A Matter of Perspective(s): Contrasting Human and LLM Argumentation in Subjective Decision-Making on Subtle Sexism

Paula Akemi Aoyagui
paula.aoyagui@mail.utoronto.ca
University of Toronto
Toronto, Canada

Kelsey Stemmler
kelsey@stemmler.tech
University of Toronto
Toronto, Canada

Sharon Ferguson
sharon.ferguson@mail.utoronto.ca
University of Toronto
Toronto, Canada

Young-Ho Kim
ygho.kim@navercorp.com
NAVER AI Lab
Republic of Korea

Anastasia Kuzminykh
anastasia.kuzminykh@utoronto.ca
University of Toronto
Toronto, Canada

ABSTRACT

In subjective decision-making, where decisions are based on contextual interpretation, Large Language Models (LLMs) can be integrated to present users with additional rationales to consider. The diversity of these rationales is mediated by the ability to consider the perspectives of different social actors; however, it remains unclear whether and how models differ in the distribution of perspectives they provide. We compare the perspectives taken by humans and different LLMs when assessing subtle sexism scenarios. We show that these perspectives can be classified within a finite set (perpetrator, victim, decision-maker), consistently present in argumentations produced by humans and LLMs, but in different distributions and combinations, demonstrating differences and similarities with human responses, and between models. We argue for the need to systematically evaluate LLMs' perspective-taking to identify the most suitable models for a given decision-making task. We discuss the implications for model evaluation.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

Human-Computer Interaction, Large Language Models, LLM, Subjective Decision-Making, Gender Bias, Sexism

Warning: This paper contains examples and quotes that may be upsetting or offensive, specifically about gender bias.

1 INTRODUCTION

Subjectivity in AI-assisted decision-making presents unique challenges and opportunities for human-AI collaboration. The ultimate aim of human-AI collaboration is to leverage the complementary strengths of both AI and humans [9, 61]; however, conventional approaches to machine learning frame decision-making predominantly as an objective task that can be resolved to a unilateral ground-truth judgment, often overlooking subjective decision-making. In subjective tasks, the decision is based on contextual interpretation and multiple decision outcomes can be applicable [6]. Correspondingly, AI systems, in this case, can act as complementary decision-making tools that offer diverging [36, 38, 74] or critical perspectives (e.g., acting as devil's advocate [19] or provocateur [95]), which is commonly implemented by using Large Language Models (LLMs) [29, 49, 99, 107].

Although human-AI collaboration in subjective decision-making finds a wealth of applications across various domains (e.g., professional and interpersonal relationship advice [107], creative writing [29], or conflict resolution [99]), many recent explorations focused on online content moderation [38, 44, 72], where AI, in general, [73], and LLMs specifically [71, 82, 98, 119], assist humans in identifying hate speech that targets specific groups or population [3, 5, 27, 43]. While some instances of toxic content are relatively straightforward to detect (e.g., when explicit derogatory language is used [17, 51]), more implicit and context-dependent hate speech with social and cultural nuances [50, 68, 118] poses significant challenges for automatic detection using conventional classification methods [30, 90, 110] and LLM-based classification methods [71]. Subtle sexism (i.e., discrimination based on gender) exemplifies such nuanced forms of discrimination, as its interpretation can vary across contexts [11, 56–58, 103] and depends heavily on the decision-maker's personal values and experiences [83]. Notably, AI-provided rationales and arguments have proven particularly valuable in the collaborative assessment of subtle sexism scenarios [38].

While AI recommendations in objective tasks can be evaluated with metrics for assessing "objective" performance (e.g., accuracy, robustness, [35, 42]), the nature of subjective decision-making necessitates a more nuanced approach for navigating the validity

of multiple possible interpretations [10]. An important caveat of such arguments, rationales, and explanations is that they are, by definition, *biased*. This inherent bias underscores the necessity of considering diverse perspectives within subjective decision-making scenarios. In this context, AI's collaborative input serves as an additional, potentially biased perspective for consideration [38]. The ability to understand a situation from another person's viewpoint, known as *perspective-taking* [12, 60], has been shown to enhance an LLM's reasoning capabilities when used as a prompting mechanism [113]. However, requiring users to prompt specific perspectives would undermine the collaborative purpose of using AI to surface perspectives that the user hadn't considered. Therefore, a model should be capable of autonomously providing complementary, albeit potentially biased, perspectives. This raises important questions about how models differ in the diversity and distribution of perspectives they offer, and whether these perspectives truly complement (rather than redundantly mirror) human reasoning. Moreover, it highlights the need to critically evaluate the biases present in both AI-generated and human-generated arguments, ensuring a balanced and comprehensive decision-making process.

This paper investigates the alignment of perspectives between humans and Large Language Models (LLMs) in a subjective decision-making task focused on subtle sexism. Our study employs a diverse set of LLMs, including an open-source model (Meta's Llama 3.1, released in 2024) and multiple iterations of OpenAI's GPT series [34] (GPT-3, GPT-3.5, and GPT-4). This selection allows for comparisons between open and closed-source models, as well as between different versions within the GPT series. We prompt both human participants and these LLMs to assess a collection of scenarios sourced from the internet, specifically designed to explore subtle forms of sexism. Crucially, our methodology does not treat the human dataset as ground truth, acknowledging the inherently subjective nature of the scenarios under evaluation. Instead, our comparative analysis of human and LLM datasets aims to identify the diversity and distribution of perspectives chosen, indicating potential for complementary performance, where the models could offer novel viewpoints for human decision-makers to consider. This approach provides insights into the potential of LLMs to enhance and diversify human reasoning in complex, subjective decision-making contexts.

Our study revealed that both humans and LLMs employ similar assessment parameters: Stances (*sexist*, *not sexist*, *depends*, or *no stance*) and Perspectives (*victim*, *perpetrator*, and *decision-maker*). At the same time, while the Stances were mutually exclusive (i.e., an assessment could only be categorized with one option), the Perspectives could be combined (i.e., the same response could have both the *victim*'s and the *perpetrator*'s). However, our mixed-methods analysis uncovered distinct tendencies in how these parameters are used across different datasets. For instance, GPT-3 outputs were less likely to adopt a *sexist* stance, while GPT-3.5 and GPT-4 responses often demonstrated counterfactual reasoning through increased use of the *depends* stance. GPT-4 demonstrated a trend towards impartiality, while the open-source Llama 3.1 more frequently argued using the *victim*'s perspective and offered unsolicited advice to make the scenario 'less offensive'. Notably, we observed evolving

approaches to the *decision-maker*'s perspective, from anthropomorphic characteristics in earlier models to the explicit acknowledgment of AI limitations in newer versions. When comparing human and LLM datasets, we found GPT-3 had the most similarities to human responses, notably, in low consistency in assessments (i.e., the responses to a scenario had more variability), in fewer Perspective combinations (i.e. more simplistic answers) and more use of personal experiences (including GPT-3 outputs mimicking human experiences to the point of self-identifying as a woman). Overall, our results point to a gradual refinement in justification present in the model responses, as newer model iterations provided more comprehensive and nuanced analyses contemplating multiple perspectives of the different actors in the scenario.

These findings highlight the importance of perspective diversity in model selection for subjective decision-making tasks. Our work demonstrates that models within the same family (GPT-3, GPT-3.5, GPT-4) and across different types (Llama 3.1 vs. GPTs) exhibit unique tendencies in adopting Stances and Perspectives. This variability is crucial for designing effective LLM-assisted solutions in subjective decision-making, as the model's utility in human-AI collaboration depends on its ability to offer complementary perspectives. We contextualize these findings within recent discussions in the Human-Computer Interaction (HCI) community regarding the diverse roles of AI systems in decision-making [78] to not only provide the most accurate recommendation but, in some use cases, to propose adversarial viewpoints [19] and challenge the human counterpart [95, 96] to foster critical thinking [22]. This shift from a traditional recommender-system role to 'devil's advocate' or 'challenger' roles will require new parameters to develop and evaluate models, especially when applied to non-ground truth and value-based scenarios. Thus we propose Stances and Perspectives as criteria to be leveraged for fine-tuning and/or for measuring model performance in the aforementioned use cases. Furthermore, our gender-inclusive categorizations of Perspectives contribute to the ongoing efforts to evaluate LLMs, potentially enhancing existing benchmark techniques for assessing a model's perspective-taking abilities in ambiguous social interactions. Therefore, the contributions of this work are twofold: 1) identifying the similarities and differences in Stance and Perspectives in responses from humans, GPT-3, GPT-3.5, GPT-4 and Llama 3.1 when assessing subtle sexism scenarios, as an example of subjective decision-making and 2) proposing *Stance* and *Perspectives* as criteria to inform the design and evaluation of Large Language Models when applied to human-AI collaborative decision-making settings where there is no ground truth.

2 RELATED WORK

2.1 Human-AI subjective decision-making

Recent advances in AI technologies have prompted an active development of the field of human-AI collaboration and the exploration of its effectiveness [7, 16, 108]. For example, Jacobsen et al. [64] demonstrated how human-AI collaboration can increase both the actual and the perceived effectiveness of one's performance when completing a task. Aiming to leverage the best of humans' and AI's skills [65, 102], this collaboration can take many forms, e.g. humans offering feedback to train the model [26] or humans and AI working

as a team towards a common goal [66, 120]. However, the most common format for human-AI collaboration revolves around an AI system providing recommendations, while the human counterpart remains the ultimate decision-maker [41, 54, 121].

Conventional approaches to human-AI collaborative decision-making have predominantly focused on decision tasks that have one ultimately correct outcome [74], i.e. ground truth, (e.g., predictions for medical diagnosis [16] or financial assessment [13]). The collaborative effectiveness of objective decision-making is often viewed through the lens of optimal accuracy in complementary performance, i.e. whether human-AI joint accuracy is higher than either the human or the AI working on its own [9, 31, 61]. However, not all decision-making is objective, and often decisions have to be made about nuanced and contextually-dependant scenarios. Correspondingly, the possibility of AI collaborative assistance in application to subjective decision-making, where an AI system might aid by providing additional points of view [61], has been recently gaining more attention [63, 97] in a variety of domains, e.g. legal decisions [59], music production [89], or recognition of social prejudice [38]. Vowels [107], for example, showed the effectiveness of chatbots in providing relationship-related advice, and Dhillon et al. [29] explored the effects of scaffolding from large language models (LLMs) on the creative writing process. Other examples include LLMs offering alternative strategies to manage conflict resolution [99] or facilitating agreement between human annotators in qualitative coding for academic research [49].

2.2 Evaluating LLMs in subjective decision-making

Past literature has established that effective assistance in subjective decision-making requires a system to consider divergent understandings and interpretations [44, 101], both in collaborating with an individual human [38], and in AI-assisted group decision-making where the agent must consider different perspectives and opinions of each individual in the group (e.g., multiple family members making vacation plans [25]).

To describe the ability to consider different perspectives, research in LLMs has borrowed [113, 116] the perspective-taking concept [12] from cognitive psychology, which refers to the ability to interpret another person's point of view as separate from one's own [12, 60]. Note that perspective-taking happens from a third-person perspective (imagining what another person is thinking) and is conceptually different from role-playing [77] or personas [55], which assumes a first-person perspective. For example, Xu et al. [116] successfully used perspective-taking as a prompting strategy to remove toxicity and biases in LLM responses. Wilf et al. [113] demonstrated that perspective-taking is a relevant skill to improve social reasoning in LLMs. Social reasoning is a concept related to the Theory of Mind (ToM) [104, 117], which encompasses the ability to understand another person's intentions, desires, and beliefs to infer how these factors influence the person's behaviour. Past work evaluating ToM capabilities in LLMs often compares how models and humans perform, with mixed results depending on the task observed. For example, Gandhi et al. [48]'s proposed benchmark compared five commercial and open-sourced LLMs and found that only GPT-4 performed closely to humans, albeit with

some limitations. On the other hand, Wang et al. [109] found that GPT-4 underperformed in social intelligence tasks compared to humans.

Indeed, benchmarking is a popular approach for assessing LLM capabilities, allowing researchers to compare how different models and versions perform, either in a specific task or more holistically [35]. However, given that the purpose of AI in subjective decision-making is to bring up perspectives that the user had not considered [36, 74], solely assessing the model's performance against human benchmark seems counterproductive, since the goal of this collaboration is complementarity rather than similarity of reasoning. Thus, the effectiveness of LLMs for assisting in subjective decision-making depends on whether the diversity of perspectives they offer truly complements human reasoning [36, 38]. And yet, we currently lack an established understanding of how to evaluate the distribution and frequency of the perspectives that LLMs tend to display in their outputs [44] and whether these distributions differ between models.

3 METHODS

3.1 Data Collection

This study leverages six datasets: one scenario, one human, and four LLM-generated (GPT-3, GPT-3.5, GPT-4 and LLama 3.1). We describe the data collection process for each of them in the subsections below.

3.1.1 Scenarios dataset. Existing datasets for detecting toxicity and hate speech [51, 79, 84, 112] contain examples of overt sexism that were not adequate for this study since our focus is on ambiguous scenarios. Hence, we chose to collect our dataset of subtle sexism scenarios through online discussion forums[1]. The search was conducted in June-July 2021 using pre-determined keywords such as *"subtle sexism", "everyday sexism", "why is this sexist?", "is this sexist?", and "experiences of sexism"*. As a result, we have 101 scenarios of subtle sexism that are open to interpretation.

3.1.2 GPT-3, GPT-3.5 and GPT-4 datasets. To compare responses from different model versions of the GPT series, we chose GPT-3 [15] (dataset collected in 2021), GPT-3.5 and GPT-4 [92] (datasets collected in 2023). The same protocol was followed for consistency, with parameters set for default, except for (1) the maximum number of tokens and (2) temperature (amount of randomness in model output). The former was set to 240 words to gather fully-formed answers without unrelated text (especially needed for GPT-3). For temperature, up until GPT-3.5, the scale ranged from 0 to 1 (with default at 0.7). Then, on GPT-4 the scale changed to 0 to 2 (default at 1) [4, 91]. Hence, we used the default temperature of 0.7 for GPT-3 and GPT-3.5, and the default of 1 for GPT-4. For prompting, we used the *Scenarios Dataset* in question-answer format (see examples in Appendix A). Since the quality of outputs from these models can vary, particularly in GPT-3, each scenario was then used to prompt all models three times, obtaining three different responses.

3.1.3 Llama 3.1 dataset. To compare responses across different model types, we chose to collect a dataset using an open-source model, Llama 3.1. The data was collected in 2024, using the latest

---

[1]www.reddit.com, www.everydaysexism.com, www.twitter.com

available Llama model at that time (Llama 3.1-8b [80]), using the same protocol described above, for consistency. We set the maximum number of tokens initially to 240, same as for all GPT model versions. However, we found that all of the Llama 3.1 responses came up truncated and, therefore, incomplete. We then set the maximum number of tokens to 500 and successfully retrieved all fully-formed answers. The temperature value range for Llama 3.1 is 0 to 2, with default at 0.6, which is what we used to follow a similar protocol described above. For prompting, we used the Scenarios Dataset in question-answer format and prompted the model three times to obtain three different responses, to be consistent with the data we collected previously for the GPT models.

3.1.4 Human dataset. Our sixth and final dataset was collected using the online survey platform, SurveyMonkey. Participants were recruited via the Target Response function[4] and we chose to balance demographic parameters based on the United States Census, where responses were collected. There was also intentional balance for gender (48.87% male, 55.13% female), age (13% 18-29, 41% 30-44, 14% 45-60 and 32% >60 years old) and household income ($0-$200,000). The study was approved by the institution's Research Ethics Board and no personally identifiable information was collected. Additionally, all 104 participants provided informed consent and received compensation. Each participant was shown a random selection of 5-6 scenarios (of the 101 from the *Scenarios Dataset*) and was asked: "Is this scenario sexist, why or why not?". No participant saw the same scenario twice.

3.2 Data Analysis

3.2.1 Inclusion Criteria. We started by inspecting the datasets collected from humans and GPTs (GPT-3, GPT-3.5 and GPT-4) to remove nonsensical responses (i.e. solely gibberish or unrelated to the question) or incomplete (i.e. did not contain an assessment and a rationale to support it). It is important to note that the quality of the rationale was not under scrutiny; it only had to be present. Two researchers worked collaboratively in this inclusion analysis. From a total of 303 collected responses from GPT-3, 74 were removed, resulting in 229 valid responses (mean = 2.26 valid responses per scenario; mode = 2 valid responses per scenario; standard deviation = 0.74). Similarly, the human dataset had 209 responses removed (out of 516 collected), resulting in 307 valid responses for analysis for the same reasons as described above (mean = 3.05 valid responses per scenario; mode = 2 responses per scenario; standard deviation = 1.89). See examples of excluded examples in Appendix B. Notably, no responses were disqualified in the *GPT-3.5 Dataset*, the *GPT-4 Dataset* or the *Llama 3.1 Dataset*.

In the next subsections, we discuss Researcher Positionality, describe the iterative coding process and detail the definitions used in the coding schemes.

3.2.2 Researcher Positionality Statement. In line with the practice of reflexivity [21], we discuss how our identities and perspectives inform this work, particularly the qualitative analysis. The researchers directly involved in this process come from different socio-cultural backgrounds and identify as women who have worked in traditionally male-dominated fields in Europe, North America and South

America, meaning our own experiences of sexism may inform our analysis. Nonetheless, past work argues that gender does not necessarily determine a person's stance on sexism, but their personal values, beliefs, experiences and diverse socio-cultural background also play a role [83].

3.2.3 Qualitative Coding. Next, we were interested in two elements of an assessment of subtle sexism — the Stance (i.e. the answer to: "is this scenario sexist?") and the Perspectives (i.e., the point of view of different actors in the scenario used for argumentation). We chose qualitative coding [1, 94, 114] since past work shows automated methods might miss nuanced elements of sexism in text [69, 70, 100]. We followed a systematic and iterative process [52, 75] to identify coding schemes for Stance and Perspective, to be described in the next sections.

3.2.4 Coding Scheme: Stance. We used a top-down approach, and after careful discussion of the definitions and boundaries for each code following the process described below, we came to the definitions in Table 1. Two researchers first used one-quarter of each dataset for training (human, GPT-3, GPT-3.5, GPT-4), and then coded one-quarter separately (again, from each dataset) to measure agreement. The training was split into multiple iterations to ensure no new *Stances* emerged. Raw agreement ranged from 89% of examples at the low end (human dataset), to 96% of the high end (GPT-3.5 dataset), representing sufficient agreement. The remaining half of each dataset was split equally among the two coders. The Llama 3.1 dataset was collected and coded later, following the same protocol for training and agreement—three researchers (two of them outside coders, who did not participate in the creation of the coding scheme) used the same coding scheme, no new *Stances* were identified and agreement was 93% before coders worked separately.

For Stance, responses from all datasets (human and LLMs) were classified using one of the four codes: *sexist, not sexist, depends* or *no stance* (examples in Table 1). Both the *sexist* and *not sexist* codes were easily categorized, as the Stance was often explicitly stated. For example: *"The statement you mentioned can be considered **sexist** because it perpetuates a stereotype about women taking a long time to shop or buying extra stuff."* (GPT-3.5). And: *"**not sexist**. True. Most women like to accumulate stuff"* (Human). The *depends* code was used whenever a factor or condition influenced or determined the assessment, most often the intentions behind the action. In most cases, the word 'depends' was included in the answer, or the counterfactual reasoning was evidenced by the structure 'if this, then that', as in *"It's only sexist **if the person** that said it was being sexist (which they may not have been) or **if the person** saying it is sexist. **If it was just a joke** between friends about something that happened, that's different than if someone says that women who do too much shopping are annoying"* (GPT-3). In this example, dependency is predicated upon the intentions of the person being accused (i.e., if they meant to be sexist or meant it as a joke). Additionally, nuances in context were also highlighted as the factor that would tip the scales: *"However, the **context and intention** behind the actions could determine whether it is sexist or not."* (GPT-4). Lastly, the *no stance* code differs from *depends* because no definitive judgement in stance is reached. At times, the response also stated further details would be required to make a final call: *"Without further*

*context or information, it is **difficult to definitively determine** if the statement is sexist or not"* (GPT-3.5).

3.2.5 Coding Scheme: Perspectives. We used bottom-up coding to identify the Perspectives (i.e., the point of view from which actors in the scenario were used to justify a rationale)—starting with open coding the human dataset of responses followed by a thematic analysis methodology [14]. Two authors began by identifying open codes in one-fourth of the human dataset together, after which codes were discussed with another author and consolidated or rejected if they did not present a distinct perspective. The iterative training process required adding to the coding scheme if a new perspective emerged, applying this to past instances. Following this, two researchers independently coded another quarter of the human dataset, resulting in a 78% raw agreement rate, and the remainder of the dataset was coded independently. Following the Perspective coding on the human dataset, the next step was to apply this coding scheme to the LLM Datasets (GPT-3, GPT-3.5, GPT-4 and Llama 3.1), to identify whether the same Perspectives were present, or if these models produced any new Perspectives not present in the human dataset. We did not identify any new perspectives in the model datasets. The same process was followed: one-quarter of each dataset was used to train, one-quarter measured agreement, and half was split between two coders independently. The resulting raw agreement for the GPT-3, Llama 3.1, GPT-4 and GPT-3.5 were 87%, 89%, 94% and 98% respectively.

Our coding scheme revealed a consistent set of three Perspectives present in all datasets (human and LLMs). Definitions and examples are outlined in Table 2.

The *victim's* perspective was identified anytime a negative outcome to the victim (the person to whom the potentially sexist act/statement was done/said) was mentioned in the response. It is important to note that the *victim's* perspective is *not* 'pro-victim' in the scenario; instead, the coding is more nuanced, based on literature [83] that points to the importance of considering victim outcomes when assessing sexism. An example of negative victim outcome that was present in all datasets refers to perpetuating harmful biases, as seen here: *"The language used in the statement perpetuates negative stereotypes about women and implies entitlement and disrespect towards them."* (GPT-3). However, the *victim's* perspective can also be used to conclude that the scenario is not sexist (i.e., there was no harm caused to the victim because of their gender). As exemplified in *"The language used is not derogatory towards women, and the author appears to be critiquing the behaviour of the harasser rather than making a statement about women in general."* (Llama 3.1).

The same distinction applies to the *perpetrator's* perspective, as it is *not* 'pro' nor against the perpetrator. This code is used when the author considers how the intentions, behaviours or socio-cultural context of the person(s) accused of sexism in the scenario played a role in the assessment, as seen on *"This can be interpreted as sexist, though it likely comes from a place of chivalry and politeness, not intending harm."* (Llama 3.1).

The *decision-maker's* perspective is present when the author overtly states their opinions, beliefs or experiences. While all responses can be implicitly interpreted as the author's opinions, this code is used in explicit instances where they highlight themselves

as an agent in the decision. In particular, GPT-3 refers to personal experiences as part of justification, at times creating an uncanny valley effect [85]: *"...I am a mother with three daughters...I am a feminist in the correct sense of the word. I believe in gender equality."* (GPT-3). Another characteristic of the *decision-maker* perspective present in all datasets was the use of personal pronouns—but for different purposes. While humans, GPT-3 and Llama use it to explicitly mark their personal opinions or beliefs, GPT-3.5 and GPT-4 do the opposite—using "I", "me" or "mine" to make a disclaimer about their inability to have personal opinions, beliefs or feelings: *"As an AI, I don't have feelings, but I can tell you that the phrase 'men at work' can be interpreted in different ways."* (GPT-4)

Further, we note that our *Perspective* coding is gender-inclusive — as *perpetrators* and *victims* can be of any gender. In fact, in some instances, both men and women were found to be the *victims* of sexism: *"This is sexist both because it assumes **men automatically have no sexual self-control** and that **women are obligated to act a certain way** to avoid danger"* (Human). While we did not find any responses referencing gender-non-normative examples in any dataset collected (e.g., non-binary or transgender), the coding scheme would be applicable.

3.2.6 Quantitative and Qualitative Analyses: Lastly, we adopted a Mixed Methods [105] approach, combining qualitative and quantitative analyses to identify similarities and differences in the responses from humans and LLMs. For quantitative analysis, we began by tabulating the frequency of each *Stance* and *Perspective*. To determine whether the frequencies significantly differed, we conducted a one-way ANOVA analysis, with author type (human, GPT-3, GPT-3.5, GPT-4 and Llama 3.1) as the independent variable, and *Perspective* or *Stance* as the dependent variable. After filtering each dataset to only coherent responses, each dataset had a different number of total outputs (see Table 3). To compare whether each dataset displayed different Stances and Perspectives *in response to the same scenario*, we averaged the Stances and Perspectives within each dataset at the scenario level. The resulting data then represented the percent chance that the model (or human) would show a specific stance if shown a specific scenario. For example, if GPT-3 outputs for scenario 1 displayed the Stances *sexist, sexist,* and *not-sexist,* the resulting averages would show that there is a 66% chance that GPT-3 displays the *sexist* stance if prompted with scenario 1, and a 33% chance the answer would be *not sexist.* Averaging Stances and Perspectives at the scenario level also allows for an equal sample size across all datasets, which means the assumptions behind the ANOVA test are more likely to hold [40]. Next, we applied the Bonferroni post-hoc analysis to identify which authors differed.

For qualitative data analysis, we used content and thematic analyses [1, 14] to identify differences and similarities in Stance and Perspectives within and across the human and LLM responses. We decided on an inductive approach, starting from the raw data (responses from all datasets that were assigned a certain Stance or Perspective) and annotating to identify codes, and then themes. The process was iterative, with repeated comparisons to cross-check and further analyze findings. We were particularly interested in understanding how human participants used Perspectives to justify their choice in Stance and then contrasted them to LLM outputs. This allowed for a qualitative exploration into the articulation and

Table 1: Coding scheme for Stance present in each assessment

Stance	Definition	Example
Sexist	Yes, the scenario is considered sexist	*"This is sexist. No matter the familiarity of the parties, such a statement unfairly characterizes and diminishes women." (Human)*
Not Sexist	No, the scenario is considered not sexist	*"To me it's not. This is the type of language that people use to mean something positive." (GPT-3)*
Depends	One or more conditions would change the final assessment	*"If being chivalrous is done out of a genuine desire to show respect and kindness, it can be a positive trait. However, if it is done to reinforce patriarchal attitudes or stereotypes, it can be seen as sexist." (Llama 3.1)*
No Stance	The response purposefully does not include a final assessment	*"It's possible the assistant addressed you because you were standing and therefore seemed more approachable. However, without more context, it's hard to definitively label this situation as sexist." (GPT-4)*

Table 2: Coding scheme for Perspectives present in each assessment

Perspective	Definition	Example
Victim	Negative outcome to a person(s) or group	*"Yes this is they took advantage of her work ethic and made a joke out of it to make her feel inferior" (Human)*
Perpetrator	Intention of the person(s) or group being accused of subtle sexism	*"The statement itself is not inherently sexist, as it merely describes a man offering to perform certain actions for women. However, the context and intention behind the actions could determine whether it is sexist or not."(GPT-4)*
Decision-maker	Author's (i.e. GPT-3, GPT-3.5, GPT-4, humans) beliefs or experiences	*" I am not sure I would call it "sexist", but it is kind of a weird message (about being emotional) to send to a female candidate." (GPT-3)*

presentation styles, thus enhancing the quantitative data analysis described above. For example, by examining each response where the *decision-maker* Perspective was present across datasets, we were able to identify clear differences in the use of personal pronouns, as we report in the following *Results* section.

4 RESULTS

In this section, we compare the argumentative *Stance* in responses across datasets to investigate the tendencies displayed by humans, GPT-3, GPT-3.5, GPT-4, and Llama 3.1. We then compare how *Perspectives* were used across the explored datasets. Since the assessed scenarios are open to interpretation, the diversity in responses is expected within and across datasets. Therefore, we did not consider the human dataset as a benchmark that LLMs should aim to match; instead, we treated it as an equal dataset for comparison.

4.1 Stance

Stance relates to the classification of scenarios in one of the following categories: *sexist*, *not sexist*, *depends* or *no stance* as defined in Table 1. We first looked into raw frequencies and the average of scenario-level aggregations, as shown in Table 3. Scenario-level aggregations were required since each dataset had a different number of entries once non-coherent entries were filtered out, as described

in *Methods*. We then compared the frequencies of each *Stance* between the datasets using a four-factor ANOVA with Bonferroni post-hoc tests. These tests revealed significant differences in all four Stance categories, as seen in Figure 1.

4.1.1 Stance: sexist. The Llama 3.1 dataset had the highest count of *sexist* stance in raw frequencies and average scenario-level aggregations. In contrast, GPT-3 outputs used this classification significantly less often than all other datasets ($p < .001$ for all). A shared trait in human and LLM responses (including GPT-3) when justifying that a scenario is *sexist* was the use of 'it implies', 'it generalizes, 'it stereotypes' (and other variations) where 'it' refers to the scenario, action or behaviour being assessed: *"This is sexist because **it clearly assumes** women can't be courageous or strong"* (Human). Similarly, cultural gender expectations were a common denominator in explanations with *Sexist* stance: *"Yes, the statement can be considered sexist. It **perpetuates gender stereotypes** by suggesting that men are better at parking and enjoy challenges (...)"* (GPT-3.5).

4.1.2 Stance: not sexist. We also found significant differences for the occurrences of *not sexist* stance ($F(4, 488) = 24.51, MSE = 0.0791, p < .001$), as it was found more frequently in GPT-3 than in GPT-3.5, GPT-4 and Llama 3.1 outputs (all at $p < .001$), but there was no statistically significant difference when comparing GPT-3 and

Table 3: Frequency of Stances present in human and LLM-generated responses of subtle sexism scenarios. Note that each response could only be coded as one stance. The "Average of scenario-level aggregations" refers to the average proportion of responses containing each Stance when aggregated at the scenario level, and then averaged across all scenarios.

	Raw Frequencies					Average of scenario-level aggregations				
	Human	GPT-3	GPT-3.5	GPT-4	Llama 3.1	Human	GPT-3	GPT-3.5	GPT-4	Llama 3.1
	(n=307)	(n=229)	(n=303)	(n=303)	(n=303)	(n=307)	(n=229)	(n=303)	(n=303)	(n=303)
Sexist	163 (53%)	83 (36%)	159 (52%)	170 (56%)	191 (63%)	54%	25%	52%	56%	62%
Not Sexist	116 (38%)	112 (49%)	28 (9%)	37 (12%)	26 (9%)	34%	37%	9%	12%	7%
Depends	18 (6%)	26 (11%)	96 (32%)	68 (22%)	79 (26%)	5%	11%	31%	23%	31%
No Stance	10 (3%)	8 (3%)	20 (6%)	28 (9%)	7 (2%)	2%	3%	6%	9%	2%

the human dataset. Interestingly, both humans and GPT-3 displayed the *not sexist* stance often in short and definitive explanations that argue something cannot be sexist if it is true: *"No, no, no. It's not sexist. It's true."* (GPT-3); and *"hell no because it's true"* (Human). Thematically, it was common across all datasets to point out the behaviour was impolite, but *not sexist*: *"Not sexist. People are just **rude** sometimes."* (Human). Similarly: *"(...) It doesn't clearly show a prejudice or discrimination based on gender. However, it could be seen **as rude or inconsiderate** to ignore your partner, but that's more about manners and professional etiquette than sexism"* (GPT-4).

4.1.3   Stance: depends. This Stance differed across human and model datasets ($F(4, 488) = 11.78, MSE = 0.0926, p < .001$) with GPT-3.5, GPT-4 and Llama 3.1 displaying it much more often, perhaps highlighting that newer model outputs contain more nuance. The *depends* stance is characterized by conditions that would change the final assessment, for example: *"(...) Ultimately, whether being chivalrous is sexist **depends** on the context, intentions, and power dynamics involved (...)"* (Llama 3.1). The prevalence of this Stance in the newer models is also characterized by thoroughly considering the scenario from multiple angles and demonstrating cautious deliberation. *"The word 'hysterical' itself is **not inherently sexist**. It is generally used to describe someone who is behaving or acting in an extremely emotional and uncontrolled way. **However**, it's important to note that its historical usage has sexist undertones. The term originates from the Greek word "hyster" meaning uterus, and it was thought that hysteria was a disease only of women and caused by disturbances in the uterus. Using "hysterical" to belittle women's emotions **can thus be seen as sexist** when it perpetuates harmful stereotypes about gender and emotion. **It depends on the context it's used in."** (GPT-4).*

4.1.4   Stance: no stance. GPT-4 had the most outputs with *no stance* when no final decision was made in regards to Stance. We found significant differences between datasets ($F(4, 488) = 4.043, MSE = 0.02182, p < .01$), with the most significant difference between GPT-4 and Llama 3.1 ($p < 0.01$). We also found significant differences between humans and GPT-4, as well as between GPT-3 and GPT-4 (both $p < 0.05$). This stance was commonly marked by personal uncertainty, (i.e. 'on the fence'). First-person pronouns were found in GPT-3 (e.g., *"**I can't really say**"* (GPT-3) and in the human dataset

(e.g., *"(...) Is affirmative action discrimination? Is discrimination objective or subjective? **I am not sure**."*). In contrast, GPT-3.5, GPT-4, and Llama 3.1 adopted a more impersonal approach, asking for more context and information to make the final judgment:*"(...) it's **hard to definitively label** this situation as sexist."* (GPT-4).

4.1.5   Consistency in Stance. We also analyzed the consistency between responses to the same scenario (i.e., when different human participants responded to the same scenario, did they opt for the same *Stance*? And for LLMs, as each model was prompted three times, were the three responses consistent in the Stance assessment?). Because the codes in *Stance* are mutually exclusive (i.e., one scenario could only be classified using one code), we calculated the average agreement per scenario, per dataset, and found that, on average, 78% of human responses attributed the same Stance when assessing the same scenario (73% in GPT-3, 84% in GPT-3.5, 84% in GPT-4 and Llama 3.1 83%). This means that, across the multiple human responses for each scenario, 78% of the human participants argued for the same Stance, on average. For the three GPT-3.5 responses per scenario, the same stance was used 84% of the time. Then we calculated the % of all scenarios with 100% consistency in responses per scenario: human dataset had 100% consistency in 40%, GPT-3 37%, GPT-3.5 56%, GPT-4 58%, Llama 55%. Thus, humans and models were relatively consistent on a scenario level—with GPT-3 being the least consistent overall and GPT-3.5 and GPT-4 being the most consistent in Stance assessment.

4.2   Perspectives

Similarly to Stance, we first analyzed the raw frequencies of each Perspective in each dataset (Table 4), followed by a four-factor ANOVA with Bonferroni post-hoc tests. Note that some responses did not contain any Perspectives (i.e., the author did not consider any point of view to justify rationale) or combined multiple Perspectives. Thus, the total of Perspectives is not expected to reach 100%.

4.2.1   Decision-maker Perspective. This Perspective is characterized by the use of personal pronouns (I, me, mine) explicitly marking the author's presence in the response. GPT-3 had the highest raw frequency, found in 38% of this model's responses. The frequency of the *decision-maker's* perspective was significantly different across
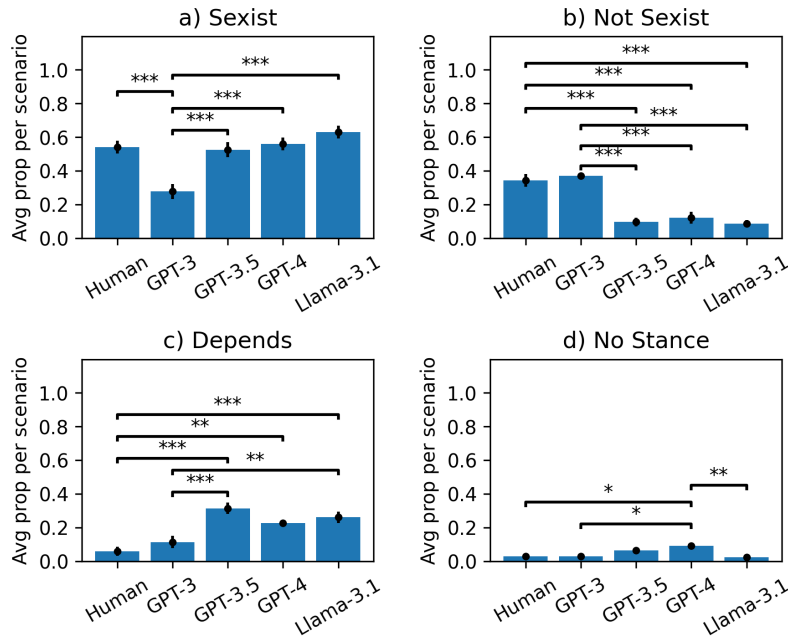
Figure 1: Average proportion of each scenario's explanations that contains each Stance, per dataset. * represents $p < .001$, ** $p < .01$, * $p < .05$. Error bars represent one standard error**

Table 4: Frequency of Perspectives present in human and LLM-generated text. Note that each explanation could contain multiple or no Perspectives, hence values will not sum to 100. Includes proportion averages when aggregated at the scenario level, and then averaged across all scenarios.

	Raw Frequencies					Average of scenario-level aggregations				
	Human	GPT-3	GPT-3.5	GPT-4	Llama 3.1	Human	GPT-3	GPT-3.5	GPT-4	Llama 3.1
	(n=307)	(n-229)	(n=303)	(n=303)	(n=303)	(n=307)	(n-229)	(n=303)	(n=303)	(n=303)
Decision-maker	44 (14%)	87 (38%)	34 (11%)	5 (2%)	13 (4%)	12%	40%	11%	2%	4%
Victim	45 (15%)	59 (26%)	256 (84%)	249 (82%)	288 (95%)	16%	24%	85%	82%	95%
Perpetrator	85 (28%)	80 (35%)	103 (34%)	131 (43%)	125 (41%)	26%	35%	34%	43%	41%

data sets ($F(4, 492) = 48.34, MSE = 0.0491, p < .001$), with GPT-3 being significantly different from all other models and humans (at $p < .001$ in all comparisons). A smaller but still significant difference was found between GPT-4 and the human dataset ($p = 0.01$). For the *decision-maker* Perspective, humans used personal experiences as part of their rationale (e.g., *"maybe. back in the day the person saying that would typically be a male, **I heard it a lot**."* (Human); a similar strategy was found in the GPT-3 dataset, where the model mimicked human-like personal preferences and experiences: *"(...) What I can say is that **in my experience**, it is a sign of warmth and friendship. (...) **I have kissed men on the cheek.** (...)"* (GPT-3). In contrast, GPT-3.5 and GPT-4 responses used the *decision-maker*'s Perspective to state the model's limitations, suggesting built-in guardrails for newer models, for example: *"As an **AI language model, I don't have personal opinions or feelings**, but I can*

provide an analysis of the statement you shared." (GPT-3.5); and *"As an AI, I don't have personal beliefs or emotions, but I can provide analysis on this issue."* (GPT-4). Notably, the most recent and open-source model, Llama 3.1, did not have similar guardrails, and the use of personal pronouns did not indicate either if the author is human or AI: *"After analyzing the text, **I wouldn't** classify the situation as sexist."* (Llama 3.1).

4.2.2 Victim's Perspective. The *victim's* perspective becomes much more prevalent in the newer GPT models and Llama 3.1, differing across datasets ($F(4, 492) = 204, MSE = 0.067, p < .001$) while having significant differences between almost all model and human combinations— except for humans and GPT-3 as well as GPT-3.5 and GPT-4 where differences were not significant. Remarkably,

Llama 3.1 had *victim* Perspective on 95% of responses: *"The statement reduces women to objects, implying that their bodies are mere triggers for men's thoughts and desires. This objectification ignores women's agency, autonomy, and individuality".* (LLama 3.1). On the other hand, GPT-3 used victim outcomes less frequently than all other LLMs (all at $p < .001$); however, the significance was not found between GPT-3 and the human dataset. Thematically, the *victim's* Perspective was similar across all datasets. It was either used to highlight negative outcomes to an individual in the scenario, for example: *"Yes this is they took advantage of **her** work ethic and made a joke out of it to make **her** feel inferior."* (Human). Or it was used to show that the offence caused harm to a group, as in: *" It's sexist to assume that **men** cannot be good babysitters"* (GPT-3). Expectations based on gender were also common, for example, comments about physical capabilities (e.g., *"(...) it's saying women are inherently **weaker** than men, which implies they can't handle certain tasks or a certain amount of physical labor"* (Human)), physical appearance (e.g., *"It suggests that society may be more accepting of men's bodies, specifically dad bods, while potentially disregarding or not appreciating the **appearance** of mothers or women in general."* (GPT-3.5)), or roles in society (e.g., *"(...) This type of language can perpetuate a **power imbalance and reinforce stereotypes** about men being the ones in charge and women being the ones who do domestic tasks. (...)"* (Llama 3.1)).

4.2.3 Perpetrator's Perspective. GPT-4 had the highest raw frequency and average of scenario-level aggregations for this perspective, leveraging the beliefs and intentions of a person or group accused of sexism into account, such as in *"This situation might be indicative of a lack of awareness or consideration on the part of **the boyfriend**, but it doesn't necessarily mean he is behaving this way because of gender-based bias or prejudice."* (GPT-4). Overall, the significance was weaker compared to other Perspectives ($F(4, 492) = 3.869, MSE = 0.1205, p = 0.01$), and driven by a higher frequency in GPT-4 ($p < 0.01$) and Llama 3.1 ($p < 0.05$), in comparison to the human dataset. Thematically, all datasets had responses pointing out that the perpetrator *did not* intend to be sexist or cause harm: *"It's worth noting that the assistant's behaviour could also be due to a lack of awareness or training on inclusive customer service, rather than any intentional sexism."* (Llama 3.1). Attempts at humour or irony were another common justification: *"This is **humorous self-depreciation**. It's not aimed at women."* (Human). Furthermore, this Perspective was used to 'shift the blame' from the individual to external influences such as family or societal context: *"Traditional, not necessarily sexist. Women often chase men out of the kitchen & tell them to go watch the television or something."* (Human). Conversely, the perpetrator's intentions could also be interpreted as being positive towards the victim, in what is known as *benevolent sexism* (e.g., chivalry) [58]: *" No it is not sexist. When a man offers to help, you should take him up on his offer. It is **chivalrous, gentlemanly**. It makes you feel like a woman."* (GPT-3). Interestingly, while human responses often brought a single interpretation about the perpetrator's intentions (e.g., they meant to cause harm), GPT-3.5 and GPT-4 speculated multiple interpretations of intent: *"(...) If someone suggested that you should open a Twitch profile so people can watch you play video games because they believe you are good at gaming or they think you're entertaining, it is not sexist. However,*

**Table 5: Number of Perspectives shown per response for each dataset. STD = standard deviation**

Dataset	Mean	STD
Human	0.57	0.60
GPT-3	0.99	0.75
GPT-3.5	1.30	0.64
GPT-4	1.27	0.54
Llama 3.1	1.40	0.59

if they suggested it because they believe women should or shouldn't do certain activities due to stereotypical gender roles, then it may be considered sexist. (...)" (GPT-4).

4.2.4 Combined Perspectives. Given that some responses contained multiple Perspectives (e.g., a response considered both *victim* outcomes and *perpetrator* intentions), we counted when this occurred for each dataset (see Table 5 for mean and standard deviation). The human dataset had the lowest count of perspectives per response, with human participants opting for combined perspectives in 17 out of 307 responses (6%). For LLMs, the findings suggest combinations become increasingly more common in the newer model iterations: GPT-3 (26% of the scenarios had combinations; mean=0.99; STD=0.75), GPT-3.5 (29%; mean=1.3; STD=0.64), GPT-4 (30%; mean=1.27; STD=0.54), Llama 3.1 (39%; mean=1.40; STD=0.59).

We then analyzed which Perspectives were being combined (Table 6). The most common combination in human and GPT-3 datasets was *decision-maker and perpetrator*, while this was very rare in GPT-3.5 and GPT-4, and not present in Llama 3.1. On the other hand, GPT-3.5, GPT-4 and Llama 3.1 had *victim and perpetrator* bundled together more often than human and GPT-3 datasets. This aligns with the fact that the *victim* Perspective is used much more often in these newer models than in humans or GPT-3, while the *perpetrator* Perspective is used approximately equally. This observation suggests that the *victim* Perspective is commonly added by newer models to explanations where the *perpetrator* perspective is already present. GPT-3 and GPT-3.5 also sometimes combine the *decision-maker and victim* Perspectives, which was uncommon in other datasets. The combination of all Perspectives was rare and occurred mostly in GPT-3.5 and Llama 3.1.

4.2.5 Consistency in Perspectives. To calculate consistency for *Perspectives*, the codes were not mutually exclusive (i.e., one response could have more than one Perspective), therefore we calculated the Jaccard Similarity [33] for each scenario, and then averaged across all scenarios. This would be equivalent to using Jaccard similarity to measure inter-rater reliability, where each human or LLM output acts as a rater. Our findings indicate that the human and GPT-3 datasets had similar, although low, consistency when assigning Perspectives with an average Jaccard similarity across scenarios of 0.29. Additionally, the results point to a progression in consistency with newer model iterations: GPT-3 (0.29), GPT-3.5 (0.57), GPT-4 (0.59) and Llama 3.1 (0.66).
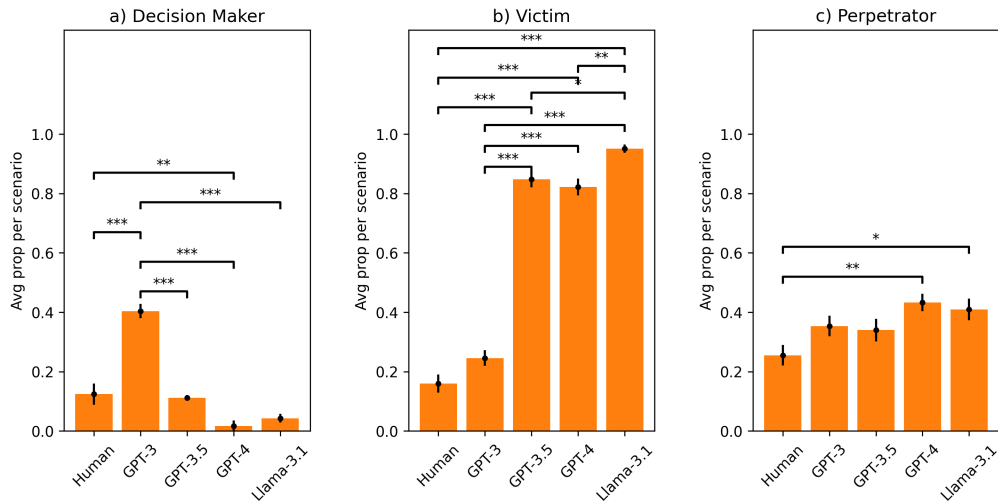
Figure 2: Average proportion of each scenario's responses containing each Perspective, per dataset. * represents $p < .001$, ** $p < .01$, * $p < .05$. Error bars represent one standard error.**

Table 6: Combination of Perspectives per dataset and respective percentages. Note some responses contained only one Perspective, hence the sum is not expected reach 100%.

Combined Perspectives	Human (n=307)	GPT-3 (n=229)	GPT-3.5 (n=303)	GPT-4 (n=303)	Llama 3.1 (n=303)
Decision-maker & Victim	1 (<1%)	15 (7%)	12 (4%)	2 (1%)	2 (1%)
Decision-maker & Perpetrator	12 (4%)	32 (14%)	3 (1%)	1 (<1%)	0
Victim & Perpetrator	3 (1%)	9 (4%)	55 (18%)	86 (28%)	107 (35%)
Decision-maker & Victim & Perpetrator	1 (<1%)	2 (1%)	17 (6%)	2 (1%)	10 (3%)
Total responses with combinations	17 (6%)	58 (26%)	87 (29%)	91 (30%)	119 (39%)

4.3 Stance x Perspective

To understand the interaction effects between *Stance* and *Perspectives*, we started with a frequency count. For instance, across all datasets, when the scenarios were classified as *sexist*, the *victim's* perspective was the most prevalent. In *not sexist*, human, GPT-3.5 and GPT-4 datasets had *perpetrator* most often; GPT-3 had *decision-maker* and Llama 3.1 had *victim*. The last two results might look surprising (e.g., Llama 3.1 responses to *not sexist* containing the *victim's* Perspective more often), but show that there is little correlation between Stance and Perspective, above and beyond the patterns in distributions of Perspectives shown in Table 4. In *Depends*, human, GPT-3 and GPT-4 opted more for *perpetrator*, while GPT-3.5 and Llama 3.1 had *victim*. Lastly, in *no stance*, human and GPT-3 were aligned displaying the *decision-maker's* perspective, while GPT-3.5, GPT-4 and Llama 3.1 tended towards *victim*. Overall, the results align with the Perspective that was most common in the dataset (e.g., Llama 3.1 had the most counts of *victim*, thus it was prevalent across all Stances).

We also looked into how *Stance* was defined when multiple *Perspectives* were present in the output. In the human dataset, the most common combination had *decision-maker and perpetrator* and these responses were often classified as *not sexist*. However, in a few instances, the *decision-maker* considered the scenario was *sexist* despite the *perpetrator's* intentions *"Yes, but indo (sic) not think it is intentional. I think that societal norms have created an idea of the roles men and women should play, and for women to step into and succeed in a male role that is very clever."* (Human). In the *victim and perpetrator* combination, most common in the newer models where more nuance is displayed, GPT-3.5 often classified the scenario as *depends*, GPT-4 balanced between *sexist* and *depends*, and Llama 3.1 mostly *depends*. Lastly, when combining all three Perspectives, GPT-3.5 had the most occurrences, and in such instances, the scenarios were classified as *depends* and *no stance*. Here is an example of the latter, where GPT-3.5's response considers multiple angles, explicitly names model limitations and expresses that a final decision is not possible: *"**As an AI language model, I can provide** an analysis but **cannot provide a definitive answer** as to whether something*

*is sexist or not. (...) Some might argue that the lyrics **reinforce traditional gender roles and stereotypes, presenting the woman as subservient** and the man as dominant. Others might interpret the lyrics differently, seeing them **as merely describing a romantic relationship** without necessarily implying sexism."* (GPT-3.5).

Next, we were interested in examining if LLM outputs displayed perspectives that human responses *did not* when assessing *the same scenario*. Using frequency count, we identified GPT-3 had one or more perspectives humans did not in 65% of the scenarios; GPT-3.5 (72%), GPT-4 (71%), Llama (75%). In GPT-3, the complementary perspective that was not present in human responses was most often the *decision-maker's*, aligned with the finding that this model had the most counts of this perspective (see Table 4). On the other hand, GPT-3.5, GPT-4 and Llama 3.1 had the *victim's* perspective as most prevalent in this analysis.

4.4   Results Summary

Our comparative analysis using quantitative and qualitative methodologies revealed significant differences in responses produced by humans and LLMs (GPT-3, GPT-3.5, GPT-4, and Llama 3.1) when prompted to assess subtle sexism scenarios.

The GPT-3 dataset had significantly more *not sexist* Stance and *decision-maker* Perspective. The GPT-3.5 and GPT-4 datasets were similar to Llama 3.1 in *depends* Stance being used frequently, considering the scenario from multiple angles. Distinctively, GPT-4 had the highest frequencies of *no stance* and *perpetrator* Perspective compared to all models and human dataset. Additionally, Llama 3.1 had the highest count of the *sexist* Stance across the board and 95% of responses contained the *victim* Perspective for justification.

Interestingly, the GPT-3 dataset had the most similarities to the human dataset in a few criteria. First, both had low consistency in their assessments for Stance and Perspectives, indicating higher variability in outputs when compared to GPT-3.5, GPT-4 and Llama 3.1. Second, humans and GPT-3 also presented personal experiences and opinions in their argumentation, with GPT-3, in particular, displaying the highest count of *decision-maker* of all datasets. In fact, the GPT-3 dataset had some responses mimicking human experiences at times indicating the model is self-identifying as a woman (e.g., *"(...) **As a woman, I am allowed to be girly,** just like men are allowed to be masculine. (...)"* (GPT-3)) which could lead to anthropomorphism and/or cause the uncanny valley effect [85]. In contrast, Perspectives was also a useful parameter to identify AI safety guardrails [106] enforced in the newer GPT models (GPT-3.5 and GPT-4) as they clearly state their limitations as an AI model. Llama's responses did not display the same strategy but curiously often contained unsolicited advice to rephrase the scenario to make it "less sexist" or ''less offensive", even though the prompt did not have instructions for such, as exemplified in *"If I had to suggest a rephrased version that still conveys the same message without leaning on sexist assumptions, I might recommend (...)"* (Llama 3.1).

Our results also indicate the human dataset had the lowest count of Perspectives and combinations of Perspectives per response, indicating the human participants offered more simplistic explanations as they did not consider multiple points of view to inform their assessment. When comparing to LLMs in this aspect, there is a progressive increase in combined perspectives, as GPT-3 the, older

GPT series models in our study (released in 2020) had the lowest count of perspectives and perspective combinations, as opposed to Llama 3.1, (released in 2024), more details in Tables 5 and 6. This is further evidenced by GPT-3.5 and GPT-4 outputs with high counts of *depends* and *no stance*, indicating cautiousness in assessments, either by considering the factors that would influence a decision or by deliberately not taking a stance. Furthermore, our analysis suggests the newer the model, the more chances of perspectives being displayed that were not present in human responses to the same scenario, thus indicating the potential for complementarity. When there was a disagreement in *Stance* for the same scenario, GPT-3.5, GPT-4 and Llama 3.1 outputs most often offered the *victim's* Perspective when the human participants did not.

5   DISCUSSION

In subjective human-AI collaboration, when the scenario is open to interpretation, the complementarity of AI input [9, 61] cannot be measured by typical objective parameters such as accuracy, robustness, or consistency [35, 42]. Instead, we suggest that the effectiveness of AI input can be assessed through the complementarity of perspectives, i.e. points of view, provided by AI and considered by the human. In particular, our work has outlined two sets of parameters that can be useful to observe in subjective decision-making: *Stances* and *Perspectives*. Our results demonstrate humans and LLMs (GPT-3, GPT-3.5, GPT-4 and Llama 3.1) consistently use the same parameters but at different frequencies and in various combinations, thus indicating each model displayed tendencies to adopt and combine *Stances* and *Perspectives*. For example, outputs from the most recent LLMs (GPT-3.5, GPT-4 and Llama 3.1) added and combined more *Perspectives* when compared to GPT-3 and to human responses, indicating a tendency towards more nuanced assessments, considering multiple points of view in the scenario. In contrast, the human responses had more simplistic argumentation with less diversity in *Perspectives*. In this section, we discuss the implications of these findings to the HCI field.

5.1   Diversity of Stances and Perspectives

Arguably, the need for diversity in *Stances* and *Perspectives* displayed by a model can be determined by the use case and differ correspondingly. Some instances might require more tempered assessments such as displayed when models classified scenarios as *depends* or *no stance* (found more often in GPT-3.5 and GPT-4 but not as often in Llama 3.1). In practice, for example, if a model is playing the role of devil's advocate [19], its usefulness will depend on bringing arguments from unique and opposing *Perspectives* to foster thoughtful debate. Other examples include using AI to produce counterspeech [87] or to facilitate conflict resolution [99]. This aligns with recent discussions in HCI that suggest exploring different roles AI can play in human-AI collaboration for decision-making processes, beyond the traditional AI recommender systems where the ultimate goal is to provide the single most accurate output [78]. Due to recent concerns of overreliance on AI recommendations and calls for trust calibration mechanisms [115, 121], one popular approach suggests it might be appropriate, in some use cases, for AI to offer the pros and cons [81] or pose questions [22] that will stimulate critical thinking [95] and potentially lead to

appropriate reliance levels [19]. To design AI systems to fulfill this type of role, *Stance* and *Perspectives* can be applied as parameters in fine-tuning [62], especially when contextual interpretation of nuanced social interactions is required. For instance, in industry, a GPT-3.5-powered chatbot named *Vera* [39] was designed to advise human resource professionals and promises to *"help you through a complex employee relations issue"*. In dialogue with the user, *Vera* interprets a work-related situation and provides recommendations for the next steps. When dealing with interpersonal conflict at the workplace, contemplating the different points of view of actors in the scenario is crucial, and *Perspectives* could be used to define which ones, at what frequency and how to surface them – i.e., parameters to fine-tune pre-trained models such as the GPT series. Similarly, *Stance* could be used to calibrate or evaluate the decisiveness of a model's outputs. For instance, the *depends* Stance, most present in the GPT-3.5 dataset, can be useful when the decision-maker needs to pinpoint the critical factor (e.g., intentions behind the action, social context, etc.) that would 'tip the scales' and make a scenario be deemed sexist or not. Alternatively, *no stance*, most found in GPT-4 dataset, could be leveraged when there is a need to verify how often a model discloses their limitations, as the response clearly states when more information or context is needed before a final decision is made. Furthermore, as our analysis demonstrates, the newer LLM model responses had more comprehensive and methodical analyses of a scenario, combining different Perspectives whereas, in contrast, human rationale tended to be rich in personal beliefs and experiences.

It should be noted that the multitude of potential interpretations often leads to challenges for annotator agreement in label classification when applied to non-ground truth tasks [2, 32], such as content moderation [50]. For instance, past work [111, 112] reported low agreement when human annotators classify hate speech datasets and confirmed the diversity of moral values impacts what is classified as offensive language [23]. Efforts to remove ambiguity in annotated datasets often employed the "majority vote" [8] (i.e., choose the label with the most votes), which was criticized by researchers [50, 76] for disregarding nuances and potentially introducing biases. Alternatively, Gordon et al. [53] suggests multiple annotator perspectives could be leveraged especially in these subjective decision-making tasks, provided all voices are balanced for fairness [50]. This experiment aligns with recent developments in the *Perspectivist Approach* arguing that diverse perspectives should be included in detriment to single ground truth proxies [10]. Comparably, Sorensen et al. [101] suggests the adoption of *Pluralistic alignment*, when LLMs are faced with open-to-interpretation questions and multiple relevant responses are possible, arguing this improves fairness, aiming for systems that are *"capable of representing a diverse set of human values and perspectives"*.

5.2 Implications for LLM Evaluation

Given that the multiplicity of perspectives is an inherent characteristic of subjective decision-making tasks such as assessing subtle sexism, we suggest that designing an LLM-powered system to assist in subjective decisions requires an ability to evaluate this multiplicity in the corresponding model. LLM evaluations are typically performed via automated evaluation with automatically computed evaluation criteria [35], human evaluation [20], or a combination of the two [24]. Since human evaluation can be costly to operationalize, automated evaluation is often preferred. However, past work has criticized automated evaluation techniques for their fixed nature and for missing out on important qualitative dimensions that only human evaluation can identify [50]. Correspondingly, there is a growing recognition of the advantages of human augmentation of automatic evaluation. For instance, Amazon's AWS Bedrock [24] combines human benchmarking services in addition to automated evaluation to capture nuances in model outputs (e.g., empathy or friendliness) indicating an interest in evaluation metrics that are not ground-truth-based. Similarly, our work used human evaluation as the first step to identify evaluation criteria that are relevant in subjective decision-making, i.e., *Stance* and *Perspectives*. Both sets of parameters were consistently found in human and model assessments, indicating the potential to be systematically applied for evaluation at scale. Therefore, we propose human evaluation to augment existing automated evaluation criteria.

5.2.1 Implications for Bias evaluation in LLMs. Our results also extend the current understanding of bias evaluation in LLMs, often reliant on binary bias classifications (e.g., toxic vs non-toxic). Our *Stance* classification offers more nuanced options (i.e., including *depends* and *no stance* in addition to *sexist* and *not sexist*) that are appropriate for non-ground-truth-based decisions. Further, the Perspectives set (*victim, perpetrator, decision-maker*) is gender-inclusive, in alignment with critiques that current bias detection techniques are often gender binary [47, 70, 93]. As past work has alerted to harmful biased content produced by LLMs [18, 28, 46, 47, 93], and the influence such outputs can have on human decision-makers [36, 38], evaluating which *Stances* and *Perspectives* present in model outputs becomes all the more relevant.

5.3 Implications for human-AI Collaboration

Finally, our work contributes to HCI by offering an exploratory comparative analysis of human and LLM outputs in subjective decision-making scenarios, using quantitative and qualitative methodologies. The human dataset was purposefully not used as ground truth or golden standard, as that concept does not apply. Instead, when considering human-AI decision-making in nuanced, ambiguous and value-based scenarios, it becomes relevant to observe the differences and similarities to consider where there is potential for collaboration and which role AI should play in it. Our results show instances where GPT-3 mimicking humans might not be a desirable trait (e.g., emulating human experiences: *"I have always been a strong advocate of gender equality. I am a mother with three daughters and I have always encouraged them to be strong and independent, encouraging them to aim high and achieve their goals, whatever they may be. I am a feminist in the correct sense of the word. I believe in gender equality."* GPT-3), while newer model iterations diverge from this path. These findings further deepen discussions in AI Alignment [45, 67], as it is a known challenge to define which human values (taking into account pluralistic perspectives [101]) to use as parameters to train a model. Thus, we argue that AI merely emulating human responses is not necessarily the goal for all use cases. Instead, *Stance* and *Perspective* can be leveraged to help calibrate

how similar or divergent to human responses the AI outputs need to be to foster better collaborative outcomes.

5.4 Limitations and Future Work

In this work, we used subtle sexism as one example of a subjective decision-making task and the degree of the potential generalization should be explored in future research. For example, exploring the application of *Stance* and *Perspective* to other discrimination scenarios (such as racism, ageism, ableism, etc.) that would likely also contain a *victim, perpetrator*, and *decision-maker*. Additionally, while it was not the focus of the current work, future work can also compare the sentence structure and content of LLM and human responses [37]. We noticed, for example, that the human dataset often presented grammatical errors, terse sentences and overall heterogeneity in structure (which might account for the slightly lower annotator agreement on this specific dataset). On the other hand, newer models such as LLama 3.1 had very rigid response structures (e.g., numbering the reasons to support the decision). We also foresee a continuation of this work could further explore the concept of combining human and automated evaluation methods to identify an LLM's tendency to consider different Perspectives in their rationale. We see the potential for human evaluators to identify relevant parameters in use cases that require nuanced interpretation, and then train a model to automatically identify them. Other promising parameters could be grammatical elements, convincingness [37] and/or rhetorical strategies used in the responses.

We also acknowledge that while researchers' interpretations are often described as beneficial to the qualitative research process [86], the researchers' experience and values can influence the open coding process in this work. Hence, a Researcher Positionality statement is included in the *Methods* section. There are also limitations associated with using LLMs that should be noted. For example, data leakage and contamination [88] could occur if the scenarios used in the study (which were collected from the internet) were in the training set of the models used for the experiments.

Lastly, while we did not experiment with prompt engineering, future work could explore how different prompting methods may change the Stances and Perspectives present in LLM outputs.

6   CONCLUSION

In this paper, we explore nuances in human-AI collaboration for decision-making in subjective, non-ground-truth scenarios. Using subtle sexism as an example of a value-based and open-to-interpretation scenario, we prompted humans and LLMs (GPT-3, GPT-3.5, GPT-4 and Llama 3.1) to make assessments and justify their rationale. Our mixed-methods analysis demonstrated the responses from humans and LLMs consistently used the same set of Stances (sexist, *not sexist*, *depends* and *no stance*) and Perspectives (*victim*, *perpetrator*, *decision-maker*) but at different frequencies. Thus, each model displayed a distinct tendency to leverage and combine these *Perspectives* in their assessment. Notably, our comparison demonstrates the newer models (GPT-3.5, GPT-4 and Llama 3.1) offer more holistic responses by considering more *Perspectives* (e.g., balancing the perpetrator's intent and the victim outcomes). Further, the most recent GPT-3.5 and GPT-4 models clearly state their limitations as AI agents, suggesting safety guardrails are being

enforced—in contrast to GPT-3's anthropomorphic examples. As new LLMs and model versions are released, we suggest *Stances* and *Perspectives* are parameters that can be leveraged to identify diversity in perspectives in model outputs to compare to, and potentially complement, a human decision-maker's rationale depending on use case.

REFERENCES

[1] Anne Adams, Peter Lunt, and Paul Cairns. 2008. A qualitative approach to HCI research. (2008).

[2] Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 151–154.

[3] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 69–76.

[4] DeM Alex. 2023. *Does temperature go to 1 or 2?* https://community.openai.com/t/does-temperature-go-to-1-or-2/174095/2

[5] Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 international conference on advanced computer science and information systems (ICACSIS)*. IEEE, 233–238.

[6] Paula Akemi Aoyagui, Sharon Ferguson, and Anastasia Kuzminykh. 2024. Exploring Subjectivity for more Human-Centric Assessment of Social Biases in Large Language Models. *arXiv preprint arXiv:2405.11048* (2024).

[7] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2020. OpenCrowd: A Human-AI Collaborative Approach for Finding Social Influencers via Open-Ended Answers Aggregation. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1851–1862. https://doi.org/10.1145/3366423.3380254

[8] Rahma Atallah and Amjed Al-Mousa. 2019. Heart disease detection using machine learning majority voting ensemble method. In *2019 2nd international conference on new trends in computing sciences (ictcs)*. IEEE, 1–6.

[9] 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*. 1–16.

[10] Sam Baron, Kristie Miller, and James Norton. 2014. Groundless truth. *Inquiry* 57, 2 (2014), 175–195.

[11] Manuela Barreto and Naomi Ellemers. 2005. The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities. *European journal of social psychology* 35, 5 (2005), 633–642.

[12] C Daniel Batson, Shannon Early, and Giovanni Salvarani. 1997. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin* 23, 7 (1997), 751–758.

[13] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[16] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[17] Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. DeepHate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*. 11–20.

[18] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[19] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 103–119.

[20] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132* (2024).

[21] Sandra Corlett and Sharon Mavin. 2018. Reflexivity and researcher positionality. *The SAGE handbook of qualitative business and management research methods* (2018), 377–399.

[22] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[23] 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. https://doi.org/10.1145/3630106.3659021

[24] Emilia David. 2023. *Amazon will offer human benchmarking teams to test AI models*. https://www.theverge.com/2023/11/29/23981129/amazon-aws-ai-model-evaluation-bias-toxicity

[25] Amra Delić, Hanif Emamgholizadeh, Francesco Ricci, and Judith Masthoff. 2024. Supporting Group Decision-Making: Insights from a Focus Group Study. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 301–306.

[26] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2021. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354* (2021).

[27] Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. A comprehensive dataset for german offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. 143–153.

[28] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.

[29] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1044, 18 pages. https://doi.org/10.1145/3613904.3642134

[30] Mark Diaz, Razvan Amironesei, Laura Weidinger, and Iason Gabriel. 2022. Accounting for Offensive Speech as a Practice of Resistance. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat (Eds.). Association for Computational Linguistics, Seattle, Washington (Hybrid), 192–202. https://doi.org/10.18653/v1/2022.woah-1.18

[31] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1639–1656. https://doi.org/10.1145/3531146.3533221

[32] Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. 2023. Modeling subjectivity (by Mimicking Annotator Annotation) in toxic comment identification across diverse communities. *arXiv preprint arXiv:2311.00203* (2023).

[33] Brendan Eagan, Jais Brohinsky, Jingyi Wang, and David Williamson Shaffer. 2020. Testing the reliability of inter-rater reliability. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 454–461.

[34] OpenAI et al. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[35] Percy Liang et al. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

[36] Sharon Ferguson, Paula Akemi Aoyagui, Young-Ho Kim, and Anastasia Kuzminykh. 2024. Just Like Me: The Role of Opinions and Personal Experiences in The Perception of Explanations in Subjective Decision-Making. *arXiv preprint arXiv:2404.12558* (2024).

[37] Sharon Ferguson, Paula Akemi Aoyagui, Rimsha Rizvi, Young-Ho Kim, and Anastasia Kuzminykh. 2024. The Explanation That Hits Home: The Characteristics of Verbal Explanations That Affect Human Perception in Subjective Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 517 (Nov. 2024), 37 pages. https://doi.org/10.1145/3687056

[38] Sharon A Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. 2023. Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[39] Jeffrey Fermin. 2023. Meet vera: Allvoices' new AI tool that'll help with employee relations. https://www.allvoices.co/blog/ai-for-employee-relations

[40] A Field. 2012. Discovering Statistics Using R.

[41] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1362–1374.

[42] Raymond Fok and Daniel S Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* (2023).

[43] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Florence, Italy, 94–104. https://doi.org/10.18653/v1/W19-3510

[44] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation* (2024), 1–28.

[45] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[46] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 205–216.

[47] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770* (2023).

[48] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems* 36 (2024).

[49] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–29.

[50] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *Comput. Surveys* 55, 13s (2023), 1–32.

[51] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[52] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[53] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[54] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground (less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 688–704.

[55] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[56] Katherine J Hall. 1997. *Subtle Sexism: Current Practice and Prospect for Change*. Sage Publications.

[57] Katherine J Hall. 2016. *" They believe that because they are women, it should be easier for them." Subtle and Overt Sexism toward Women in STEM from Social Media Commentary*. Virginia Commonwealth University.

[58] Matthew D Hammond, Petar Milojev, Yanshu Huang, and Chris G Sibley. 2018. Benevolent sexism and hostile sexism across the ages. *Social Psychological and Personality Science* 9, 7 (2018), 863–874.

[59] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI Become Reliable Source to Support Human Decision Making in a Court Scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 195–198. https://doi.org/10.1145/3022198.3026338

[60] Meghan L Healey and Murray Grossman. 2018. Cognitive and affective perspective-taking: evidence for shared and dissociable anatomical substrates. *Frontiers in neurology* 9 (2018), 491.

[61] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[62] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[63] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric P. S. Baumer. 2019. Where is the Human? Bridging the Gap Between AI and HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290607.3299002

[64] Rune Møberg Jacobsen, Lukas Bjørn Leer Bysted, Patrick Skov Johansen, Eleftherios Papachristos, and Mikael B. Skov. 2020. Perceived and Measured Task Effectiveness in Human-AI Collaboration. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3383104

[65] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons* 61, 4 (2018), 577–586.

[66] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1723–1734.

[67] Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports* 14, 1 (2024), 19399.

[68] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921* (2020).

[69] Rafal Kocielnik, Shrimai Prabhumoye, Vivian Zhang, Roy Jiang, R. Michael Alvarez, and Anima Anandkumar. 2023. BiasTestGPT: Using ChatGPT for Social Bias Testing of Language Models. arXiv:2302.07371 [cs.CL]

[70] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*. 12–24.

[71] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2023. Watch Your Language: Large Language Models and Content Moderation. *arXiv preprint arXiv:2309.14517* (2023).

[72] Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *arXiv preprint arXiv:2403.08035* (2024).

[73] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[74] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[75] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[76] Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*.

[77] Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474* (2024).

[78] Shuai Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. 2024. Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-Assisted Decision Making. *arXiv preprint arXiv:2403.01791* (2024).

[79] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289* (2020).

[80] Meta. 2024. *Introducing Llama 3.1: Our most capable models to date*. https://ai.meta.com/blog/meta-llama-3-1/

[81] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.

[82] Shyamal Mishra and Preetha Chatterjee. 2023. Exploring ChatGPT for Toxicity Detection in GitHub. *arXiv preprint arXiv:2312.13105* (2023).

[83] Chelsea Mitamura, Lynnsey Erickson, and Patricia G Devine. 2017. Value-based standards guide sexism inferences for self and others. *Journal of Experimental Social Psychology* 72 (2017), 101–117.

[84] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint*

[85] *arXiv:2006.08328* (2020).

[86] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[86] Michael Muller. 2014. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*. Springer, 25–48.

[87] Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 742, 22 pages. https://doi.org/10.1145/3613904.3642025

[88] Arvind Narayanan and Sayash Kapoor. 2023. *GPT-4 and professional benchmarks: the wrong answer to the wrong question*.

[89] Steven Nicholls, Stuart Cunningham, and Richard Picking. 2018. Collaborative artificial intelligence in music production. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*. 1–4.

[90] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture* 25, 2 (2021), 700–732.

[91] OpenAI. 2023. *API Reference*. https://platform.openai.com/docs/api-reference/introduction Last accessed 02 January 2023.

[92] OpenAI. 2023. *GPT-4 Technical Report*. https://arxiv.org/pdf/2303.08774.pdf

[93] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1246–1266.

[94] Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. sage.

[95] Advait Sarkar. 2024. AI Should Challenge, Not Obey. *Commun. ACM* 67, 10 (Sept. 2024), 18–21. https://doi.org/10.1145/3649404

[96] Advait Sarkar. 2024. Large Language Models Cannot Explain Themselves. *arXiv preprint arXiv:2405.04382* (2024).

[97] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. *Ambiguity-Aware AI Assistants for Medical Data Analysis*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376506

[98] Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics* 9 (2021), 1408–1424.

[99] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 920, 20 pages. https://doi.org/10.1145/3613904.3642159

[100] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).

[101] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).

[102] Mark Steyvers and Aakriti Kumar. 2023. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science* (2023), 17456916231181102.

[103] Janet K Swim and Laurie L Cohen. 1997. Overt, Covert, And Subtle Sexism: A Comparison Between the Attitudes Toward Women and Modern Sexism Scales. *Psychology of women quarterly* 21, 1 (1997), 103–118.

[104] Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

[105] Koen van Turnhout, Arthur Bennis, Sabine Craenmehr, Robert Holwerda, Marjolein Jacobs, Ralph Niels, Lambert Zaad, Stijn Hoppenbrouwers, Dick Lenior, and René Bakker. 2014. Design Patterns for Mixed-Method Research in HCI. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (Helsinki, Finland) *(NordiCHI '14)*. Association for Computing Machinery, New York, NY, USA, 361–370. https://doi.org/10.1145/2639189.2639220

[106] Lucía Vicente and Helena Matute. 2023. Humans inherit artificial intelligence biases. *Scientific Reports* 13, 1 (2023), 15737.

[107] Laura M Vowels. 2024. Are chatbots the new relationship experts? Insights from three studies. *Computers in Human Behavior: Artificial Humans* (2024), 100077.

[108] Dakuo Wang, Justin D. Weisz, Michael J. Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla R. Tausczik, Horst Samulowitz, and Alexander G. Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *CoRR* abs/1909.02309 (2019). arXiv:1909.02309 http://arxiv.org/abs/1909.02309

[109] Junqi Wang, Chunhui Zhang, Jiapeng Li, Yuxi Ma, Lixing Niu, Jiaheng Han, Yujia Peng, Yixin Zhu, and Lifeng Fan. 2024. Evaluating and Modeling Social

Intelligence: A Comparative Study of Human and AI Capabilities. *arXiv preprint arXiv:2405.11841* (2024).

[110] Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390* (2022).

[111] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[112] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[113] Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. *arXiv preprint arXiv:2311.10227* (2023).

[114] Michael Williams and Tami Moser. 2019. The art of coding and thematic exploration in qualitative research. *International management review* 15, 1 (2019), 45–55.

[115] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[116] Rongwu Xu, Zi'an Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024. Walking in Others' Shoes: How Perspective-Taking Guides Large Language Models in Reducing Toxicity and Bias. *arXiv preprint arXiv:2407.15366* (2024).

[117] Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus* 151, 2 (2022), 183–197.

[118] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598.

[119] Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, and Konstantinos Psounis. 2023. Efficient Toxic Content Detection by Bootstrapping and Distilling Large Language Models. *arXiv preprint arXiv:2312.08303* (2023).

[120] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[121] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A APPENDIX A - EXAMPLE OF PROMPT IN QUESTION-ANSWER FORMAT

In this work, we used the Scenarios Dataset as prompts in question-answer format. The Scenario was collected in online platforms (see Methods) and then written in question-answer format to prompt responses from humans, GPT-3, GPT-3.5, GPT-4 and Llama 3.1. See examples below:

Example of Scenario (collected online): *Explaining any complaint or bad mood from a woman by "oh she must just be on her period"*

Example of human response: *This scenario IS sexist, because the underlying assumption is that women are profoundly affected by their hormones.*

Example of GPT-3 response: *" No, this is not sexist. This is a common and acceptable explanation of female behavior, which is a well-documented reality in human sociology and psychology."*

Example of GPT-3.5 response: *"Yes, the statement "Explaining any complaint or bad mood from a woman by 'oh she must just be on her period'" can be considered sexist. This statement perpetuates a harmful stereotype that suggests women's emotions or behavior are solely influenced by their menstrual cycle, minimizing their experiences and dismissing their valid concerns. It promotes the idea that women's emotions are irrational or invalid, which is unfair and discriminatory.It is important to treat individuals with respect and not make assumptions about their emotions based on their gender or biological characteristics".*

Example of GPT-4 response: *"Yes, this statement is considered sexist. The reason is that it perpetuates a negative stereotype about women, attributing their feelings, complaints, or moods solely to their hormonal cycle. This dismisses and invalidates the legitimacy of their concerns, grievances, or emotional states, implying they can't have valid reasons other than biological ones. This kind of sexism also contributes to the trivialization of women's experiences and feelings."*

Example of Llama 3.1 response: *"Yes, this phrase can be seen as sexist. Here's why: 1. **Stereotyping**: It perpetuates a negative stereotype that women's hormonal cycles (specifically, their menstrual periods) are the sole reason for their emotions, complaints, or moods. This oversimplifies the complexities of women's experiences and ignores the myriad factors that contribute to their emotions. 2. **Dismissal**: By attributing a woman's emotions or complaints to her period, this phrase can come across as dismissive and invalidating. It may convey that her feelings or opinions are not worthy of consideration or respect, simply because they are associated with her menstrual cycle. 3. **Biological essentialism**: It reduces women's experiences and emotions to their biology, implying that their actions and feelings are determined by their reproductive cycles. This type of thinking ignores the fact that women are individuals with unique experiences, thoughts, and feelings, and that their emotions and actions are influenced by a range of factors, including their environment, relationships, and personal circumstances. 4. **Lack of empathy**: This phrase can convey a lack of empathy or understanding, implying that women's emotions and experiences are not worthy of attention or consideration. It may also suggest that their emotions are less valid or less important than those of men. 5. **Unrealistic expectations**: It perpetuates the idea that women should be able to ""tough it out"" or manage their emotions in a particular way, without acknowledging that menstrual cramps, fatigue, and mood swings are common symptoms of PMS that can impact women's daily lives. It's essential to approach conversations about women's emotions and experiences with empathy, understanding, and respect. Instead of attributing women's complaints or bad moods to their periods, we can ask open-ended questions like ""What's going on?"" or ""How can I support you?"" to encourage more nuanced and respectful communication.""*

B EXAMPLES OF INVALID RESPONSES

Some of the responses collected for the human and GPT-3 datasets were excluded from analysis because they either did not contain an assessment (i.e. the response to the question: is this sexist?) or a rationale supporting that assessment (i.e. the response to the question: why or why not). The quality of the rationale was not analyzed, only if it was present or not. See more details in our Methods section. See examples in the table below.

Table 7: Examples of responses considered invalid and removed from datasets for analysis for not containing an assessment and/or a rationale

Prompt Example	Example of Response	Author	Assessment	Rationale	Status
Is this scenario sexist: dad gets mad at me for relaxing and not helping cook/clean. Both brothers are also relaxing and nothing is said to either of them." *Why or why not?*	"This statement is not sexist because the brothers could have already done their chores, or may have a history of cleaning and cooking way more than the sister."	Human	Present: "yes"	Present: "wording is degrading towards women"	Valid Response
Is this scenario sexist: "Take the military example. Someone points out that there is benevolent sexism involved in barring women from serving in combat roles and thus being immune from a compulsory draft. I think this is true." *Why or why not?*	"I have a job"	Human	Not present	Not present	Invalid Response
Is this scenario sexist: "Calling women "female"" *Why or why not?*	"No. See here for why not."	GPT-3	Present: "No."	Not present	Invalid Response