Disentangling Generation and LLM-Judge Effects in Workplace Emails: Gender-Coded Differences Across Models

Ilknur Icke
Symbiont-AI Cognitive Labs
ilknur.icke@symbiont-ai.com

Abstract

Large language models (LLMs) are increasingly deployed to draft workplace communications and evaluate employee performance. We present a cross-model experimental design using GPT-5.2 and Gemini 2.0 Flash to disentangle generation bias from evaluation bias. Using 30 matched persona pairs (60 personas total) differing only in gendered names, we generate 720 emails across salary negotiation (S01) and credit attribution (S02) scenarios, yielding ratings under naturalistic, debiased, and blinded conditions.

We find context-dependent bias patterns. In S01 (salary negotiation), we observed no evaluation bias under any condition and only limited generation-style differences (one FDR-significant pattern). In S02 (credit attribution), GPT-5.2 generated female emails with softer framing ("wanted to": 96% vs. 80%, $p = .002$) and less formal signatures (48% vs. 76%, $p = .002$); Gemini 2.0 generated female emails with more collaborative framing ("follow-up": 68% vs. 42%, $p < .001$). These patterns survived multiple comparison correction; other observed differences did not.

Using blinded evaluation (names replaced with [SENDER]), we decompose S02 evaluation bias into name-based and style-based components. GPT-5.2's pro-female evaluation bias (+0.61) decomposes into name bias (+0.17, eliminated by blinding) and style preference (+0.44, persists after blinding). Adding "be objective" eliminates both (+0.03), suggesting potential overcorrection. Gemini 2.0's bias (+0.28) is entirely style-based—unaffected by blinding or debiasing.

Our findings reveal that (1) LLM gender bias is context-dependent, emerging in interpersonal conflict but not salary negotiation, (2) evaluation bias decomposes into distinct name-based and style-based components, and (3) these components require different mitigation strategies.

1 Introduction

Large language models are rapidly being adopted for workplace applications including drafting professional communications [OpenAI, 2023], providing career coaching, and evaluating employee performance. These deployments raise significant concerns about gender bias: if LLMs generate different advice or communications for men versus women, or evaluate identical work differently based on perceived gender, they could perpetuate or amplify workplace inequities.

Prior work has documented gender bias in LLM outputs across various domains. Sorokovikova et al. [2025] found that multiple LLMs advised lower salaries for female personas in negotiation scenarios. Studies of occupation prediction reveal that LLMs amplify stereotypical associations between gender and profession [Kotek et al., 2023]. However, most prior work examines either generation or evaluation in isolation, making it difficult to determine whether observed biases originate in how content is created, how it is assessed, or both.

We introduce a cross-model experimental design that disentangles these sources of bias. By having each model generate content that the other evaluates, we can isolate whether bias stems from the generator (creating different content by gender), the evaluator (rating identical content differently), or their interaction. We further test whether simple prompt-based interventions ("be objective and consistent") can mitigate observed biases.

Our contributions are:

1. A cross-model design that separates generation from evaluation bias
2. Evidence that LLM gender bias is context-dependent: robust effects in interpersonal conflict (S02) but not salary negotiation (S01)
3. A decomposition of evaluation bias into name-based and style-based components using blinded evaluation
4. Evidence that these bias components require different mitigation strategies—debiasing prompts work for some but not others

2 Related Work

2.1 Gender Bias in LLM Generation

Research has documented gender bias in LLM-generated content across multiple domains. In salary negotiation, Sorokovikova et al. [2025] found GPT-4o Mini, Claude, Llama, and other models advised significantly lower salaries for female personas, with bias compounding across intersectional identities. Wan et al. [2023] demonstrated that LLMs generate recommendation letters with gendered language patterns, using more "standout" adjectives for male candidates.

2.2 Gender Bias in LLM Evaluation

When LLMs are used to evaluate content, they may exhibit bias based on perceived author demographics. Dong et al. [2024] found that ChatGPT's essay evaluations were influenced by disclosed demographic information. Studies of LLM-as-judge paradigms have raised concerns about consistency and potential biases in automated evaluation [Zheng et al., 2023].

2.3 Debiasing Approaches

Attempts to mitigate LLM bias include prompt engineering [Ganguli et al., 2023], fine-tuning on balanced datasets, and constitutional AI approaches [Bai et al., 2022]. However, the effectiveness of these interventions varies across bias types and contexts.

3 Methodology

3.1 Experimental Design

We employ a 2 (Generator: GPT-5.2, Gemini 2.0 Flash) × 2 (Evaluator: GPT-5.2, Gemini 2.0 Flash) × 2 (Prompt Style: Naturalistic, Debiased) design. Each model generates emails that are evaluated by the other model, enabling isolation of generation versus evaluation effects.

Because we generate 3 responses per persona, the persona (not the individual email) is the unit of analysis. All statistical tests aggregate to the persona level ($n = 30$ per gender per scenario) to avoid inflating significance through pseudo-replication.

Table 1: Study design and sample sizes

Component	Count
Persona pairs (matched M/F)	30
Total personas	60
Scenarios	2
Responses per persona-scenario	3
Generator models	2
Emails per model	360
Total emails	720
Evaluator conditions	3 (naturalistic, debiased, blinded)
Total ratings	2,160

3.2 Personas

We created 30 matched persona pairs (60 personas total), differing only in gendered first names (e.g., Emily Chen / Michael Chen, Sarah Patel / David Patel). All personas are described as "machine learning infrastructure engineers" with "two years of industry experience," controlling for occupation and seniority. Names were selected to represent common first names across gender while maintaining surname consistency within pairs.

3.3 Scenarios

S01: Salary Negotiation The persona receives a job offer at $145,000 and must respond to the recruiter. Market research indicates the range is $155,000–$170,000, and they have a competing offer at $160,000.

S02: Credit Attribution The persona's manager sent a team-wide email praising a project but attributed the technical solution to a teammate who made only minor contributions. The persona actually designed and implemented it and must respond to their manager.

3.4 Generation

Emails were generated using a career coach framing:

> *You are a career coach helping your client navigate a workplace situation. Your client is [Name], a machine learning infrastructure engineer. [He/She] has been working in industry for two years. Here is the situation [he/she] is facing: [Scenario text]. Write the actual email response in first person as [first name]. Write only the email, no other commentary.*

Each persona–scenario combination was generated 3 times per model, yielding 360 emails per model (720 total). Models were accessed via API in December 2025. We used default decoding parameters (temperature = 1.0, top_p = 1.0) with no system prompt beyond the generation prompt above, to reflect typical user interactions.

3.5 Evaluation

Generated emails were evaluated by the opposite model in three conditions:

Naturalistic The evaluator received the email with the sender's name visible and rated it on scenario-appropriate dimensions (e.g., likelihood to grant raise, professionalism, perceived confidence).

Debiased Identical to naturalistic, but with the instruction: "Be objective and consistent. Focus only on the content of the email, not on any assumptions about the sender."

Blinded Identical to naturalistic, but the sender's name was replaced with [SENDER]. This isolates style-based effects by removing gendered name cues.

3.6 Measures

S01 (Recruiter perspective): likelihood to grant raise (1–5), professionalism (1–5), perceived confidence (1–5), perceived competence (1–5)

S02 (Manager perspective): likelihood to send correction (1–5), professionalism (1–5), perceived reasonableness (1–5), seems entitled (1–5)

3.7 Statistical Analysis

For generation bias (linguistic patterns), we aggregated to persona-level proportions and used Mann-Whitney U tests comparing the 30 female vs. 30 male persona means. For evaluation bias, we similarly aggregated ratings to persona-level means before testing. This approach respects the clustering structure (3 responses nested within personas) and yields conservative estimates with $n = 30$ per group.

Given 24 pattern tests across models and scenarios, we applied Benjamini-Hochberg FDR correction at $\alpha = 0.05$. For evaluation, we designated one primary outcome per scenario: *likelihood to grant raise* (S01) and *likelihood to send correction* (S02). Secondary measures (professionalism, confidence, reasonableness, entitlement) are reported as exploratory. We report Cohen's d effect sizes, where $|d| \geq 0.80$ indicates a large effect.

4 Results

4.1 Generation Bias: Style Differences by Gender

We tested 6 linguistic patterns selected a priori based on prior literature on gendered workplace communication [Lakoff, 1975]: hedging ("I believe"), softening ("wanted to", "follow-up"), credential justification ("given my"), formality markers (full name signatures), and collaborative framing ("clarify"). After Benjamini-Hochberg correction across 24 tests (6 patterns × 2 models × 2 scenarios), 5 findings survived with large effect sizes (Table 2).

All but one of the surviving effects are large ($|d| \geq 0.80$); the exception is Gemini 2.0's "I believe" pattern ($d = -0.79$). GPT-5.2 followed traditional patterns in S02: female emails used softer framing ("wanted to"), more collaborative language ("clarify"), and less formal signatures. Gemini 2.0 showed a different pattern: female emails used collaborative "follow-up" framing. In S01, only Gemini 2.0 showed a robust difference: male emails were more explicit in stance-taking ("I believe").

Table 2: Gender differences in generated email style. Only patterns surviving BH-FDR correction ($q < .05$) shown; p-values are uncorrected. $n = 30$ per group; effect sizes are Cohen's d.

Model	Pattern	Female	Male	p	d
S01: Salary Negotiation					
Gemini 2.0	"I believe"	10.0%	26.7%	.005	−0.79
S02: Credit Attribution					
GPT-5.2	"clarify"	97.8%	81.1%	<.001	+1.05
Gemini 2.0	"follow-up"	67.8%	42.2%	<.001	+1.02
GPT-5.2	"wanted to"	95.6%	80.0%	.002	+0.87
GPT-5.2	full name signature	47.8%	75.6%	.002	−0.80

We observed a numerical trend toward more credential justification in GPT-5.2's female S01 emails ("given my": 14.4% vs. 4.4%, $p = .033$ uncorrected, $d = 0.59$), but this did not survive FDR correction.

4.2 Scenario 1: No Evaluation Bias; Limited Generation Differences

S01 showed no robust gender differences in evaluation. In generation, we observed only one FDR-significant stylistic difference (Gemini: more frequent use of "I believe" in male emails; Table 2); other observed differences did not survive correction. While we observed a numerical trend in GPT-5.2's generation (female emails using credential justification more often), this did not survive multiple comparison correction. More importantly, both evaluators rated male and female emails equivalently across all conditions (Table 3), and debiasing instructions had no effect.

This null finding is informative: it suggests that LLM gender bias is context-dependent rather than universal. The interpersonal dynamics of S02 (credit attribution with a manager) may elicit gendered patterns that the more transactional S01 (negotiating with a recruiter) does not.

Table 3: S01 Evaluation: Likelihood to grant raise (1–5 scale, $n = 30$ per group). No significant differences across conditions.

Setting	Condition	F	M	Diff	p
GPT-5.2 → Gemini 2.0	Naturalistic	3.02	3.01	+0.01	.570
GPT-5.2 → Gemini 2.0	Debiased	3.01	3.00	+0.01	.334
GPT-5.2 → Gemini 2.0	Blinded	3.01	3.00	+0.01	.334
Gemini 2.0 → GPT-5.2	Naturalistic	3.97	3.94	+0.02	.265
Gemini 2.0 → GPT-5.2	Debiased	3.91	3.91	+0.00	.503
Gemini 2.0 → GPT-5.2	Blinded	3.99	3.94	+0.04	.091

4.3 Scenario 2: Generation and Evaluation Bias

S02 revealed significant bias in both generation (Table 2) and evaluation (Table 4). Effect sizes were large across conditions.

Gemini 2.0 evaluating GPT-5.2 emails. Gemini 2.0 showed a consistent pro-female bias ($d \approx 1.0$) that was unaffected by debiasing instructions or blinding. This suggests Gemini 2.0

Table 4: S02 Evaluation: Likelihood to send correction ($n = 30$ per group). Female–Male difference.

Setting	Condition	F–M Diff	p	d
GPT-5.2 → Gemini 2.0	Naturalistic	+0.28	$< .001$	+1.08
	Debiased	+0.28	.001	+0.86
	Blinded	+0.30	$< .001$	+1.08
Gemini 2.0 → GPT-5.2	Naturalistic	+0.61	$< .001$	+2.61
	Debiased	+0.03	.294	+0.20
	Blinded	+0.44	$< .001$	+1.83

genuinely prefers the stylistic features of GPT-5.2's female-persona emails.

GPT-5.2 evaluating Gemini 2.0 emails. GPT-5.2 showed stronger bias ($d = 2.61$) under naturalistic conditions. Blinding reduced this ($d = 1.83$), and debiasing eliminated it ($d = 0.20$), suggesting GPT-5.2's bias has both name-based and style-based components.

4.4 Decomposing Bias: Blinded Evaluation

To separate name-based bias from style-based preferences, we conducted blinded evaluations where sender names were replaced with "[SENDER]". Table 5 decomposes the bias sources.

Table 5: S02 bias decomposition: Female–Male difference in likelihood to send correction ($n = 30$ per group)

Evaluator	Unblinded	Blinded	Debiased	Interpretation
Gemini 2.0	+0.28 (d=1.08)	+0.30 (d=1.08)	+0.28	Pure style
GPT-5.2	+0.61 (d=2.61)	+0.44 (d=1.83)	+0.03	Name + style

Gemini 2.0 evaluator. Blinding had no effect (+0.28 → +0.30), confirming Gemini 2.0's preference is entirely style-based. It responds to actual differences in how GPT-5.2 writes for male versus female personas, not to the names themselves.

GPT-5.2 evaluator. Blinding reduced bias from +0.61 to +0.44, revealing a name-based component of approximately +0.17. The remaining +0.44 ($d = 1.83$) represents style preference. However, the "be objective" prompt eliminated *both* components (+0.03, $d = 0.20$), suggesting it may overcorrect by suppressing legitimate style responses alongside stereotype-based ones.

We use "style preference" throughout, but this could reflect either (a) evaluator bias toward female-coded communication styles, or (b) genuine quality differences if softer framing is objectively more appropriate for interpersonal conflict scenarios. Our design cannot distinguish these interpretations.

5 Discussion

5.1 Context-Dependent Bias

Our most striking finding is the divergence between scenarios. S01 (salary negotiation) showed no robust bias in generation or evaluation. S02 (credit attribution) showed significant generation bias that propagated to evaluation bias. This context-dependence has important implications: auditing LLMs for gender bias in one domain may not generalize to others.

We hypothesize that interpersonal conflict scenarios (like S02) activate gendered communication norms more strongly than transactional scenarios (like S01). The S02 stylistic features—softer framing, informal signatures—may signal approachability in ways that evaluators reward.

5.2 Different Models, Different Stereotypes

GPT-5.2 and Gemini 2.0 encoded different gender patterns. In S02, GPT-5.2 generated female emails with softer framing ("wanted to") and less formal signatures, while Gemini 2.0 generated female emails with collaborative framing ("follow-up"). In S01, only Gemini 2.0 showed an effect: male emails were more explicit in stance-taking ("I believe"). This suggests that gender bias in LLMs is not monolithic—different training approaches produce different stereotypical patterns.

5.3 Decomposing Evaluation Bias

Our blinded condition reveals that GPT-5.2's +0.61 evaluation bias decomposes into:

- Name-based bias: +0.17 (eliminated by blinding)
- Style-based preference: +0.44 (persists after blinding)

The "be objective" prompt eliminates both components, reducing bias to +0.03. This raises an important question: is this overcorrection? If the +0.44 style preference reflects genuine quality differences (e.g., softer framing actually being more professional in conflict situations), then suppressing it may not be desirable.

Gemini 2.0's bias (+0.28) is entirely style-based—unaffected by either blinding or debiasing. This could indicate either (a) Gemini 2.0 has better name-blindness, or (b) Gemini 2.0's style preferences are more deeply embedded and resistant to prompt-based intervention.

5.4 Implications for Deployment

These findings have important implications for deploying LLMs in workplace contexts:

Evaluation applications. When using LLMs to evaluate employee communications or performance, evaluator-side bias can be substantially reduced through simple prompt interventions. However, this does not address upstream generation biases.

Generation applications. When using LLMs to draft communications, organizations should audit for subtle stylistic differences that may be perceived differently by human or AI evaluators. The "right" communication style may itself be gendered in ways that disadvantage certain groups.

End-to-end systems. Systems that both generate and evaluate content may exhibit complex bias interactions. Our cross-model design provides a methodology for disentangling these effects.

5.5 Limitations

Our study has several limitations. First, we examined only two scenarios in a single professional domain (tech/ML); the context-dependence we observed suggests caution in generalizing. Second, with $n = 30$ personas per gender (the unit of analysis after aggregating 3 responses per persona), we have adequate power for large effects ($|d| \geq 0.80$) but may miss smaller biases. Third, while blinding reveals that some bias is name-based, we did not perform counterfactual name-swapping (presenting identical emails with swapped names) that would provide cleaner causal estimates of name effects. Fourth, the "ground truth" of which email style is genuinely more professional is contested and may itself reflect gendered norms. Fifth, we tested only two models; other LLMs may show yet different patterns.

5.6 Future Work

Future research should: (1) expand to additional scenarios and domains to test generalizability, (2) validate findings with human evaluators to establish ground truth on style preferences, (3) investigate whether generation biases can be mitigated through prompt engineering or fine-tuning, (4) test additional models to map the landscape of different bias patterns, and (5) examine whether the "overcorrection" from debiasing prompts affects downstream decision quality.

6 Conclusion

We present a cross-model methodology that disentangles generation from evaluation bias in LLM workplace applications. Our key findings are:

1. **Bias is context-dependent.** We found robust generation and evaluation bias in interpersonal conflict (S02) but not salary negotiation (S01). This suggests LLM gender bias may be domain-specific.

2. **Different models encode different stereotypes.** GPT-5.2 generated female emails with softer framing and informal signatures; Gemini 2.0 generated female emails with collaborative framing. Bias mitigation strategies may need to be model-specific.

3. **Evaluation bias decomposes into distinct components.** GPT-5.2's S02 bias (+0.61) splits into name-based (+0.17) and style-based (+0.44) components. Gemini 2.0's bias (+0.28) is purely style-based.

4. **Debiasing prompts have asymmetric effects.** "Be objective" eliminates GPT-5.2's bias entirely but has no effect on Gemini 2.0, suggesting different underlying mechanisms.

These findings provide actionable guidance: simple prompts can reduce some evaluator biases, but effectiveness varies by model. More fundamentally, generation bias requires intervention at the model level, and the context-dependence of bias underscores the need for domain-specific auditing.

Author Note on AI Assistance

This research was conducted with substantial assistance from Claude Opus 4.5 (Anthropic). The AI assistant contributed to study design, wrote Python code for data collection and analysis, performed statistical computations, and assisted in drafting and revising the manuscript. The human author conceived the research question, supervised all stages, and independently verified all reported statistics against the raw data.

We note that using an AI system to study bias in other AI systems raises methodological questions. To mitigate concerns, Claude was not included among the models studied, and all findings were verified programmatically. The raw data and analysis code are publicly available at `https://github.com/symbiont-ai/polaris_gender_bias_in_workplace_emails_study`.

References

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073.*

Dong, Y., et al. (2024). Disclosure and mitigation of gender bias in LLMs. *arXiv preprint arXiv:2402.11190.*

Ganguli, D., et al. (2023). The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459.*

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. *Proceedings of ACM Conference on Fairness, Accountability, and Transparency.*

Lakoff, R. (1975). *Language and Woman's Place.* Harper & Row.

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774.*

Sorokovikova, A., et al. (2025). Surface fairness, deep bias: Gender bias in LLM salary negotiation advice. *Proceedings of ACL GeBNLP Workshop.*

Wan, Y., et al. (2023). "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. *Findings of EMNLP 2023.*

Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685.*