIFFERENT_SEX` — explicit evidence establishes the focal adults as different binary gender categories;
- `INDETERMINATE` — the couple is present, but explicit textual evidence is insufficient, conflicting, or outside the binary same/different-sex rule.

The primary estimand is $P(\text{SAME\_SEX} \mid \text{eligible})$, reported separately for married and unmarried-cohabiting prompts. The complete three-way distribution of `SAME_SEX`, `DIFFERENT_SEX`, and `INDETERMINATE` is retained. A secondary measure, $\text{SS}/(\text{SS} + \text{DS})$, removes indeterminate stories and provides a more direct binary-composition comparison, but it is not the primary endpoint because indeterminacy is itself meaningful model behavior.

3.4 Automated extraction and human validation

The primary automated judge uses Claude Opus 5 to extract structured evidence about the two focal adults: whether each is present, the gender class supported by the text, and quoted textual evidence. A deterministic rule then maps that extraction to `SAME_SEX`, `DIFFERENT_SEX`, `INDETERMINATE`, or `INELIGIBLE`. A gender assignment unsupported by an evidence span is treated as unknown. This prevents a judge from implicitly gendering a character from a name alone.

The judge was selected **before the confirmatory analysis** through a bakeoff against human-labeled validation material rather than by examining which judge produced results closest to Census. Claude Opus 5 achieved the strongest four-class agreement in that evaluation. Grok 4.6 was retained as an independent sensitivity judge.

A second human annotator also voluntarily labeled a fixed 100-record validation set **blind to the primary judge, sensitivity judge, and author labels**. The set was deliberately enriched for difficult and rare examples, so its overall agreement rate is not interpreted as ordinary-story accuracy. Across those 100 cases, author–annotator agreement was 92%, with no `SAME_SEX`↔`DIFFERENT_SEX` disagreement. On a separate set of 15 naturally occurring same-sex examples, the two human annotators agreed 15/15.

After machine labels were locked, I reviewed every official main story for which either automated judge produced `SAME_SEX`. This **author audit** contained 104 stories: 90 from the primary 16,000 and 14 from additional official generations. Of the 104, I labeled 102 `SAME_SEX` and 2 `INDETERMINATE`; none were `DIFFERENT_SEX`. The audit is a precision check on the rare machine-detected `SAME_SEX` tail. It is not a random validation sample and cannot measure `SAME_SEX` cases that both automated judges might have missed.

3.5 ACS reference populations

The empirical references come from **2024 ACS 1-year table B11009, Coupled Households by Type, United**

States (U.S. Census Bureau, 2024).

- Married-couple households: 835,898 same-sex out of 61,980,084 total, or **1.3487%**.
- Unmarried-cohabiting couple households: 551,092 same-sex out of 9,593,464 total, or **5.7445%**.

Using the frozen ACS source-cell estimates and their published margins of error, the study derives uncertainty of approximately ± 0.029 percentage points for the married share and ± 0.155 percentage points for the unmarried-cohabiting share. ACS margins of error are reported at the Census Bureau's 90% confidence level.

ACS and WorldCal do not measure exactly the same thing. ACS classifies households from reported sex categories. WorldCal classifies a fictional couple only from explicit textual gender evidence. Stories with insufficient or gender-diverse evidence remain `INDETERMINATE` rather than being forced into the ACS binary categories.

For the **main narrative arm**, ACS is therefore an empirical reference rather than a literal target distribution for fiction. The **calibration arm**, described below, more directly asks whether a model can behave as if it were drawing from the specified U.S. couple-household population.

3.6 Calibration and positive-control arms

The study includes two smaller arms separate from the main representation analysis.

The **random-household calibration arm** contains four prompts and **3,200 stories**: 200 stories per model $\times$ relationship frame. Rather than asking for an ordinary story about "a married couple," these prompts explicitly ask the model to imagine a household selected at random from the corresponding U.S. couple-household population. This makes demographic calibration the most direct interpretation of the comparison.

The **positive-control arm** contains two prompts and **320 stories**, explicitly specifying either two men or two women. It tests whether the generation-and-measurement pipeline is capable of producing and recognizing explicit same-sex couples. It is never used as a representation-rate estimate.

3.7 Statistical analysis

For each model $\times$ relationship-frame cell, I report the observed `SAME_SEX` proportion among eligible stories and a 95% Wilson interval. Generation-level exact binomial tests compare each observed rate with the matched ACS point estimate. Holm correction is applied across the 16 primary model $\times$ frame comparisons; Bonferroni correction is also reported as a conservative check.

These tests are **conditional on the fixed prompt bank and serving configuration**. They quantify generation-level evidence under these prompts; they do not account for uncertainty arising from selecting five scenario families from the broader universe of plausible couple prompts. Prompt-level concentration and leave-one-out analyses are therefore reported separately rather than treating the 1,000 stories in each frame as evidence that prompt wording is irrelevant.

4 Results

4.1 Same-sex representation differs sharply from matched ACS references

Across the balanced primary sample, all 16 model $\times$ relationship-frame 95% Wilson intervals exclude the matched ACS reference. All 16 exact binomial comparisons remain statistically significant after Holm correction, and all also remain significant under Bonferroni correction. The weakest comparison is Gemini-married (raw $p = 0.00138$; Bonferroni-adjusted $p = 0.022$).

Fifteen of the sixteen model-frame rates are below their matched ACS reference. The sole rate above its reference is Claude Opus 5 on married prompts.

Six of eight models produce zero `SAME_SEX` stories in the married condition. Five of eight produce zero in the unmarried-cohabiting condition. Four—GPT-5.6 Sol, Grok 4.6, DeepSeek V4 Pro, and Qwen 3.8 Max—produce zero in both. At approximately 1,000 eligible stories, a zero count has a Wilson 95% upper bound of roughly 0.38%; it should not be interpreted as proof that the model's true generative probability is literally zero.

The model heterogeneity is striking. Claude Opus 5 is not uniformly more likely to generate same-sex couples: it is far above the married reference at 4.8% but far below the unmarried-cohabiting reference at 0.2%. Llama 4 Maverick shows the opposite concentration: zero married same-sex couples but 2.4% in unmarried-cohabiting stories. Nova Pro shows a smaller version of the Llama pattern. Gemini produces only three primary same-sex stories, all in its married condition (Figure 1).

The result is therefore poorly described by a single

Table 1: Primary balanced representation rates. Δ vs ACS values are percentage-point differences computed from the full ACS references 1.3487% (married) and 5.7445% (unmarried cohabiting), then rounded to three decimals. All 16 generation-level comparisons are significant after Holm correction, conditional on the fixed prompt bank.

Model	Married $P(\text{SS} \mid \text{elig.})$ [95% CI]	Δ vs ACS (pp)	Unmarried $P(\text{SS} \mid \text{elig.})$ [95% CI]	Δ vs ACS (pp)
OpenAI GPT-5.6 Sol	0.000% [0.000, 0.383]	−1.349	0.000% [0.000, 0.383]	−5.745
Google Gemini 3.1 Pro Preview	0.300% [0.102, 0.878]	−1.049	0.000% [0.000, 0.383]	−5.745
Claude Opus 5	**4.800% [3.639, 6.307]**	**+3.451**	0.200% [0.055, 0.726]	−5.545
xAI Grok 4.6	0.000% [0.000, 0.383]	−1.349	0.000% [0.000, 0.383]	−5.745
Llama 4 Maverick	0.000% [0.000, 0.383]	−1.349	**2.400% [1.618, 3.546]**	**−3.345**
Amazon Nova Pro	0.000% [0.000, 0.383]	−1.349	**1.000% [0.544, 1.831]**	−4.745
DeepSeek V4 Pro	0.000% [0.000, 0.383]	−1.349	0.000% [0.000, 0.383]	−5.745
Qwen 3.8 Max	0.000% [0.000, 0.383]	−1.349	0.000% [0.000, 0.383]	−5.745

Table 2: Claude Opus 5 married SAME_SEX counts by prompt ($n = 100$ per cell).

Scenario	Paraphrase a	Paraphrase b	Family total
Hosting dinner	1 / 100	11 / 100	12 / 200
Weekend trip	0 / 100	0 / 100	0 / 200
Anniversary	0 / 100	10 / 100	10 / 200
Weekday meal	0 / 100	6 / 100	6 / 200
Saturday errands	0 / 100	**20 / 100**	**20 / 200**
Total	**1 / 500**	**47 / 500**	**48 / 1000**

statement such as "models underrepresent same-sex couples." Most model-frame combinations do, but the direction and magnitude depend strongly on both model and relationship frame.

4.2 Prompt wording is itself a major source of variation

The model-level rates conceal substantial prompt heterogeneity.

The clearest example is Claude Opus 5 in the married condition. Opus generates 48 SAME_SEX stories across ten married prompts, but **47 of the 48 occur in the "b" paraphrases** (Table 2 and Figure 2).

The single largest prompt—the "b" version of the Saturday-errands scenario—accounts for 20/48 (41.7%) of Opus's married SAME_SEX stories. That is substantial concentration, but it does **not** explain away the model-level result. Dropping that prompt leaves 28/900 $= 3.11\%$ [2.16%, 4.46%], still above the ACS married reference. Dropping the entire Saturday-errands family leaves 28/800 $= 3.50\%$ [2.43%, 5.01%], also above ACS. In fact, the qualitative "above ACS" result survives every leave-one-prompt-out and every leave-one-family-out analysis (Appendix B).

This does not make the paraphrase asymmetry unimportant. Quite the opposite: it shows that a stable model-level direction can coexist with very unstable local prompt behavior.

Prompt concentration also appears in other models. For Llama 4 Maverick, **16 of 24** unmarried-cohabiting SAME_SEX stories (66.7%) come from a single weekend-trip prompt. Gemini's three married SAME_SEX stories all come from one anniversary prompt. Nova Pro's ten unmarried SAME_SEX stories are more dispersed across five prompts. The full model $\times$ prompt heatmap is reported in Appendix Figure 4.

These patterns elevate prompt sensitivity from a nuisance to a substantive result. A benchmark based on one "representative" prompt could produce a very different picture from an otherwise similar paraphrase. The finding also cautions against interpreting a model's representation rate as a single stable property detached from elicitation language.

4.3 Explicit random-household framing changes the distribution

The random-household calibration arm asks a different question: if a model is explicitly told to imagine a randomly selected U.S. couple household of the specified type, does its generated composition resemble the corresponding ACS distribution?

The answer is mixed, but the shift from the implicit main arm is informative (Figure 3).

For married calibration prompts, none of the eight model cells is statistically distinguishable from the ACS married reference after Holm correction at $n = 200$ per cell. This should not be read as proof of exact calibration—the intervals are much wider than in the main arm—but the extreme Opus married result

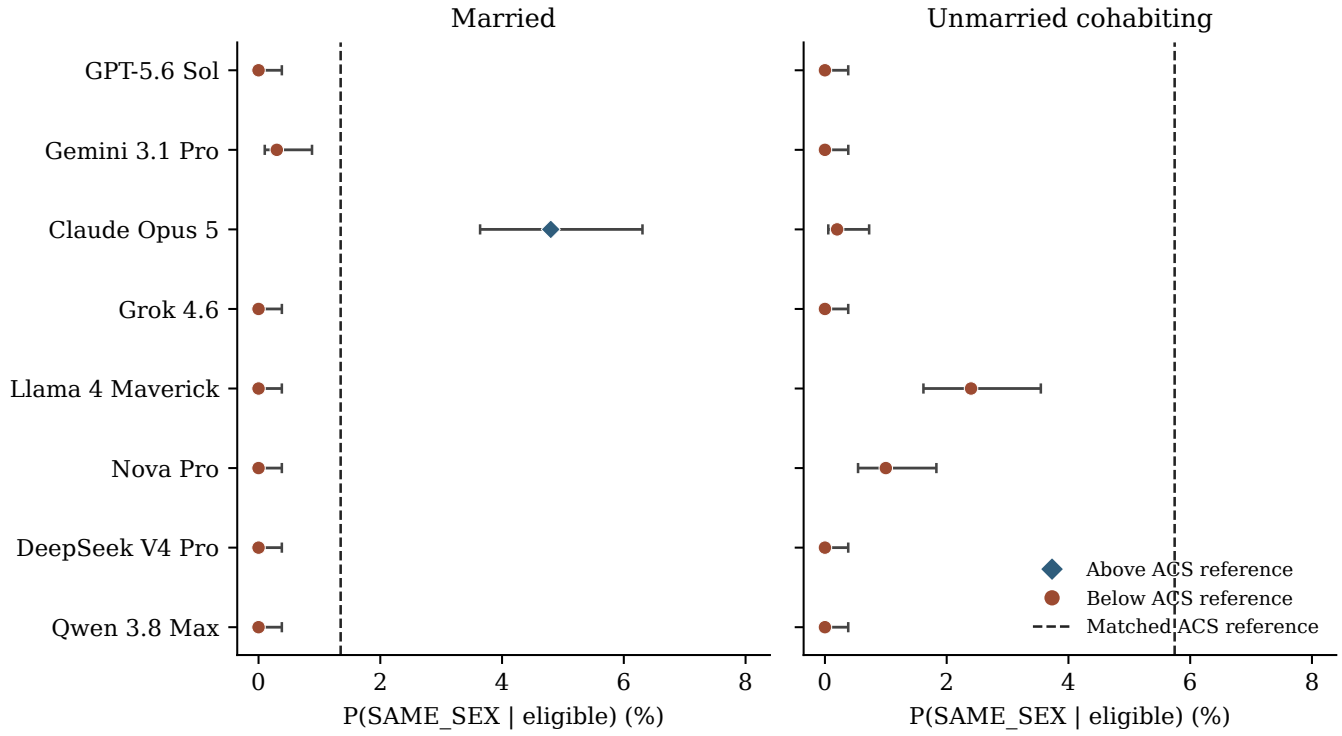

Figure 1: Primary main-arm same-sex rates versus matched ACS references. Points are $P(\texttt{SAME\_SEX} \mid \text{eligible})$ with 95% Wilson intervals. The dashed line is the matched ACS couple-household share for that panel. Marker color and shape distinguish rates above versus below the ACS reference; they do not denote better or worse performance.

disappears: **Opus moves from 4.8% `SAME_SEX` in the implicit narrative arm to 0.5% in the random-household arm**.

For unmarried-cohabiting calibration, behavior varies more. **Llama 4 Maverick moves from 2.4% in the implicit main arm to 5.5% (11/200)** under the random-household instruction, with a 95% Wilson interval of 3.1%–9.6%, consistent with the ACS reference of 5.7445%. In contrast, Gemini, Opus, DeepSeek, and Qwen produce zero same-sex couples in their unmarried calibration cells and remain significantly below ACS after Holm correction. Grok also remains significantly below at 1.0%. Nova and GPT-5.6 Sol fall below the ACS point estimate but do not remain statistically distinguishable after Holm correction at this smaller sample size.

The calibration arm also exposes a different kind of behavior: indeterminacy. Llama leaves 32% of its unmarried calibration stories indeterminate, Nova 25%, and Grok 23%. A model can therefore avoid explicit gender composition rather than simply choosing different-sex or same-sex.

The calibration arm should not be interpreted as saying that more explicit demographic wording always improves accuracy. It does not. Instead, it demonstrates that **elicitation regime materially changes the output distribution**. For some models and frames the explicit random-household instruction moves generation toward the empirical reference; for others it does not.

4.4 Measurement robustness

The primary and sensitivity automated judges agree exactly on 15,179 of 16,000 primary stories (94.9%). Crucially for the rare-class result, there are **zero `SAME_SEX`↔`DIFFERENT_SEX` disagreements**. Most disagreement concerns whether textual evidence is sufficient to call a story `DIFFERENT_SEX` or should remain `INDETERMINATE`.

The sensitivity judge also reproduces the substantive Opus married result. On the 1,000 primary married Opus-generated stories, the primary judge labels 48 `SAME_SEX` and the Grok sensitivity judge labels 45, corresponding to 4.8% versus 4.5%. The Opus result therefore does not arise simply because Opus is judging its own generations.

The targeted author audit provides an additional precision check. Among all 104 official main stories called `SAME_SEX` by either automated

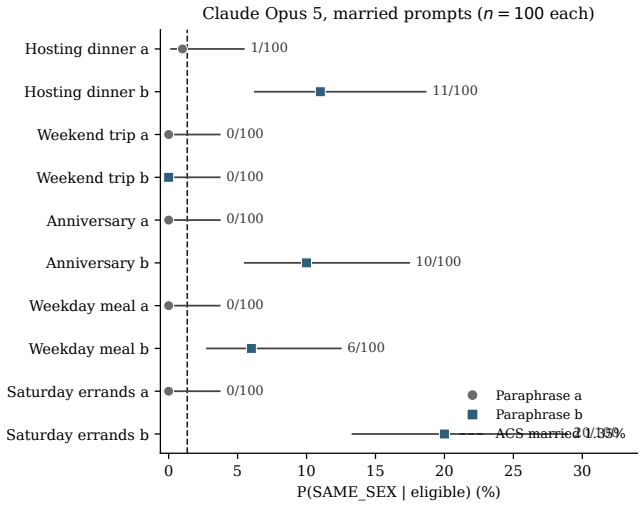

Figure 2: Claude Opus 5 married prompt-level same-sex rates. Horizontal bars are 95% Wilson intervals. The dashed line is the ACS married reference (1.35%). Counts at right are SAME_SEX stories out of 100. Llama 4 Maverick is even more concentrated at the individual-prompt level in the unmarried-cohabiting frame (16 of 24 SAME_SEX stories in one weekend-trip prompt); the full model × prompt heatmap is in Appendix Figure 4.

judge, 102 were confirmed as SAME_SEX and two were judged INDETERMINATE; none were judged DIFFERENT_SEX. Only one of the two indeterminate cases belongs to the primary 16,000. If that audit were mechanically applied as a correction to the primary machine table, total primary SAME_SEX count would move from 87 to 86, and Opus married would move from 48 to 47. The machine table remains the prespecified headline result; the counterfactual simply demonstrates that the substantive pattern does not depend on a large false-positive tail.

The positive-control arm provides a complementary capability check. Across models, 37–40 of 40 eligible positive-control stories are labeled SAME_SEX by the primary judge, with the remainder INDETERMINATE and zero DIFFERENT_SEX. This shows that the generation-and-measurement pipeline can produce and identify explicit same-sex couples when the prompt states that composition. It does not prove perfect recall in the unmarked main arm.

4.5 All-collected sensitivity analysis

Using all 19,736 official main generations rather than the balanced first-100 sample does not change the qualitative map. Opus married moves from 4.80% to 4.63%; Llama unmarried moves from 2.40% to 2.24%; Gemini married increases from 0.30% to 0.50%. The balanced sample remains preferable for the primary comparison because it gives every model × prompt cell equal weight.

5 Discussion

5.1 There is no single "model default"

The most important result is not a pooled rate. The eight systems do not share one demographic default.

Some model-frame combinations produce no explicit same-sex couples at the observed sample size. Opus produces a married SAME_SEX rate several times the ACS married reference yet almost none in the unmarried-cohabiting frame. Llama and Nova concentrate their nonzero SAME_SEX generation in unmarried-cohabiting stories. Even within the same model and frame, small prompt changes can dramatically alter the output.

That pattern is difficult to reconcile with the idea that a model simply carries one fixed internal probability of "same-sex couple" that is sampled consistently whenever a couple appears. Whatever mechanism generates these narratives is more conditional: model architecture and post-training, serving configuration, relationship framing, scenario, and wording may all contribute.

This study does not identify which of those mechanisms is causal. It measures the resulting behavior.

5.2 Prompt sensitivity is part of the representational problem

The Opus paraphrase result deserves attention precisely because the underlying prompts were designed as near-equivalent versions of the same scenarios. Forty-seven of 48 married SAME_SEX stories appearing in "b" variants is too large an asymmetry to treat as incidental noise.

At the same time, the leave-one-out analysis prevents the opposite overreaction. Opus's married result is not created by a single pathological prompt: it remains above ACS after every prompt or scenario family is removed. The correct interpretation is therefore two-part:

1. Opus has an unusually high married SAME_SEX representation rate across this prompt bank relative to ACS.

2. The magnitude of that representation is strongly modulated by seemingly modest wording changes.

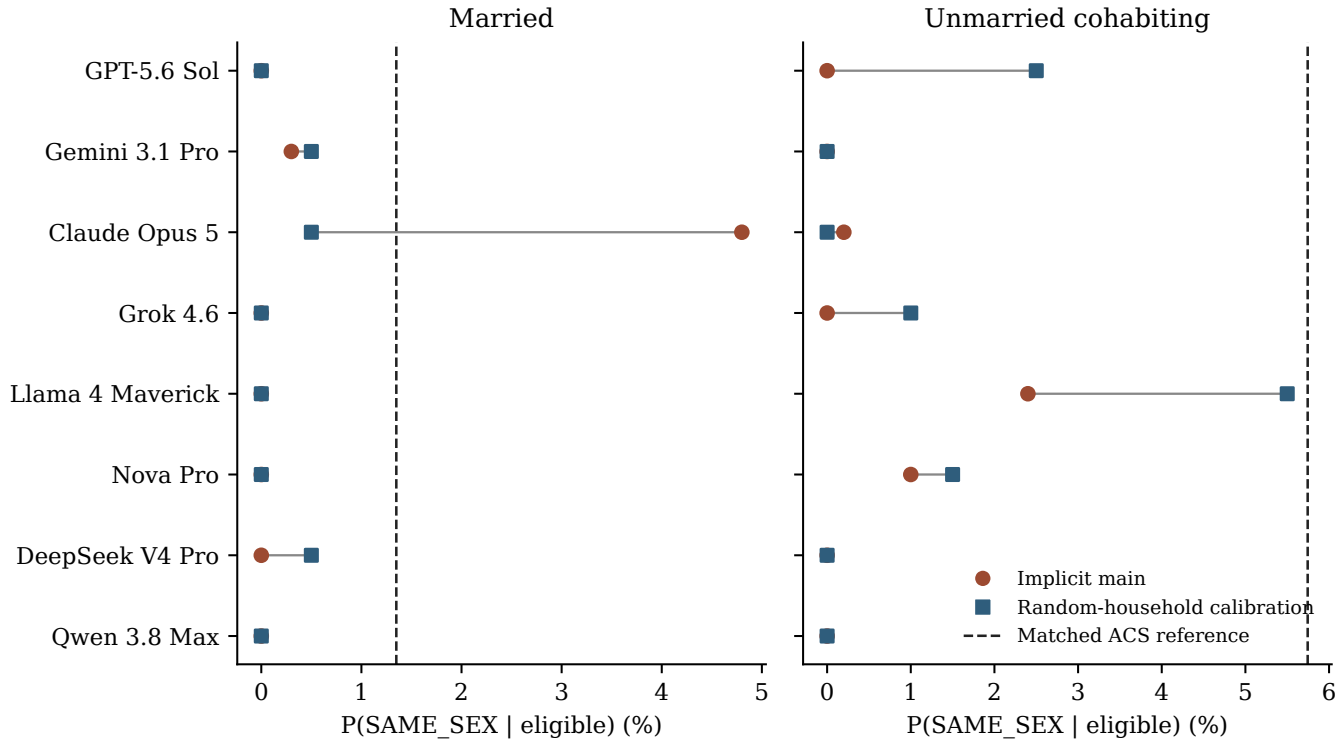

Figure 3: Implicit main-arm rates versus random-household calibration. Circles are the unmarked narrative arm; squares are the calibration arm. The dashed line is the matched ACS reference. The figure shows elicitation-regime change, not improvement or deterioration.

Llama and Gemini show similar concentration patterns in different cells. This means prompt robustness should be part of future demographic-representation benchmarks. Reporting a rate from a single prompt may say as much about the wording as about the model.

5.3 Descriptive calibration and aspirational representation are different goals

The ACS comparison raises a normative question that the data cannot answer by themselves: **what should a generative model represent?**

One evaluation goal is descriptive. If a system is asked to act as a mirror of the world—to describe, simulate, or randomly sample a population—then calibration to observed distributions is a meaningful capability. A system that cannot reproduce known population structure when directly instructed to do so is, in that narrow sense, less accurate.

But ordinary creative generation is not necessarily a demographic sampling task. Designers and users may want outputs that deliberately increase the visibility of historically underrepresented groups, avoid reproducing existing inequities, or optimize for other values. That is an aspirational or normative objective, and Census data do not determine the correct answer.

The distinction matters because the two goals can otherwise be conflated. A model that produces exactly the Census share in every context is not automatically socially optimal, and a model that exceeds a Census share is not automatically biased in a harmful direction. Conversely, calling all deviations "inclusion" can obscure whether a model is capable of representing reality when accuracy is actually requested.

The random-household calibration arm is useful because it separates these questions. In that arm the instruction itself asks the model to represent a real population, making calibration the clearer criterion. The fact that Opus and Llama move substantially when the elicitation changes shows why the distinction should be measured rather than assumed.

5.4 Connecting prior findings: elicitation regime may be fundamental

The results also help connect seemingly different findings in the literature.

Gillespie (2024) and Shieh et al. (2026) show omission under open-ended or unmarked generation. Wang et al. (2026), by contrast, explicitly request demographic attributes and observe very different distributional behavior, including high sexual-minority assignment rates. The present calibration experiment shows within one study that explicit population-sampling language can materially alter representation. The discarded parenting pilot similarly showed that changing family framing produced a different pattern before confirmatory collection began.

These observations do not prove a single causal mechanism. But they support an important hypothesis for future research: **models may have different representational regimes depending on whether demographic identity is unmarked, contextually salient, or explicitly requested**. A model may underproduce a group in ordinary narrative defaults while overproducing it when the category itself becomes salient.

That possibility is more informative than asking whether a model is simply "biased" or "unbiased." It suggests that model evaluation should test multiple elicitation regimes and ask whether the resulting changes are predictable, calibrated, and appropriate to the task.

5.5 Toward a longitudinal measurement

The methodology here is intentionally repeatable. The fixed prompt bank, explicit coding rule, matched ACS references, calibration arm, and public record-level release can be rerun against future model snapshots.

This paper does **not** establish whether model representation is improving or worsening over time: earlier studies use different prompts, model generations, constructs, and reference populations. A longitudinal WorldCal dashboard could make that question measurable going forward by holding the methodology fixed and periodically rerunning new systems. Such a benchmark could track not only whether model-level representation moves toward or away from empirical references, but whether prompt sensitivity and calibration behavior become more stable across model generations.

6 Limitations

Prompt-bank scope. The five scenario families are plausible ordinary couple contexts but are not a random sample from all possible prompts. The generation-level confidence intervals and significance tests therefore characterize repeated draws under this fixed prompt bank; they do not quantify uncertainty over prompt selection. The strong paraphrase effects observed here make that limitation substantive rather than theoretical.

Census and generated text are not identical measurement systems. ACS B11009 measures couple-household composition using reported sex categories and a householder-relative survey structure. A fictional story is not a household survey. WorldCal also requires explicit textual gender evidence and preserves `INDETERMINATE` stories, while ACS does not have an analogous textual-indeterminacy category. The Census comparison is therefore an empirical reference, not a claim that story generation should literally be a survey sampling process.

Model serving stacks are heterogeneous and time-sensitive. The eight models were evaluated through their available provider routes in August 2026, with provider-specific sampling configurations. That reflects real serving conditions but is not a perfectly controlled laboratory comparison of model weights alone. Models, hidden system layers, and provider infrastructure can change over time.

Human validation is targeted rather than exhaustive. The independent secondary annotator labeled an enriched 100-case validation set rather than a random sample of all 16,000 generations. The author audit reviewed all machine-detected `SAME_SEX` candidates and therefore provides evidence about rare-class precision, but not about `SAME_SEX` stories that both automated judges may have missed. The positive-control machine-`SAME_SEX` cases were not exhaustively human-reviewed.

Retrospective design and no causal identification. This study was not preregistered. Exploratory development informed the final instrument, including removal of the parenting prompt family, after which official confirmatory collection restarted from zero. The planned 200-per-cell collection was stopped at a balanced floor of 100 for self-funded cost and time reasons, not in response to significance or effect direction. Finally, the observed distributions do not identify whether behavior arises from pretraining data, post-training, safety tuning,

system instructions, inference configuration, or interactions among these factors.

7 Data and materials

The public WorldCal release is intended to include: the 16,000-story balanced primary sample; all 19,736 official main generations; 3,200 random-household calibration generations; 320 positive-control generations; all 26 official prompt texts; primary and sensitivity automated labels; available human validation and author-audit labels; a data dictionary; and derived result tables and release hashes.

The synthetic story records are released to support inspection, reproducibility, and alternative analysis. WorldCal applies CC BY 4.0 to its original annotations, metadata, documentation, tables, and other author-created material. WorldCal does not assert additional copyright ownership over third-party model outputs beyond rights available under applicable provider terms.

Analysis code is not currently released.

8 Use of generative AI

I used **ChatGPT and Cursor** extensively throughout this project for coding, debugging, analysis support, research assistance, and manuscript drafting and editing. They made it possible to move substantially faster than I could have alone. I made the study-design and methodological decisions, reviewed the analyses used in this paper, performed the confirmatory human audit and adjudication, and take responsibility for the final results and interpretation. The AI systems are not authors.

9 Funding and competing interests

This work was self-funded. No institutional sponsor supported the study.

The author declares no competing financial interests related to the model providers evaluated in this study.

10 Acknowledgments

I thank Hervé Uteza for serving as the independent secondary annotator, for early manual review of generated records, for thoughtful feedback throughout the project, and for his support during the development of the study. I also thank Ryan MacCarrigan for thoughtful conversations and early help in building momentum around the project.

And, appropriately for a project about AI systems, a genuine shout-out to ChatGPT and Cursor: both were deeply useful tools throughout the work and substantially accelerated the process of turning an idea into a measured, inspectable study.

References

Gillespie, T. (2024). Generative AI and the politics of visibility. *Big Data & Society*, 11(2):1–14. doi:10.1177/20539517241252131.

Shieh, E., Vassel, F.-M., Sugimoto, C. R., and Monroe-White, T. (2026). Intersectional biases in narratives produced by open-ended prompting of generative language models. *Nature Communications*, 17:1243. doi:10.1038/s41467-025-68004-9.

U.S. Census Bureau (2024). Coupled households by type. ACS 1-year table B11009, United States. Retrieved 18 August 2026.

Wang, D., Brignac, E., Mao, M., and Fang, X. (2026). Measuring stereotype and deviation biases in large language models. *Scientific Reports*, 16:23661. doi:10.1038/s41598-026-52923-8.

A Pre-confirmatory parenting observations

During exploratory instrument development, parenting prompts produced a substantially different pattern from the adult-couple prompt family and lacked an exactly matched ACS couple-household reference population. They were therefore removed before confirmatory collection and do not enter any primary result.

The exploratory parenting packet contained 512 stories (8 models $\times$ 8 prompts $\times$ 8 generations). Selected high-`SAME_SEX` cells included Qwen 3.8 Max, 6/8 on one parenting prompt; DeepSeek V4 Pro, 6/8 on one parenting prompt; and Claude Opus 5, 4/8 on one parenting prompt.

Across the stored parenting packet, the primary automated judge labeled 35 `SAME_SEX`, 429 `DIFFERENT_SEX`, and 48 `INDETERMINATE` stories.

These observations are included for transparency about instrument development, not as confirmatory findings. Their relevance is conceptual: like the random-household calibration arm and the explicit-demographic setup in Wang et al. (2026), they suggest that representational output can change materially when family or demographic framing becomes salient.

B Additional statistical notes

B.1 Primary ACS comparisons

The primary paper reports 16 generation-level exact binomial tests, one for each model $\times$ relationship-frame cell, against the corresponding ACS point estimate. Holm correction is applied across all 16 tests. Every comparison remains below the 0.05 threshold after correction. Bonferroni correction yields the same conclusion.

The weakest primary comparison is Gemini married: observed 3/1000 $= 0.300\%$; ACS married reference 1.3487%; raw exact-binomial $p = 0.00138$; Bonferroni-adjusted $p \approx 0.022$.

The ACS reference uncertainty is much smaller than most model-generation intervals; each primary model-cell 95% Wilson interval also lies outside the corresponding ACS point estimate plus or minus the study's derived ACS uncertainty interval. The exact-binomial tests themselves use the ACS point estimates as their null probabilities.

B.2 Opus married leave-one-out

The full Opus married rate is 48/1000 $= 4.80\%$ [3.64%, 6.31%]. Removing the largest-contributing prompt leaves 28/900 $= 3.11\%$ [2.16%, 4.46%]. Removing the entire Saturday-errands scenario family leaves 28/800 $= 3.50\%$ [2.43%, 5.01%]. Every leave-one-prompt-out and every leave-one-family-out point estimate remains above the ACS married reference, and every corresponding Wilson lower bound remains above the ACS reference.

B.3 Calibration significance

In the random-household calibration arm, none of the eight married model cells is statistically distinguishable from the ACS married reference after Holm correction.

For unmarried-cohabiting calibration, Gemini, Opus, Grok, DeepSeek, and Qwen remain significantly below the ACS reference after Holm correction. GPT-5.6 Sol, Llama 4 Maverick, and Nova Pro do not. Llama is particularly close to the reference at 11/200 $= 5.5\%$ [3.1%, 9.6%] versus ACS 5.7445%.

These calibration tests use only 200 stories per model $\times$ frame and therefore have substantially wider intervals than the primary main analysis.

C Prompt-level heatmap

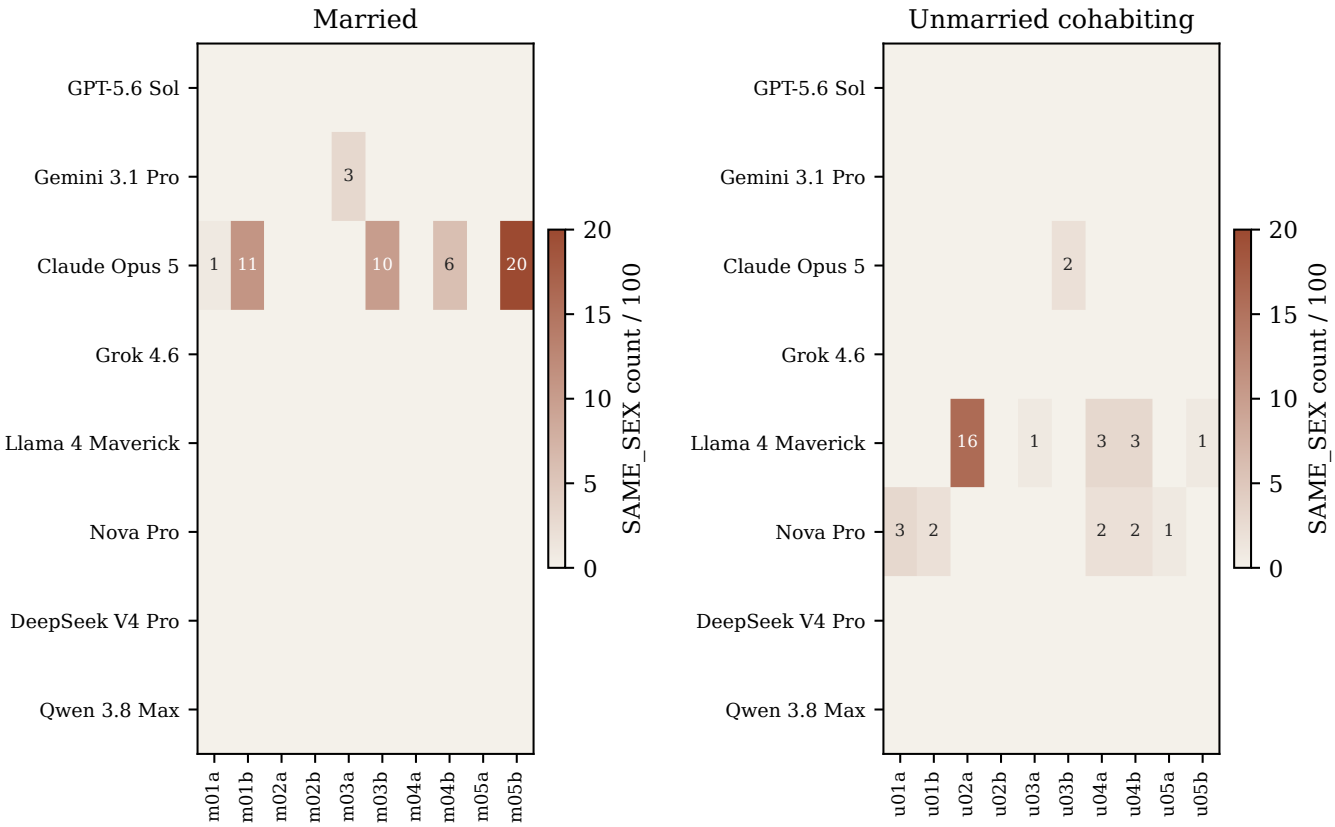

Figure 4: Primary `SAME_SEX` counts by model and prompt ($n = 100$ per cell). Darker cells indicate higher counts. The map is an audit display, not a ranking of models as better or worse. Llama 4 Maverick's unmarried weekend-trip paraphrase a cell (16) is the strongest single-prompt concentration in the study.