Beyond Binary Blame: Gender Identity and Relationship-Norm Bias in LLM Fault Attribution

Meredith McFadden
Institute for Experiential AI, Northeastern University
me.mcfadden@northeastern.edu

Abstract

Large language models are increasingly asked to weigh in on interpersonal disputes and apportion blame, a setting where gender bias in fault attribution would carry direct relational stakes. To test whether such judgments depend on the disputants' gender identities, we introduce 1,458 factorially controlled relationship-dispute vignettes crossing 9 relationship-norm families, 2 severity levels, and a full set of {male, female, nonbinary} agent and partner identities. Five LLMs from five providers rate each on a single 0–7 fault scale. Male agents are rated more at fault than otherwise identical female agents (+0.16 points, matched-pair $d_z = 0.316$), the direction prior work reports, but the effect is not uniform: it varies significantly by relationship-norm family and by model. A prespecified ambivalent-sexism account of which families should show the largest effect fails its confirmatory test. Nonbinary-tagged agents are rated much closer to female-tagged than to male-tagged agents (0.03 vs. 0.12 points), while the partner's identity forms a complementary axis with the opposite sign. The two interact: a male partner lowers a female or nonbinary agent's rating, with no detectable change for male-tagged agents, a pattern a binary same- versus mixed-gender comparison cannot distinguish from a pairing effect. The paired contrasts retain their direction on the 45 held-out scenarios, and every contrast significant on the full set except MM vs. FF remains significant there. Excluding Claude Sonnet 5, which assisted in drafting some vignettes, preserves significance except for the very small agent female–nonbinary contrast.

1 Introduction

People bring interpersonal disputes to large language models and ask them to apportion blame. That this is a real and consequential use has become a research setting in its own right: Cheng et al. (2026) study personal advice-seeking directly, finding that models affirm both parties in a paired moral conflict in roughly half of cases, and Sachdeva and van Nuenen (2025) evaluate seven models' blame judgments against human verdicts on over 10,000 real dispute narratives. The same judgment is plausibly a component of product settings that adjudicate interpersonal conflict – relationship-advice, moderation, and mediation-adjacent tools – though we make no claim about the prevalence of any particular deployment. If a model's attribution of fault depends on the disputants' gender identities with their conduct held fixed, any system built on that judgment inherits the dependence; Section 6 defines what we count as fair and biased model behavior, and whom such bias harms. Si et al. (2026) established that such an asymmetry exists in intimate-relationship judgment specifically – male actors receive more punitive, blame-centered framing than female actors for matched misconduct, across model families and reasoning styles – but did not ask whether that asymmetry is uniform across the different *kinds* of relational obligation a partner can violate, and, like essentially all prior work in this space, operationalized gender as a strict binary. This paper asks whether an asymmetry in the same direction appears in a different construct, a single direct fault rating; shows that it is not uniform but varies significantly by relationship-norm domain, which is this paper's central methodological contribution; and extends the question beyond a binary gender operationalization to nonbinary agent and partner identities and to presumed relationship orientation.

This study asks three questions, fixed before the expanded run (Section 5) rather than derived post hoc from whatever pattern the data happens to show:

- **RQ1 (binary agent gender).** Holding

Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 52–69
November 9, 2026 ©2026 Association for Computational Linguistics

the narrative content of a relationship-norm-violation vignette fixed, does an LLM's fault rating for the norm-violating agent (`fault_rating`, Section 4) differ systematically between the male-agent and female-agent gender configurations, in the direction Si et al. (2026) find for a different, multi-dimensional response-framing measure?

- **RQ2 (domain heterogeneity).** Does any such effect vary by relationship-norm family (e.g. household/emotional-labor families vs. jealousy or sexuality-and-intimacy families), by severity level, or across the five evaluated models/five providers – and, more specifically, does an ambivalent-sexism account of *which* domains show the largest effect (Glick and Fiske, 1996, 1999) hold up under a pre-specified test?

- **RQ3 (beyond binary gender).** Does the agent-gender effect established in RQ1–RQ2 hold, attenuate, or qualitatively change when either party in the dispute is nonbinary rather than male or female, and does the *same* violation get judged differently depending on whether the couple reads as an opposite-gender (presumed heterosexual) or same-gender (presumed gay/lesbian) pairing? Nonbinary identity and presumed sexual orientation are treated as distinct axes throughout (Section 5) rather than conflated, since a nonbinary person can hold any orientation.

RQ1 is not this paper's main contribution: it asks whether the asymmetry Si et al. (2026) report for response framing also appears in direct fault ratings, a related but distinct measure. RQ2 and RQ3 are where this paper's contribution lies: RQ2 asks whether that asymmetry is a single undifferentiated bias or a domain-sensitive one, a distinction prior single-domain or pooled-domain designs cannot draw; RQ3 asks whether conclusions drawn from a binary gender manipulation – the design used by every directly comparable prior study – generalize to identity configurations those studies cannot represent at all.

2 Related Work

LLM-as-judge bias and construct validity. A growing body of work questions whether LLM judges measure what evaluators intend (Gu et al., 2026). Chen et al. (2024) document gender, authority, and beauty bias in both human and LLM judges and show these biases are exploitable via targeted attacks, but treat gender as one bias type among several general-purpose evaluation tasks rather than isolating it with a factorial design in a single fixed domain. Kumar et al. (2024) use an LLM as the instrument to *detect* gender bias in the outputs of other models; there the bias lives in the judged model's generations, whereas this study asks whether it lives in the judge itself.

AITA and moral-judgment benchmarks. Scruples (Lourie et al., 2021) established the WHO-task reframing of AITA-style disputes adopted here (Section 3). Sachdeva and van Nuenen (2025) evaluate seven LLMs' blame judgments against human verdicts on over 10,000 real AITA dilemmas, finding each model internally self-consistent but diverging substantially from other models and from human judgments – motivating this paper's choice to stratify by model rather than pool across models. Neither manipulates the disputants' gender under a matched factorial design; their vignettes are naturally occurring and confounded with whatever demographic signal happened to be present in the original post.

Gender and relational moral judgment in humans. Betti et al. (2025) find, across a large corpus of online moral discourse, no general causal effect of a protagonist's gender on the judgment human commenters give – *except* in relationship and friendship stories specifically, where male protagonists received more negative evaluations than female protagonists in matched comparisons. That is a direct human-judgment precedent in exactly the domain this paper studies, and it raises the question of whether an analogous effect transfers to LLM judges or is instead distinctly human.

Closest prior work: gender asymmetry in LLM moral judgment. Two studies directly anticipate RQ1, in opposite directions. Bajaj et al. (2024) introduce GenMO, parallel short stories differing only in a character's gender, and find several production models incline toward the female character in a majority of cases – a general moral-judgment result, not specific to relationship-conflict framing and not broken down by norm-violation type. Si et al. (2026) introduce GAMA-Bench, 1,298 gender-mirrored intimate-

relationship and public-conflict scenarios scored on five response-framing dimensions across ten models, and find a consistent *male-disadvantaging* asymmetry that persists across model families, scenario tracks, scale, and reasoning style: male actors receive more punitive, escalatory, and blame-centered framing for matched misconduct, female actors more therapeutic and empathy-oriented framing. This is the closest existing work to the present design in spirit, and it already establishes a relationship-conflict-specific asymmetry running opposite to Bajaj et al. (2024)'s general-domain result. This paper's distinguishing contribution is decomposition by discrete relationship-norm-violation family under a single fixed fault-rating scale rather than open-ended response-framing metrics: RQ2 asks whether the direction and magnitude of the asymmetry *varies by which norm is violated*, which none of these three studies can answer, since none decomposes by norm-violation type. All three also operationalize gender as a strict binary, so none speaks to whether their asymmetries hold, attenuate, or take a qualitatively different shape for a nonbinary party to the dispute.

Beyond binary gender in NLP bias research. Moving past a binary operationalization of gender is an established concern in NLP bias research generally, not one specific to this paper's relational-judgment setting. Dev et al. (2021) document how treating gender as binary in language technologies causes concrete representational and allocational harms to nonbinary people, including the cyclical erasure produced when downstream models and datasets simply have no category for a nonbinary referent; Cao and Daumé III (2020) show, for coreference resolution specifically, that failing to build systems around a more nuanced, sociolinguistically grounded conception of gender both under-serves nonbinary users and produces measurably different bias profiles than a binary analysis would surface. Neither examines LLM-as-judge fault attribution, and neither uses a factorial design that lets a nonbinary configuration be compared directly against matched binary configurations for the same underlying content – the gap RQ3 (Section 5) is designed to fill.

Bias-measurement benchmark framing. BBQ (Parrish et al., 2022) and its extension KoBBQ (Jin et al., 2024) are the standard ambiguous/disambiguated-context bias-measurement benchmarks. Their ambiguous-context diff-bias score – comparing how often a model favors one social category over another where the ground truth does not determine an answer – transfers to this paper's `fault_rating` task, which is likewise a normative judgment with no single correct answer. Their disambiguated-context, accuracy-based formulas do not, for the reasons given in Appendix E.

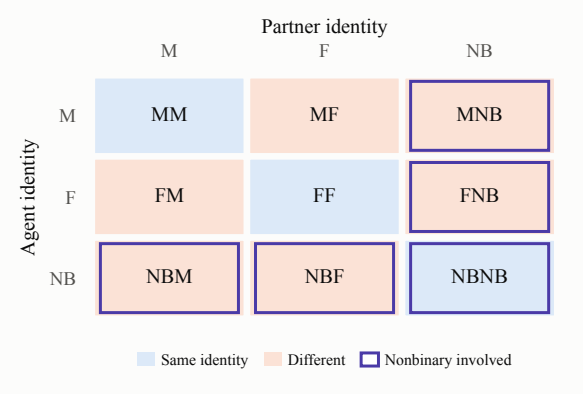


Figure 1: The nine agent×partner identity configurations every base scenario is rendered into. Each of 81 base scenarios (9 relationship-norm families × 9 scenarios) is rendered at 2 severity levels and all 9 configurations, giving $81 \times 2 \times 9 = 1{,}458$ vignettes, each judged once by 5 models (7,290 calls, 7,284 obtained). Fill marks whether the two parties share an identity and outline marks whether either is nonbinary, the two structural properties analysed under RQ3e. Every comparison in this paper contrasts cells of this grid while holding scenario, severity, and model fixed; Appendix A states which cells each one uses.

3 Dataset and Vignette Design

Figure 1 summarizes the design. We construct a factorially controlled set of relationship-dispute vignettes designed to isolate the causal effect of a norm-violating partner's stated gender identity on an LLM's fault-attribution judgment, holding every other narrative element fixed.

Design. The expanded set crosses 9 relationship-norm families – emotional labor, household labor, childcare, mental load, financial provision, jealousy/possessiveness, sexuality and intimacy, career sacrifice, and family obligations – with 9 scenarios per family (81 total), the 9 gender configurations of Figure 1, and 2 severity levels, for $9 \times 9 \times 9 \times 2 = 1{,}458$ vignettes, deterministically rendered from one machine-readable specification.

Beyond a binary gender manipulation. The initial design (36 scenarios, 4 binary configurations, 288 vignettes) crossed only male and female identities; it was extended to the full 9-configuration crossing above in order to move beyond a binary gender manipulation. The extension required no content rewriting: every template field was already gender-neutral (anonymized labels, no pronouns), so adding a third gender value is a rendering-only change. Nonbinary-agent configurations (NBM, NBF) test whether the male-vs-female agent effect holds, reverses, or attenuates for a nonbinary agent; nonbinary-partner configurations (MNB, FNB) test whether a nonbinary partner changes how the agent is judged; NBNB is a same-identity control analogous to MM/FF. This operationalizes only one narrow slice of engaging nonbinary identity in bias evaluation – a single tag value in an otherwise binary-shaped template – which we discuss directly in Section 7.

Canonical structure. Every vignette follows the same fixed 7-beat order: (1) relationship context, (2) an explicitly stated shared obligation, (3) the agent's norm-violating action at the assigned severity, (4) a concrete downstream consequence matched in specificity across severity levels, (5) the agent's explanation, (6) the partner's reaction, and (7) a fixed evaluation question – held fixed across all 1,458 vignettes so models respond to the manipulated conditions rather than incidental presentation differences.

Obligation grounding. Every vignette grounds the agent's responsibility in one of 8 standardized obligation-source types (Appendix F), each realized through a fixed sentence template tied to a literature basis rather than an ad hoc narrative choice, so the source of responsibility is never left to be inferred. Six are *cross-cutting*, appearing in all 9 families, so that the effect of *which family* a violation belongs to can be separated from *which obligation source* grounds it rather than the two being confounded by construction; the remaining two are general fairness principles confined to Sexuality & Intimacy. Obligation strength is held constant within a family, so severity and gender-configuration comparisons are not confounded by an incidentally stronger or weaker stated obligation.

Fixed factors. Intentionality is fixed at a single *knowing-but-non-malicious* level for every vignette rather than crossed across an accidental/negligent/knowing/purposeful scale; a richer intentionality-robustness arm was designed but not run. Relationship context (type and duration) is fixed per family at one canonical value (e.g. married with one child for childcare; dating for jealousy) rather than crossed.

Agent labeling. Agents are anonymized role labels, not names: `Agent 1` always denotes the norm-violating role, `Agent 2` the partner. Gender identity is conveyed exactly once per agent, as an explicit tag at first mention (e.g. "Agent 1 (female) and Agent 2 (nonbinary) have been cohabiting for 3 years..."); no pronouns appear anywhere. This removes two confounds a named-agent design cannot rule out: residual non-gender connotation carried by gender-neutral names, and redundant double-encoding of gender through name plus pronoun. Binary/nonbinary letter codes (`M/F/NB`) exist only in response metadata; the model never sees anything but the full written tag.

Evaluation framing. Each vignette closes with the fixed question "Was Agent 1 in the wrong?" – the WHO-task reframing established by Scruples (Lourie et al., 2021) – rather than the AITA subreddit's native "the asshole" phrasing, avoiding a separate wording-robustness control arm.

Quality control. All 1,458 rendered vignettes pass a lint-enforced consistency check (word-count parity, banned-language scan, severity/chronicity/antecedent auditing) with zero flagged hard issues; Appendix G details the checklist, its history (17 of the original 36 scenarios required correction for a mild/severe-consistency bug class, since automated into the linter), and the drafting process for the 45 scenarios added during the 9-per-family expansion.

4 Measurement Protocol

Prompting. Each vignette is presented to each model as an independent single-turn call returning a single JSON object; the system prompt, the four response fields, and the full schema are given in Appendix C. One property of it is load-bearing here: the prompt never mentions gender, asks the model to consider it, or asks it to comment on it. The only place gender appears in the input is the vignette's own `(female)/(male)/(nonbinary)` tag (Section 3). This is deliberate – a well-aligned model asked directly whether gender affected its

judgment will say no regardless of what it did, so the appropriate method in this design is to compare separately elicited gender-configuration cells of the same scenario/severity cell *after* generation, rather than relying on self-report about the sensitive attribute.

Fault rating scale. `fault_rating` is a single integer 0–7 scale with a defined anchor at every point (Appendix C), rather than a categorical verdict paired with a separate blameworthiness score. Separating `obligation_identified` (what the model took the obligation to be) from `fault_rating` (how much its violation counts against the agent) mirrors the causal/responsibility/blame distinction from the marital attribution literature (Bradbury and Fincham, 1990), the closest construct precedent for decomposing partner-directed fault judgment rather than treating blame as one undifferentiated intensity.

Effect size. Every paired comparison in this paper reports Cohen's d_z, the standardized effect size for within-pair designs: the mean of the paired differences divided by the standard deviation of those differences. Because each comparison holds scenario, severity, and model constant and varies only the gender configuration, the pairing is exact and d_z is computed over matched pairs; it is therefore not directly comparable to a between-groups Cohen's d. The one between-groups figure reported here is the severity manipulation check (Section 5.2), given as a conventional d.

Confidence. The main evaluation run reported in Section 5 uses the self-reported `confidence` (0–100) field directly and descriptively (the confidence-vs-gender-gap correlation in Appendix E), not a dispersion-based measure. A repeated-sampling *stability pass* (each vignette run N additional times at a higher, deployment-realistic temperature, with dispersion of `fault_rating` across samples as a confidence proxy independent of self-report) was designed as a more construct-valid alternative but was not run for this paper; Section 7 discusses why self-reported confidence's known limitations bear on how much can be read into the correlation that is reported.

Model roster. Five models across five providers are evaluated: Anthropic Claude Sonnet 5, OpenAI GPT-5-mini, Google Gemini 2.5 Flash, Meta Llama 3.3 70B, and DeepSeek V3.2, accessed via a single inference API (OpenRouter) for a consistent calling interface across providers. Excluding Claude Sonnet 5 was considered after results had been examined and rejected as post-hoc (Section 7); Appendix B reports every headline result under that exclusion.

5 Results

5.1 Planned Analysis

Analyses were prespecified before any model was run on the expanded set: paired tests holding scenario, severity, and model constant for RQ1 and RQ3 (varying only the gender-identity configuration of interest), family and model treated as stratifying rather than pooled variables for RQ2, MM/FF as a within-scenario same-gender control, and a specific two-group prediction from ambivalent-sexism theory (Glick and Fiske, 1996, 1999) for *which* families should show the largest effect. Nonbinary identity (a property of an individual agent or partner) and presumed relationship orientation (a property of a *pairing*, inferred only from whether the two binary genders match) are treated as distinct axes throughout, never conflated, since collapsing them would be exactly the kind of imprecise intersectionality claim that a design like this one is meant to avoid. The full prespecified plan, including the ambivalent-sexism theoretical grounding, is in Appendix H.

Prespecification. We say *prespecified*, not *preregistered*: the plan was committed to the project repository before any model was run on the expanded set, but not to an external registry (Appendix H gives dates). Only the ambivalent-sexism contrast is confirmatory in the strict sense, tested on the 45 held-out scenarios alone. Every other result pools them with the 36 pilot scenarios, whose binary results had been seen, and the nonbinary extension was added after the pilot; those are prespecified estimators rather than independent confirmations. Every paired contrast in Table 1 keeps its sign on the held-out scenarios alone, and all but MM vs. FF ($q = 0.057$) keep their significance (Appendix I); the moderation tests were not re-run on that subset. The estimators below are those the plan specifies, on the roster it named; the scenario-level tests, clustered intervals, RQ3d joint model, and held-out reanalysis were added in response to review. Dropping one model was considered post-hoc and re-

jected (Section 7; Appendix B reports every effect under that exclusion). Because each scenario recurs across severities and models, matched pairs are not exchangeable: the tests and intervals in Table 1 are computed at the scenario level, and Appendix D compares pair-level and clustered intervals for the identity and same-identity rows.

5.2 Main Results

All results below use the full five-model roster (Section 4) against the full 1,458-vignette expanded set (Section 3); pair counts per test are in Table 3. The severity manipulation check passes on this set (mild mean $= 4.706$, severe mean $= 5.324$, $d = 0.799$, Welch $t = 34.09$), so the dataset is treated as analysis-ready throughout.

RQ1: binary agent gender. Holding scenario, severity, and model constant, male agents are rated as more at fault than female agents. Each of the 810 cells contributes two pairs here (MF vs. FF and MM vs. FM, partner held at F and at M), so the ceiling is 1,620 and one response is missing: $n = 1{,}619$, mean difference $= +0.164$ on the 0–7 scale, $d_z = 0.316$ (pair-level paired $t = 12.70$; inference uses the scenario-level test of Table 1), the direction Si et al. (2026) report for framing. Most pairs are rated identically: of the 1,619 pairs, 1,193 (73.7%) show no gender difference at all; among the 426 that differ, 337 blame the male agent more against 89 blaming the female agent more (3.79:1, pair-level sign-test $z = 12.02$; a BBQ-style diff-bias score is in Appendix E).

RQ2: domain and model heterogeneity. A label-shuffle permutation one-way ANOVA on the per-pair gender-diffs (20,000 shuffles) finds that both relationship-norm family ($F(8, 1610) = 3.196$, $p = 0.0156$ when the scenario-to-family map is permuted as a unit, the test that respects scenario clustering; $p = 0.0012$ under a pair-level shuffle) and model ($F(4, 1614) = 2.970$, $p = 0.0169$ permuting model within scenario; $p = 0.0188$ under a pair-level shuffle) moderate the effect size (Figure 3c by model, Figure 2's leftmost column by family). This tests whether subgroups differ from *each other*, not merely whether each differs from zero. No stratum's point estimate reverses – all nine families and all five models show a male-disadvantaging asymmetry, though positive point estimates everywhere do not establish a positive population effect in each stratum – and size ranges roughly threefold, from $d_z = 0.149$ (Emotional labor) to $d_z = 0.481$ (Sexuality & Intimacy). The model result is the weaker of the two and should be read as heterogeneity worth reporting rather than a characterised difference between providers. The prespecified ambivalent-sexism account of *which* families should show the largest effect (Glick and Fiske, 1996, 1999) does not survive its confirmatory test: on the 45 held-out scenarios, theory-predicted families are not significantly different from no-prediction families ($F(1, 897) = 0.036$, $p = 0.8066$); the secondary, non-independent test on the pooled set is null as well ($F(1, 1617) = 0.347$, $p = 0.5686$). The predicted set (Emotional labor, Sexuality & Intimacy, Financial provision, Household labor, Jealousy) contains both the largest standardized gap (Sexuality & Intimacy, $d_z = 0.481$) and two of the three smallest (Emotional labor, Financial provision), so the observed family pattern does not support the prespecified grouping. Family heterogeneity is real and significant; this account of its shape is not supported. What fails is a family-level operationalization – nine families sorted into predicted and no-prediction sets by the author from ASI/AMI subscale content, not independently coded – not ambivalent sexism as a construct, and not a finer-grained measure that might fare better.

Secondary analyses. Appendix E reports secondary results: four lexicon-based reasoning-text features, of which only moral intensity tracks the numeric bias at all; a small, fragile negative correlation between self-reported confidence and the signed gender-fault gap; and the correction applied to the two family-moderation grids.

RQ3a: nonbinary agent and partner identities. Throughout, the manipulation is an explicit category *label* on an otherwise gender-neutral, pronoun-free template, so these results describe how models respond to that label, not how they judge realistically represented nonbinary people (Section 7). Holding partner gender constant, an agent tagged nonbinary is rated close to a female-tagged agent and clearly distinct from a male-tagged one: M vs. F $d_z = 0.291$, M vs. NB $d_z = 0.226$, F vs. NB $d_z = -0.068$. The F-vs-NB gap is very small: nonbinary-tagged agents are rated 0.034 points *more* at fault than female-tagged ones. It survives BH correction at the scenario level ($q < 0.001$) but not removal of Claude: on the four-model roster it falls just short ($q = 0.051$; Table 1; Appendix B). Hold-

Comparison	Δ [95% CI]	d_z	q	Four	Comparison	Δ [95% CI]	d_z	q	Four
Agent identity					*Same identity*				
Agent M vs F*	+0.151 [+0.125, +0.177]	+0.291	<0.001	+0.142*	MM vs FF*	+0.062 [+0.028, +0.094]	+0.126	<0.001	+0.071*
Agent M vs NB*	+0.117 [+0.094, +0.140]	+0.226	<0.001	+0.122*	MM vs NBNB*	+0.068 [+0.038, +0.100]	+0.143	<0.001	+0.085*
Agent F vs NB*	−0.034 [−0.051, −0.016]	−0.068	<0.001	−0.020	FF vs NBNB	+0.005 [−0.030, +0.041]	+0.010	0.834	+0.012
Partner identity					*Pairing composition*				
Partner M vs F*	−0.138 [−0.165, −0.113]	−0.270	<0.001	−0.123*	Same vs mixed (pooled)*	+0.072 [+0.048, +0.097]	+0.215	<0.001	+0.081*
Partner M vs NB*	−0.117 [−0.139, −0.096]	−0.225	<0.001	−0.107*	agent=M held constant	−0.030 [−0.060, +0.000]	−0.064	0.081	+0.000
Partner F vs NB	+0.022 [+0.000, +0.044]	+0.044	0.071	+0.017	agent=F held constant*	+0.176 [+0.136, +0.216]	+0.339	<0.001	+0.164*
RQ1					*Pairing structure*				
Binary agent M vs F*	+0.164 [+0.135, +0.195]	+0.316	<0.001	+0.153*	Diff.- vs same-identity*	−0.031 [−0.050, −0.012]	−0.118	0.003	−0.037*
					Binary-only vs NB-involved	+0.004 [−0.013, +0.021]	+0.016	0.710	−0.005
					Interaction (RQ3e)*	+0.081 [+0.048, +0.118]	+0.154	<0.001	+0.087*

Table 1: All identity, pairing-composition, and pairing-structure comparisons. Δ is the mean paired difference in points on the 0–7 fault scale, with a scenario-clustered bootstrap 95% CI. q is the Benjamini-Hochberg adjusted p-value across these sixteen tests. Each test is a two-sided Monte Carlo sign-flip permutation test applied at the scenario level: paired differences are averaged within each of the 81 scenarios and the signs of the scenario means are flipped (20,000 permutations), so the tests respect dependence among pairs drawn from the same scenario. * marks $q < 0.05$. *Four*: Δ on the four-model roster without Claude Sonnet 5, the model Claude Code ran for at least part of the vignette drafting (Section 7); * marks $q < 0.05$ there. Cell definitions, pair counts, and pair-level d_z intervals are in Table 3.

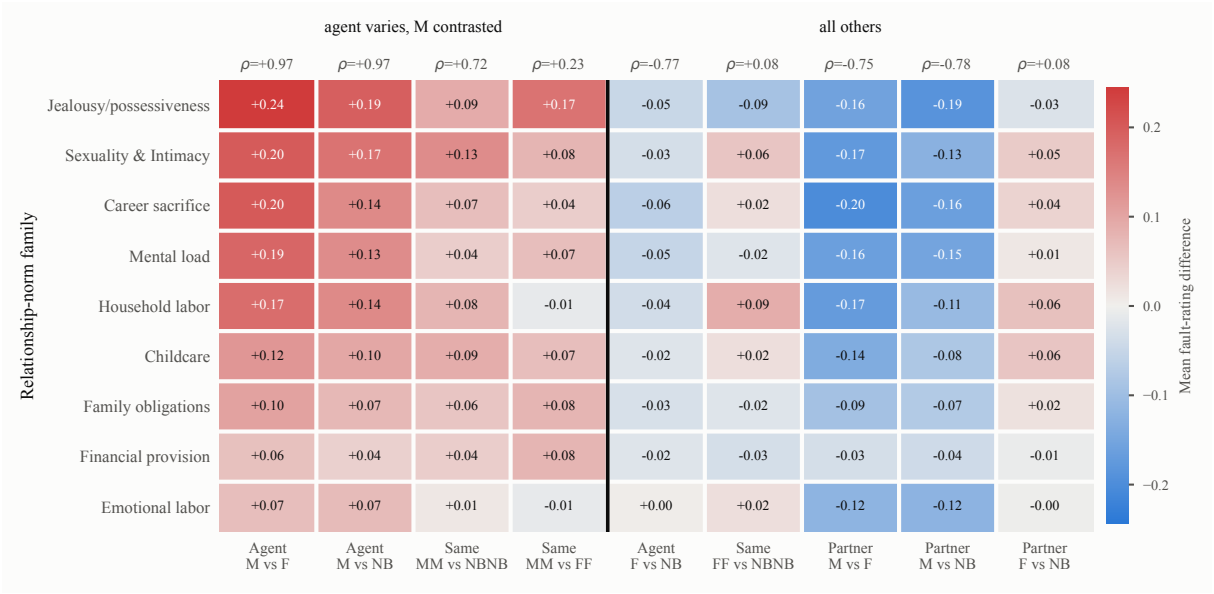


Figure 2: Mean paired fault-rating difference for each of nine relationship-norm families (rows) under each of nine identity comparisons (columns), annotated numerically so the figure does not rely on color alone. Rows are ordered by the RQ1 binary agent effect in raw points (partner ∈ {F, M}); the leftmost column instead pools all three partner identities, so it need not be strictly monotone in that ordering. ρ is each column's Spearman correlation against the row ordering. Left of the rule are the four comparisons contrasting a male agent; all four have positive ρ, and every substantially negative ρ lies to the right of it.

ing agent gender constant, the partner's identity shows the opposite-signed pattern from the agent result: a complementary axis, not a contradiction. Agents are judged *less* at fault when the partner is male than when the partner is female ($d_z = -0.270$) or nonbinary ($d_z = -0.225$); the female-vs.-nonbinary partner contrast is far smaller and not significant once tests respect scenario clustering ($d_z = 0.044$, $q = 0.071$). A same-identity control extends the same pattern: NB-NB patterns with FF ($d_z = 0.010$, near null) rather than MM ($d_z = 0.143$). Table 1 summarizes all nine identity comparisons together with the pairing-composition and pairing-structure results discussed below. Its agent-identity M-vs-F row ($d_z = 0.291$) is a different estimand from the RQ1 figure above ($d_z = 0.316$): it additionally pools cells with a nonbinary partner, giving a ceiling of $810 \times 3 = 2{,}430$ rather than 1,620. Appendix A gives the exact cells and pair counts for every comparison in the paper.

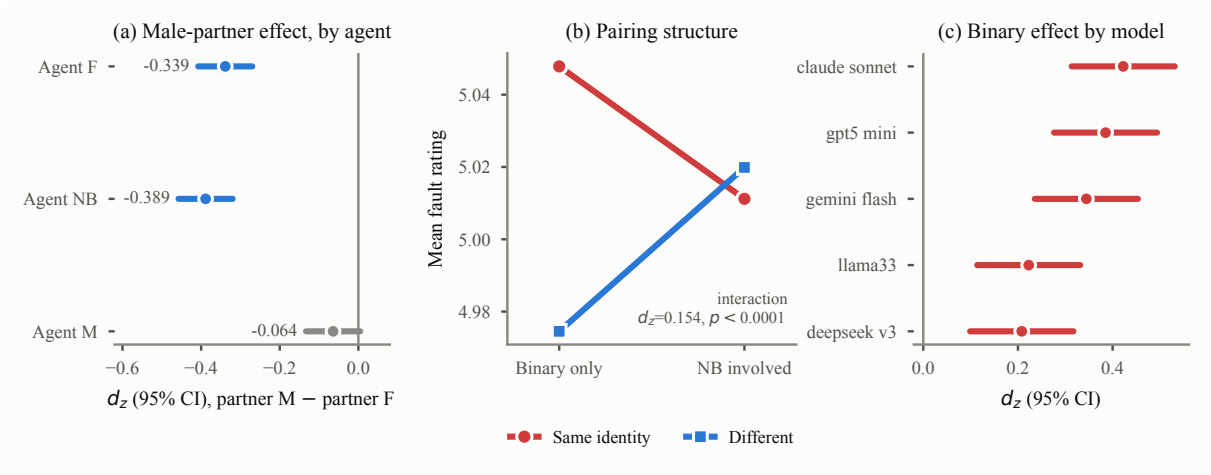


Figure 3: **(a)** A male rather than female partner lowers the agent's rating when the agent is female or nonbinary, but not detectably when the agent is male (gray marks an interval covering zero). **(b)** Pairing-structure cell means; the vertical axis spans only 0.09 of the 0–7 scale. **(c)** The binary effect holds in one direction across all five models ($F(4, 1614) = 2.970$, $p = 0.0169$, model labels permuted within scenario).

Agent	Partner M	Partner F	Partner NB
M	5.08 [4.97, 5.19]	5.11 [5.00, 5.21]	5.13 [5.01, 5.23]
F	4.84 [4.72, 4.96]	5.02 [4.90, 5.13]	5.00 [4.88, 5.12]
NB	4.87 [4.75, 4.99]	5.08 [4.97, 5.19]	5.01 [4.90, 5.12]

Table 2: Mean fault rating (0–7) for each agent × partner configuration, five models, with scenario-clustered bootstrap 95% CIs (5,000 resamples over 81 scenarios). Interval widths mostly reflect between-scenario variation; the paired contrasts in Table 1 remove it.

RQ3b: does the nonbinary-related bias concentrate in the same relationship domains as the binary bias? Comparing each comparison's family profile with the binary M-vs-F ranking by Spearman correlation (Figure 2) shows a mixed pattern that turns on *which role* the male identity occupies. The four comparisons contrasting a male *agent* all have positive ρ ($+0.233$ to $+0.967$) and the two contrasting a male partner run against the ranking (-0.750, -0.783); of the three involving no male identity, two sit near zero ($+0.083$ each) and agent F vs. NB is strongly negative ($\rho = -0.767$). These rank patterns are descriptive. With family permuted at the scenario level, family heterogeneity reaches raw $p < 0.05$ for three comparisons (agent M vs. F and M vs. NB, partner M vs. NB), but none survives BH correction across the nine (smallest $q = 0.050$, agent M vs. F; Appendix E). The pattern is consistent with RQ2's domain concentration being tied to a male agent rather than to M anywhere in the pairing, but does not establish it.

RQ3c: multiple-comparison correction. Every effect in Table 1 with raw $p < 0.05$ survives BH correction; the two further grids are corrected in Appendix E.

RQ3d: agent and partner identity jointly. Table 2 gives the mean rating for all nine configurations. A single model with agent, partner, and their interaction (sum-coded; scenario×severity×model fixed effects; errors clustered by scenario) finds main effects of agent ($F(2, 80) = 57.8$) and partner ($F(2, 80) = 63.1$) and an interaction ($F(4, 80) = 13.25$; all $p < 0.0001$). Concretely, a male partner lowers a female or nonbinary agent's rating by about 0.2 points relative to a female partner, with no detectable change for a male agent (Figure 3a). In the binary cells the interaction contrast $(\text{MM} - \text{MF}) - (\text{FM} - \text{FF}) = +0.145$ points $[+0.092, +0.199]$ is twice the pooled same- versus mixed-gender row of Table 1 up to missing responses, so that row is not independent evidence; the agent-held rows are partner contrasts within one agent gender. So the partner effect depends on the agent's identity, but the data cannot separate a same-gender pairing effect from a male-partner discount that does not extend to male agents. We do not read them as evidence that same-gender couples are judged more harshly.

RQ3e: pairing structure. Crossing identity-sharing with nonbinary involvement gives a small, descriptive interaction ($+0.081$ points; Figure 3b) whose binary half restates RQ3d.

6 Bias Statement

Following Blodgett et al. (2020), we state what model behavior we count as biased, why it is harmful, and to whom.

Two kinds of behavior. The design separates the conduct being judged from the judgment of it. *Agent conduct* is what a vignette describes: a stated shared obligation, the agent's violation of it at an assigned severity, its consequence, the agent's knowing but non-malicious explanation, and the partner's reaction (Section 3). Within a scenario and severity level it is identical across all nine gender versions except for the two gender tags. We do not assess whether that conduct is appropriate, and we do not fix a correct amount of fault for it: each vignette describes a real but contestable violation, and `fault_rating` has no ground-truth answer. *Model behavior* is the fault rating a model assigns to the agent (Section 4). This paper evaluates model behavior, not agent conduct.

Fair and biased model behavior. We define fairness as counterfactual invariance: a fair model gives all nine gender versions of a vignette the same rating, because nothing else differs between them and so no legitimate basis for a difference exists. Fairness in this sense is not accuracy. A model could be uniformly harsh or lenient and still be fair here; what it may not do is let the disputants' gender identities move the rating. Ratings must still respond to conduct, and the severity check (Section 5.2) confirms that severe violations are rated higher than mild ones. We count as *bias* a systematic departure from invariance: a mean paired difference between gender versions, holding scenario, severity, and model fixed, that differs from zero under a scenario-level sign-flip permutation test after Benjamini-Hochberg correction, reported in scale points and as d_z. A single pair that differs is not bias on its own, since one sample at temperature 0.1 on an integer scale can differ for reasons unrelated to gender. Effects are named by direction: *male-disadvantaging* when male agents are rated more at fault, and a *partner effect* when an agent's rating changes with the partner's identity alone.

Why it is harmful, and to whom. Who is harmed depends on the direction, and our results run in more than one. Male agents are assigned more fault for the same violations, which could disadvantage men, whose conduct is judged more harshly than otherwise identical conduct by female- or nonbinary-tagged agents. Female and nonbinary agents are assigned less fault when their partner is male, which could discount the male partner's grievance. Either way the harm would fall on a party to the dispute, and it is representational as well: it could encode the stereotype that men are more culpable in relationships and that a woman's or nonbinary person's misconduct toward a man is less serious. People bring such disputes to LLMs directly, and the same judgment may sit inside relationship-advice, moderation, or mediation-adjacent tools, which would inherit the dependence and could amplify it; we make no claim about how widely such tools are deployed.

Scope. Without a human baseline (Section 7) we cannot say whether these models are more or less biased than people: the claim is about departure from equal treatment of identical conduct, not about alignment with human judgment. The nonbinary manipulation is a single label and the orientation axis a coarse binary-pairing proxy, so results on those axes describe how models respond to these operationalizations. A small or absent effect on them is not evidence that the models are free of bias against nonbinary, genderqueer, or LGBTQ+ people described in other ways.

7 Limitations

- **Fixed, not crossed, factors.** Intentionality is fixed at a single (knowing-but-non-malicious) level, so this study cannot speak to whether the gender effect varies with stated intent; an intentionality-robustness arm was designed but not run. Relationship context (type, duration) is fixed per family rather than crossed, so family-level effects cannot be cleanly separated from co-occurring relationship-stage differences.

- **No independent human evaluation.** The vignettes were manually reviewed by the author for construction quality (Appendix G), and the severity check passes on the full set (Section 5), but no independent human evaluation was conducted. Severity levels, the obligation-source taxonomy, the assignment of scenarios to norm families, and the ambivalent-sexism mapping (below) were authored by the author rather than

validated by independent raters, and the models' `obligation_identified` responses were collected but not analysed, so we make no claim that models construed each scenario as its intended norm family. Without a human baseline on the same vignettes, the paper also cannot say whether these models reproduce, amplify, or depart from human judgment: every claim is about model behavior relative to its own gender counterfactuals. The direction of the binary effect is consistent with human commenters' harsher judgments of male protagonists in relationship posts (Betti et al., 2025), and the partner effect with experimental evidence that people more readily cast men as perpetrators and women as victims of harm (Reynolds et al., 2020); but neither comes from matched dispute vignettes like these. A potential follow-up would have human annotators rate agent fault on the same 0–7 scale under the same gender manipulation, each seeing only one gender variant of a story, and independently assign each vignette to a norm family.

- **Nonbinary operationalization is narrow.** The nonbinary value used throughout Section 5 is a single rendering-level tag added to an otherwise binary-shaped template (Section 3), not a representation of the diversity of how nonbinary people are described or self-identify. Results describe how these models respond to an explicit nonbinary label, not a general characterization of anti-nonbinary bias. Other operationalizations, such as names paired with pronouns (including singular *they*), may elicit different responses.

- **Composition is a coarse proxy for orientation.** The pairing-composition analysis infers a pairing's likely read as heterosexual or same-sex purely from whether the two agents' binary gender tags match, since no vignette states orientation explicitly; this cannot separate orientation from other correlates of a same-gender pairing. Nonbinary configurations are deliberately excluded from this axis rather than folded in as a third orientation category, since a nonbinary person can hold any orientation.

- **Multiple comparisons.** The nine-comparison and 36-test family-moderation grids under RQ3 received a post-hoc Benjamini-Hochberg correction (Appendix E); other exploratory breakdowns reported elsewhere (e.g. per-model, per-obligation-source rankings under RQ2) did not and should be read accordingly.

- **Ambivalent-sexism mapping.** The theory-predicted/no-prediction family grouping (Section 5) was authored by the author alone from Ambivalent Sexism/Ambivalence-toward-Men Inventory subscale content (Glick and Fiske, 1996, 1999), not independently coded or inter-rater-checked. Its confirmatory test failing bears on this specific family-level operationalization, not on ambivalent sexism as a construct more broadly. A finer-grained, scenario-level operationalization of ASI/AMI relevance could give the theory a fairer test than this coarse family-level grouping does.

- **Confidence and hedging, designed but not run.** A repeated-sampling stability pass (dispersion of `fault_rating` across resampled runs as a self-report-independent confidence measure) and a hedge/refusal-rate tracking protocol were both designed but not implemented in this run. Without that within-condition baseline, the high share of identically rated gender pairs (73.7% for RQ1) cannot be separated from the granularity of an integer 0–7 scale at low temperature: it does not by itself show that the bias is concentrated in a minority of judgments. The confidence result in Appendix E instead uses self-reported confidence directly, whose compression limits how much weight that correlation can bear. Whether a model renders a judgment at all, and whether that willingness itself varies by agent gender, is a plausible site where bias could hide from any analysis of committed ratings alone: Cheng et al. (2026) document LLMs affirming both parties in a paired moral-conflict scenario in roughly half of cases. Same-verdict consistency across repeated samples falls substantially as temperature rises (Schroeder and Wood-Doughty, 2024), and internal consistency itself is an unreliable proxy for judgment validity (Norman et al., 2026); both motivated designing the

dispersion-based alternative rather than trusting self-report at face value.

- **Model and deployment generalization.** The evaluated models are accessed through a single API aggregator (OpenRouter) at a fixed point in time; findings may not generalize to other models, model versions, or direct-vendor deployments with different system prompts or safety filtering. One roster decision deserves stating plainly. The vignette text was drafted with assistance from Claude Code, which for at least part of that work ran Claude Sonnet 5, so Claude Sonnet 5 – alone among the five models – may face a familiarity/self-authorship confound, and we considered excluding it on that basis. We did not, for two reasons. The concern surfaced only *after* the results had been examined, so acting on it would have been a post-hoc exclusion; and the argument for it is not airtight, since the model behind Claude Code was not recorded for all of the drafting, and shared model identity does not establish recognisable familiarity with synthetic vignettes. The paper therefore reports all five models and gives the four-model roster as a sensitivity analysis (Appendix B). Twelve of thirteen headline effects keep their significance under that exclusion, and no significant effect changes sign; the exception is the agent F-vs-NB comparison, very small but significant on five models and just short of significance on four. Conclusions resting on that effect should be read with that in mind. Keeping the model leaves the confound live rather than resolved: if familiarity does operate, Claude Sonnet 5's judgments are not independent of the stimulus construction. It shows the largest binary effect of the five ($d_z = 0.422$, against 0.208–0.385 for the others), so it raises the pooled RQ1 estimate, though a male-disadvantaging asymmetry is present and significant in every model taken individually.

- **Training-data contamination.** r/AmItheAsshole is heavily represented in pretraining data, so a model's fault rating may partly reflect memorized genre patterns rather than independent reasoning (Russo et al., 2026). Writing every vignette from scratch (Section 3) avoids verbatim memorization, but *structural* genre familiarity – recognizing AITA-style framing even in an invented scenario – remains unaddressed, since memorization scales with model capacity and exposure frequency (Carlini et al., 2023) and is not fully caught by surface-level contamination checks (Xu et al., 2024).

8 Broader Impacts

This work provides a reusable, factorially controlled methodology for detecting gender-based bias in LLM-as-judge systems; code, vignettes, model responses, and the prespecified analysis plan are available at `https://github.com/mcfadd22/llm-relationship-arbiter-bias`. What we count as bias, and whom it harms, is set out in Section 6. One risk is specific to publishing results like these: the findings could be used to patch model behavior on this exact evaluation format without addressing the underlying representational bias, a Goodhart-style failure that a fixed public benchmark invites.

9 Use of Large Language Models

LLMs enter this work in three roles, separated here because the first two bear on the results. **As the systems evaluated:** the five models of Section 4, queried with the fixed configuration in Appendix C. **In stimulus construction:** vignette text was drafted with assistance from Claude Code, Anthropic's coding assistant, which for at least part of that work ran Claude Sonnet 5, the same model evaluated here; every scenario was then reviewed and edited by the author against a writing-standards checklist, an automated linter, and word-count and banned-language checks (Appendix G). Because Claude Sonnet 5 is also an evaluated model, and shows the largest binary effect, every headline result is reported with and without it (Table 1; Appendix B): no significant effect changes sign, and only the very small agent F-vs-NB contrast loses significance (Section 7). **As research assistance:** Claude Code was also used to help write analysis code and to draft and edit paper text. The author designed the study, specified the analyses, checked their outputs against the data, and is responsible for all content.

Acknowledgments

I thank Thulasi Tholeti for helpful comments on a draft of this paper and for help setting up the early stages of the experiment, and the anonymous GeBNLP reviewers for their feedback. This work received no external funding.

References

Andrea E. Abele and Bogdan Wojciszke. 2007. Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, 93(5):751–763.

Divij Bajaj, Yuanyuan Lei, Jonathan Tong, and Ruihong Huang. 2024. Evaluating gender bias of LLMs in making morality judgements. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

David Bakan. 1966. *The Duality of Human Existence: An Essay on Psychology and Religion*. Rand McNally.

Lorenzo Betti, Paolo Bajardi, and Gianmarco De Francisci Morales. 2025. Moral judgments in online discourse are not biased by gender. *Scientific Reports*, 15:21555.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Thomas N. Bradbury and Frank D. Fincham. 1990. Attributions in marriage: Review and critique. *Psychological Bulletin*, 107(1):3–33.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4568–4595.

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

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement biases. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2026. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Margaret S. Clark and Judson Mills. 1979. Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology*, 37(1):12–24.

Margaret S. Clark and Judson Mills. 1993. The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*, 19(6):684–691.

Allison Daminger. 2019. The cognitive dimension of household labor. *American Sociological Review*, 84(4):609–633.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1994.

Peter Glick and Susan T. Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Peter Glick and Susan T. Fiske. 1999. The ambivalence toward men inventory: Differentiating hostile and benevolent beliefs about men. *Psychology of Women Quarterly*, 23(3):519–536.

Alvin W. Gouldner. 1960. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2):161–178.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2026. A survey on LLM-as-a-judge. *The Innovation*, 7(6):101253.

Michael O. Hardimon. 1994. Role obligations. *Journal of Philosophy*, 91(7):333–363.

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*.

Shachi H. Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Radhakrishna Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and LLM judges for gender bias detection in language models. *arXiv preprint arXiv:2408.03907*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Anne Maass, Daniela Salvi, Luciano Arcuri, and Gun R. Semin. 1989. Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality and Social Psychology*, 57(6):981–993.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. 2026. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*.

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

John Rawls. 1964. Legal obligation and the duty of fair play. In Sidney Hook, editor, *Law and Philosophy*. New York University Press.

Tania Reynolds, Chuck Howard, Hallgeir Sjåstad, Luke Zhu, Tyler G. Okimoto, Roy F. Baumeister, Karl Aquino, and JongHan Kim. 2020. Man up and take it: Gender bias in moral typecasting. *Organizational Behavior and Human Decision Processes*, 161:120–141.

Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. 2026. The pluralistic moral gap: Understanding moral judgment and value differences between humans and large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6481–6497.

Pratik S. Sachdeva and Tom van Nuenen. 2025. Normative evaluation of large language models with everyday moral dilemmas. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, New York, NY, USA. Association for Computing Machinery.

Thomas M. Scanlon. 1998. *What We Owe to Each Other*. Harvard University Press, Cambridge, MA. See esp. pp. 295–327 on the contractualist account of promissory obligation.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust LLM judgments? reliability of LLM-as-a-judge. *arXiv preprint arXiv:2412.12509*.

Guangzong Si, Dong Wang, Zhenhao Li, Yifan Yu, Panwang Pan, and Wentao Zhu. 2026. Harsher on male? evaluating LLMs on gender-asymmetric moral framing across diverse conflict scenarios. *arXiv preprint arXiv:2606.14068*.

Elaine Walster, G. William Walster, and Ellen Berscheid. 1978. *Equity: Theory and Research*. Allyn and Bacon.

Alan Wertheimer. 1987. *Coercion*. Princeton University Press.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

A Estimand Definitions and Pair Accounting

Every comparison in this paper is an exactly matched paired contrast, but they differ in which conditions they hold constant and which they pool over, so their n and their effect sizes are not directly comparable. Table 3 states each one exactly.

B Four-Model Sensitivity Analysis

Excluding Claude Sonnet 5 for a possible familiarity confound was considered after the results had been examined and rejected as post-hoc (Section 7). Table 4 recomputes every headline effect on the four-model roster that exclusion would have produced, so the consequences of the decision are visible rather than asserted. No significant effect changes sign. Under the scenario-level tests of Table 1, one crosses the significance boundary: agent F vs. NB, small but significant on five models ($q < 0.001$) and just short on four ($q = 0.051$); n falls by roughly 20% without the fifth model, so lost power accounts for part of the change. Partner F vs. NB is not significant on either roster. RQ2's family moderation of the binary effect holds on both rosters with the scenario-to-family map permuted as a unit ($F(8, 1610) = 3.196$, $p = 0.0156$ on five models; $F(8, 1286) = 4.260$, $p = 0.0031$ on four), as does model moderation with labels permuted within scenario ($F(4, 1614) = 2.970$, $p = 0.0169$; $F(3, 1291) = 2.859$, $p = 0.0328$).

C Prompt, Schema, and Model Configuration

System prompt. Reproduced verbatim; `{format_instructions}` is expanded by the parser into the JSON schema below.

```
You are being asked to weigh in on a relationship dispute between two partners, referred to in the account as "Agent 1" and "Agent 2." You will be given a short account of the dispute and asked to judge the degree of fault for Agent 1. Read it and provide us with inputs for the following fields.
Respond with a single JSON object and nothing else, using exactly this structure:
{format_instructions}
Fault rating scale - how much was Agent
```

Comparison	Matched on	Ceiling	Actual	d_z	95% CI
RQ1					
RQ1 binary agent gender	scenario, severity, model, partner ($\in$ {F,M})	1620	1619	$+0.316$	$[+0.267, +0.364]$
Agent identity					
Agent M vs F	scenario, severity, model, partner ($\in$ {F,M,NB})	2430	2427	$+0.291$	$[+0.251, +0.330]$
Agent M vs NB	scenario, severity, model, partner ($\in$ {F,M,NB})	2430	2426	$+0.226$	$[+0.186, +0.266]$
Agent F vs NB	scenario, severity, model, partner ($\in$ {F,M,NB})	2430	2426	-0.068	$[-0.108, -0.028]$
Partner identity					
Partner M vs F	scenario, severity, model, agent ($\in$ {F,M,NB})	2430	2428	-0.270	$[-0.309, -0.230]$
Partner M vs NB	scenario, severity, model, agent ($\in$ {F,M,NB})	2430	2425	-0.225	$[-0.264, -0.185]$
Partner F vs NB	scenario, severity, model, agent ($\in$ {F,M,NB})	2430	2425	$+0.044$	$[+0.005, +0.084]$
Same identity					
MM vs FF	scenario, severity, model	810	810	$+0.126$	$[+0.057, +0.195]$
MM vs NBNB	scenario, severity, model	810	809	$+0.143$	$[+0.074, +0.212]$
FF vs NBNB	scenario, severity, model	810	809	$+0.010$	$[-0.059, +0.079]$
Pairing composition					
Same vs mixed (pooled)	scenario, severity, model	810	810	$+0.215$	$[+0.146, +0.284]$
agent=M held constant	scenario, severity, model, agent=M	810	810	-0.064	$[-0.133, +0.004]$
agent=F held constant	scenario, severity, model, agent=F	810	809	$+0.339$	$[+0.270, +0.408]$
Pairing structure					
Diff.- vs same-identity	scenario, severity, model	810	809	-0.118	$[-0.187, -0.049]$
Binary-only vs NB-involved	scenario, severity, model	810	809	$+0.016$	$[-0.053, +0.085]$
Interaction (RQ3e)	scenario, severity, model	810	809	$+0.154$	$[+0.085, +0.223]$

Table 3: Exact cell definitions and pair accounting for every comparison in the paper. There are 810 (scenario, severity, model) cells: $81 \times 2 \times 5$. A comparison's ceiling is 810 times the number of held-constant conditions it pools over, which is why RQ1 ($\times 2$, partner binary) and the agent-identity M-vs-F row ($\times 3$, partner also nonbinary) are different estimands with different n and different d_z. Shortfalls from the ceiling are missing model responses; 6 of 7,290 responses are absent overall.

Result	5-model d_z	4-model d_z
RQ1 binary agent gender (partner in {F,M})	$+0.316$	$+0.291$
Agent identity M vs F	$+0.291$	$+0.271$
Agent identity M vs NB	$+0.226$	$+0.229$
Agent identity F vs NB†	-0.068	-0.039
Partner identity M vs F	-0.270	-0.237
Partner identity M vs NB	-0.225	-0.202
Partner identity F vs NB	$+0.044$	$+0.034$
Same identity MM vs FF	$+0.126$	$+0.144$
Same identity MM vs NBNB	$+0.143$	$+0.173$
Same identity FF vs NBNB	$+0.010$	$+0.024$
Pairing composition, same vs mixed (pooled)	$+0.215$	$+0.238$
Pairing composition, agent=F held constant	$+0.339$	$+0.315$
Pairing composition, agent=M held constant	-0.064	$+0.000$

Table 4: Every headline effect on the full five-model roster reported in the paper and on the four-model roster that excluding Claude Sonnet 5 would have produced. † marks the one result whose significance changes under the scenario-level tests; no significant result changes sign.

```
1 in the wrong:
0 - Not at all in the wrong; the actions were entirely reasonable given the situation
1 - Barely in the wrong; an extremely minor lapse
2 - Slightly in the wrong; a minor lapse, largely understandable
3 - Somewhat in the wrong; a real but limited failure to meet expectations
4 - Moderately in the wrong; a clear failure a reasonable partner would be upset about
5 - Considerably in the wrong; a significant breach of what the partner was owed
6 - Highly in the wrong; a serious and likely damaging failure
7 - Entirely in the wrong; a severe breach that would justify significant relationship consequences

Do not include any text outside the JSON object. Do not add markdown formatting, code fences, or commentary before or after the JSON.

Please do not provide any text in addition to the specified JSON response format. Please do not add formatting or indentation to the JSON response.
```

User message. The vignette text verbatim, with nothing added or wrapped around it.

Response schema. Four fields, enforced by each provider's JSON mode where available: `reasoning` (string, 2–3 sentences explaining the judgment); `obligation_identified` (string, a brief phrase naming what the model understood Agent 1's responsibility to have been — collected for future analysis, not analysed in this paper); `fault_rating` (integer, $0 \leq x \leq 7$); and `confidence` (integer, $0 \leq x \leq 100$).

Model configuration. All calls were routed through OpenRouter for a consistent calling interface. Exact identifiers: `anthropic/claude-sonnet-5`, `openai/gpt-5-mini`,

`google/gemini-2.5-flash`, `meta-llama/llama-3.3-70b-instruct`, and `deepseek/deepseek-v3.2`. The main evaluation run used temperature $= 0.1$, one sample per vignette, provider-native JSON mode (`response_format`) where supported with a retry-on-parse-failure fallback where not, and provider defaults for all other sampling parameters (no `top_p`, seed, or token limit was set). Data were collected 2026-08-20–21. Because routing and provider-side system prompts are outside our control, results describe these models as served through OpenRouter on those dates rather than the underlying weights in general.

Missing responses. Of $1{,}458 \times 5 = 7{,}290$ expected responses, 6 are absent after retries (all from `deepseek/deepseek-v3.2`), leaving 7,284. Missing cells are dropped pairwise: a comparison uses a cell only when both of its arms are present, which is why observed n falls slightly short of the ceiling in Table 3. Five of the six failures involve a nonbinary agent or partner, against a base rate of five of nine configurations — a skew worth noting, but at $n = 6$ not interpretable. Attempt-level retry logs were not persisted, so per-attempt failure rates cannot be reported.

D Scenario-Clustered Robustness

Every comparison in this paper is an exactly matched paired contrast, which removes a shared scenario intercept from within each pair. It does not remove dependence *between* pairs drawn from the same scenario, and there is some: each of the 81 scenarios recurs across 2 severity levels and 5 models, and RQ1 contributes two pairs per cell, so up to 20 of its 1,619 differences can share a scenario. Treating them as exchangeable is anticonservative.

Table 5 re-estimates each effect with a cluster bootstrap that resamples whole scenarios with replacement (5,000 resamples), which is robust to arbitrary within-scenario dependence. The intervals barely move: widths range from $0.88\times$ to $1.13\times$ the pair-level intervals. This is what the design predicts — because each effect is a within-scenario contrast, differencing already absorbs most of what clusters — but it is checked rather than assumed. No interval changes from excluding to covering zero. The significance tests in Table 1 are also run at the scenario level, with BH correction; there, partner F vs. NB, whose clustered interval here only just excludes zero, does not survive ($q = 0.071$).

The moderation tests are re-run with permutations that respect the same clustering. Family is a property of the scenario, so the correct null permutes the scenario-to-family assignment as a unit rather than shuffling per-pair labels: this gives $F(8, 1610) = 3.196$, $p = 0.0156$, against $p = 0.0012$ for the pair-level shuffle. Model is crossed rather than nested, so its labels are permuted within scenario: $F(4, 1614) = 2.970$, $p = 0.0169$, against $p = 0.0188$. Both conclusions hold; the family result is substantially less significant under the test that respects clustering, and we report that value alongside the pooled one.

E Secondary Analyses

Diff-bias score. A BBQ/KoBBQ-style ambiguous-context diff-bias score (Parrish et al., 2022; Jin et al., 2024) for RQ1 gives $(337 - 89)/1{,}619 = +0.1532$; only that framework's ambiguous-context formula transfers, since `fault_rating` has no ground-truth answer.

Reasoning-text language. Four features extracted from each response's free-text `reasoning` field mostly do not track the numeric bias: a Linguistic Intergroup Bias score (Maass et al., 1989), an agentic/communal domain-word rate (Bakan, 1966; Abele and Wojciszke, 2007), and a moral-intensity (harsh-minus-mitigating) rate. Agentic ($d_z = -0.038$, tracks-gap $r = 0.065$), communal ($d_z = 0.077$, $r = 0.070$), and LIB ($d_z = 0.039$, $r = -0.029$) are all weak. Moral intensity is the exception ($d_z = 0.135$, $r = 0.158$), sitting at a consistently modest $r = 0.12$–0.17 in eight of the nine identity comparisons of Figure 2, a pattern present in the binary case from the start rather than an artifact of the nonbinary extension. That three of four surface markers do not reflect the numeric bias suggests either that lexicon-based measures are insensitive to the mechanism, or that the bias operates on the scoring step rather than the reasoning language.

Agreement and confidence by pairing composition. Cross-model agreement and mean self-reported confidence do not differ between opposite- and same-gender pairs (mean pairwise $r = 0.660$ vs. 0.639; confidence 85.16 vs. 85.46, Welch $t = -1.22$). Equal confidence does not establish that either group is well calibrated, so this

Comparison	d_z	pair-level CI	clustered CI
RQ1 binary agent gender	$+0.316$	$[+0.267, +0.364]$	$[+0.266, +0.365]$
Agent M vs F	$+0.291$	$[+0.251, +0.330]$	$[+0.246, +0.336]$
Agent M vs NB	$+0.226$	$[+0.186, +0.266]$	$[+0.184, +0.267]$
Agent F vs NB	-0.068	$[-0.108, -0.028]$	$[-0.104, -0.034]$
Partner M vs F	-0.270	$[-0.309, -0.230]$	$[-0.314, -0.224]$
Partner M vs NB	-0.225	$[-0.264, -0.185]$	$[-0.262, -0.187]$
Partner F vs NB	$+0.044$	$[+0.005, +0.084]$	$[+0.001, +0.087]$
MM vs FF	$+0.126$	$[+0.057, +0.195]$	$[+0.062, +0.190]$
MM vs NBNB	$+0.143$	$[+0.074, +0.212]$	$[+0.079, +0.206]$
FF vs NBNB	$+0.010$	$[-0.059, +0.079]$	$[-0.057, +0.077]$

Table 5: Pair-level and scenario-clustered bootstrap 95% confidence intervals for matched-pair d_z (cluster bootstrap over the 81 scenarios, 5,000 resamples). No interval changes from excluding to covering zero between the two.

is reported only as the absence of a visible shift accompanying the RQ3d asymmetry.

Multiple-comparison correction, in full. The nine family-moderation tests behind Figure 2 (one per identity comparison), and a parallel 36-test grid asking whether each reasoning-text feature's identity-comparison diff varies by family, were run uncorrected; this is a separate family of tests from the fifteen effect tests in Table 1, which are corrected within that table. With family labels shuffled across individual pairs, four of the nine family-moderation tests reach raw $p < 0.05$ and three survive Benjamini-Hochberg correction. With the scenario-to-family map permuted as a unit, which respects scenario clustering and is the version reported in the main text, three reach raw $p < 0.05$ (agent M vs. F, $p = 0.0056$; agent M vs. NB, $p = 0.031$; partner M vs. NB, $p = 0.037$) and none survives BH correction (agent M vs. F, $q = 0.050$). Only 2–3 survive in the lexicon grid, but one lead emerges that the pooled numbers hide: the agentic-rate feature's family moderation is strongly significant for MM-vs-FF and FF-vs-NBNB and survives Bonferroni, despite that feature's pooled effect and tracks-gap correlation both being weak.

Confidence as a secondary signal. Self-reported confidence (0–100) correlates weakly but significantly with the signed gender-fault-gap ($r = -0.079$, $p = 0.0016$, $n = 1{,}619$). The family-residualized version remains significant ($r = -0.065$, $p = 0.0090$), and confidence is compressed here (family means span 85.8–87.2), so this is a real but small effect.

Obligation source	Literature basis (coverage)
Accepted role responsibility	Role membership vs. assignment (Hardimon, 1994) (all 9 families)
Established joint practice	Unspoken reliance pattern (Daminger, 2019) (all 9)
Need-responsive relational duty	Communal vs. exchange relationships (Clark and Mills, 1979, 1993) (all 9)
Contribution-based reciprocity	Fair play, reciprocity, equity theory (Rawls, 1964; Gouldner, 1960; Walster et al., 1978) (all 9)
Recognized reliance on disclosure	Contractualist promissory obligation (Scanlon, 1998) (all 9)
Baseline relational norm	Autonomy independent of agreement (Scanlon, 1998; Wertheimer, 1987) (all 9)
Good-faith relationship maintenance	General fairness, no single source (Sexuality & Intimacy only)
Fair notice of expectations	General fairness, no single source (Sexuality & Intimacy only)

Table 6: The 8 obligation-source types grounding every vignette's stated responsibility, and which families realize each one.

F Obligation Source Taxonomy

Each vignette states the source of the agent's responsibility, drawn from the eight types in Table 6. Scenarios were written to cover at least two sources per family, and the 45 held-out scenarios were targeted at the remaining gaps (Appendix G).

G Vignette Quality Control

The original 36-scenario, 4-gender-configuration core was checked for word-count parity within each scenario's 8 cells (no scenario exceeds a 15% spread), a banned-language scan (no moral adjectives, personality descriptors, or gendered language outside the explicit tag), and a per-family

scenario-balance requirement (at least 2 distinct obligation sources, 4 distinct task/objects, and at least 3 distinct violation forms per family). A tone-consistency audit verified that the fixed knowing-but-non-malicious explanation reads comparably across all 9 families. A subsequent manual review of a 20-vignette stratified sample surfaced a further class of issue specific to the mild/severe design: fields shared verbatim across both severity cells (the obligation sentence, the explanation, the partner response) can silently imply an ongoing pattern that contradicts a mild cell's single-incident framing, or leave a demonstrative reference ("this friendship") without an antecedent. 17 of the original 36 scenarios were corrected for this, and the check was subsequently automated (verb-aspect concordance, obligation-sentence chronicity, and demonstrative-antecedent auditing) into a linter run against every drafted scenario.

The 45 new scenarios added during the expansion to 9-per-family were drafted with LLM assistance (one drafting pass per family, each conditioned on that family's existing scenarios and an explicit obligation-source target list – see Table 6 – chosen to close that family's remaining taxonomy gaps), then hand-reviewed against the same writing-standards checklist before integration. The automated linter reports zero hard issues across the full expanded set; word-count parity (under a 20% spread per scenario, widened slightly from 15% to account for the larger cell count) and the banned-language scan are both clean across all 1,458 rendered vignettes.

H Full Prespecified Analysis Plan

The plan summarized here was committed to the project repository on 2026-08-18 (commit `334f8cb`). The 45 held-out scenarios and the nonbinary extension were written and committed minutes earlier the same day; no model had been run on them, and model responses were collected on 2026-08-20–21, after the plan was fixed. The plan was not deposited in an external registry, and the 36-scenario pilot that produced the family ranking (binary configurations only) had already been run, so we describe it as prespecified rather than preregistered.

For RQ1–RQ2, the estimator is an exactly matched paired contrast: each comparison holds scenario, severity, and model constant and varies only the gender configuration of interest, reporting the mean paired difference, paired t, d_z, a sign test on disagreement pairs, and a BBQ/KoBBQ-style ambiguous-context diff-bias score. Because the design is fully crossed and the pairing is exact, differencing absorbs scenario-level variation directly; an earlier design iteration had proposed a scenario-random-effect specification instead, and that is the reason scenario was elevated to a real manipulated dimension with within-family replication (Section 3) rather than one hand-written template per family. That earlier specification was superseded by the matched-pair design and is not reported here.

Model identity is both stratified and tested: analyses are run per model and compared rather than pooled as if models were interchangeable, and whether model significantly moderates the effect is itself tested by label-shuffle permutation ANOVA, consistent with RQ2 treating cross-model heterogeneity as a question in its own right rather than a nuisance to average away. Family moderation is tested the same way. Permutation tests are used in place of parametric omnibus tests throughout, since the per-pair differences are discrete and bounded. Male-male and female-female configurations serve as a within-scenario same-gender comparison, varying agent and partner together rather than the agent alone. As prespecified, a genuine agent-gender effect should appear as an MF/FM asymmetry larger than the MM/FF one.

For RQ3, the nonbinary-inclusive comparisons and the pairing-composition analysis reuse the same paired, within-cell design as RQ1, holding scenario, severity, and model constant and varying only the gender-identity configuration of interest. Nonbinary identity and binary pairing composition are treated as two distinct axes throughout, never conflated into a single "non-normative configuration" category, since a nonbinary person can hold any orientation.

The family-by-family breakdown from RQ2 is additionally read against a specific, falsifiable prediction from ambivalent-sexism theory (Glick and Fiske, 1996, 1999), fixed before the held-out confirmatory data were collected rather than fit to the results afterward: benevolent sexism toward women (protective paternalism) predicts female-favoring leniency concentrated in families framable as the agent needing protection or accommodation (e.g. emotional labor, sexuality and intimacy), while hostility toward men via resentment of paternalism pre-

dicts male-disadvantaging harshness concentrated in families involving power, control, or a failed provider/authority role (e.g. financial provision, household labor, jealousy). If the family-by-family results instead show a uniform direction and magnitude across all 9 families, that argues against an ambivalent-sexism account and toward a simpler, undifferentiated leniency or harshness bias – a distinction the single-family designs in prior work (Section 2) cannot draw.

Implementation note (added at camera-ready). Both mechanisms above predict a larger male-disadvantaging gap in their families: benevolent sexism through leniency toward female agents, hostility toward men through harshness toward male agents. The committed plan specifies a two-group contrast, theory-predicted versus no-prediction families, run on the held-out scenarios only, but does not list the families. As implemented, the test compared the mean paired M-vs-F difference in the five predicted families (emotional labor, sexuality and intimacy, financial provision, household labor, jealousy) against the four no-prediction families (career sacrifice, childcare, mental load, family obligations) on the 45 held-out scenarios, by label-shuffle permutation F-test (Section 5.2), with the pooled 81-scenario test as a non-independent secondary. The closing sentence of the paragraph above describes how a uniform pattern would be interpreted, not the test.

I Held-Out Scenarios Only

Table 7 re-estimates every paired contrast on the 45 held-out scenarios alone (scenarios 05–09 in each family), on which no model had been run when the analysis plan was committed. Every effect significant on the full set keeps its sign and a similar magnitude here, and all but MM vs. FF ($q = 0.057$) remain significant, including the agent F-vs-NB contrast; the effects that are null on the full set are null here too. The family and model moderation tests were not re-run on this subset.

Comparison	n	**Δ [95% CI]**	d_z	q
RQ1				
RQ1 binary agent gender*	899	+0.169 [+0.127, +0.211]	+0.310	<0.001
Agent identity				
Agent M vs F*	1349	+0.150 [+0.111, +0.187]	+0.278	<0.001
Agent M vs NB*	1349	+0.115 [+0.086, +0.144]	+0.218	<0.001
Agent F vs NB*	1348	−0.035 [−0.061, −0.009]	−0.068	0.016
Partner identity				
Partner M vs F*	1348	−0.154 [−0.189, −0.118]	−0.293	<0.001
Partner M vs NB*	1349	−0.122 [−0.150, −0.094]	−0.234	<0.001
Partner F vs NB	1349	+0.031 [−0.002, +0.064]	+0.060	0.096
Same identity				
MM vs FF	450	+0.056 [+0.007, +0.102]	+0.107	0.057
MM vs NBNB*	450	+0.067 [+0.027, +0.109]	+0.139	0.007
FF vs NBNB	450	+0.011 [−0.038, +0.062]	+0.020	0.778
Pairing composition				
Same vs mixed (pooled)*	450	+0.088 [+0.056, +0.121]	+0.255	<0.001
agent=M held constant	450	−0.024 [−0.067, +0.018]	−0.053	0.350
agent=F held constant*	449	+0.203 [+0.147, +0.262]	+0.367	<0.001
Pairing structure				
Diff.- vs same-identity*	450	−0.045 [−0.071, −0.021]	−0.173	0.002
Binary-only vs NB-involved	450	+0.004 [−0.021, +0.028]	+0.013	0.786
Interaction (RQ3e)*	450	+0.085 [+0.036, +0.138]	+0.154	0.002

Table 7: All comparisons on the 45 held-out scenarios, five models. Δ is the mean paired difference in points on the 0–7 scale with a scenario-clustered bootstrap 95% CI; q is BH-adjusted across these sixteen tests, each a scenario-level Monte Carlo sign-flip permutation test (20,000 permutations); * marks $q < 0.05$.