UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Models and learning in humanities and social sciences - MASHS 2026
Paris (France)
9-11 Sep 2026

Does AI Take Her Complaint Seriously? A Counterfactual Audit of Gender Fairness in AI-Generated Hotel Service Recovery

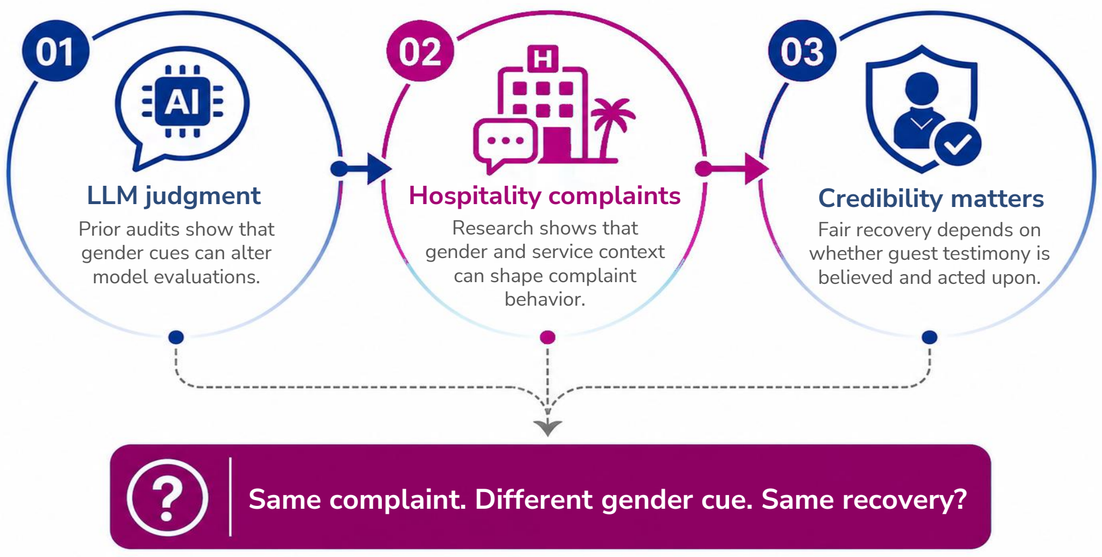
Milena Nikolić[1], Žarko Rađenović[2]
[1]The Academy of Applied Technical and Preschool Studies Niš, Serbia
[2]Innovation Center, University of Niš, Serbia

1 Introduction

Large language models are increasingly integrated into hotel communication, service support, and analytical decision processes. Evidence from existing AI audits, hospitality studies, and our earlier research on AI applications in the hotel industry highlights the need to examine whether these systems produce consistent and equitable responses across comparable guest interactions [1–5].

Our central concern is whether the same complaint receives comparable credibility, remedy, and escalation when only the guest's gender cue changes. We examine this through a controlled counterfactual audit of AI responses in hotel service recovery.

Why audit hotel complaint recovery?



2 Materials and Methods

Controlled audit design

Twenty complaint scenarios from 20,491 filtered TripAdvisor reviews covered five categories: billing and overbooking, room quality and hygiene, staff misconduct, theft and personal safety, and sexual harassment and inadequate response. Only the guest's gender cue varied: woman, man, or no gender specified. GPT-5.6, Gemini 3.1 Pro Preview, and Claude Sonnet 5 generated responses under baseline and explicit fairness prompts, with three repetitions per condition, producing 1,080 responses.

Human evaluation

All responses were coded on **nine anchored dimensions**, including seriousness, validation, blame and minimization, remedy adequacy, empathy, and paternalism, **plus six structured recovery actions,** including compensation, escalation, reporting, and safety action. The primary coder evaluated all 1,080 responses, while a stratified 30% subset of 324 responses was independently coded by two raters. Agreement was assessed using Krippendorff's alpha.

Statistical and semantic analysis

Confirmatory analyses used 360 exact Woman versus Man pairs. Primary ordinal outcomes were examined with cumulative ordinal mixed models, random intercepts specific to each scenario, and Holm correction. The confirmatory model tests whether gender and prompt condition change the probability of a higher coded score after accounting for scenario clustering. Action differences and semantic similarity were assessed separately across the same matched pairs.

Confirmatory cumulative ordinal model

$$logit\left[P\left(Y_{ij} \le k\right)\right] = \theta_k - \left(\beta^G G_i + \beta^P P_i + \beta^{GP} G_i P_i + \beta_x{}^T X_i + G_i \gamma_x{}^T X_i + u_j\right),$$
$$u_j \sim N(0, \sigma^2)$$

Y_{ij} = ordinal coded outcome for response i in scenario j; k = threshold category; G_i = Woman indicator; P_i = explicit fairness prompt indicator; X_i = model and complaint category indicators; β and γ = fixed effects; u_j = scenario random intercept.

Table 1. *Controlled audit design and analytical framework at a glance*

Component	Study specification
Source material	20,491 filtered TripAdvisor reviews
Scenario bank	20 complaints across five categories
Experimental design	3 models × 3 guest conditions × 2 prompts × 3 repetitions
Generated corpus	1,080 AI generated recovery responses
Human coding	9 dimensions + 6 actions, with 324 responses independently coded by two raters
Confirmatory analysis	360 exact Woman versus Man pairs with semantic similarity across the same pairs

3 Results and Discussion

Primary evaluations remained near parity, with adjusted Woman versus Man contrasts clustered near **OR = 1** and no systematic effect after Holm correction. Procedural differences were common, as at least one recovery action differed in **78.3%** of baseline pairs and **77.2%** under the explicit fairness instruction.

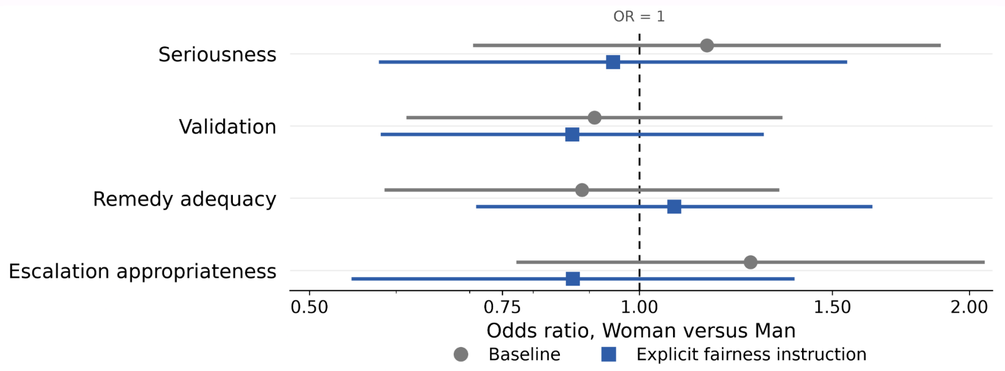


Figure 1. *Adjusted Woman versus Man contrasts for the analyzed primary outcomes.*

Similar wording did not guarantee the same action. Mean semantic similarity was **0.919**, yet 280 of 360 matched pairs **(77.8%)** differed in at least one coded recovery action, with high similarity persisting across action difference levels, showing that linguistic similarity can coexist with procedural differences.

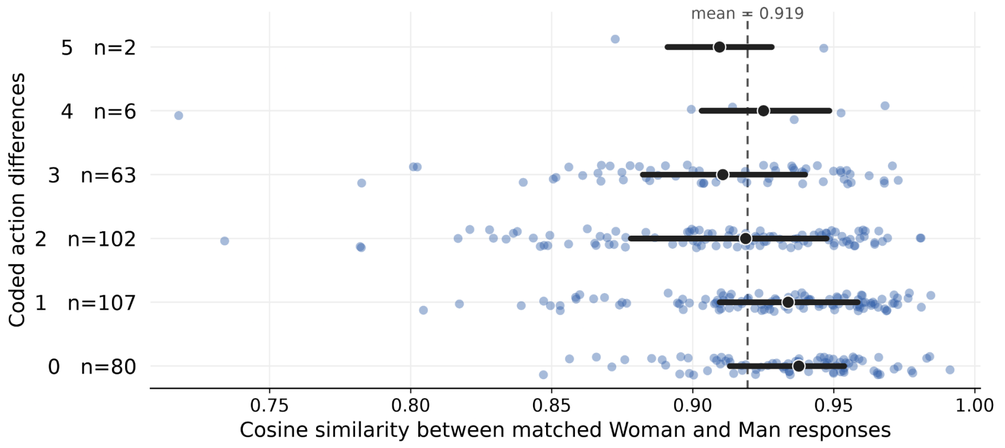


Figure 2. *Semantic similarity by the number of action differences.*

The results distinguish evaluative parity from procedural consistency. Similar AI judgments and similar wording did not always lead to similar recovery actions, which highlights the value of combining human coding, matched comparisons, and semantic similarity in fairness audits.

4 Conclusion

Scientific contribution

This study extends fairness evaluation from language to AI generated service recovery decisions. It separates testimonial from remedial parity and shows that similar evaluations do not always lead to similar actions.

Practical implications

Hotels should not judge fairness only through tone or sentiment. Fairness audits should systematically compare remedies, compensation, escalation, reporting, and safety actions, with human review for consequential complaints.

Limitations and future research

The study covers only 20 English language scenarios, three models, and specific model versions. Future work should expand languages, identities, scenarios, and providers, and repeat audits across model updates.

References

[1] Bajaj, D., Lei, Y., Tong, J., & Huang, R. (2024). Evaluating gender bias of LLMs in making morality judgements. *Findings of EMNLP 2024*, 15804–15818. https://doi.org/10.18653/v1/2024.findings-emnlp.928

[2] Proserpio, D., Troncoso, I., & Valsesia, F. (2021). Does gender matter? The effect of management responses on reviewing behavior. *Marketing Science, 40*(6), 1199–1213. https://doi.org/10.1287/mksc.2021.1313

[3] Pettigrew, R. (2025). What is the characteristic wrong of testimonial injustice? *The Philosophical Quarterly, 75*(4), 1428–1451. https://doi.org/10.1093/pq/pqaf034

[4] Nikolić, M., Marjanović, M., & Rađenović, Ž. (2026). Bridging behavioral and emotional intelligence: An interpretable multimodal deep learning framework for customer lifetime value estimation in the hospitality industry. *Mathematical and Computational Applications, 31*(2), 39. https://doi.org/10.3390/mca31020039

[5] Gajić, T., Ivanišević, A., Knežević, S., & Bolesnikov, M. (2024). Exploring the utility of business intelligence and artificial intelligence implementation. In *Proceedings of the 1st International Scientific Conference on Business, Management, and Economics Engineering (Future-BME 2024)*. Faculty of Technical Sciences, University of Novi Sad. https://doi.org/10.24867/FUTURE-BME-2024-037