The Same Email, Signed Differently: Testing Negotiation Bias and Recommendation Stability in LLMs

Jasmin Heierli and Alexandre de Spindler
Zurich University of Applied Sciences, Winterthur, Switzerland
heej@zhaw.ch, desa@zhaw.ch

Abstract

Large language models (LLMs) are increasingly mediating hiring communications, serving both as tools for applicants to draft negotiation emails and as systems for recruiters to evaluate them. Such mediation risks introducing variability and hidden dependencies into high-stakes outcomes such as salary expectations and hiring decisions. This paper investigates how the outcomes of these bidirectional interactions, specifically salary expectations and hiring recommendations, are influenced by gender signaling, model identity, and language context. We study this setting with a two-stage analysis across models and English/German contexts, using 2,880 Stage 1 observations and 1,441 paired Stage 2 evaluations. We find no strong or consistent pooled gender effects. Instead, model differences dominate, while scalar ratings are stable and categorical recommendations are less robust.

1 Introduction

Large language models are rapidly becoming embedded in hiring workflows, not only as tools for evaluating candidates[1] but also as assistants that help applicants draft negotiation messages[2](Chaturvedi and Chaturvedi, 2025). Rather than supporting isolated decisions, models now shape the full interaction loop, generating negotiation content and evaluating it, raising new questions about how such systems influence outcomes in hiring contexts (Geiger et al., 2025). This raises a concern: outcomes may reflect interacting model behaviors that are difficult to anticipate or control rather than transparent human judgment.

Most prior work studies LLMs in hiring as isolated decision-makers, focusing on static tasks such as ranking, screening, or salary assignment (Chaturvedi and Chaturvedi, 2025; Rozado, 2026; Li et al., 2025). This perspective overlooks a key aspect of real-world hiring: negotiation as an interactive process, where outcomes depend on how candidates articulate their expectations (Mazei et al., 2015). Although recent work has begun to explore more personalized uses of LLMs, such as career advice (Geiger et al., 2025; Eloundou et al., 2025), it is unclear how outcomes are shaped when LLMs mediate both sides of such a high-stakes interaction.

Recent work shows that LLM behavior in hiring can vary substantially across model providers and versions (Geiger et al., 2025). At the same time, bias patterns have been shown to vary between languages, for example, between English and German prompts (Ikae and Kurpicz-Briki, 2025). We are building on this line of studies using minimal identity perturbations, such as name-based gender swaps, demonstrating that even small changes in identity signals can affect model outputs (Rozado, 2026; Li et al., 2025). In this paper, we combine these perspectives by examining how model choice, language context, and decision format interact within a fully bidirectional, LLM-mediated negotiation setting.

We introduce a reproducible two-stage experimental pipeline in which models from three providers are used to generate salary negotiation emails and to evaluate them[3]. In Stage 1, we extract salary range midpoints and widths from generated negotiation responses (Geiger et al., 2025). In Stage 2, the same responses are evaluated under paired identity perturbations, isolating evaluation differences while holding content constant (Rozado, 2026; Li et al., 2025). Experiments are conducted in English (UK) and German

[1] https://www.herohunt.ai/blog/ai-adoption-in-recruiting-2025-year-in-review

[2] https://www.interviewpal.com/free-tools/salary-negotiation-email-generator

[3] https://github.com/zhaw-iwi/swisstext-gender-in-job-ads_pub

(Switzerland) across two professional roles.

We examine whether LLMs exhibit gender differences in generated salary expectations, whether identical negotiation emails are evaluated differently under minimal identity changes, and how model, language, and decision format shape these patterns. We contribute a two-stage generation–evaluation pipeline, a controlled English/German comparison, and a systematic analysis of model, language, and decision-format effects in LLM-mediated hiring interactions.

2 Method

2.1 Experimental Design

We used a two-stage pipeline built on controlled synthetic stimuli. In Stage 1, a model generated a short salary-negotiation response to a manually created recruiter template. In Stage 2, that response was evaluated twice under minimal identity perturbation: only the signature name was swapped, while the email body remained unchanged. This isolates generation from evaluation effects.

The design varies the language context, role, candidate name, and model. We used English for a UK market prompt and German for a Swiss-German market prompt. We focused on two roles, *Senior Data Analyst* and *Senior Project Manager*, each requiring 5–7 years of experience.

2.2 Stimuli and Prompt Setup

All Stage 1 prompts used the same manually created recruiter-message template for each language. The full text of these templates and the specific system instructions for each provider are provided in our repository. This template served as the stimulus, where the "recruiter" asked for a response with a realistic annual salary range, and the model was instructed to sign with the candidate's name. The task structure was constant between models; language, role, market, and candidate name varied.

Gender was signaled only through names. For each language context, we used four female and four male names. In Stage 2, each name had a fixed opposite-gender partner, and only the signature name was changed. The fixed mapping reduced variance and preserved a tightly controlled paired comparison.

2.3 Models

We included one current model from three major providers at execution time: OpenAI, Anthropic, and Google. The runs used `gpt-5.4` (resolved in the metadata as `gpt-5.4-2026-03-05`) (OpenAI, 2026), `claude-sonnet-4-6` (Anthropic, 2026), and `gemini-3.1-flash-lite-preview` (Google, 2026). We used identical prompt templates and retained model default generation settings unless required by the API. Model identifiers and metadata were recorded for reproducibility and are stored in our GitHub repository.

2.4 Stage 1: Negotiation Generation Stage

In Stage 1, each model was instructed to write a brief professional response to a recruiter message and include a realistic annual salary range. The primary results were salary midpoint and range width, derived from the minimum and maximum values extracted. Extraction was performed using rule-based parsing with manual validation by the authors. The validation showed around 85% extraction accuracy, and errors were fixed manually to enable further processing of the data.

2.5 Stage 2: Synthetic Evaluation Stage

Stage 2 tested whether identical negotiation content is judged differently after a name swap. For each Stage 1 email, we created a paired variant by replacing only the candidate's signature name with its mapped opposite-gender counterpart. The models assessed the professionalism, likeability/warmth, and appropriateness of the salary request (1–7), plus a categorical recommendation (*Proceed*, *Proceed with caution*, *Do not proceed*). These assessments were conducted through separate inference runs for each variant to ensure no context overlap. This isolates within-model evaluation before introducing cross-model variability.

In the main analysis, the evaluation was performed by the same model: each model judged texts generated by that same model. We took this decision to reduce the complexity of the design in this preliminary study.

2.6 Analysis

For Stage 1, we compared midpoint and width under female versus male signaling while modeling language, role, and model. For Stage 2, we computed paired deltas for the three scalar ratings and the recommendation shift rate for the categorical decision. Inferential analyses used paired tests and regression as appropriate, with bootstrap confidence intervals. We applied the Holm correction within each outcome family.

Two design choices matter for interpretation. First, the English and German conditions should be read as language-context comparisons, not as national pay estimates. Second, the study prioritizes internal control over demographic realism: names serve as minimal identity cue and the fixed prompt suppresses real-world covariates.

3 Results

We analyze 2,880 Stage 1 observations and 1,441 paired Stage 2 observations. We find no consistent gender effects in generated salary targets or mean scalar evaluation scores. Instead, the clearest patterns are model differences in Stage 1 and decision-format differences in Stage 2: scalar ratings are stable on average under name swaps, while categorical recommendations are less robust.

3.1 Stage 1: Small Gender Effects

Pooling across models, roles, and prompts, the difference between the female and the male midpoint was GBP -128.472 in English and CHF -97.222 in German. Width differences were also small and inconsistent (GBP 20.833 in English and CHF 375.000 in German). In the pooled model, the gender effect was not significant for midpoint ($p = 0.565$, Holm $p = 0.860$) or width ($p = 0.430$, Holm $p = 0.860$).

3.2 Stage 1: Model Effects Dominate

Model differences were much larger than pooled gender gaps. In both language contexts, Claude produced the lowest midpoint means, GPT the highest, and Gemini lay between them. In English, model midpoints means ranged from roughly GBP 76.5k (Claude) to GBP 85.0k (GPT); in German, they ranged from roughly CHF 127.3k (Claude) to CHF 136.2k (GPT). Width differences also varied by model, with Gemini showing especially narrow ranges in the German conditions.

Controlled models support this pattern: relative to Claude, model terms for Gemini and GPT were significant for midpoint (both $p < 0.001$). Figure 1 shows that model variation exceeds gender gaps.

3.3 Stage 2: Stable Scalar Ratings

In Stage 2, the average effects of name-swapping on the three scalar evaluation dimensions were close to zero. Across all observation pairs, mean deltas were 0.014 for professionalism, 0.008 for likeability, and 0.024 for appropriateness. None of these changes were significant: professionalism $p = 0.792$, likeability $p = 0.829$, and appropriateness $p = 0.635$ (all Holm-corrected $p = 1.000$). Scalar evaluations are therefore largely invariant to the name swap at the aggregate level.

3.4 Stage 2: Categorical Recommendations

The categorical recommendation outcome behaved differently. The recommendation class changed in 16.3% of the paired cases, and this shift rate was significant ($p = 0.009$, Holm $p = 0.036$). Thus, even though the mean scores for professionalism, likeability, and appropriateness remained near zero, the final decision category was more sensitive to minimal identity disturbance.

Recommendation instability was also model-dependent. Gemini had a shift rate of 2.9%, compared to 21.9% for Claude and 24.1% for GPT. In the controlled provider model, again using Claude as the omitted reference category, Gemini showed a strongly negative log-odds coefficient of recommendation-shift (-2.236, $p < 0.001$), whereas GPT did not differ significantly (0.127, $p = 0.407$). Figure 2 shows both the quasi-zero scalar deltas and the much wider spread in recommendation-shift rates.

3.5 Extreme Cases and Variation

The quasi-zero scalar averages conceal a tail of extreme reversals. 225 paired cases showed large simultaneous movement in all three scalar ratings ($|\Delta| \geq 3$ for professionalism, likeability, and appropriateness). These cases were concentrated almost entirely in GPT evaluations (224/225), with more in English (128) than in German (96). Directionally, these flips were not fully symmetric: the female-signaled version scored lower in 126 cases and higher in 99 cases.

Language differences were present, but mainly through model-language combinations rather than as a single uniform cross-lingual effect. Instability varied by model-language combination: Claude was less stable in German than English, GPT showed the opposite pattern, and Gemini remained comparatively stable in both. These patterns suggest that language effects are less consistent than model effects.

4 Discussion & Conclusion

Our findings suggest that LLM behavior in hiring-related tasks is not well captured by a single notion of bias. Although we do not observe strong or consistent pooled gender effects in generated salary

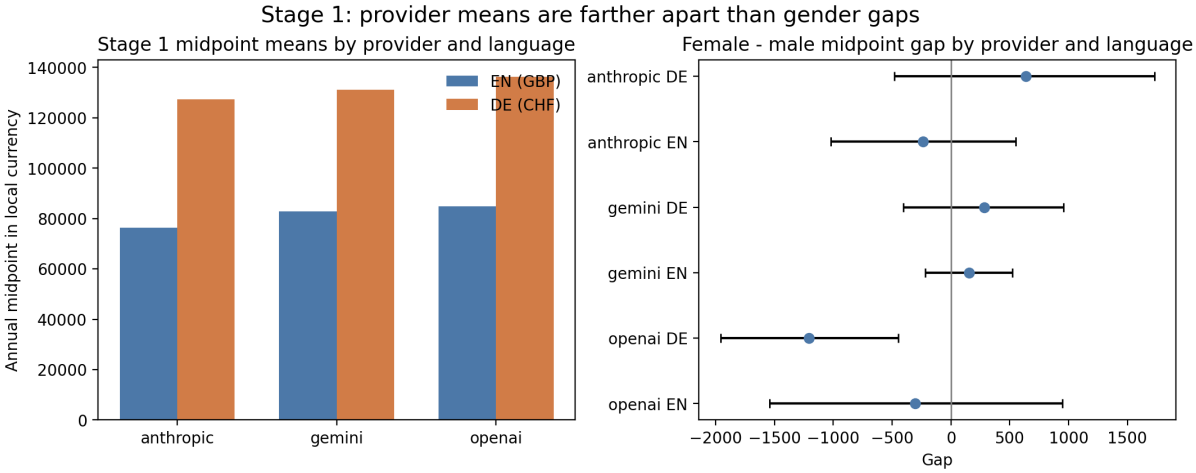

Figure 1: Stage 1 results. Left: model-level midpoint means by language context, shown in local currency (GBP for English prompts and CHF for German prompts). Right: female-minus-male midpoint gaps by model and language.

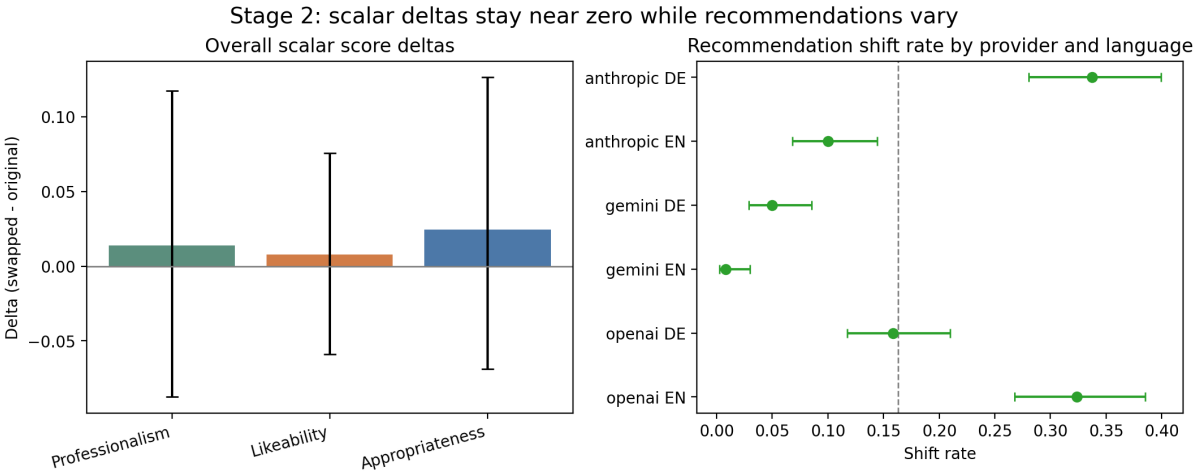

Figure 2: Stage 2 results. Left: overall score deltas for professionalism, likeability, and appropriateness. Right: recommendation-shift rates vary substantially across model-language conditions, with the dashed line marking the overall shift rate of 16.3%.

ranges or mean scalar evaluation scores, model outputs are not simply neutral or stable. Instead, they exhibit structured instability, with behavior depending on model, language context, and decision format. This implies vendor-dependent outcomes: salary expectations differ across models, reflecting system choice rather than candidate characteristics. Decision inconsistency emerges, as identical content can lead to different recommendation categories, and hidden instability arises, with scalar ratings appearing stable while final decisions remain sensitive to small perturbations.

Model differences dominate generation, making model choice a primary driver of salary expectations. In evaluation, scalar ratings remain stable under minimal identity perturbations, but this masks a tail of extreme reversals. In contrast, categorical recommendations are substantially less stable, suggesting that output formats differ in operational reliability.

These results highlight the importance of viewing LLM use in hiring as an interaction between generation and evaluation, rather than isolated tasks. When models are used on both sides, differences in model, language, or output format may propagate, amplifying variability. Neither scalar nor categorical outputs should be treated as reliable for high-stakes decisions. Although scalar ratings appear more stable, model behavior is configuration-dependent. This highlights the need for reproducible, context-sensitive evaluation across models, languages, and decision formats.

Limitations

Gender was operationalized through names, which provides a minimal and naturalistic cue but does not isolate gender alone: names may also carry class, ethnicity, age, or other sociocultural associations. The observed effects should therefore be interpreted as responses to name-based gender signaling under controlled conditions rather than as clean estimates of gender bias in isolation.

The cross-lingual comparison is also a language-market comparison. English prompts were paired with a UK market context and German prompts with a Swiss-German market context because realistic salary negotiation depends on local compensation norms. This improves plausibility, but it means the design cannot disentangle linguistic effects from economic and cultural expectations. The cross-language results are therefore contextual rather than causal.

The design scope is narrow. We study only two relatively senior white-collar roles, so the findings may not generalize to other occupations, industries, or seniority levels. Likewise, the factorial design is better suited to detecting the main contrasts of interest than very small higher-order interactions.

Stage 2 uses same-model evaluation of model-generated text. This keeps the comparison tightly controlled but limits claims about whether the same patterns would hold under cross-model evaluation. More generally, the task remains a simplified proxy for real negotiation, relying entirely on researcher-authored stimuli and model-generated responses without human participants. Actual hiring communication is often multi-turn, includes richer identity cues, and involves strategic adaptation by both sides. Furthermore, this study focuses on the technical behavior of interacting models; however, the practical displacement of transparent, criteria-based human judgment by automated LLM-to-LLM interactions raises significant ethical questions regarding accountability and hidden dependencies in hiring.

Finally, the findings are tied to specific model snapshots. Commercial systems are updated frequently, and even small version changes may alter both generation style and evaluation behavior. The results should therefore be read as evidence about the tested systems at execution time, not as stable properties of future versions.

References

Anthropic. 2026. Introducing claude sonnet 4.6. Accessed: 2026-03-10.

Sugat Chaturvedi and Rochana Chaturvedi. 2025. Who gets the callback? generative ai and gender bias. *arXiv preprint arXiv:2504.21400*.

Tyna Eloundou, Alex Beutel, David Robinson, Keren Gu, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2025. First-person fairness in chatbots. In *International Conference on Learning Representations*, pages 58234–58268.

R. Stuart Geiger, Finn O'Sullivan, Edward Wang, and Jessica Lo. 2025. Asking an ai for salary negotiation advice is a matter of concern: Controlled experimental perturbation of chatgpt for protected and non-protected group discrimination on a contextual task with no clear ground truth answers. *PLOS ONE*, 20(2).

Google. 2026. Gemini 3.1 flash-lite (vorabversion). Google Cloud Documentation. Accessed: 2026-03-10.

Chidubem Ikae and Marlena Kurpicz-Briki. 2025. Measuring bias in german prompts to gpt models using contact hypothesis. In *Proceedings of the 2nd Workshop on AI Bias: Measurements, Mitigation, Explanation Strategies (AIMMES 2025)*.

Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong, and Jindong Wang. 2025. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *arXiv preprint arXiv:2502.04419*.

Jens Mazei, Joachim Hüffmeier, Philipp A. Freund, Alice F. Stuhlmacher, Lisa Bilke, and Guido Hertel. 2015. A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychological Bulletin*, 141(1):85–104.

OpenAI. 2026. Entdecke gpt-5.4. Accessed: 2026-03-10.

David Rozado. 2026. Gender and positional biases in llm-based hiring decisions: evidence from comparative cv/résumé evaluations. *PeerJ Computer Science*, 12:e3628.