TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY-NC-ND

Citation (APA)
Lagos Rojas, C., Genç, H. U., Bozzon, A., & Colombo, S. (2026). "are Compliments Bad Now?": Comparing LLMs and Human Interpretations of Gender Microaggressions in the Workplace. In N. Oliver, D. A. Shamma, H. Candello, P. Cesar, P. Lopes, A. Bozzon, T. Kosch, V. Liao, X. Ma, V. Artizzu, F. Draxler, G. Lopez, A. V. Reinschluessel, X. Tong, & P. O. Toups Dugas (Eds.), *CHI 2026 - Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* Article 1519 (Conference on Human Factors in Computing Systems - Proceedings). Association for Computing Machinery (ACM). https://doi.org/10.1145/3772318.3790436

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

“Are Compliments Bad Now?”: Comparing LLMs and Human Interpretations of Gender Microaggressions in the Workplace

Catalina Lagos Rojas*
Knowledge and Intelligence Design
Delft University of Technology
Delft, Netherlands
c.p.lagosrojas@tudelft.nl

Hüseyin Uğur Genç
Knowledge and Intelligence Design
Delft University of Technology
Delft, Netherlands
u.genc@tudelft.nl

Alessandro Bozzon
Knowledge and Intelligence Design
Delft University of Technology
Delft, Netherlands
a.bozzon@tudelft.nl

Sara Colombo
Knowledge and Intelligence Design
Delft University of Technology
Delft, Netherlands
sara.colombo@tudelft.nl

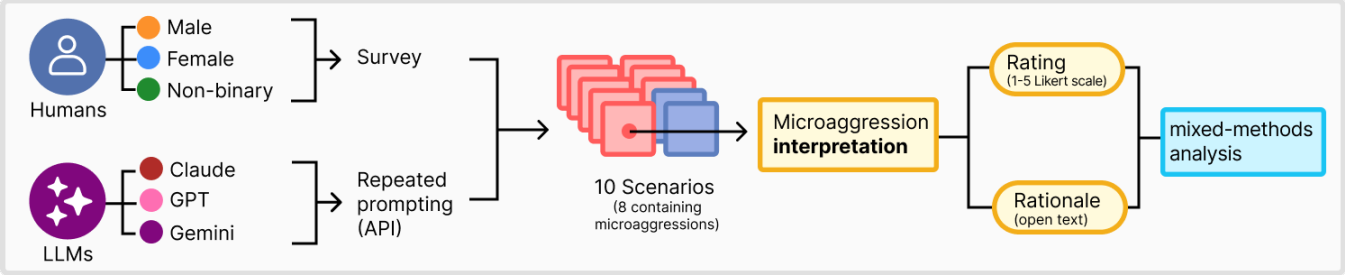


Figure 1: Study pipeline for comparing human and LLM interpretations of microaggressions. Humans (via surveys) and LLMs (via repeated prompting) evaluated ten dialogue scenarios, providing ratings and rationales for perceived microaggressions. Mixed-methods analysis examined similarities and differences across groups.

Abstract

Gender microaggressions are subtle yet persistent forms of discrimination in workplace interactions. While LLMs can detect them in written texts, it remains poorly understood how their interpretations align or diverge from human perspectives and experiences. We present a mixed-method study comparing how LLMs and humans differing in gender identity and lived experience, interpret gender microaggressions in the workplace. Using short dialogues adapted from real-world accounts, we asked 141 participants to rate the likelihood that a scenario contains a microaggression and provide a rationale for their answers. The same tasks were completed by 7 different LLM models. Our analysis reveals significant differences in how humans and LLMs interpret microaggressions, captured in both ratings and rationales, and more interestingly, the effect of gender and lived experience on human interpretations. These findings highlight the need for systems detecting microaggressions to embrace interpretive plurality, and support reflection and awareness while accounting for ambiguity.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790436

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Large Language Model, Human-AI alignment, Microaggressions, Gender

ACM Reference Format:
Catalina Lagos Rojas, Hüseyin Uğur Genç, Alessandro Bozzon, and Sara Colombo. 2026. “Are Compliments Bad Now?”: Comparing LLMs and Human Interpretations of Gender Microaggressions in the Workplace. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3790436

1 Introduction

Gender bias remains widespread and persistent in professional environments [77]. While overt sexism may have declined, discriminatory practices have not decreased [53]; rather, they manifest in subtler forms such as microaggressions [20, 24, 48, 78, 84]. These can range from interruptions and dismissive remarks to being overlooked for promotions and more structural manifestations [49], with a cumulative effect that can undermine women's professional identities, leadership trajectories, and sense of belonging [48, 62]. While these effects have been most commonly studied in women, they also affect other groups, such as non-binary individuals [25].

Microaggressions also have organizational consequences: people who experience them report higher levels of burnout and are more likely to consider leaving their organizations [2, 24, 53], thus creating a loss of key talent and leadership potential [60]. This has motivated interest in automated tools to help identify and address workplace microaggressions.

While microaggressions can target different aspects of a person's identities, we focus on gender-based microaggressions because they are the most common form reported in workplaces, strongly linked to burnout and attrition, and central to ongoing equity and inclusion efforts [53].

Yet microaggressions are especially challenging to address, as they are often perpetrated unconsciously, their manifestations are ambiguous, and not everyone agrees on whether they are harmful [59]. These disagreements highlight that interpretation is socially situated, meaning that it depends on context, positionality, and lived experience rather than objective classifications [69].

A growing body of work has approached microaggression detection as a text classification problem. Researchers have deployed machine learning (ML) systems to detect sexism, toxic speech, as well as microaggressions, in online contexts [3, 8, 28, 37, 38, 82]. While reported detection accuracy can be high, these pipelines face consistent limitations associated with microaggressions' ambiguous and context-dependent nature [3, 79]. Annotators that help create these systems provide labels for what "counts" as microaggression, but recent work demonstrates that their social identity and context significantly shape the labels they provide [27, 38]. This challenges the assumption that microaggressions can be objectively classified using a "gold standard" [18, 27].

More recently, researchers have begun exploring large language models (LLMs) for microaggression detection, as they can incorporate contextual details and generate explanations for their judgments [4]. In the workplace, LLMs are increasingly integrated into work flows to support decision-making [7, 31]. Yet, recent work shows that LLMs are not neutral observers, but rather their outputs often reflect dominant cultural norms and values, potentially marginalising perspectives that differ from these defaults [4, 19, 71, 72]. This raises critical questions about how LLMs interpret socially sensitive interactions, and how their interpretations may diverge from human perspectives, particularly those of people who experience gender microaggressions first-hand.

There is little empirical evidence on how LLM interpretations compare to those of humans, especially when accounting for differences in gender identity and lived experience with discrimination, that research shows shape microaggression perception [10, 27, 36]. In this paper, we aim to address this knowledge gap by seeking answers to the following research questions:

(1) **RQ1: How do LLMs and humans differ in detecting gender microaggressions in workplace scenarios?**
(2) **RQ2: What argumentation patterns and interpretive strategies do humans and LLMs use when providing rationales for their microaggression judgements?**
(3) **RQ3: How do human and LLM rationales differ in the ways they explain or justify their judgments?**

Here, we use *interpretation* to refer to the overall judgment of whether a scenario constitutes a gender microaggression and why. We analyse it through a numerical rating and a rationale (explanation of that rating). We emphasize interpretation rather than detection accuracy, as the latter would presuppose an objective ground truth. To examine these questions, we stratified human participants into six subgroups based on gender identity (male, female, non-binary) and lived experience with workplace gender microaggressions (with or without). This stratification enables us to examine whether the patterns observed map onto gender and lived experience differences.

Our analysis reveals a core distinction that we introduce to characterize differences in interpretation: LLMs exhibit what we term *categorical sensitivity*, identifying microaggressions based on what type of harm they represent, often abstracted from context, while humans (particularly those with lived experience) demonstrate *situated sensitivity*, where their interpretations are grounded in specific contextual factors, including attention to ambiguity. This distinction has significant implications for the design of automated detection systems, suggesting that current LLM approaches may flatten the interpretive plurality that makes human sensitivity meaningful, and highlighting the need for designs that preserve rather than override situated knowledge.

Our work makes several contributions to the growing literature on LLM-supported systems for microaggression detection. First, we provide empirical evidence of significant differences between LLMs and humans in interpreting gender microaggressions in the workplace. Second, we identify patterns of alignment between LLMs and human subgroups defined by gender identity and lived experience. Third, we offer a mixed-methods evaluation approach that pairs numerical judgments with rationale analysis, moving beyond treating microaggression interpretation as a classification task. Finally, we propose design implications emphasizing interpretive plurality, power dynamics, and the potential exclusion of non-dominant perspectives.

2 Background and Related Work

Our work is situated at the intersection of three areas: research on gender microaggressions (what they are and how they manifest), Human-Computer Interaction (HCI) and Computer Science efforts to detect and mitigate microaggressions with computational tools, and feminist/critical perspectives providing a lens to better understand the complexities of the social contexts in which they emerge. Bringing these areas together clarifies the technical challenges in designing LLM-based systems that detect and mitigate microaggressions in the workplace.

2.1 Gender Microaggressions in the Workplace

Microaggressions are common manifestations of bias or discrimination directed at individuals or groups positioned within systems of oppression [81]. They are typically brief and subtle, and manifest verbally, behaviorally, or environmentally [78]. Microaggressions can be conscious or unconscious, do not require intent, and often can pass as jokes, compliments, or misunderstandings [61, 70, 78].

The workplace is a distinct setting for these dynamics. Employees are subject to formal rules and hierarchies they must comply with, and spend a substantial amount of time in these environments

[49]. In practice, overt sexism may have declined, yet gender discriminatory practices and behaviors persist in more ambiguous forms that are harder to identify and contest [20, 24, 48, 78, 84]. The prefix *micro* refers to the scale or granularity of such acts and not the severity of their effects [69, 81]. The accumulated effect can result in many negative impacts on women such as limitation of professional trajectories and diminishing sense of belonging, higher rates of burnout, poorer well-being, and increased intentions to leave their current roles [2, 24, 46, 48, 49, 53, 62].

There is no single definition or taxonomy of gender microaggressions as boundaries remain debated [59]. A widely used starting point is Sue's distinction between microassaults (conscious derogations), microinsults (demeaning slights), and microinvalidations (communications that negate lived experience) [49, 62, 78]. Domain-specific work extends this taxonomy to capture how gender microaggressions emerge in particular contexts. Kim and Meister identify five types of recurrent workplace microaggression in STEM: devaluation of technical competence, devaluation of physical presence, denial of one's reality, pathologizing character, and pathologizing gender, which in their data structure map to an invisible versus visible distinction (devaluation and denial vs. pathologizing) [48]. In a campus-based study of undergraduate women, Gartner identifies seven themes: invisibility, presumed incompetence, sexual objectification, caretaker/nurturer expectations, women-dominated occupations, environmental invalidations, and intersectionality [35], that are also relevant to the workplace.

Ambiguity is a defining feature of microaggressions. The same interaction may be read as harmless by some and harmful by others. Targets often need to exert cognitive and emotional labor to interpret their subtle manifestations [49]. This ambiguity enables for plausible deniability and undermines recognition [46, 68]. Interpretations vary with situational and cultural context, and with the positionalities of those involved and interpreting [10, 55, 61].

Consistent with this, some scholars argue that microaggressions are best understood as observable events whose classification does not depend on the perpetrator's intent or the target's immediate appraisal [68, 81]. Rather, what constitutes a microaggression is better defined by members of the marginalized group affected [61]. An empirical study on perceptions of gender microaggressions in the workplace determined that women were significantly more likely to perceive workplace gender microaggressions than men when microaggressions are not explicit [10]. This positional variation in interpretation provides empirical grounding for our focus on these factors in comparing human and LLM interpretations.

Finally, gender microaggressions frequently intertwine with race, ethnicity, age, class, disability, and sexuality. These intersections produce harms that cannot be understood by analysing a single identity dimension in isolation [55, 70]. In this work we focus on gender for analytic tractability and scope, but we recognise this as a limitation and return to it in our discussion section.

2.2 Approaches for Automatic Detection of Microaggressions

A growing body of work has focused on automating detection of microaggressions, sexism and toxic speech in online text using Machine Learning (ML) models. Most computational work treats them as text classification problems, training models using RNNs or transformers to predict binary labels or assign categories [3, 8, 28, 37, 38, 82]. To address data scarcity, some pipelines rely on LLM-assisted annotation and synthetic examples [79].

While reported detection scores can be high [8, 37, 82], several consistent limitations surface: lower recall for detecting microaggressions [3, 79], domain shift penalties when moving across platforms [3], and category confusions when statements span multiple harm types [8]. Crucially, recent work shows that annotation is positional, and annotators' identity and context determine labels [38], challenging the assumption that crowdsourced annotations represent objective "gold labels" [58].

These detection pipelines rely heavily on textual cues and statement explicitness, making it difficult to capture subtle manifestations, relational dynamics, and power asymmetries encountered in workplace microaggressions. Even multimodal approaches struggle with nuanced interactions [21].

In the workplace, automated microaggression detection could be deployed in several contexts, such as real-time flagging in communication platforms like Slack or Teams, retrospective analysis of meeting transcripts, or monitoring offline interactions with appropriate transcription pipelines. Some organizations already use artificial intelligence (AI) tools for harassment detection in workplace online communications [32]. Extending these applications to microaggressions is a likely trajectory.

The rise of LLMs introduces new possibilities for microaggression detection [4, 38]. Unlike traditional classifiers, LLMs can incorporate contextual details and generate textual rationales. They are already entering workplace flows to provide humans with recommendations and support decision-making [7, 31], underscoring the need to examine how they interpret socially sensitive interactions. However, LLM integration introduces new challenges. Research shows that LLM explanations can increase user reliance without improving awareness [7], and that people consider AI systems fairer when the systems' biases align with users' own biases [86].

A growing body of work examining LLMs as evaluators or judges operates largely within a positivist framework [44, 83, 85], treating human-LLM agreement as the validity benchmark and disagreement as noise to be minimized. But human annotation reflects social position, and annotator identity and beliefs influence the ratings annotators give [73]. For socially contested phenomena like microaggressions, disagreement among humans is something to be expected.

Recent work illustrates this tension: Movva et al. [64] found that while GPT-4 can match overall human consensus on conversational safety, it cannot anticipate which conversations will be perceived differently by different demographic groups. They also found GPT-4 tends to be more stringent than human annotators, rating some conversations as unsafe that annotators found benign, particularly cases involving "positive" stereotypes (e.g., complimenting a group's perceived traits, yet still based on stereotypes) or neutral responses to hateful statements.

Moreover, LLMs do not represent any particular population uniformly. Schäfer et al. [74] found that LLMs align more closely with annotations from White individuals than from Black/African American individuals, and were progressively less accurate for older individuals in offensiveness rating tasks. Dominguez-Olmedo et

al. [30] show that models are closer to uniform distributions than to any human population, better representing subgroups whose aggregate statistics happen to be closest to uniform. These patterns may partly reflect safety-oriented alignment training, which optimizes models to choose "least harmful" options [26, 39, 42, 57], a process that can flatten interpretive diversity.

These limitations have prompted alternative approaches. Recent scholarship advances a perspectivist approach that challenges gold-standard annotation practices [18, 33]. Cabitza et al. [18] argue that multiple perspectives should be preserved in annotation practices rather than collapsed through majority voting, as disagreement is unavoidable for complex, ambiguous phenomena. Similarly, Fleisig et al. [33] extend this argument, showing that lived experience constitutes a legitimate form of expertise systematically overlooked when majority-vote aggregation discards minority annotators' interpretations.

This perspectivist turn in computational work echoes longstanding feminist epistemological arguments about situated knowledge and the value of marginalized perspectives [40, 41], a connection we develop in the following sub-section. Our work extends this perspectivist paradigm to human-LLM comparison. Rather than asking whether LLMs judge microaggressions accurately (which presupposes a ground truth), we examine how their interpretations differ from humans whose social positions shape their understanding of contested workplace interactions.

2.3 Feminist and Critical Perspectives

Feminist and critical perspectives offer essential tools for navigating these complexities because they start from the premise that knowledge is always situated and partial [40]. In the case of microaggressions, interpretation is highly context-dependent, and harm is often invisible to those not experiencing it, leading to systematic disagreement about what "counts" as a microaggression [59, 69, 78].

Rather than viewing disagreement as a problem to be solved, feminist frameworks treat multiple perspectives as analytically valuable. Fricker's [34] concept of epistemic injustice helps explain why targets' interpretations are often dismissed, both through testimonial injustice (not being believed) and hermeneutical injustice (lacking shared conceptual resources to articulate harm). Similarly, feminist standpoint theory argues that those positioned at social margins often develop sharper insights into power operations, suggesting that lived experience with discrimination provides epistemic advantages rather than bias.

These theoretical insights have direct implications for computational approaches to bias detection. Feminist perspectives push beyond narrow technical views of fairness toward intersectional, power-aware understanding of how datasets and models reproduce inequality [58]. Knowledge-enhanced language models risk reproducing epistemic injustice because knowledge itself is socially situated: researchers must be aware of the social nature of knowledge, and avoid assuming content labeled "knowledge" to be objective and neutral [52].

Moreover, AI systems are not neutral [19, 72] observers but are built within existing power structures. Recent work demonstrates that large language models often reflect dominant cultural perspectives, potentially marginalizing viewpoints that diverge from these defaults [71]. This poses fundamental questions for microaggression detection systems: whose perspectives are centered when defining harm, and whose interpretations are excluded or overridden?

The stakes of these questions are heightened by how humans interact with AI-generated explanations. Research shows that people are prone to automation bias, or mistaking LLM outputs as meaningful or authoritative [12, 13], and that users consider AI systems fairer when their biases align with users' own expectations [86]. When AI systems provide confident-sounding explanations about contested social phenomena like microaggressions, they risk creating an illusion of objectivity that obscures the situated nature of interpretation. This theoretical grounding leads us to examine not whether interpretations are correct, but how they emerge from different social and technical positions.

2.4 Situated Comparison Between Humans and LLMs

Building on this body of work, this study compares how LLMs and human participants interpret short workplace dialogues containing gender microaggressions. Unlike detection approaches that frame microaggression identification as binary classification, we examine *interpretations*, both ratings (judgments) and rationales (justifications), to understand the interpretive stances of both human participants and LLMs, rather than to evaluate correctness against a ground truth. Our analysis examines how interpretations converge or diverge across LLMs and human subgroups, and how their rationales frame microaggressions differently.

Our focus on gender identity and lived experience as key demographic variables is grounded in empirical evidence showing systematic differences in microaggression sensitivity [35, 49]. Research demonstrates that women are significantly more likely than men to perceive workplace gender microaggressions when they are not explicit [10], and that lived experience with discrimination shapes interpretive frameworks in ways that provide epistemic advantages for recognizing subtle bias. While much of this literature focuses on cisgender women's experiences, gender microaggressions are also experienced by other groups, such as trans and non-binary people [25]. Our study includes non-binary participants and examines how these gender patterns compare to LLM interpretations.

Importantly, we are not seeking to resolve interpretive ambiguity or determine who is "right" or "wrong" in their assessments. Instead, we aim to understand how different interpretive positions approach the same nuanced interactions. In this approach we treat interpretive diversity as analytically valuable instead of an error.

These theoretical insights have direct implications for human-computer interaction (HCI) design. If microaggression interpretation is inherently situated and contested, then LLM systems designed to detect or respond to workplace microaggressions cannot simply optimize for accuracy against a ground truth. Instead, researchers must grapple with fundamental questions about whose perspectives these systems should reflect, how they might support rather than replace human judgment, and whether computational approaches risk flattening the interpretive plurality that feminist theory suggests we should preserve.

3 Methods

We conducted a mixed-methods study comparing how humans and LLMs interpret workplace gender microaggressions. We presented 141 participants and 7 LLM models with 10 workplace dialogue scenarios, 8 containing different types of gender microaggressions and 2 presenting no microaggressions (as controls). Through an online survey for humans and API prompting for LLMs, we collected numerical ratings (1-5 Likert scale) on the perception of a scenario containing a microaggression, and an open-text rationale for their ratings, for each scenario (see Fig. 1). We analyzed the data using non-parametric statistical tests for ratings and mixed qualitative methods for rationales. This study was preregistered prior to data collection, study materials and data are available in OSF public repository[1].

Drawing on feminist perspectives, we approach the interpretation of microaggressions as context-dependent, situational, and relational [40, 59, 69]. Thus, we acknowledge that there is no *gold standard* in this matter to which LLMs or humans should align. This shapes the methodological choices we made: we are not studying LLMs' capability to detect gender microaggressions, but rather how participants and LLMs rate scenarios as microaggressive or non-microaggressive, and the justifications they provide for those ratings. We acknowledge that a numerical rating does not always map directly to the accompanying written rationale (for either humans or LLMs), yet we treat these paired responses as proxies for interpretive stance, enabling us to examine their perspectives in more nuanced ways.

3.1 Scenarios

We created 8 workplace dialogue scenarios (vignettes [5, 6]), adapted from self-reported accounts of workplace microaggressions submitted to *microaggressions.com*[2], a public website where people anonymously share accounts of microaggressions they have experienced. We chose to ground our scenarios in lived experiences rather than create hypothetical examples because microaggressions are inherently ambiguous phenomena that are best understood through narratives situated in real social contexts.

We filtered the submissions using the website's "gender" tag, alongside workplace-relevant keywords ("boss," "work," "office," "coworker," "colleague," "manager," "job," "career", "interview"), resulting in 401 posts. We excluded intersectional cases (entries with multiple tags), in order to focus specifically on gender-based dynamics, yielding 187 posts.

We then reviewed these posts individually, applying inclusion criteria to retain only those that: (1) were understandable as standalone narratives, (2) could be adapted into a dialogue format, and (3) were clearly related to a workplace context. This process resulted in 60 posts.

To standardize the presentation format of each submission, we transformed them into short dialogues that include gender markers (e.g.,"[Woman]:", "[Man]:") while preserving the original content and meaning (see Table 1). This transformation was performed manually with occasional assistance from GPT-4o for initial formatting, followed by careful review and editing to ensure accuracy, natural conversational flow, and preservation of the original meaning.

Because participants could only be exposed to a limited number of vignettes without risking cognitive overload [1], we used ten final scenarios (see Appendix B): eight containing gender-based microaggressions and two neutral workplace interactions [65] that served as controls and attention checks, in accordance with similar studies [10].

To ensure diversity in our selection, we developed an expanded classification system building on existing taxonomies. We began with Sue's [78] foundational three-category taxonomy (microassaults, microinsults, and microinvalidations), and incorporated domain-specific categories from Gartner's [35] campus-based framework, which identifies forms such as sexual objectification and undermining of competence, and Kim's [48] workplace STEM taxonomy, which highlights dynamics like gender as professional liability and exclusion from networks. This integration resulted in eight categories: sexual objectification, gender hostility, undermining competence, pathologizing character, restrictive gender roles, gender as liability, denial of experience or invalidation, and exclusion (see Appendix A for complete taxonomy). We selected scenarios to represent a balance across these categories, with some scenarios exemplifying multiple types.

3.2 Pilots and Explorations

We conducted several pilot studies to refine our methodology and study design. No pilot participants were included in the final study. Initial pilots (n=7 via Prolific) tested whether dialogue formats were clear enough for participants to assess. No participants reported comprehension issues.

A preliminary pilot used binary yes/no responses for microaggression judgments, but this failed to capture the inherent ambiguity in how microaggressions are interpreted. Based on pilot feedback, we adopted a 5-point Likert scale (1=Definitely not, 3=Uncertain, 5=Definitely yes) to better accommodate interpretive uncertainty. We focused on perceived likelihood rather than severity because our interest was in interpretive disagreement about what constitutes a microaggression, not in comparing harm magnitude. Additionally, initial pilots asked about lived experience using yes/no questions, which yielded limited insight and failed to capture nuanced perspectives. We refined this to multiple-choice options (personal experience, witnessing, uncertainty, no experience) followed by open-text descriptions for those reporting personal experience.

Finally, we conducted systematic tests of *temperature* and *top-p* settings across model families. While parameter variations affected phrasing, thematic coding remained consistent. Given GPT-5's lack of parameter control (as of August 2025), we standardized on default settings (temperature=1) across all models. Early testing showed high consistency in LLM outputs across multiple generations, leading us to expand from three to ten repetitions per scenario [4].

3.3 Tasks

The study consisted of two tasks directed to both human participants and LLMs. Participants completed an online survey (Qualtrics), and LLMs were prompted via their respective APIs. Our working definition of gender-based microaggression was given

[1] https://osf.io/37qgh/overview?view_only=58d9e950f6af4b60a1de5ffd6e920655
[2] https://www.microaggressions.com

Table 1: Example of dialogue construction from original microaggression account to structured scenario format

Original snippet	Transformed dialogue
"Why don't you fax him a picture of yourself? I'm sure he'll get that memo to you faster." A male colleague, when I tell him his co-worker has not yet faxed me the memo I requested.	*[Woman]:* "I still haven't received that memo, could you check with him?" *[Man]:* "Why don't you fax him a picture of yourself? I'm sure he'll send it faster."

to both groups: *"Subtle and often ambiguous forms of discrimination based on a person's gender. They are often normalized or dismissed as minor, but their harm is cumulative. It emerges through patterns of marginalization, misrecognition, and social exclusion over time. They may take the form of jokes, comments, or actions that convey hostility, reinforce stereotypes, or make individuals feel unsafe, undermined, excluded, or dismissed".* Then both humans and LLMs were presented with the same two tasks for each scenario:

(1) **Microaggression rating (Likert)**: "Do you think this interaction contains a gender-based microaggression?"
Responses were collected on a 5-point Likert scale: 1 (Definitely not), 2 (Probably not), 3 (Uncertain), 4 (Probably yes), 5 (Definitely yes).
(2) **Rationale (open text)**: "Why did you choose that answer?"
Participants provided a free-text explanation for their rating. For LLMs, we acknowledge we cannot claim that these accounts represent the true "reasoning" to assign a rating, but they are treated as outputs that reflect interpretive positioning that is assessed in complement to the rating.

3.4 Participants

The study was approved by our institution's Human Research Ethics Committee (application ID 5560). Participants were informed about the sensitive nature of the content, and informed consent was obtained before any data collection. They were recruited via Prolific[3]. We used crowd workers because they provide access to a diverse, geographically distributed sample. To ensure participants had sufficient professional exposure to workplace dynamics, we required a minimum of one year of work experience as a screening criterion.

To attend to diversity of perspectives, we stratified human participants by self-reported gender identity (47 male, 47 female, and 47 non-binary). In the survey we also asked participants to report their lived experience with workplace gender microaggressions.

We focused on gender identity and lived experience rather than geographic or cultural background for two reasons: first, recent work suggests that social attitudes and lived experience predict annotation behavior more reliably than demographic categories alone [11, 45, 66, 73]; second, this focus kept the study design manageable while maintaining theoretical coherence with our emphasis on situated interpretation.

To capture participants' self-reported lived experience, we first provided our working definition of gender microaggressions and then asked whether they had experienced or witnessed such incidents at work through a multiple choice question. Response options included personal experience, witnessing it happening to others, uncertainty, or no experience.

Participants who reported personal experience were asked an open-ended follow-up at the end of the survey describing, in their own words, how they had experienced this in the workplace. This question was designed to provide context for interpreting results rather than to systematically analyze experience characteristics; we did not collect structured data on frequency or severity. The distribution of lived experience across gender groups reflected documented prevalence patterns: 7 of 44 male participants, 30 of 44 female participants, and 27 of 47 non-binary participants reported lived experience with workplace gender microaggressions (See Table 4). The low number of males with lived experience is consistent with research showing gender microaggressions disproportionately target women and gender minorities [53], and we interpret findings for this subgroup (ME, n=7) with appropriate caution given constraints on statistical power.

During data analysis, close reading of the rationales revealed six participants whose responses showed distinctly different patterns that suggested they were AI generated: responses were substantially longer, followed identical structures across all scenarios, and consistently incorporated verbatim phrasing from the question prompt. One response explicitly included "Ask ChatGPT" and was immediately excluded. The remaining five participants were contacted through Prolific and self-reported AI assistance. All six were excluded from analysis, yielding a final sample of 135 participants (44 male, 44 female, 47 non-binary).

3.5 LLM Setup

We evaluated three families of large language models with API access: GPT (OpenAI)[4], Claude (Anthropic)[5], and Gemini (Google)[6]. These models were selected because they are widely available in "ready-to-use" form and are actively being integrated into commercial and research applications, making them relevant for understanding how LLMs interpret microaggressions.

Within each model family, we selected both the most powerful and most cost-efficient variants to capture potential performance differences across computational resources. This dual selection criterion ensures our findings are relevant for both research applications (which may use premium models) and practical deployments (which often rely on efficient models). As of August 2025, this resulted in Claude Opus 4 (CO), Claude Sonnet 4 (CS), GPT-5 (G5), GPT-5 nano (G5n), Gemini 2.5 Pro (GMp), and Gemini 2.5 Flash (GMf). We also included GPT-4o (G4o), as GPT-5 had been released recently.

[3]https://www.prolific.co

[4]https://openai.com/api
[5]https://www.anthropic.com
[6]https://ai.google.dev

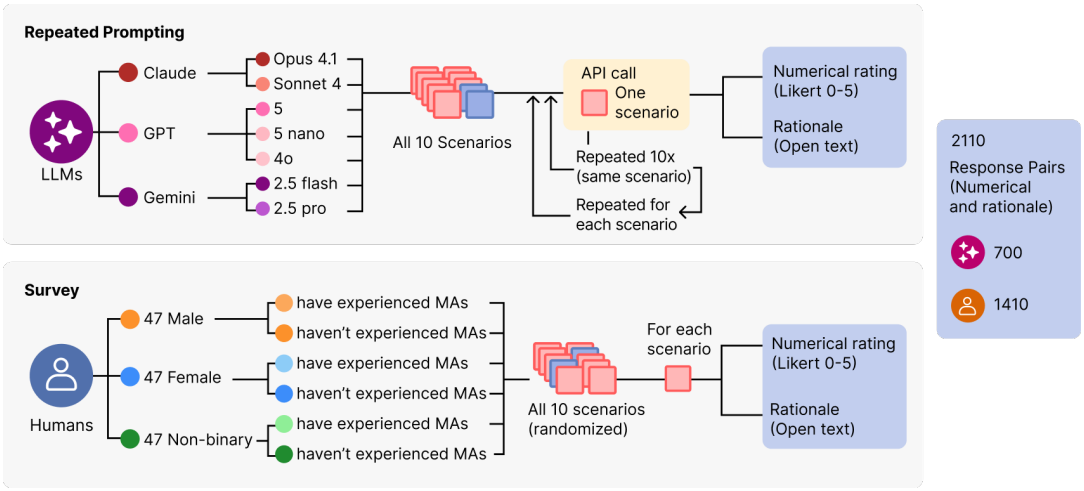


Figure 2: Data collection method showing participant stratification and response collection across human subgroups and LLM model variants.

We used default settings (temperature = 1) across all models to ensure consistency, particularly given GPT-5's lack of parameter control. Each model received the same ten scenarios and was prompted with identical instructions given to participants (see Appendix C for complete prompt). The instructions did not include examples. We used ten repetitions per scenario to increase robustness of comparisons [4]. For each run, both the Likert rating and open-text rationale were stored in csv format for analysis.

Our prompt included a definition that described microaggressions as involving cumulative harm, marginalization, and misrecognition (see Appendix C). Prior work shows that including definitions affects LLM detection sensitivity; Kumar et al. [54] found that removing definitions from toxicity detection prompts substantially reduced model recall. Informed by this, we included a definition in our prompt, and provided the same definition to human participants to ensure identical task framing across all raters.

3.6 Data Analysis

We conducted both quantitative and qualitative analyses to examine similarities and differences between human participants and LLMs in their interpretations of gender-based workplace microaggressions.

3.6.1 A quick note on our chosen approach. We are working within a disciplinary tension: most computational approaches to bias detection follow positivist framings oriented toward quantifiable metrics and algorithmic solutions. Conversely, critical HCI scholarship acknowledges that meaning is socially constructed and requires interpretive, non-positivist approaches.

We deliberately inhabit this tension: adopting some familiar structures from CS/AI evaluation to enable comparison, while grounding our interpretation in reflexive, constructivist principles. Our analysis includes both reflexive thematic analysis and descriptive coding with frequency counts (what might appear as quantitative content analysis)[15]. We begin with structures familiar to CS/HCI audiences (ratings, frequencies) while fundamentally grounding our interpretation in constructivist principles. While our primary analytical focus remains on examining rationales, constructing themes, and reflexively engaging with meaning-making, the ratings offer additional analytical dimensions that enrich our interpretation.

3.6.2 Quantitative Analysis. We analyzed the Likert-scale ratings of the ten scenarios by computing descriptive statistics (mean, standard deviation) for each rater group across microaggression and control scenarios. This included human subgroups based on gender identity and lived experience and each LLM model. Data analysis was performed using IBM SPSS Statistics 20.

A variety of statistical tests were employed to determine the statistical significance of observed effects and to extract meaningful insights. We used non-parametric procedures (i.e., Kruskal–Wallis H tests for multi-group comparisons and Mann–Whitney U tests for two-group contrasts), because normality and homoscedasticity were not consistently satisfied across variables, following the assumption guidance in Harwell [43]. For brevity, we omit assumption test outputs. To interpret the magnitude of these effects beyond statistical significance, we computed effect-size statistics using Epsilon-squared (ϵ^2) for Kruskal-Wallis tests and r for Mann-Whitney U tests. We interpreted ϵ^2 values based on guidelines by Tomczak and Tomczak [80] and King et al. [50], where values in the range $[0.01, 0.08)$ indicate a *small*, $[0.08, 0.26)$ a *medium*, and ≥ 0.26 a *large* effect. For pairwise contrasts, we interpreted r following Cohen [23]'s benchmarks: $r < 0.1$ (very small), $0.1 \leq r < 0.3$ (small), $0.3 \leq r < 0.5$ (moderate), and $r \geq 0.5$ (large).

Finally, when an omnibus Kruskal–Wallis test was significant, we conducted Dunn's post-hoc pairwise comparisons with Bonferroni correction and report adjusted p-values.

3.6.3 Qualitative Analysis. Our approach for the qualitative analysis of rationales was to conduct reflexive Thematic Analysis (TA), following Braun and Clarke's six-phase process[14, 16]. For this, we used the software ATLAS.ti. We familiarized ourselves with the

data, and coded all rationales (both from humans and LLMs) iteratively. We started coding closely to the data, then two researchers discussed the codes and developed initial themes. After this we analyzed the coded data and iterated on the themes. Then a split approach was adopted: we developed descriptive codes, meaning codes that resembled topics or categories, contrary to what is expected from reflexive TA [17]. In parallel we also created themes based on meaning, or interpretative stories, aligning more closely with reflexive TA approaches [16].

We adopted a dual coding approach to bridge quantitative and qualitative insights. First, we developed category codes (descriptive), that diverge from what is expected from reflexive TA practice [15]. However, this allowed us to quantify patterns and create analytical bridges to ratings data. Second, we constructed interpretive themes through reflexive meaning making, aligning with reflexive TA principles[17]. These themes were not discovered or proven by code counts, as their meaning or significance does not depend on their frequency.

Category codes included stance, argumentation elements such as ambiguity cues, framing context, stating impact, stating intent, and describing a microaggression mechanism. The codes were generated inductively, and in a second iteration, microaggression mechanisms were coded in alignment to the expanded taxonomy described in Appendix A.

4 Results

In this section we present the main differences found in the interpretations given by humans and LLMs. As explained in the previous section, we consider interpretations paired answers of a numerical rating, and an open rationale justifying the rating. We begin this section with the ratings, providing an overview of humans and LLMs answers, and explore more in depth the differences in human responses based on gender and lived experience, and in LLMs based on model family and type. We then present the differences in how the rationales from humans and LLMs are constructed, showing overall distributions and main differences. Finally, we present 3 emerging themes that were constructed through reflexive thematic analysis of the rationales.

4.1 Ratings

As explained in the method section, ratings were captured using a 5-point Likert scale to capture if the rater thought a scenario contained a microaggression. Higher ratings meant the scenario more likely contained a microaggression. To assess the consistency of these interpretations, we calculated inter-rater reliability using Kendall's coefficient of concordance (W). We found moderate agreement levels within both groups, with Human participants ($W = 0.55$, $p < .001$) showing slightly higher concordance than AI outputs ($W = 0.51$, $p < .001$). This variability is further illustrated by the rating distributions presented in Figure 3, which vary notably between humans and LLMs and between scenarios.

In the scenarios that contained microaggressions (S1-8), human ratings (in yellow) show higher diversity than LLMs' (in purple). Particularly for scenarios 4, 6, and 8. Scenario means (Table 2) show LLM averages between 4.66–5.00 vs. human 3.04–4.73, with notably smaller standard deviations for LLMs indicating more uniform responses. Inferentially, LLMs and humans differed significantly on nearly all scenarios containing microaggressions (S1–S8; all $p < .001$). Specifically, **S1** ($U = 3604$, $z = -3.85$, $p < .001$, $r = -0.19$; *small effect*), **S2** ($U = 3275$, $z = -4.67$, $p < .001$, $r = -0.25$; *small effect*), **S3** ($U = 3201$, $z = -4.84$, $p < .001$, $r = -0.26$; *small effect*), **S4** ($U = 3353$, $z = -3.73$, $p < .001$, $r = -0.23$; *small effect*), **S5** ($U = 3920$, $z = -3.58$, $p < .001$, $r = -0.13$; *small effect*), **S6** ($U = 1596$, $z = -8.04$, $p < .001$, $r = -0.54$; *large effect*), **S7** ($U = 3570$, $z = -4.43$, $p < .001$, $r = -0.20$; *small effect*), and **S8** ($U = 2481$, $z = -5.94$, $p < .001$, $r = -0.39$; *medium effect*).

For the two control scenarios, the differences did not reach significance (S9: $U = 4900$, $z = 1.79$, $p = .073$; S10: $U = 4725$, $z = 1.83$, $p = .068$), supporting that neutral interactions were not over-classified as microaggressive by either group. In what follows, the analysis focuses on Scenarios 1–8 (those containing gender microaggressions).

While differences were statistically significant across all microaggression scenarios, the magnitude of this difference varied. Scenarios such as S1 and S5 showed small effect sizes (r between -0.13 and -0.19), and it suggests a directional agreement despite intensity differences. In contrast, Scenario 6 yielded a large effect size ($r = -0.54$), and Scenario 8 a medium effect ($r = -0.39$), highlighting instances where human and LLM interpretations fundamentally diverge rather than simply differ in rating intensity.

4.1.1 Gender Effects Among Human Participants. Gender identity significantly influenced ratings on four of the eight microaggression scenarios: **S1** (Kruskal–Wallis H: $\chi^2(2) = 7.30$, $p = .026$, $\epsilon^2 = .04$; *small effect*), **S2** ($\chi^2(2) = 6.02$, $p = .049$, $\epsilon^2 = .03$; *small effect*), **S3** ($\chi^2(2) = 14.33$, $p < .001$, $\epsilon^2 = .09$; *medium effect*), and **S6** ($\chi^2(2) = 20.59$, $p < .001$, $\epsilon^2 = .14$; *medium effect*). These omnibus effects were modest-to-moderate in size. No significant gender effects were found for **S4** ($p = .823$), **S5** ($p = .552$), **S7** ($p = .251$), or **S8** ($p = .357$). The presence of medium effects for S3 and S6, compared to small effects for S1 and S2, suggests that gender identity played a more substantial role in shaping interpretations for scenarios involving 'gender as liability' and 'restrictive gender roles' compared to other categories.

Table 3 presents the means and SD when disaggregating by gender. Results reveal that non-binary participants gave slightly higher means than male/female participants in all scenarios. Male participants rated lowest in S1, S2, S3, S5, and S6, rating particularly low in the later ($M = 2.32$, $SD = 1.29$). Post-hoc pairwise comparisons with Bonferroni correction showed that for **S1**, non-binary participants rated the scenario significantly higher than male participants ($adj.p = .034$). For **S3**, the difference between non-binary and male participants was also significant ($adj.p < .001$). **S6** showed significant differences between male participants and both female ($adj.p = .043$) and non-binary participants ($adj.p < .001$). For **S2**, no pairwise comparisons were significant after correction.

4.1.2 Effects of Lived Experience. When looking at the differences in lived experiences, we expand the 3 gender groups to 6 (see Fig. 4). Participants with lived experience consistently provided higher ratings than those without such experience. Independent-samples Mann–Whitney tests showed significant differences for **S1** ($U = 1894.0$, $z = 2.48$, $p = .013$, $r = -0.17$; *small effect*), **S2** ($U = 1245.5$, $z = 3.94$, $p < .001$, $r = -0.29$; *small effect*), **S3** ($U = 1250.5$, $z = 4.69$,

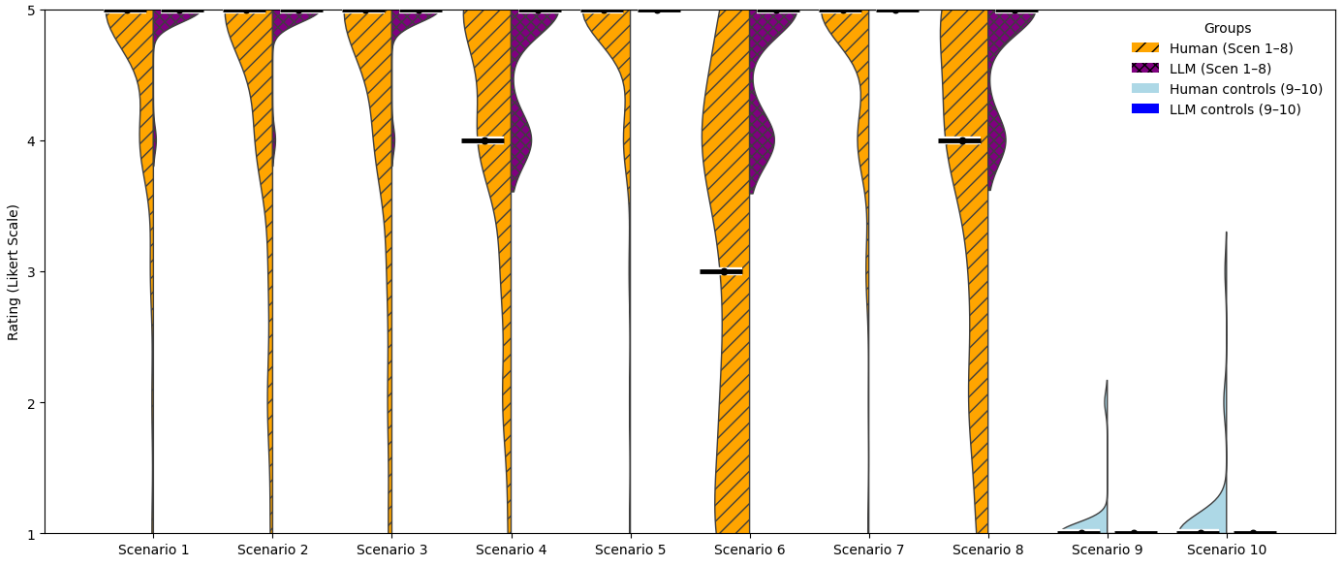


Figure 3: Rating distributions showing greater interpretive diversity among humans compared to LLMs across microaggression scenarios (S1-S8) and control scenarios (S9-S10)

Table 2: Mean ratings and standard deviations by scenario, showing consistently higher LLM ratings with lower variance compared to humans across microaggression scenarios (S1-S8)

Group	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9	Scenario 10
Human	4.56 ± 0.86	4.36 ± 1.05	4.33 ± 1.09	4.13 ± 1.08	4.73 ± 0.71	3.04 ± 1.43	4.65 ± 0.74	3.71 ± 1.30	1.05 ± 0.22	1.11 ± 0.40
LLM	4.94 ± 0.23	4.94 ± 0.23	4.94 ± 0.23	4.70 ± 0.46	5.00 ± 0.00	4.66 ± 0.48	5.00 ± 0.00	4.73 ± 0.45	1.00 ± 0.00	1.00 ± 0.00

Table 3: Mean microaggression ratings by gender identity, showing systematic differences with non-binary participants rating highest and male participants lowest across most scenarios

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
Male	4.32 ± 1.09	4.07 ± 1.25	3.93 ± 1.25	4.11 ± 1.13	4.64 ± 0.78	2.32 ± 1.29	4.61 ± 0.75	3.61 ± 1.42
Female	4.50 ± 0.88	4.41 ± 1.02	4.32 ± 1.20	4.05 ± 1.12	4.77 ± 0.68	3.07 ± 1.37	4.61 ± 0.65	3.55 ± 1.37
Non-Binary	4.83 ± 0.43	4.60 ± 0.80	4.72 ± 0.62	4.21 ± 1.00	4.79 ± 0.69	3.68 ± 1.32	4.72 ± 0.80	3.96 ± 1.10

$p < .001$, $r = -0.35$; *medium effect*), and **S6** ($U = 1255.5$, $z = 3.67$, $p < .001$, $r = -0.31$; *medium effect*). Differences were not significant for **S4** ($p = .101$), **S5** ($p = .126$), **S7** ($p = .820$), or **S8** ($p = .153$). All means and SD values can be seen in Table 2, among all LLM rater types. Notably, while lived experience had a small impact on ratings for S1 and S2, it showed a medium effect size for S3 and S6, indicating that for certain subtle microaggressions, possessing lived experience significantly alters the perception of harm.

These results show that within each gender category, those with lived experience (ME, FE, NBE) rated scenarios higher than their counterparts without lived experience (MX, FX, NBX). Notably, males without experience (MX) consistently showed the lowest means, while non-binary participants with experience (NBE) often showed the highest ratings (e.g., S1: $M = 5.00$, $SD = 0.00$). Post-hoc analyses of the six distinct human rater types revealed significant differences between MX and NBE on **S1** ($adj.p = .006$), S2 ($adj.p = .006$), **S3** ($adj.p < .001$), and **S6** ($adj.p < .001$), representing the extremes of the spectrum.

4.1.3 LLM Model Family and Model Version Differences. In general, all LLM families had ceiling effects ($M = 5.00$, $SD = 0.00$), as reported in Table 5. For Gemini models this happened across all 8 scenarios, while for Claude models this happened in 6 scenarios and for GPT models in 2 scenarios. SDs were usually low, with the highest one being 0.51 (Claude, Sc 4). Kruskal–Wallis tests indicated significant model family differences in **S4** ($\chi^2(2) = 12.83$, $p = .002$, $\epsilon^2 = .16$; *medium effect*), **S6** ($\chi^2(2) = 53.83$, $p < .001$, $\epsilon^2 = .77$; *large effect*), and **S8** ($\chi^2(2) = 34.28$, $p < .001$, $\epsilon^2 = .48$; *large effect*). No

Table 4: Participant stratification by gender identity and lived experience with workplace gender microaggressions

Abbreviation	Rater type	*n*
MX	Male without lived experience with gender microaggressions	37
ME	Male with lived experience with gender microaggressions	7
FX	Female without lived experience with gender microaggressions	14
FE	Female with lived experience with gender microaggressions	30
NBX	Non-binary without lived experience with gender microaggressions	20
NBE	Non-binary with lived experience with gender microaggressions	27

model family effects were found in **S1** ($p = .062$), **S2** ($p = 1.000$), **S3** ($p = 1.000$), **S5** ($p = 1.000$), and **S7** ($p = 1.000$). Dunn–Bonferroni tests showed that in **S4**, Claude and GPT were lower than Gemini (adj. $p = .002$ and $p = .018$). In **S6**, Claude differed from both GPT and Gemini (both adj. $p < .001$). In **S8**, GPT differed from Claude and Gemini (both adj. $p < .001$).

When examining specific model versions, significant differences emerged for **S2** ($\chi^2(6) = 14.11, p = .028, \epsilon^2 = .13$; *medium effect*), **S3** ($\chi^2(6) = 14.11, p = .028, \epsilon^2 = .13$; *medium effect*), **S4** ($\chi^2(6) = 57.27, p < .001, \epsilon^2 = .81$; *large effect*), **S6** ($\chi^2(6) = 55.88, p < .001, \epsilon^2 = .79$; *large effect*), and **S8** ($\chi^2(6) = 44.58, p < .001, \epsilon^2 = .61$; *large effect*). No version effects were found in **S1** ($p = .107$), **S5** ($p = 1.000$), **S7** ($p = 1.000$), **S9** ($p = 1.000$), or **S10** ($p = 1.000$). The large effect sizes observed across both model families (up to $\epsilon^2 = .77$) and specific versions (up to $\epsilon^2 = .81$) in scenarios S4, S6, and S8 highlight that both broad different models and granular within-model family updates can alter detection thresholds for ambiguous scenarios.

For S2 and S3, no pair survived Bonferroni correction; S4, S6, and S8 had multiple significant pairs (see Appendix **??** for full post-hoc tables).

4.1.4 All Rater Types Comparison. Considering all 13 rater types (six human subgroups + seven LLM variants), Kruskal–Wallis tests showed significant differences in S1, S2, S3, S4, S6, S7, and S8 (all $p < .05$) while there was no difference in S5 ($p = .447$) (see Table 6). This analysis reveals the full spectrum of interpretive variation, from males without lived experience (often lowest ratings) to various LLM models (often ceiling ratings).

4.2 Rationales: comparison of code frequencies in rationales among humans and LLMs

In this subsection we first compare rationales between humans and LLMs, considering both stance and argumentation elements. Following the same structure as with ratings, we then go more in depth into gender, lived experience, model family and model version differences.

When providing rationales, raters would often give a *stance*. Either stating that the interaction contains a microaggression (Yes-MA), or stating it does not contain a microaggression (No-MA). When none of them was provided, rationales where coded with "no stance". Table 7 compares stance frequencies and ratings. Qualitative codes for stance align closely with the ratings they gave. We grouped ratings into pairs: 5 & 4 (indicating the scenario likely contains a microaggression) and 2 & 1 (indicating it likely does not), as these pairs differ only in certainty level ("definitely" versus "probably"). Humans expressed a mix of stances 81.29% of the rationales contained stance affirming there was a microaggression in the scenario (Yes-MA), 13.14% denied there was a microaggression (No-MA), and 5.55% provided no stance, whereas LLMs rationales uniformly contained Yes-MA stances (100.00%).

Figure 4 summarizes argumentation elements present in the rationales. From all human rationales ($n = 1080$), 7.87% contained ambiguity cues, meaning instances in which either the rater was unsure or pointed out that context was not sufficient to be certain. In contrast, among LLM rationales ($n = 560$), only 0.18% contained ambiguity cues (equivalent to only one answer). The same is true for stating intent, although much less frequent (1.94% in humans vs 0.18% in LLMs). Humans also provided specific framings to the context of the dialogue (e.g., "feedback", "advice", "backhanded compliment") more often than LLMs (11.67% vs 4.82%).

Alternatively, both stating an impact and describing a microaggression mechanism, where found more often in LLM rationales than humans. Impact was stated in 47.32% of LLM rationales vs. 5.83% of human rationales, and mechanisms were found in almost all LLM rationales (92.50%) vs. 57.13% in humans. Together, these distributions indicate that human reasoning often acknowledges uncertainty and context, while LLMs rationales often focus on impact and mechanisms.

4.2.1 Genders vs LLM Model families. To deepen the analysis, we break down the results by gender and model family. Table 8 presents the stance frequencies found in rationales compared to the ratings. In the same way as Human-LLM comparison, we found that the qualitative codes for stance align to ratings. Among gender groups, non-binary participants showed the lowest no-MA stance rationales (Table 8), and greater use of ambiguity cues and microaggression mechanisms than other human groups (Fig. 5). Additionally, male participants' rationales showed the highest percentages for framing context, compared to other genders.

LLM families remain highly consistent in mentioning microaggression mechanisms in rationales. GPT shows lower use of mechanisms and higher use of impact statement than other LLM families.

4.2.2 All rater types. Comparing the six human rater types (defined in Table 4: MX/ME/FX/FE/NBX/NBE), Table 9 shows how groups with lived experience more often provided "Yes-MA" stances and fewer "no-stance" rationales than their same-gender counterparts without such experience, consistent with the ratings analysis. In terms of argumentation elements (Fig. 6), they also referenced microaggression mechanisms more frequently.

Table 5: Mean microaggression ratings by LLM model family, showing ceiling effects across all families with minimal variance.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
Claude	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.50 ± 0.51	5.00 ± 0.00	4.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
GPT	4.87 ± 0.35	4.87 ± 0.35	4.87 ± 0.35	4.63 ± 0.49	5.00 ± 0.00	4.87 ± 0.35	5.00 ± 0.00	4.37 ± 0.49
Gemini	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00

Table 6: Mean microaggression ratings across all participant types, revealing systematic variation from male participants without experience (lowest) to LLM models (highest with ceiling effects). Above the divider are human subgroups (see Table 4). Below the divider are LLM models (CO = Claude Opus, CS = Claude Sonnet, G5 = GPT-5, G5n = GPT-5 nano, G4o = GPT-4o, GMf = Gemini Flash, GMp = Gemini Pro)

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
MX	4.24 ± 1.16	3.97 ± 1.32	3.78 ± 1.29	4.11 ± 1.22	4.65 ± 0.72	2.38 ± 1.32	4.59 ± 0.80	3.51 ± 1.50
ME	4.71 ± 0.49	4.57 ± 0.53	4.71 ± 0.49	4.14 ± 0.38	4.57 ± 1.13	2.00 ± 1.15	4.71 ± 0.49	4.14 ± 0.69
FX	4.50 ± 0.76	4.00 ± 1.24	3.86 ± 1.35	3.64 ± 1.15	4.86 ± 0.36	2.36 ± 1.15	4.50 ± 0.76	3.14 ± 1.46
FE	4.50 ± 0.94	4.60 ± 0.86	4.53 ± 1.07	4.23 ± 1.07	4.73 ± 0.78	3.40 ± 1.35	4.67 ± 0.61	3.73 ± 1.31
NBX	4.60 ± 0.60	4.25 ± 0.97	4.45 ± 0.83	3.90 ± 1.07	4.70 ± 0.57	3.20 ± 1.54	4.75 ± 0.79	3.85 ± 1.18
NBE	5.00 ± 0.00	4.85 ± 0.53	4.93 ± 0.27	4.44 ± 0.89	4.85 ± 0.77	4.04 ± 1.02	4.70 ± 0.82	4.04 ± 1.06
CO	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
CS	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.00 ± 0.00	5.00 ± 0.00	4.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
G5	4.80 ± 0.42	4.90 ± 0.32	5.00 ± 0.00	4.80 ± 0.42	5.00 ± 0.00	4.90 ± 0.32	5.00 ± 0.00	4.50 ± 0.53
G5n	4.80 ± 0.42	5.00 ± 0.00	4.90 ± 0.32	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.60 ± 0.52
G4o	5.00 ± 0.00	4.70 ± 0.48	4.70 ± 0.48	4.10 ± 0.32	5.00 ± 0.00	4.70 ± 0.48	5.00 ± 0.00	4.00 ± 0.00
GMf	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
GMp	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00

Table 7: Comparison of stance frequencies from rationales and corresponding rating distributions, showing alignment between qualitative codes and quantitative measures

Rater	n	Yes-MA	5-4	No stance	3	No-MA	2-1
Humans	1080	81.29%	80.18%	5.55%	7.22%	13.14%	12.59%
LLMs	560	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%

Regarding LLM model differences, GPT4o less often provided microaggression mechanism in rationales, and was the one that most often stated impact. It is interesting to note also how Gemini models almost always state mechanisms, but a difference appears between GMf and GMp when stating impact (60.00% vs 27.50%).

4.3 Themes in rationales

We constructed two main themes through reflexive TA. The first one, *Anchored Tensions in Human Rationales*, refers to how humans ground their rationales in specific framings or contextual anchors, often creating some tensions in their reasoning. The second one, *Centering Harm in LLM Rationales*, refers to how LLMs outputs center harm in their rationales, often generalizing or detaching

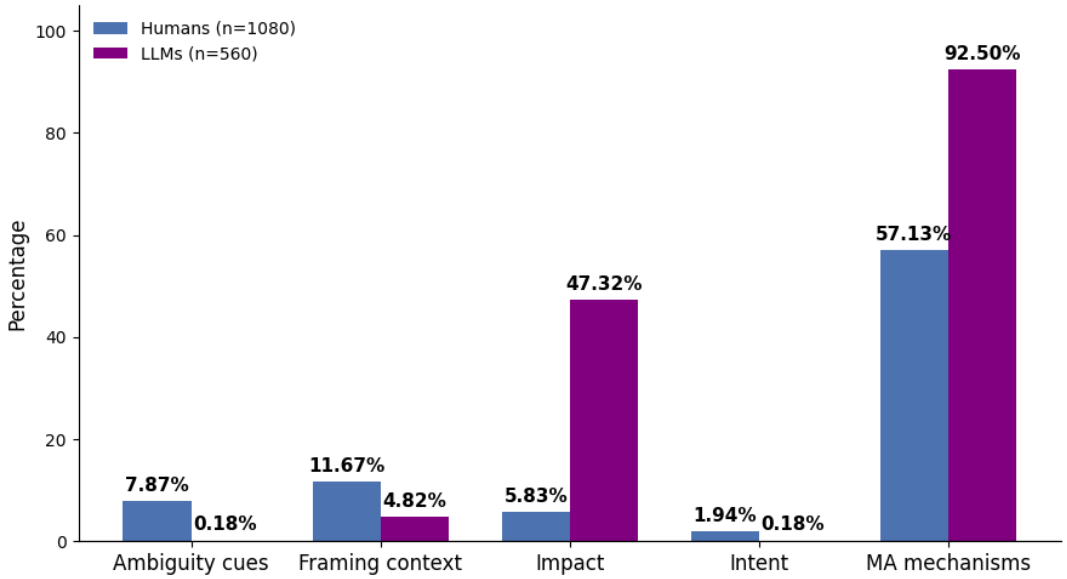


Figure 4: Argumentation style distribution showing humans employ diverse reasoning strategies while LLMs predominantly focus on mechanism-based explanations

Table 8: Stance frequencies by gender and model family, revealing systematic differences in interpretive certainty across human subgroups while LLM families show uniform agreement

Rater	n	Yes-MA	5-4	No stance	3	No-MA	2-1
Male	352	75.40%	74.10%	5.40%	7.70%	19.00%	18.20%
Female	352	80.70%	79.50%	4.80%	6.80%	14.50%	13.60%
Non-Binary	376	87.20%	86.40%	6.40%	7.20%	6.40%	6.40%
Claude	160	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
GPT	240	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
Gemini	160	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%

them from the dialogue's context. We describe these two themes below.

4.3.1 Anchored Tensions in Human Rationales. Human rationales were frequently characterized by an effort to "anchor" their interpretations in specific parts of the context. In other words, participants grounded their rationales in elements, norms, or conditions that they would point to. In some cases, these anchors took the form of specific framings of the context in which the interaction took place. Some participants described the context of the interaction as compliments, jokes or feedback: *"To me it sounds like a joke so I won't take it as a gender-based microaggression"* (FX; Sc2), *"Are compliments bad now?"* (MX; Sc1), *"ha, absolutely! highlighting someones gender like this is a weird backhanded compliment. just say 'one of my best mechanics' if you must"* (NBX; Sc1).

At other times, participants anchored their reasoning in specific elements of the dialogue itself. For instance, one respondent explained: *"Without the phrase 'I expect more from you' I would have said probably not, but this feels like she is saying she expects it specifically from this person"* (NBE; Sc6). Another highlighted the absence of potential targets: *"There is no one present to be offended by the statement, since both of the people speaking are identified as men… If this had been said in front of women, the answer would be different"* (NBX; Sc2). Building on this, some participants used anchors to articulate their boundaries for what "counts" as a microaggression. One participant noted: *"If the second person was a man I would probably say yes but in this case I am unsure"* (FX; Sc6). Similarly, for a different scenario another participant explained: *"This seems like a dialogue where men are acknowledging that they regularly engage in inappropriate workplace conversation. However, because no bystanders are mentioned, is this really a microaggression if it not aimed at a specific target?"* (NBE; Sc2).

In cases where participants felt they lacked sufficient anchors, they often pointed to the need for additional context. For example: *"I'd have to know the context as people are generally ignorant, but it is a female term"* (ME; Sc4), *"The word 'bitch' is kind of sexist in origin, but also some people are definitely reclaiming it, including some gay men… so it's unclear from this interaction"* (NBE; Sc4), or *"Ok I get that the implication here is that she's attractive and that will speed things up, but it's also such a bizarre suggestion that I can also*

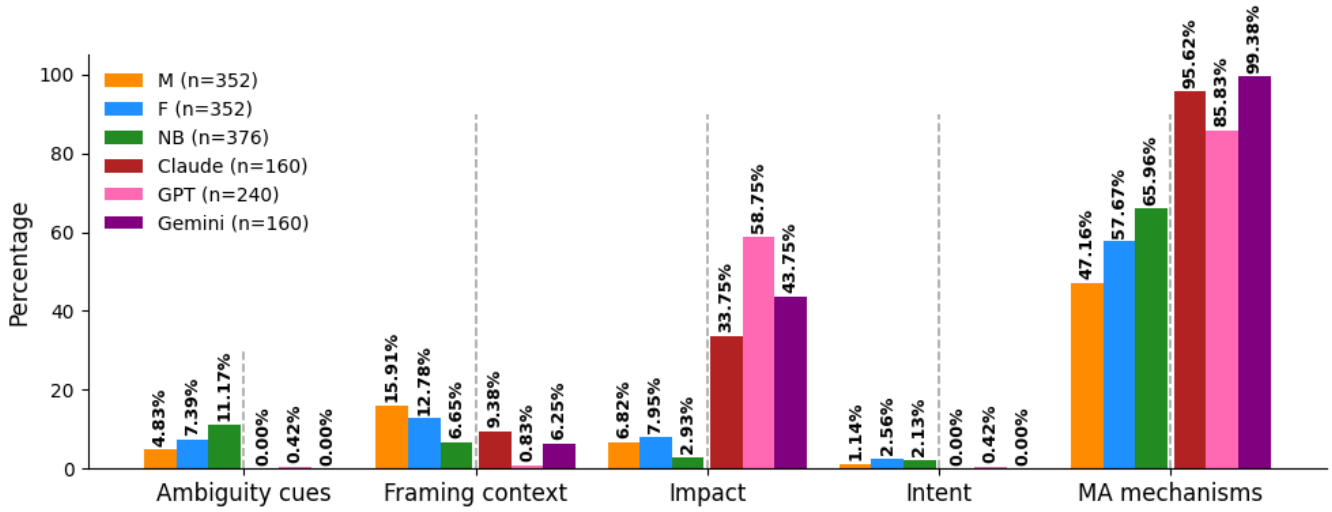


Figure 5: Argumentation elements by gender and model family, showing non-binary participants use more ambiguity cues and microaggression mechanisms while LLM families demonstrate uniform high mechanism usage

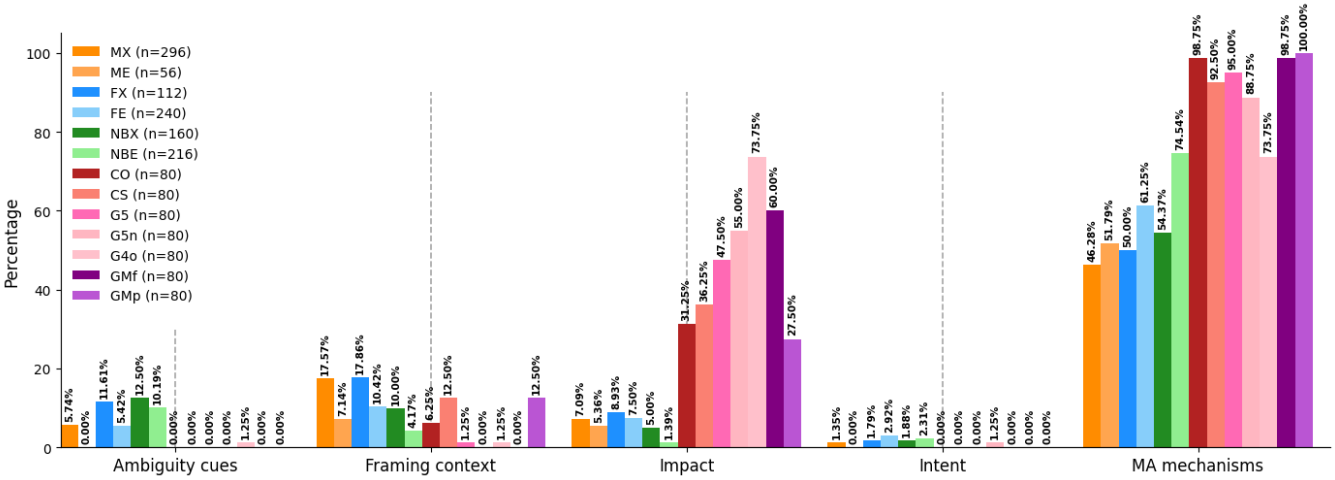


Figure 6: Argumentation styles across all participant types, revealing systematic patterns from human interpretive diversity to uniform LLM mechanism-focused explanations

kind of see the possibility of him thinking she looks really stern and scary" (NBE; Sc7). In that line, it is worth mentioning that these anchors often make tensions visible in participants' reasoning. For instance: *"Probably yes, but honestly valid. I call everyone a bitch too. Not good for the workplace but if they were friends I'd say not a microaggression"* (NBE; Sc4). Or: *"It's certainly rude and would make me angry if someone said it to me, I am unsure because I wouldn't be able to state whether this is gender-based. It depends on the profession of the people speaking"* (NBX; Sc6).

Together, these examples illustrate how participants incorporated context and ambiguity into their interpretations by anchoring them in relational, normative, or contextual cues. Rather than treating the scenarios as self-evident, human rationales often account for uncertainty and conditionality. The rationales were not centered on the linguistic content of the dialogues, but how situated context defines meaning.

4.3.2 Centering Harm in LLM Rationales. The second theme appeared in LLM rationales, which tended to be centered around harm framed at a generalized level. While models sometimes referenced parts of the dialogue, these were not central for explaining their rationales. Instead, the rationales positioned harm itself as the main explanation for why a comment was or was not considered a microaggression. For example, one rationale explained: *"It targets a woman with a gendered insult; the reply tries to normalize it, reinforcing hostility and misrecognition"* (G5n; Sc4).

Table 9: Stance frequencies across all participant types, demonstrating the full spectrum from conservative human interpretations to uniform LLM agreement

Rater	n	Yes-MA	5-4	No stance	3	No-MA	2-1
MX	296	74%	72%	6%	8%	20%	20%
ME	56	86%	86%	2%	4%	12%	11%
FX	112	71%	69%	8%	12%	21%	20%
FE	240	85%	85%	3%	5%	12%	11%
NBX	160	79%	80%	8%	9%	13%	11%
NBE	216	93%	91%	6%	6%	1%	3%
CO	80	100%	100%	0%	0%	0%	0%
CS	80	100%	100%	0%	0%	0%	0%
G5	80	100%	100%	0%	0%	0%	0%
G5n	80	100%	100%	0%	0%	0%	0%
G4o	80	100%	100%	0%	0%	0%	0%
GMf	80	100%	100%	0%	0%	0%	0%
GMp	80	100%	100%	0%	0%	0%	0%

Another noted: *"The man dismisses her programming skills as unearned luck due to diversity hiring, implying she's not qualified and reinforcing harmful stereotypes"* (CS; Sc8). These rationales can be seen as categorical statements, for instance, *"Labeling her by gender and dismissing her discomfort signals bias; it misrecognizes her feelings and reinforces marginalization"* (G5; Sc1). Although some rationales sometimes pointed to elements of the dialogue, these details were not used to situate the interpretation in context, but rather to support broader claims about harm, as in *"The man highlights gender unnecessarily ('female mechanic') and dismisses her discomfort ('too easily offended'), reinforcing stereotypes and invalidating her feelings"* (Gmf; Sc1).

These generalizations often manifested in two ways that often overlapped. First, many statements described harm in broad terms, without anchoring it in the specifics of the interaction, or to a specific target (individual or group). For example, *"The man's comment sexualizes the woman, implying her appearance, not professional merit, would expedite a task. This reinforces harmful gender stereotypes"* (Gmf; Sc7). Second, some rationales invoked harms that sounded correct in general, but did not clearly align with the scenario at hand. For example, *"Focuses on gender stereotypes (smile/warmth) over performance, undermining credibility"* (G5; Sc6), *"The feedback dismisses performance to demand gendered emotional labor (smiling, warmth), reinforcing stereotypes."* (Gmf; Sc6), or *"reinforcing unequal power, implying her work depends on appearance"* (G5; Sc7). In these cases, the harms identified (undermined credibility, demands of gendered emotional labor, and unequal power) appear detached from the immediate context of the vignette.

The interpretive significance of this theme is that LLMs frame their rationales by centering and abstracting harm. They drew on generalized categories such as "reinforcing stereotypes" or "undermining autonomy", and expressed these as definitive statements. In doing so, the rationales moved beyond the particulars of the dialogues and projected harms at the level of accumulated or patterned interactions.

5 Discussion

Our findings reveal a fundamental tension in how microaggressions are interpreted: LLMs consistently assign high ratings (4.7-5.0) with minimal variance, while humans show substantially more interpretive diversity that systematically varies by gender identity and lived experience. This pattern suggests that LLMs' uniform responses miss the contextual anchoring and situated knowledge that human participants demonstrated. LLMs' high ratings appear similar to the higher ratings from participants with lived experience (NBE), but this apparent alignment masks important differences in their rationales that have significant implications for automated detection systems. We discuss the significance of these results and their implications for the design of automated systems for microaggression detection, adopting a feminist lens.

Additionally, the systematic differences between human subgroups constitute significant findings in relation to the interpretative capabilities of LLMs. The fact that MX participants consistently rated scenarios lowest while NBE participants rated them highest illustrates standpoint theory [40]: those with lived experience of marginalization develop epistemic advantages in recognizing subtle discrimination. While LLMs achieve high ratings providing generalized rationales, in the next sections we raise concerns about how automated systems might override the interpretive frameworks of humans without providing the situated knowledge that makes sensitivity meaningful.

Finally, building on our findings and considerations, we provide design implications for the future development and adoption of LLMs for automated detection and interpretation of microaggression in the workplace.

5.1 High ratings without situatedness do not equal "greater sensitivity"

Safety-oriented alignment and harm-aversion training appear to push models toward conservative thresholds that prioritize avoiding false negatives (missing harm) over false positives [39, 42, 57, 63]. This produces high ratings even in scenarios where humans disagree because they are particularly ambiguous. In our data, LLMs' high means ($4.7-5.0$) occur with minimal variance, whereas human means are lower ($3.04-4.73$) with substantially larger spread. We observed that the magnitude of the difference between human and LLM ratings was not uniform; rather, it scaled with the ambiguity of the scenario. In straightforward scenarios like S1, where the insult is overt, the effect size was small ($r = -0.19$), indicating that humans and LLMs largely agreed on the presence of a microaggression, differing mostly in certainty. However, in highly context-dependent scenarios like S6 ("Smile more"), we observed a large effect size ($r = -0.54$). Humans may have downgraded their ratings due to contextual ambiguity, whereas LLMs maintained maximum ratings because the text might have triggered a "gender stereotyping" rule. The large effect size does not just show disagreement; it quantifies the models' inability to account for the situational nuance that significantly lowered human ratings.

Small effect sizes in straightforward scenarios (S1, S5) indicate that humans and LLMs can reach similar conclusions when microaggressions are overt, differing mainly in certainty. The medium-to-large effects in ambiguous scenarios (S6: $r = -0.54$; S8: $r = -0.39$)

reveal where categorical and situated sensitivity fundamentally diverge. These cases are precisely when contextual judgment is required.

Notably, we initially expected LLMs to align more closely with dominant groups' perspectives (particularly male participants without lived experience), based on research showing AI systems often reflect viewpoints of privileged cultures [71, 72]. However, findings on demographic alignment are mixed: some studies find alignment with Western, white, or younger populations [71, 74], while others suggest models are closer to uniform distributions than to any specific population [30]. When certain demographic groups appear "better represented" by LLMs, this may simply reflect which groups happen to have more uniform response patterns. We caution against interpreting higher ratings as "greater sensitivity". To clarify why, we propose a distinction between two types of "sensitivity":

Categorical sensitivity Involves consistently interpreting gender microaggressions through identifying whether interactions fit into predefined harm categories (such as "reinforcing stereotypes" or "undermining competence") and applying these labels consistently across similar cases, regardless of specific contextual factors.

Situated sensitivity emerges from lived experience and contextual understanding of how microaggressions operate within specific social and organizational dynamics. Human participants with lived experience demonstrated this by grounding their rationales in specific relational and situational factors.

LLMs' categorical sensitivity provides high rates for microaggression detection. However, their rationales' detachment from lived experiences makes these detections less contextually grounded and actionable. In contrast, human situated sensitivity correlates with gender identity and lived experience (NBE participants rating highest), consistent with literature showing that rater identity significantly affects annotation patterns [27, 36]. Perspectivist approaches foreground whose interpretations are centered in annotation practices [18, 33]. Our distinction between categorical and situated sensitivity complements this by characterizing *what* grounds interpretation.

The distinction between categorical sensitivity and situated sensitivity has critical implications. When participants provided rationales like *"This seems like a dialogue where men are acknowledging that they regularly engage in inappropriate workplace conversation. However, because no bystanders are mentioned, is this really a microaggression if it is not aimed at a specific target?" (NBE; Sc2)* they demonstrated situated sensitivity by using contextual anchors to articulate boundaries for what they believe constitutes a microaggression in that specific context. When LLMs provided rationales like *"The feedback to "smile more, be more warm" despite meeting all targets reinforces gender stereotypes about women needing to be pleasant/nurturing." (CS; Sc6)*, they demonstrated categorical sensitivity by identifying harm through general patterns. We are not claiming humans are universally "better" interpreters. Rather, we argue that some humans (particularly those with relevant lived experience) provide more grounded reasoning that reflects a richer understanding of context-specific microaggressions. This aligns with feminist standpoint theory's claim that marginalized positions offer epistemic advantages [40].

5.2 The Challenge of Interpretive Plurality in Automated Systems

Ambiguity is intrinsic to gender microaggressions [48, 69, 78]. Our results revealed that only humans surfaced ambiguity cues and constructed meaning around specific "anchors" (contextual elements like speaker relationships, organizational dynamics, presence of bystanders, and dialogue-specific details), while LLMs centered on harm, detaching their rationales from context, lacking nuanced reasons for their interpretations, and thus flattening ambiguity. While Aoyagui et al. [4] found that earlier GPT models (3.5 and 4) acknowledged some ambiguity through context-dependent responses, their examples reveal generalized rather than situated reasoning. For instance, responses like "it depends on the context it's used in" (GPT-4) acknowledge contextual variation without specifying the particular anchors that would shift interpretation. Our findings with newer models suggest this gap has widened, with current LLMs showing even less ambiguity acknowledgment (0.2% vs their reported 11-32%), further demonstrating how safety-oriented alignment in newer model generations may push models toward categorical identification and higher rates of detection.

The mismatch between how humans and LLMs approach interpretive challenges has significant implications. First, systems that flatten interpretive plurality may miss the nuanced reasoning that makes human sensitivity meaningful. Second, deploying LLMs to detect and explain microaggressions risks creating over-reliance on algorithmic judgments [7, 13, 31], for example by dismissing human interpretations as "subjective" while treating AI outputs as "objective", losing the situated knowledge that enables meaningful sensitivity, and increasing confidence in AI explanations without improving actual understanding of workplace dynamics.

Finally, when LLM systems are positioned as authoritative interpreters of workplace interactions, the focus shifts from supporting humans in articulating their experiences to replacing their judgment with algorithmic verdicts. This risks epistemic injustice in two forms: testimonial injustice occurs when targets' or even perpetrators' interpretations are dismissed in favor of "objective" algorithmic assessments, while hermeneutical injustice emerges when LLMs' generalized harm rationales replace the situated knowledge frameworks that targets use to make sense of their experiences [34].

Prior HCI work cautions that LLM explanations can increase reliance without improving awareness [7]. Our findings suggest this risk may be particularly acute for microaggression detection, where LLMs' objective-sounding but contextually detached outputs could create the illusion of understanding why particular interactions constitute microaggressions, without real awareness of the contextual factors that influence microaggression perception (such as power relationships, organizational norms, and interpersonal dynamics).

With *situated sensitivity*, naming what would "tip the scales" (e.g., intent, power asymmetry, organizational norms, the relationship between actors) creates the possibility for awareness in humans. If

LLMs could provide insights on what contextual factors are anchoring different possible judgments, they could help users consider alternative interpretations or expand their interpretive framework to encompass multiple viewpoints. But generalizations, by lacking plurality and conveying a seemingly objective truth, do not provide the same opportunity. Moreover, people tend to consider AI recommendations only when they align with their expectations and beliefs, not when they are "correct" or "fair" [86]. These findings suggest that if automated systems are to meaningfully support workplace interactions around microaggressions, they must account for interpretive plurality rather than flattening it.

A natural response to these challenges is to ask whether situated sensitivity could be approximated through context-aware models. However, we caution against framing this as an optimization problem. Situated sensitivity emerges from being positioned within social structures and developing interpretive frameworks through lived experience. Models exhibit what Cabitza et al. [18] would characterize as *weak perspectivism*, meaning that at best, LLMs may have been trained on diverse human judgments, but their outputs still reflect aggregated patterns that flatten interpretive plurality into categorical responses. A model can be made context-sensitive (responsive to contextual features), but not genuinely situated (interpreting from a social position). Moreover, current models cannot predict when demographic groups will differ in their perceptions of harm (i.e., find it more harmful than others) [64]. Context sensitivity should not be mistaken for the interpretive grounding that makes human sensitivity meaningful.

This limitation would likely extend to purpose-built systems fine-tuned for microaggression detection (see Section 5.3.4). While such systems might produce outputs that appear more contextually grounded, fine-tuning on particular perspectives would risk encoding those perspectives as a new ground truth [33], but still flattening interpretive plurality.

5.3 Design implications: Critical and Technical Considerations

In this section, we provide implications for designing systems that use LLMs for automatic detection of workplace gender microaggressions. Our findings reveal fundamental tensions between how humans and LLMs interpret microaggressions, raising both critical questions about whether such systems should exist and technical considerations for how they might be designed if deployed. Crucially, our critique of LLMs does not imply that humans provide an objective alternative. Human interpretations are themselves situated and partial, so if neither LLMs nor any single human perspective can serve as an objective arbiter, the question becomes what role technology should play. We argue the goal is not to find a more "accurate" judge, but to support interpretive dialogue that preserves plurality. Our approach aligns with feminist HCI's emphasis on pluralism, embodiment, and situated knowledge as foundations for design [9].

5.3.1 Questioning the Need for Automated Detection of Microaggressions. There is no singular, context-free ground truth for whether a given interaction "counts" as a microaggression [68, 69, 78, 81]; interpretations are value-laden and socially situated. Our findings show that even when LLMs appear to align with marginalized groups' ratings, their interpretations fundamentally differ from the situated knowledge that makes human sensitivity meaningful. Positioning automated systems as arbiters of truth risks policing speech and disproportionately burdening marginalized groups, as seen in pipelines that over-flag some marginalized communities' speech [42] and in ableism detection systems that make incorrect assumptions and provide explanations people with disabilities found "judgmental instead of educational" [67].

The workplace is already a site where power imbalances affect people [49]. Designers should first question why automated detection is required, who would benefit, and who might be harmed, aligning with feminist principles that interrogate technological solutionism [29, 51] and interpret technology as a medium of power [22].

This concern is particularly acute because microaggressions already involve epistemic injustice [34], as targets often struggle to have their experiences believed or validated within organizational hierarchies. As we established through feminist standpoint theory[40] and empirical evidence showing women are more likely than men to recognize non-explicit microaggressions [10], those who experience discrimination develop nuanced interpretive frameworks for understanding subtle bias. However, when automated systems are positioned as authoritative arbiters of what constitutes a microaggression, they risk institutionalizing the dismissal of these experiential perspectives in favor of algorithmic "objectivity". This could deepen existing power imbalances by replacing the situated knowledge that enables meaningful sensitivity with categorical judgments that lack contextual grounding. Therefore, we caution against adopting LLMs for microaggression detection in the workplace, while proposing alternative applications that go beyond mere detection to support situated self-reflection.

5.3.2 Supporting Self-Reflection Through Interpretive Dialogue. Our findings reveal that meaningful microaggression sensitivity emerges from situated reasoning that acknowledges contextual anchors and interpretive uncertainty. Rather than replacing human judgment with algorithmic verdicts, systems should support users in articulating and examining their own interpretive frameworks while exposing them to alternative perspectives. This approach builds directly on our finding that human participants with lived experience demonstrated contextually grounded reasoning by identifying specific contextual elements that affect their interpretations.

When LLM systems flag potential microaggressions, they could prompt reflection rather than simply providing verdicts. Instead of categorical statements about harm, systems could invite users to consider the contextual anchors required for meaningful interpretation, such as speaker relationships, organizational norms, and interaction history. Additionally, systems could frame their rationales as viewpoints from different positionalities, preserving interpretive plurality while expanding users' awareness that participants with different backgrounds often interpreted similar interactions differently. This approach would encourage the kind of situated reasoning we observed in participants with lived experience, while avoiding the risk of algorithmic verdicts overriding human interpretive frameworks.

5.3.3 Meeting Users' Interpretive Starting Points. Research on confirmation bias in AI-assisted decision-making demonstrates that

users tend to override AI recommendations when the system's implicit values conflict with their own perspectives [86], revealing a fundamental design challenge for microaggression detection systems. While some recent work has attempted to personalize explanations to individual users [47], such approaches still aim to achieve interpretive consensus rather than recognizing that disagreement itself may be meaningful and legitimate. Rather than treating disagreement among users as noise to be filtered out, it should be recognized that effective AI intervention requires understanding and working with users' interpretive frameworks. The systematic differences we found in how people interpret microaggressions based on gender identity and lived experience are legitimate starting points that systems should account for. This could involve adaptive explanation strategies that recognize users' gender identity and lived experience as relevant contextual factors, and recommendation systems that can present alternative perspectives without invalidating users' initial assessments. Systems should be designed to expand rather than override interpretive frameworks, recognizing that the goal is to foster dialogue rather than enforce consensus.

5.3.4 *Technical Recommendations: If Such Systems Are Built.*

Rationale Design for Situated Reasoning. Building on our finding that meaningful sensitivity involves contextual anchoring, systems should move beyond generalized harm rationales toward contextual specificity [4]. This requires technical approaches that surface interpretive dependencies rather than flatten them. Prompt engineering could explicitly request situated reasoning by asking models to identify contextual elements that would alter interpretations.

Training objectives could aim to move beyond categorical sensitivity toward more contextually grounded outputs, though as noted in Section 5.2, this remains a partial approximation rather than genuine situated sensitivity. Systems could be designed to present contextual factors (such as speaker relationships, organizational dynamics, or interaction setting), allowing users to explore how different contextual assumptions lead to different interpretive outcomes.

Prior work offers some support for contextual approaches: Kumar et al. [54] found that including conversational context corrected 35% of LLM errors in moderation tasks. This suggests that the contextual anchoring we observed in human rationales could potentially be approximated through prompt design. However, as we noted in Section 5.2, context-sensitivity is not the same as situated sensitivity. Providing more information may improve outputs without producing genuinely situated reasoning. Additionally, Jiang et al. [45] found that overloading prompts with additional information (such as annotator demographics) can decrease model performance, suggesting that pre-trained models may lack perspectivist foundations that cannot be easily added through prompting alone. It is important to also note that collecting such information in workplace settings raises significant privacy and surveillance concerns, and needs to account for power dynamics, tying back to our critique in Section 5.3.1.

Implementation Considerations. The aforementioned approaches could be embedded through multiple technical pathways. Fine-tuning could incorporate diverse perspective datasets, treating interpretive stance as a learnable parameter following Aoyagui et al.'s work [4]. For embedding ambiguity and plurality of perspectives [76], this could involve setting LLMs training objectives that preserve interpretive disagreement rather than forcing one-sided perspectives. Additional approaches include prompt engineering templates that explicitly request multiple plausible interpretations with uncertainty ranges, and output formats that present contextual factors and interpretive alternatives rather than single categorical judgments. Ultimately, these technical approaches should support human agency and situated reflection rather than automating judgment, recognizing that the goal is fostering awareness rather than enforcing conformity to algorithmic verdicts.

5.4 Limitations

Our findings are bounded by specific model versions and prompt design choices. Different model versions, temperature settings, or prompt formulations could yield different patterns of interpretation. LLM rationales represent stochastic text generation [12] rather than deliberative reasoning processes, limiting claims about models' actual "understanding" of microaggressions. Also, our observations about differences between model generations should be interpreted cautiously, as these differences could result from various factors including changes in training data, alignment procedures, or architectural modifications rather than representing systematic trends in model capabilities.

Our study design may have influenced the patterns we observed in LLM responses. The working definition provided to both humans and LLMs explicitly framed microaggressions in terms of cumulative harm, marginalization, and misrecognition (see Appendix C). Prior work has documented LLM sensitivity to prompt wording [56, 75], thus LLM rationales likely reflect this framing. It is possible that alternative prompt formulations would produce different LLM behaviors. Understanding to what extent prompt modifications alter the manifestation of categorical sensitivity was outside the scope of this study. Yet under identical conditions, humans and LLMs produced systematically different response patterns. This asymmetry has implications for the design of LLM-based systems that aim to automatically detect microaggressions: while we cannot exclude the possibility that LLMs could approximate situated reasoning, such behavior would require explicit programming, for example, through mechanisms such as fine-tuning that accounts for [4], with diverse annotation schemes [18], and involving annotators with lived experience [33].

The scenarios were adapted from self-reported experiences on a specific platform, which may not represent the full range of workplace microaggressions across different organizational cultures and contexts. Additionally, subgroup analysis is limited by uneven sample sizes, particularly for males with lived experience (n=7). While this small sample size limits the generalizability of findings specific to this group, the patterns observed align with theoretical expectations and provide valuable preliminary insights.

We captured lived experience as a binary variable (with/without) rather than collecting frequency or severity data. While this limits

the granularity of our analysis, this reflects our view that lived experience provides epistemic resources, without hierarchizing experiences based on intensity. Future work could explore whether frequency or intensity of lived experience correlates with interpretive patterns.

Our participants were recruited via Prolific without stratifying by geographic location or cultural background. Because gender norms and microaggression interpretation can vary across cultures, this limits the generalizability of our findings. Recent work suggests regional variation in sensitivity to offensive language [27], though the same study found that participants' moral values play a more important role than geo-cultural distinctions in shaping perceptions of offensiveness. We note that demographic variables do not always predict annotation behavior [11, 45, 66], and meaningful cross-cultural comparison would require theoretical grounding beyond demographic proxies. Future work could explore how cultural context and moral values jointly shape microaggression interpretation.

This study focuses exclusively on gender-based microaggressions, excluding intersectional cases where gender intersects with race, ethnicity, age, class, disability, or sexuality. This decision was made for analytic tractability and scope management, but it represents a substantial limitation. Our findings may not generalize to more complex intersectional dynamics, and the scenarios we studied likely underrepresent the full spectrum of workplace discrimination experiences.

6 Conclusion

This study reveals three key findings on human and LLM interpretations of workplace gender microaggressions. First, LLMs consistently assign high ratings with minimal variance, while humans show substantial interpretive diversity that varies systematically by gender identity and lived experience. Second, humans employed diverse argumentation strategies including ambiguity acknowledgment and contextual framing, while LLMs predominantly relied on mechanism-based explanations that focus on categorical harm identification. Third, we identified two distinct reasoning approaches: humans demonstrated situated sensitivity that acknowledges interpretive uncertainty and grounds judgments in specific contextual elements, whereas LLMs provided more generalized harm-centered explanations that were detached from dialogue specifics. These systematic differences support feminist standpoint theory's claim that marginalized positions offer epistemic advantages in recognizing subtle discrimination.

These findings challenge the assumption that higher detection rates indicate greater sensitivity. We distinguish between LLMs' *categorical sensitivity*, and humans' *situated sensitivity*, which is contextually grounded and acknowledges interpretive uncertainty. This distinction matters because meaningful microaggression detection requires understanding not just whether harm occurred, but how contextual factors shape interpretation.

The implications are significant for automated detection systems. Rather than treating interpretive disagreement as an error, systems should preserve and leverage interpretive plurality. If deployed, such systems must account for situated knowledge while avoiding algorithmic verdicts that override human interpretive frameworks. This work advances critical computing methodology by moving beyond accuracy metrics to examine interpretive positioning in contested social phenomena. Our mixed-methods approach, pairing numerical ratings with rationale analysis, offers a possibility for studying AI interpretation of subjective tasks. We contribute the distinction between categorical and situated sensitivity as a framework for evaluating systems handling ambiguous social interactions, while demonstrating how feminist standpoint theory can inform the design of automated detection systems that preserve interpretive plurality.

Acknowledgments

First and last author are part of the Feminist Generative AI Lab, funded by the Convergence AI, Data & Digitalisation Programme. We thank our survey participants, and the reviewers for their thoughtful and insightful feedback, which helped shape this work. We also thank our colleagues for their valuable insights on earlier drafts of this manuscript.

References

[1] Herman Aguinis and Kyle J. Bradley. 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods* 17, 4 (8 2014), 351–371. doi:10.1177/1094428114547952

[2] Mona Algner and Timo Lorenz. 2022. You're Prettier When You Smile: Construction and Validation of a Questionnaire to Assess Microaggressions Against Women in the Workplace. *Frontiers in Psychology* 13 (3 2022). doi:10.3389/fpsyg.2022.809862

[3] Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. 2020. Automated Detection of Racial Microaggressions using Machine Learning. *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (12 2020), 2477–2484. doi:10.1109/ssci47803.2020.9308569

[4] Paula Akemi Aoyagui, Kelsey Stemmler, Sharon A Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A Matter of Perspective(s): Contrasting Human and LLM Argumentation in Subjective Decision-Making on Subtle Sexism. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 529, 16 pages. doi:10.1145/3706598.3713248

[5] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology* 6, 3 (1 2010), 128–138. doi:10.1027/1614-2241/a000014

[6] Christiane Atzmüller, Dan Su, and Peter Steiner. 2017. Designing valid and reliable vignette experiments for survey Research: A case study on the fair gender income gap. *Journal of Methods and Measurement in the Social Sciences* 7, 2 (6 2017). doi:10.2458/v7i2.20321

[7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[8] Matthieu Bardal, Kaleb Chisholm, Sean Jamieson, Mark Alwast, Kasi Viswanath Nilla, Khanh Le, and Yasaman Amannejad. 2023. Automated Identification of Microaggressions. In *2023 IEEE International Humanitarian Technology Conference (IHTC)*. IEEE, Santa Marta, Colombia, 1–6. doi:10.1109/IHTC58960.2023.10508881

[9] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1301–1310. doi:10.1145/1753326.1753521

[10] Tessa E. Basford, Lynn R. Offermann, and Tara S. Behrend. 2013. Do you see what I see? Perceptions of gender microaggressions in the workplace. *Psychology of Women Quarterly* 38, 3 (11 2013), 340–349. doi:10.1177/0361684313511420

[11] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2589–2615. doi:10.18653/v1/2024.eacl-long.159

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability,*

and Transparency (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[13] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 905, 23 pages. doi:10.1145/3706598.3714097

[14] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806 arXiv:https://doi.org/10.1080/2159676X.2019.1628806

[15] Virginia Braun and Victoria Clarke. 2020. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research* 21, 1 (10 2020), 37–47. doi:10.1002/capr.12360

[16] Virginia Braun and Victoria Clarke. 2021. *Thematic analysis: a Practical Guide*. Sage Publications Limited.

[17] Virginia Braun and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *International Journal of Transgender Health* 24, 1 (2023), 1–6. doi:10.1080/26895269.2022.2129597 arXiv:https://doi.org/10.1080/26895269.2022.2129597

[18] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 771, 9 pages. doi:10.1609/aaai.v37i6.25840

[19] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 572, 19 pages. doi:10.1145/3491102.3501998

[20] Christina M. Capodilupo, Kevin L. Nadal, Lindsay Corman, Sahran Hamit, Oliver B. Lyons, and Alexa Weinberg. 2010. The Manifestation of Gender Microaggressions. In *Microaggressions and Marginality: Manifestation, Dynamics, and Impact*, Derald Wing Sue (Ed.). John Wiley & Sons, Hoboken, NJ, 193–216.

[21] Feng Chen, Manas Satish Bedmutha, Ray-Yuan Chung, Janice Sabin, Wanda Pratt, Brian R. Wood, Nadir Weibel, Andrea L. Hartzler, and Trevor Cohen. 2024. Toward Automated Detection of Biased Social Signals from the Content of Clinical Conversations. *PubMed* 2024 (7 2024), 252–261. doi:10.48550/arxiv.2407.17477

[22] Cynthia Cockburn. 1988. *Machinery of dominance*.

[23] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

[24] Lilia M. Cortina, Dana Kabat-Farr, Emily A. Leskinen, Marisela Huerta, and Vicki J. Magley. 2011. Selective incivility as modern discrimination in organizations. *Journal of Management* 39, 6 (9 2011), 1579–1605. doi:10.1177/0149206311418835

[25] Terri A. Croteau and Todd G. Morrison. 2022. Development of the nonbinary gender microaggressions (NBGM) scale. *International Journal of Transgender Health* 24, 4 (2 2022), 417–435. doi:10.1080/26895269.2022.2039339

[26] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback: Helpful, harmless, honest? Sociotechnical limits of AI alignment... *Ethics and Inf. Technol.* 27, 2 (June 2025), 13 pages. doi:10.1007/s10676-025-09837-2

[27] Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2007–2021. doi:10.1145/3630106.3659021

[28] Priom Deb, Asibur Rahman Bhuiyan, Habiba Mahrin, Marzanul Momenine, Md. Saharan Evan, Md. Sajid Ullah Sohan, and Md. Tazfiq Khan. 2024. Evaluating Online Sexism Detection: A Comparative Study of Machine Learning Models using the EDOS Dataset. *2022 IEEE 7th International conference for Convergence in Technology (I2CT)* (4 2024). doi:10.1109/i2ct61223.2024.10543680

[29] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism*. doi:10.7551/mitpress/11805.001.0001

[30] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 1458, 29 pages.

[31] Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted Decision-Making. *ArXiv* abs/2402.17385 (2024). https://api.semanticscholar.org/CorpusID:268032159

[32] Hayden Field. 2024. How Walmart, Delta, Chevron and Starbucks are using AI to monitor employee messages. https://www.cnbc.com/2024/02/09/ai-might-be-reading-your-slack-teams-messages-using-tech-from-aware.html. CNBC.

[33] Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2279–2292. doi:10.18653/v1/2024.naacl-long.126

[34] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. doi:10.1093/acprof:oso/9780198237907.001.0001

[35] Rachel E. Gartner. 2021. A New Gender Microaggressions Taxonomy for Undergraduate Women on College Campuses: A Qualitative Examination. *Violence Against Women* 27, 14 (1 2021), 2768–2790. doi:10.1177/1077801220978804

[36] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (11 2022), 1–28. doi:10.1145/3555088

[37] Dylan Grosz and Patricia Conde-Céspedes. 2020. Automatic Detection of Sexist Statements Commonly Used at the Workplace. In *Trends and Applications in Knowledge Discovery and Data Mining*, Wei Lu and Kenny Q. Zhu (Eds.). Lecture Notes in Computer Science, Vol. 12237. Springer International Publishing, Cham, 104–115. doi:10.1007/978-3-030-60470-7_11

[38] Uma Sushmitha Gunturi, Anisha Kumar, Xiaohan Ding, and Eugenia H. Rho. 2024. Linguistically differentiating acts and recalls of racial microaggressions on social media. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (4 2024), 1–36. doi:10.1145/3637366

[39] Rajdeep Haldar, Ziyi Wang, Qifan Song, Guang Lin, and Yue Xing. 2025. LLM Safety Alignment is Divergence Estimation in Disguise. *ArXiv* abs/2502.00657 (2025). https://api.semanticscholar.org/CorpusID:276095181

[40] Donna Haraway. 1988. Situated Knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Studies* 14, 3 (1 1988), 575. doi:10.2307/3178066

[41] Sandra Harding. 1992. After the neutrality ideal: science, politics, and "Strong objectivity". *Social Research: An International Quarterly* 59 (1 1992). https://philpapers.org/rec/HARATN

[42] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 175, 26 pages. doi:10.1145/3706598.3713998

[43] Michael R Harwell. 1988. Choosing between parametric and nonparametric tests. *Journal of Counseling & Development* 67, 1 (1988), 35–38.

[44] Tao Huang. 2025. Content moderation by LLM: from accuracy to legitimacy. *Artificial Intelligence Review* 58 (07 2025). doi:10.1007/s10462-025-11328-1

[45] Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining Sexism and Misogyny Classification with Annotator Attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15103–15125. doi:10.18653/v1/2024.findings-emnlp.887

[46] Veronica E. Johnson, Kevin L. Nadal, D. R. Gina Sissoko, and Rukiya King. 2021. "It's Not in Your Head": Gaslighting, 'Splaining, Victim Blaming, and Other Harmful Reactions to Microaggressions. *Perspectives on Psychological Science* 16, 5 (9 2021), 1024–1036. doi:10.1177/17456916211011963

[47] Hyojin Ju, Jungeun Lee, Seungwon Yang, Jungseul Ok, and Inseok Hwang. 2025. Toward Affective Empathy via Personalized Analogy Generation: A Case Study on Microaggression. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 379, 31 pages. doi:10.1145/3706598.3714122

[48] Jennifer Y. Kim and Alyson Meister. 2022. Microaggressions, Interrupted: The experience and effects of gender microaggressions for women in STEM. *Journal of Business Ethics* 185, 3 (8 2022), 513–531. doi:10.1007/s10551-022-05203-0

[49] Jennifer Young-Jin Kim, Duoc V. Nguyen, and Caryn J. Block. 2018. The 360-Degree Experience of Workplace Microaggressions: Who Commits Them? How Do Individuals Respond? What Are the Consequences? In *Microaggression Theory: Influence and Implications*, Gina C. Torino, David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue (Eds.). John Wiley & Sons, Hoboken, NJ, 157–177. doi:10.1002/9781119466642.ch10

[50] Bruce M King, Patrick J Rosopa, and Edward W Minium. 2018. *Statistical reasoning in the behavioral sciences*. John Wiley & Sons.

[51] Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 100–112. doi:10.1145/3630106.3658543

[52] Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New

York, NY, USA, 1433–1445. doi:10.1145/3630106.3658981

[53] Alexis Krivkovich, Emily Field, Lareina Yee, Megan McConnell, and Hannah Smith. 2024. *Women in the Workplace 2024: The 10th Anniversary Report.* Technical Report. McKinsey & Company and Lean In. https://www.mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace#/ Accessed July 31, 2025.

[54] Deepak Kumar, Yousef Abuhashem, and Zakir Durumeric. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (05 2024), 865–878. doi:10.1609/icwsm.v18i1.31358

[55] Jioni A. Lewis, Marlene G. Williams, Anahvia T. Moody, Erica J. Peppers, and Cecile A. Gadson. 2019. Intersectionality Theory and Microaggressions: Implications for Research, Teaching, and Practice. In *Microaggression Theory: Influence and Implications*, Gina C. Torino, David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue (Eds.). John Wiley & Sons, Hoboken, NJ, 48–64.

[56] Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are Prompts Different in Terms of Sensitivity?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5833–5856. doi:10.18653/v1/2024.naacl-long.325

[57] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 660–686. doi:10.1145/3630106.3658932

[58] Cass Mayeda, Arinjay Singh, Arnav Mahale, Laila Shereen Sakr, and Mai ElSherief. 2025. Applying Data Feminism Principles to Assess Bias in English and Arabic NLP Research. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1769–1792. doi:10.1145/3715275.3732119

[59] Emma McClure and Regina Rini. 2020. Microaggression: Conceptual and scientific issues. *Philosophy Compass* 15, 4 (April 2020), e12659. doi:10.1111/phc3.12659

[60] Alyson Meister, Amanda Sinclair, and Karen A. Jehn. 2017. Identities under scrutiny: How women leaders navigate feeling misidentified at work. *The Leadership Quarterly* 28, 5 (1 2017), 672–690. doi:10.1016/j.leaqua.2017.01.009

[61] Yara Mekawi and Nathan R. Todd. 2021. Focusing the lens to see more clearly: overcoming definitional challenges and identifying new directions in racial microaggressions research. *Perspectives on Psychological Science* 16, 5 (9 2021), 972–990. doi:10.1177/1745691621995181

[62] Berg Miller. 2025. Toward an empirically based definition of microaggression. *Journal of Ethnic & Cultural Diversity in Social Work* (4 2025), 1–12. doi:10.1080/15313204.2025.2495562

[63] Yutao Mou, Yuxiao Luo, Shikun Zhang, and Wei Ye. 2025. SaRO: Enhancing LLM Safety through Reasoning-based Alignment. *ArXiv* abs/2504.09420 (2025). https://api.semanticscholar.org/CorpusID:277780470

[64] Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9048–9062. doi:10.18653/v1/2024.emnlp-main.511

[65] Amy Nivette, Christof Nägel, and Andrada Stan. 2022. The use of experimental vignettes in studying police procedural justice: a systematic review. *Journal of Experimental Criminology* 20, 1 (8 2022), 151–186. doi:10.1007/s11292-022-09529-7

[66] Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The Ecological Fallacy in Annotation: Modeling Human Label Variation goes beyond Sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1017–1029. doi:10.18653/v1/2023.acl-short.88

[67] Mahika Phutane, Ananya Seelam, and Aditya Vashistha. 2025. "Cold, Calculated, and Condescending": How AI Identifies and Explains Ableism Compared to Disabled People. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1927–1941. doi:10.1145/3715275.3732128

[68] Regina Rini. 2018. How to Take Offense: Responding to Microaggression. *Journal of the American Philosophical Association* 4, 3 (2018), 332–351. doi:10.1017/apa.2018.23

[69] Regina Rini. 2020. *The Ethics of microaggression.* doi:10.4324/9781315195056

[70] Mary P. Rowe and Anna Giraldo-Kerr. 2017. Gender Microinequities. In *The SAGE Encyclopedia of Psychology and Gender*, Kevin L. Nadal (Ed.). SAGE Publications, Inc., Thousand Oaks, CA, 679–682. doi:10.4135/9781483384269.n

[71] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1244, 34 pages.

[72] Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. http://arxiv.org/abs/2306.01943 arXiv:2306.01943 [cs].

[73] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5884–5906. doi:10.18653/v1/2022.naacl-main.431

[74] Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wuehrl, Sabine Weber, and Roman Klinger. 2025. Which Demographics do LLMs Default to During Annotation?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 17331–17348. doi:10.18653/v1/2025.acl-long.848

[75] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *ArXiv* abs/2310.11324 (2023). https://api.semanticscholar.org/CorpusID:264172710

[76] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1882, 23 pages.

[77] Nicole M. Stephens, Lauren A. Rivera, and Sarah S.M. Townsend. 2020. The cycle of workplace bias and how to interrupt it. *Research in Organizational Behavior* 40 (1 2020), 100137. doi:10.1016/j.riob.2021.100137

[78] Wing Sue. [n. d.]. Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation. ([n. d.]).

[79] Asif Tareque, Harshith Hullakere Siddegowda, Denster Joseph Frank, Nicole Lee, and Rezza Moieni. 2024. Overview of machine learning algorithms for detecting microaggression in written text. *Open Journal of Social Sciences* 12, 07 (1 2024), 347–358. doi:10.4236/jss.2024.127025

[80] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. (2014).

[81] Gina C. Torino, David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue. 2019. Everything You Wanted to Know about Microaggressions but Didn't Get a Chance to Ask. In *Microaggression Theory: Influence and Implications*, Gina C. Torino, David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue (Eds.). John Wiley & Sons, Hoboken, NJ, 3–15.

[82] Malgi Nikitha Vivekananda, Vinayaka Vivekananda Malgi, and Prashant Ashok Shidlyali. 2025. Integrating Transformers and Hybrid Machine Learning Models for Automated Sexism Detection in Online Discourse. In *2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*. 1–6. doi:10.1109/AMATHE65477.2025.11081244

[83] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. doi:10.18653/v1/2024.acl-long.511

[84] Yang Yang and Doris Wright Carroll. 2018. Gendered Microaggressions in Science, Technology, Engineering, and Mathematics. *Leadership and Research in Education* 4, Special Issue (2018), 28–45.

[85] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2020, 29 pages.

[86] Domenique Zipperling, Luca Deck, Julia Lanzl, and Niklas Kühl. 2025. It's only fair when I think it's fair: How Gender Bias Alignment Undermines Distributive Fairness in Human-AI Collaboration. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1261–1274. doi:10.1145/3715275.3732084

A Expanded Taxonomy

Table 10: Expanded taxonomy of gender microaggressions combining Sue's [78] foundational framework with workplace-specific categories from Kim & Meister [48] and Gartner [35]

Subtype	Source	Definition	Example
Microassault — Overt, conscious bias or discriminatory act			
Sexual Objectification	Gartner	Being treated primarily as a sexual object or valued for appearance over intellect	Leering; comments on body/clothes; innuendo
Gendered Hostility	Self	Overt hostile behavior that explicitly targets someone's gender or gender expression	"You're just here for diversity points"; sexist jokes; "Women shouldn't be in tech"
Microinsult — Subtle snubs that convey bias or stereotyping (often unconscious).			
Undermining Competence *(Assumptions of Inferiority / Devaluation of Technical Competence)*	Kim & Meister / Gartner	Presuming women are less capable, especially intellectually or professionally; undermines identity as a skilled expert	Surprise at skill; explaining basic concepts; attributing work to male colleagues; assuming men are more competent
Pathologizing Character	Kim & Meister	Framing a woman's personality as the problem	"You're too assertive/aggressive"; judged more harshly in evaluations
Restrictive Gender Roles	Gartner	Expecting women to behave in stereotypically feminine or nurturing ways	Being told to smile; expected to take notes; criticized for being assertive
Gender as a Liability	Kim & Meister	Framing gender itself as a defect; treating gender identity or expression as incompatible with competence, leadership, or reliability	"You'll be distracted now that you have kids"; being seen as "the woman" on the team; unsolicited comments about starting a family
Microinvalidation — Negation of the thoughts, feelings, or experiential reality of a marginalized person.			
Dismissal *(Invalidation of Experience / Denial of One's Reality)*	Gartner	Dismissal or denial of one's gendered experiences or feelings; discrediting perceptions of sexism	"You're too sensitive"; "It's not that deep"; "I'm sure he didn't mean it like that"
Exclusion *(Devaluation of Presence)**	Kim & Meister (classified as microinsult)	Treating women as invisible or not legitimate; denying contribution or right to be heard	Ignoring women in meetings; excluding from decisions

B Selected Scenarios

Sc.	Label(s)	Dialogue
1	Dismissal + Gender as Liability	[Man]: "This is my best female mechanic!" [Woman]: "I've told you I don't like being introduced that way." [Man]: "You're too easily offended."
2	Exclusion	[Man 1]: "Oh great, we're getting a woman on the team." [Man 2]: "Yeah, guess we'll have to clean up our conversations now." [Man 1]: "Can't joke around like we used to."
3	Gender as Liability	[Woman]: "I don't plan on getting married or having children." [Man]: "Well, all the independent women I knew eventually settled down. Your instinct will kick in soon."

4	Gendered Hostility + Dismissal	[Woman]: "Calling her a bitch is not ok." [Man]: "When I call her a bitch, I'm not being sexist. I call men bitches too."
5	Pathologising Character + Exclusion	[Woman]: "I'd like to take the lead on this team project." [Man]: "No offense, but you're more of a mother than a team leader."
6	Restrictive Gender Roles	[Woman]: "I've had no customer complaints and met all my sales targets." [Woman]: "Yes, but you need to smile more be more warm. I expect more from you"
7	Sexual Objectification	[Woman]: "I still haven't received that memo, could you check with him?" [Man]: "Why don't you fax him a picture of yourself? I'm sure he'll send it faster."
8	Undermining Competence	[Man]: "Oh, you're a programmer? Wow, you're so lucky!" [Woman]: "Lucky?" [Man]: "Yeah, software companies are dying to hire women right now."
9	Control	[Man]: "Could you send me the updated file?" [Woman]: "Sure, I'll send it over after the meeting." [Man]: "Thanks!"
10	Control	[Woman]: "I've added the client notes to the shared document." [Man]: "Great. I'll review them this afternoon." [Woman]: "Let me know if anything's missing."

C LLM Prompt

C.1 System Definition

For this study, we define gender-based microaggressions as: subtle and often ambiguous forms of discrimination based on a person's gender. They are often normalized or dismissed as minor, but their harm is cumulative. It emerges through patterns of marginalization, misrecognition, and social exclusion over time. They may take the form of jokes, comments, or actions that convey hostility, reinforce stereotypes, or make individuals feel unsafe, undermined, excluded, or dismissed.

C.2 Question Prompt

A. Do you think this interaction contains a gender-based microaggression?
Respond with ONE digit only (1–5) on its own line. Where:
1 = Definitely not
2 = Probably not
3 = Unsure
4 = Probably yes
5 = Definitely yes

B. Why did you choose that answer?
Respond in 150 characters or fewer.

D Statistical Test Results

The following tables report full Kruskal–Wallis and Mann–Whitney test results, including omnibus effects across groups. Significance is evaluated at $\alpha = .05$.

Table 12: Kruskal–Wallis tests across Gender with Dunn–Bonferroni post-hoc comparisons.

	Kruskal–Wallis H	Post-hoc comparisons (Adj. p)		
Scenario	H	Male vs Female	Male vs Non-Binary	Female vs Non-Binary
S1	7.301*		*	
S2	6.023*			
S3	14.332***		***	
S4	0.389			
S5	1.187			
S6	20.592***	*	***	
S7	2.767			
S8	2.057			
S9	2.733			
S10	5.639			

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 13: Mann–Whitney tests comparing human vs. AI ratings.

Scenario	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p-value	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	.073	.014
Kruskal–Wallis H	14.84***	21.76***	23.43***	13.94***	12.78***	64.59***	19.61***	35.32***	3.21	6.04*

Table 14: Kruskal–Wallis tests across Model with Dunn–Bonferroni post-hoc comparisons (Scenarios S1–S10).

	Kruskal–Wallis H	Post-hoc comparisons (Adj. p)		
Scenario	H	GPT vs Claude	GPT vs Gemini	Claude vs Gemini
S1	5.576			
S2	5.576			
S3	5.576			
S4	12.830**		*	**
S5	0.000			
S6	53.833***	***		***
S7	0.000			
S8	34.275***	***	***	
S9	0.000			
S10	0.000			

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 15: Kruskal–Wallis tests across AI model versions.

Scenario	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p-value	.107	.028	.028	< .001	1.000	< .001	1.000	< .001	1.000	1.000
Kruskal–Wallis H	10.46	14.11*	14.11*	57.27***	.000	55.88***	.000	44.58	.000	.000

Table 16: Mann–Whitney tests for lived experience (with vs. without).

Scenario	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p-value	.013	< .001	< .001	.101	.126	< .001	.820	.153	.075	.022
Result	Sig.	Sig.	Sig.	n.s.	n.s.	Sig.	n.s.	n.s.	n.s.	Sig.

Table 17: Kruskal–Wallis tests across rater type with Dunn–Bonferroni post-hoc comparisons (Scenarios S1–S10).

	Kruskal–Wallis H	Post-hoc comparisons (Adj. p)			
Scenario	H	MX–NBE	MX–FE	FX–NBX	ME–NBE
S1	13.968*	**			
S2	17.898**	**			
S3	28.569***	***	**	**	
S4	8.138				
S5	4.748				
S6	29.946***	***		**	*
S7	3.560				
S8	4.674				
S9	8.935				
S10	13.182*	*			

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 18: Kruskal–Wallis tests across all rater types (humans and AI).

Scenario	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p-value	< .001	< .001	< .001	< .001	.071	< .001	.016	< .001	.164	.012
Kruskal–Wallis H	14.84***	21.76***	23.43***	13.94***	12.78	64.59***	19.61*	35.31***	3.21	6.04*