Tell Me What You Know About Sexism: Expert-LLM Interaction Strategies and Co-Created Definitions for Zero-Shot Sexism Detection

Myrthe Reuver♥[1], Indira Sen[2], Matteo Melis[3], Gabriella Lapesa[3,4]
[1]Populytics [2]University of Mannheim
[3]GESIS - Leibniz Institute for the Social Sciences [4]Heinrich-Heine University Düsseldorf
[1]myrthe@populytics.nl, [2]indira.sen@uni-mannheim.de, [3]first.last@gesis.org

Abstract

This paper investigates *hybrid intelligence* and collaboration between researchers of sexism and Large Language Models (LLMs), with a four-component pipeline. First, nine sexism researchers answer questions about their knowledge of sexism and of LLMs. They then participate in two interactive experiments involving an LLM (GPT3.5). The first experiment has experts assessing the model's knowledge about sexism and suitability for use in research. The second experiment tasks them with creating three different definitions of sexism: an expert-written definition, an LLM-written one, and a co-created definition. Lastly, zero-shot classification experiments use the three definitions from each expert in a prompt template for sexism detection, evaluating GPT4o on 2.500 texts sampled from five sexism benchmarks. We then analyze the resulting 67.500 classification decisions. The LLM interactions lead to longer and more complex definitions of sexism. Expert-written definitions on average perform poorly compared to LLM-generated definitions. However, some experts do improve classification performance with their co-created definitions of sexism, also experts who are inexperienced in using LLMs.

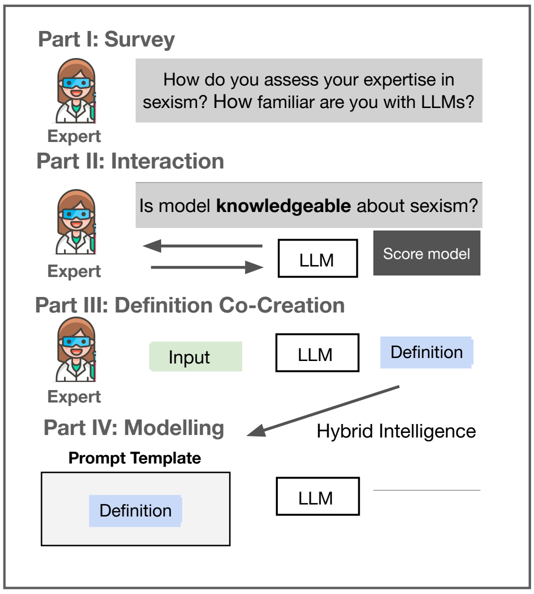

Figure 1: Experts participate in a survey (part I) as well as two interactive experiments (part II and III), after which we perform zero-shot classification experiments (part IV) with two LLMs, using the sexism definitions created during the interaction experiments.

1 Introduction

Large Language Models (LLMs) with chat interfaces are an increasingly popular tool in various scientific fields, for a variety of tasks: from writing assistance to data annotation and data analysis. These interactive models produce pleasant and convincing (while not necessarily factually correct) conversations (Ji et al., 2023), due to their training on human feedback. In social science, LLMs are used for tasks related to social science research questions (Dey et al., 2024), in particular to detect complex social constructs in text (Weber and Reichardt, 2023), including sexism (Sen et al., 2023), which is the focus of this paper.[1]

LLMs can be evaluated with computational measurements, e.g. by testing the performance of a model on datasets that have been annotated for a specific construct. However, this approach is not the only form of evaluation, and it can miss important nuances that knowledgeable domain experts want to consider when addressing their research question using these models. Combining the strengths of such nuances in human expertise with

♥ Majority of work done while employed at the Vrije Universiteit Amsterdam and visiting the Computational Social Science department at GESIS on a GESIS Visiting Junior Researcher grant.

[1]WARNING: This paper researches sexism, and includes sensitive and hateful content. The researchers in no way condone sexism or hate of any kind.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 8453–8482

the strengths of computational models is called *hybrid intelligence* (Dellermann et al., 2019). This means that each (human and model) perform better at complex tasks together than they could do separately.

Our research pipeline analyzes this connection between human expertise and LLMs. We address two high-level questions:

Q1 How do experts interact with instruction-tuned LLMs when assessing them for further use in research?

Q2 What do we learn from the expert / LLM interactions, and can we use these insights for improving computational detection?

We focus on **sexism detection** to answer these questions, and do so by analyzing interactions between sexism researchers and LLMs in a four-component pipeline. These four study components are explained below, and illustrated in Figure 1.

In **Part I: Survey**, 9 expert participants answer questions about their expertise in sexism research, as well as their habits when using and evaluating LLMs. **Part II: Interaction** then has these experts use an interactive interface to observe and record their interactions with LLMs. First, sexism experts are asked to assess the model for suitability for sexism research in any way they see fit. We then conduct a qualitative analysis of the interactions by creating a taxonomy of strategies. Some experts ask for definitions, others generate examples, or let the LLM analyze examples.

After that, **Part III: Construct Definition Co-Creation** asks experts to co-create sexism definitions with the LLMs. For each expert, we collect three definitions: one produced before interacting with the LLM, i.e., their own working definition of sexism (expert-written), the one that they deemed best among those produced by the LLM (LLM-generated), and the one that the expert co-created with the LLM, i.e., by editing and adding new aspects to the definition.

Lastly, **Part IV: Modeling** uses these definitions in zero-shot sexism classification. The three definitions (expert-written, LLM-generated, co-created) from the nine experts are used in a prompt template to detect sexism with zero-shot classification in 2.500 texts from five sexism benchmarks, using LLM `GPT4o`. In these 67,500 classification decisions, we find that expert-written definitions perform poorly, while models perform better with LLM-generated definitions, and only some experts improve performance with co-written definitions.

Our contributions are at multiple levels:[2]

- At the *methodological level*, our study is the first to combine the different methodologies (survey, interactive experiment, modeling) in one pipeline (refer to Section 2 for a detailed discussion of previous work). Our contribution is a framework and method for eliciting and recording multi-turn LLM-human interactions, specifically for researching collaboration and hybrid intelligence for construct detection;
- At the *level of novel resources*, our contribution is a dataset containing a) expert-LLM interactions on sexism b) sexism experts' ratings of `GPT4o`'s suitability for sexism research, as well as c) for each expert, three different definitions of sexism (an expert-written, LLM-generated, and co-created one). These datasets are complemented by the insights we gathered through in-depth qualitative analysis of the interactions between human experts and LLMs.
- At the level of *computational modeling*, we use the collected definitions and employ them for zero-shot LLM classification, effectively connecting findings on the expert's expertise, prompting strategies, and model performance on sexism detection, as well as researching the impact of hybrid intelligence on zero-shot classification.

2 Related Work

Our work relates to several domains of research: interaction between humans and language technology, zero-shot prompting for complex construct detection, and the use of definitions in prompting. This section identifies how our study fills a research gap at the connection between these domains.

How People Use Language Technology Earlier work has used surveys and interviews to determine how end-users of NLP technologies conceptualize and use these systems (Jakesch et al., 2023). This is especially common in cases where there is potential harm for at-risk communities, for instance with translation systems and LGTBQ+ individuals (Lauscher et al., 2023; Ungless et al., 2023). This is also done when end-users do not belong to at-risk communities: Ter Hoeve et al. (2022) surveys users on why and how they use summarization systems.

[2] All experimental artifacts, including survey templates, code, data, and other material are available in our repository: github.com/myrthereuver/ExpertInteractionsZeroShotSexism

In addition to asking end-users questions, the methodology employed in this paper also belongs to the domain of ethnographic methods. Such methods aim to study interactions of people and their surroundings, and observe their behavior in natural, open-ended, and unguided settings (Brewer, 2000). These observations are then studied to find patterns, commonalities and perhaps avenues for further experiments and research. Ethnographic analysis of language technology systems can shed light on how such systems are currently used, as well as on how the use and design of a system can be improved for contexts (Hughes et al., 1994). Such studies for instance observe interactions with chat systems by hospital personnel (Wang et al., 2020), use of news recommendation systems (Schjøtt Hansen and Hartley, 2023) or e-Commerce platforms (Kusk and Bossen, 2022). In machine translation, Désilets et al. (2008) use ethnographic techniques such as contextual inquiry to analyze usage of machine translation by professional translators. Their study highlights how such observations can be used to develop new technical approaches.

Zamfirescu-Pereira et al. (2023) analyze the prompt designs of 10 participants with no LLM experience in a no-code environment, tasking them improve a virtual cooking chatbot. They find these non-experts lack systematicity in their prompt design, and stop easily because of errors.

Generative models are also studied with ethnographic or observational analysis in the social sciences (Liu, 2023; de Seta et al., 2024). These studies analyze the inputs and outputs of models, and how humans react to them in the form of perceived usefulness or emotions their responses evoke in a user. However, these studies do not directly connect their observations to more traditional methods of evaluation in NLP.

Zero-Shot Prompting for Social Constructs Jacobs and Wallach (2021) introduce practices on measurement modeling from the social sciences to computer science. They argue that the computational operationalization of a complex construct should not only be evaluated on *predictive validity* (e.g. classifying an unseen test set), but should also involve testing the broader notion of *construct validity*. Reuver et al. (2021) show how social science theory can help looking beyond such task-based evaluation when there is a connection to a societal challenge, such as a lack of diversity in news recommendation.

Work on LLMs understanding of constructs such as hate and racism has used zero-shot prompting, that is: asking an LLM in a language-based prompt without performing any additional training, to perform a specific classification task. For instance, Shaikh et al. (2023) evaluate the Chain of Thought (CoT) prompting (Wei et al., 2022) technique for identifying and responding to harmful or toxic questions about people, and find it increases both nonsensical reasoning and biased answers based on generalizations of socio-demographic aspects, e.g. "[racial group X] is dumber". A crucial role is here also for the nuance and human context in the prompt: Beck et al. (2024) find adding sociodemographic information in prompts for subjective NLP tasks can influence performance. Jiang et al. (2024) infuse prompts with annotator information with five prompting strategies for sexism detection with LLMs, and find that models are biased by annotators' attitudes. Therefore, information in the prompt matters. Which leads us to our next component: definitions in zero-shot prompting.

Definitions in Prompting for Social Constructs Other work also researches the use of definitions in prompts for zero-shot classification with LLMs. This is done for detecting social science concepts, and also for hate and sexism specifically. Peskine et al. (2023) research expert-written vs GPT-generated definitions for classifying tweets into different categories of conspiracy theories. Their results indicate that human-written definitions are better than ones written by LLMs, but they have no examples that are co-written between experts and LLMs. They find that GPT definitions similar to human-written definitions are better in performing on unseen test sets. Khurana et al. (2025), building on earlier work that looks into different granular aspects of hate speech definitions (Khurana et al., 2022), analyze whether Transformer models actually reflect their dataset's definitions of hate speech. It introduces a method using a user-specific definition of hate speech, and quantifies to what extent a model reflects the intended definition. They find most models do not capture the aspects of hate that are defined in their dataset's definition. Most recently, Korre et al. (2025) present a dataset of hate speech definitions and analyzes the semantic properties of definitions and their classification performance in zero-shot classification with LLMs. They find that hate speech definitions and their components are culturally specific.

Research Gap Earlier work has observed or interviewed end-users of LLMs and related technologies, conducted zero-shot prompting experiments with LLMs for complex construct detection, or analyzed definitions used in computational detection of sexism or hate. Others have studied the role of different definitions in zero-shot prompting, but these either use existing definitions of the construct (Korre et al., 2025), or compare expert-made and LLM-made definitions of the construct (Peskine et al., 2023). To our knowledge, no previous work has expert participants completing a specific task together with the LLM (interactions about sexism and co-creation of definitions of sexism), and then using the outputs of this task for zero-shot learning on benchmark datasets. This study connects these aspects in a four-component pipeline.

3 Methods

Our methodology connects a qualitative, in-depth analysis of experts interacting with LLMs to computational detection of sexism, in a four-part pipeline. The following sections describe every step in this pipeline.

3.1 Part I: Pre-Interaction Survey

First, we conduct a pre-interaction survey on our participants, who are sexism experts. Our intention with the survey step is twofold: (1) find how and how confidently researchers of sexism use LLMs, and (2) collect information on the participants for analyzing their interactions later in the pipeline.

Participants We reached out to sexism experts within and outside our network. Study participation took place from June until September 2024. Initially, 11 experts completed the pre-interaction survey; however, two participants (experts 7 and 8) did not complete the interactive experiments, which resulted in nine fully completed experimental responses. A participant dropout rate of 18% is considered a low dropout rate for online experiments, with previous work mentioning 20% (Gagné and Franzen, 2023) to 30% (Galesic, 2006).

Our participants are sexism researchers connected to research universities and institutes in Europe and the United States, with varying levels of computational experience. Nine participants is a small group, but is considered a suitable number of participants for a qualitative, in-depth study (Guest et al., 2006; Hennink and Kaiser, 2022). Participants gave their informed consent about the study and its purpose. Participants were offered 6 euros for 30 minutes of participation, the minimum wage in Germany, which some declined.[3]

All participants are self-identified researchers of sexism, with a mean of 6.09 (SD = 2.9) years of expertise in research related to sexism, ranging in career stage from doctoral student to assistant professor. Only one participant was a man, the rest identified as women.[4]

Survey Design Our survey was designed with Qualtrics,[5] with responses recorded anonymously. In addition to information about their expertise and career stage, we ask experts about their previous experience with LLMs. We then use four items from a 1-7 Likert scale to assess participants' self-confidence in sexism research. These were adapted items from a validated expertise and self-confidence measurement tool called the Collective Self Esteem Scale (CSEM) (Luhtanen and Crocker, 1992), where we used items related to professional self-esteem. Validation means the scale has been tested to consistently and accuracy measure confidence (Boateng et al., 2018). This scale contains items like "I have done substantial research related to sexism, hate towards women, or related concepts" and "I feel like I have substantially less ability in detecting or researching sexism than others."

Our second set of items elicited attitudes towards LLMs in social science research. Participants responded on a 1 - 5 agreement Likert scale to statements such as "I want to use LLMs in my next research project" and "I want to learn more about LLMs". These questions were not taken from a standardized measurement instrument, and we used a standard 5-point Likert scale. After completing this pre-survey, participants were introduced to the first interaction with the LLM.

3.2 Part II: LLM-Expert Interactions

The interactions between LLMs and sexism researchers allow us to analyze how experts assess the knowledge of LLMs about their domain of expertise. The interactions also allow us to observe experts' unguided interactions with LLMs.

[3] We completed an ethics check about our study at the VU Amsterdam, and participants were explicitly debriefed about the potential of seeing harmful responses from the LLM. See Appendix A for more information on our precautions to ensure a responsible participant study.

[4] More detailed information about the participants can be found in Appendix B.

[5] https://www.qualtrics.com/

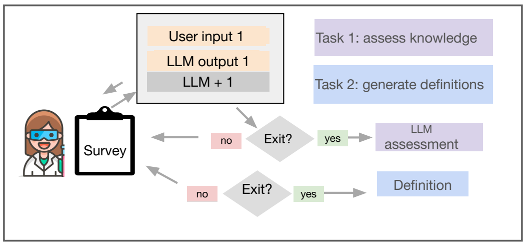

Figure 2: Explanation of the interactive experiments of Part II and Part III of our pipeline.

Experimental Design The participants were instructed to assess the LLM's knowledge of sexism in any way they saw fit, for a maximum of 10 input-output interactions, in an interactive environment. Our interaction model is `gpt-3.5-turbo`[6]. Experts could choose to end the interaction at any time by exiting a loop after indicating that they were ready to assess the model's knowledge. See Figure 2 for the basic representation of the interface and loop. We release our Qualtrics template for future use of this methodology. More information on this design is in Appendix B.1.

After the interactions, experts rate how satisfied they are with the model's knowledge and suitability for sexism detection on four items with a Likert scale range from 1 to 5 indicating from full agreement to full disagreement with the statement, which included "This model can distinguish nuances of sexism" and "I trust this model's capabilities".

3.3 Part III: Co-Creation of Definitions

Experimental Design Our second interactive experiment concerns co-creating definitions of sexism. The participants are told that the definitions will be used for sexism detection. They are then asked for their own comprehensive definition of sexism (to which we will refer in the reminder of the paper as *expert-written*). The participants also have available a maximum of 10 interactions (prompts) to co-create a definition with the same model. In this set-up, the participants can again choose to end an interaction any time. At the end of the interaction, the participants are asked to review the full interaction and perform two actions: a) select their preferred definition among those generated by the LLMs (referred to in the paper as *LLM-generated* definition) and b) copy and edit their preferred definition (we refer to this definition as *co-created*). After the interactions, the participants were asked to complete a 7-point Likert scale to rate their satisfaction with the co-created definition, as well as rate on five different aspects of the definition: validity, comprehensiveness, simplicity, covering all aspects of sexism, and its depth. Afterwards, participants could exit the survey.

3.4 Part IV: Modeling

Our modeling experiments consist of zero-shot classification with generative LLMs for sexism detection. These experiments connect the experts' strategies and definitions to benchmark performance, and measure the impact of *hybrid intelligence* (Dellermann et al., 2019) - whether co-creating definitions allows models to use the strength of both expert and model knowledge. Although LLM classification has limitations, in particular due to the reliability and reproducibility issues of these models (Reiss, 2023), we still consider it a relevant benchmark due to the growing application of LLM and the potential for applied measurement designs (Atreja et al., 2024).

Definitions and Prompts Our modeling experiments use three definitions of sexism from each of the nine experts: the *LLM-generated* definition, the *expert-written* definition, and the *co-created* definition. This leads to 27 different definitions of sexism used for prompting. Our prompt template was based on the one in Sen et al. (2023) (see Appendix C), and completed with each of the 27 definitions of sexism.[7] The LLMs were prompted to respond with "sexist", "not sexist", or "don't know".

Models We use `GPT4o`[8] for our zero-shot prompting for sexism detection. We report generated results with a temperature of 0, since this setting is the most deterministic, and a higher temperature leads to generation of less probable tokens (Renze, 2024). A temperature of 0 is commonly used by social scientists when using ChatGPT for labelling social constructs (Fatemi et al., 2023), and also by recent computational work assessing the effect of different definitions on zero-shot LLM prompting (Korre et al., 2025).[9]

[6]Version with knowledge cut-off September 2021

[7]Table 7 in the Appendix contains all these definitions.

[8]`https://openai.com/index/hello-gpt-4o/`.

[9]We completed additional experiments with a different GPT temperature and a second LLM, `LlaMa-3.1-70B-Instruct` (Dubey et al., 2024). Our prompting with LlaMa, see Appendix H, shows lower results and less variability over definitions and datasets. The same counts for a higher temperature of GPT4o, see Appendix G.

Evaluation To limit the computational costs[10] while still ensuring coverage of different sexism benchmarks, we use samples rather than full benchmark datasets for zero-shot classification. We evaluate on 500 annotated texts randomly sampled from each of the following English-language sexism datasets. These contain content from several social media platforms or content resembling social media text:

1. The "Call me Sexist But" ("CallMeSexist") dataset collected by (Samory et al., 2021), which has three types of data — tweets, survey scale items assessing sexist attitudes, and adversarial or counterfactually augmented versions of the two former data types.
2. "Explainable Detection of Sexism" (EDOS) dataset, which is based on a shared task in Semeval 2023, with sexist texts from Reddit and Gab (Kirk et al., 2023).
3. "Reddit Misogyny dataset" curated by Guest et al. (2021).
4. The EXIST sexism dataset, comprising of tweets (Rodriguez-Sanchez et al., 2021)
5. Finally, we also use a subset of the Hatecheck dataset (Röttger et al., 2021), where the target of hate is women. This dataset consists of test suites for evaluating the robustness of hate speech classifiers.

Several of these datasets have fine-grained sexism categories (Samory et al., 2021; Kirk et al., 2023). However, these categories are neither consistent across datasets and nor do they map to the dimensions invoked in the expert and LLM definitions, hence we use the binary sexism labels.

In terms of class distribution, our samples are representative of the original datasets. The RedditGuest dataset (13% sexism) has sexism as a rare class compared to the CallMeSexist and Hatecheck datasets (closer to 50% each class). The EDOS and EXIST datasets are somewhat in between these, with each around 25% sexist. We believe these samples a) fairly represent performance on the respective benchmarks and b) fairly encode that sexism vs. non-sexism content in the real-world online social media contexts that these datasets represent, where sexism is not always 50% of the data. See Appendix D for the respective distributions per full and sampled dataset.

[10]See Appendix C.2 for a specification of costs.

4 Results

4.1 Part I: Pre-survey

Below we discuss the key results of our pre-survey measurements. Detailed results for all experts and outcomes are in Appendix B.4.

Attitudes towards LLMs Experts' attitude towards LLMs, and how confident they are in using and evaluating them, are moderately positive to high (M = 3.47, SD = 1.06) for all experts, with Expert 6 the only negative outlier. Notably, this is an expert without direct prior LLM knowledge.

Self-confidence on sexism research All the experts self-report to be experienced researchers of sexism (M = 5.55, SD = 1.23) and are confident about their own knowledge of sexism (M = 5.91, SD = .67).[11].

4.2 Part II: Interactions with LLMs

Interaction Data Out of the 10 possible interactions, participants use an average of 4 interactions (min: 1 - max: 7) with a standard deviation of 1.63. Broadly, experts use different strategies, from asking questions to classification examples.

Model Suitability Rating After interacting with the LLM, participants rate its suitability for research on sexism moderately high on a 5-point Likert scale M = 3.47 (SD = 1.06). Expert 6, who previously reported low confidence in using LLMs, reports the second highest suitability (M = 4.75).

Qualitative Coding and Analysis of Interactions We follow an inductive approach to create a taxonomy of different types of interactions between experts and the LLMs. We create two disjoint, but related, taxonomies — one characterizing the experts' questions, directions, and instructions to the LLMs, and another for characterizing the LLMs' responses. For both, we use a grounded theory-based approach (Charmaz, 2015), where two annotators (both of whom are authors of the papers) independently assess the prompts (for expert taxonomy) and the responses (for the LLM response taxonomy), coming up with categories. The annotators then discuss to create a taxonomy consisting of

[11]The difference in these two averages is due to Expert 10 reporting a low score (2) on the first question, however they also reported a total of 8 years experience in the social science research, and given their answers to other items, we hypothesize this was an error possibly due to the two earlier survey items being framed negatively (i.e. phrased with a negation, such that a low score actually meant high confidence).

shared categories, merging certain categories, and resolving disagreements. The taxonomy is then applied to the LLM-expert interactions. Next, a third annotator (also a paper author) applies the taxonomy to a subset of LLM-expert interactions. This data-driven categorizations of expert prompts and LLM responses is summarized in Table 1.

4.3 Part III: Co-Creation of Definitions

Interaction Data and Definition Ratings Recall that each expert produces three definitions in these interactions: *expert-written*, *LLM-generated*, and *co-created*. This definition-related task shows more interactions with the LLM (M = 9.11 out of 10) compared to the previous task of questioning the LLM's knowledge of sexism: out of the nine participants, six use all (10 out of 10) available interactions. Overall experts rate the co-created definitions positively (M = 5.09, SD = 0.76).[12]

Qualitative Coding and Analysis of Interactions Following the same methodology illustrated in Section 4.2, i.e., a ground theory approach, we identify the strategies of the experts when co-creating a definition of sexism with the LLM, and organize them into a taxonomy of strategies. We notice strategies such as inducing personas ("You are an expert."), testing the LLM on closely related constructs (misogyny), or explicitly stating the experts' own goal ("I want to detect sexism"). See Table 2 for a full overview of these, with examples.

Definition Change and Similarity Expert-written definitions are generally shorter (M = 34.44, tokens, SD = 25.24), than the LLM-generated definitions (M = 119.89, SD = 58.29) and the co-created definitions (M = 110.55, SD = 56.44). Most experts do minimal edits on the LLM definition for their co-created definition.

Expert-written definitions align with different aspects in definitions of the five sexism benchmarks. For example, Expert 2 wrote a definition that mentions stereotypes and disrespect, similar to the definition of EXIST data set, which also mentions stereotyping and prejudice.[13] Expert 11's initial definition is relatively short and generic, and becomes much more specific to gender issues after interacting with the LLM.

[12]Figure 7 in the Appendix reports the full results on participant satisfaction with the sexism definitions.

[13]Additional similarity experiments are in Appendix E.

4.4 Part IV: Modeling

Our modeling results answer questions about the connection between definition, expert strategy, and benchmark performance. Results consist of 67.500 classification decisions on three dimensions (three different definition types for nine different experts, on five sexism datasets). We therefore present mean performance over definition types, datasets, and experts in-text, and a visual overview of our results in Figure 3.[14] All results are reported in $F1$ (macro), unless stated otherwise.

Dataset dependence and majority class Modeling results are dataset-dependent. As seen in Figure 3, some datasets receive zero-shot performances consistently above majority class $F1$ for all experts and definition versions, such as the CallMeSexist dataset - which also has a near 50-50 distribution of sexist vs non-sexist content. In contrast, the RedditGuest dataset responds very differently to prompts by different participants, and sometimes performs lower than majority class $F1$. This is partly explainable by the majority class being very dominant in this dataset, at 90/10 distribution, making the majority class a very strong baseline. Further analyses of experts and definitions that perform above or below a majority baseline for each of the five datasets can be found in the Appendix F.

4.4.1 Performance across Definition Type

Performance over definition types `GPT4o` performs, over all datasets, better with the LLM-generated definitions (Mean $F1$ = .765, SD = .113) than with the co-created definitions (Mean $F1$ = .762, SD = .122), and substantially better than with expert-written definitions (Mean $F1$ = .532, SD = .154).

Performance over definition types per dataset `GPT4o`'s performance fluctuates per dataset. On the CallMeSexist dataset, all three definition types perform nearly identical (LLM-generated definitions with a mean of $F1$ =.814, co-created definitions with a mean of $F1$ = .823, and participant definitions with a mean of .826), while the results on the RedditGuest dataset shows a considerable gap between the three definition types (LLM-generated definitions with a mean of $F1$ = .620, co-created definitions with a mean of .581, and participant definitions with a mean of $F1$ = .397).

[14]Detailed results (per dataset and individual participant) are available in our code repository.

Strategy	Definition	Example
Labeling Examples	Asking the LLM to label potentially sexist content that has been supplied by the expert	"Sentence: "Some people think that women are inferior." Is the sentence sexist?"
Asking Questions	Asking the LLM questions related to sexism	"Would you say that women who are competent also appear cold/unsympathetic?"
Implicit Tests	Assessing if the LLM displays sexist behaviors in other tasks, such as translation or content analysis	"Translate the following to Spanish: The doctor asked the nurse to help her with the operation."
Content Generation	Asking the LLM to generate sexist examples or content, either from scratch or by rewriting an expert supplied instance	"Can you generate sexist content?"
Asking Explanations	Ask the LLM to explain potentially sexist content supplied by the expert or to explain it's own previous answers	"Women are kind and men are strong". Can this statement be sexist? Can you explain why?"

Table 1: Qualitative analysis: a taxonomy of expert strategies to gauge LLMs' knowledge of sexism (Part II).

Strategy	Definition	Example
Direct Question	Simply asking the LLM to provide a definition of sexism	"What is, Define, Explain sexism"
Persona	Giving a persona to the LLM	"You are an expert in understanding slight linguistic nuances..."
Step-by-step reasoning	Asking the LLM for a step-by-step reasoning when describing or explaining something	"Define sexism step by step"
Task definition	Naming the specific task in which the definition will be employed	"I want to use a LLM to detect sexism..."
Content Generation	Asking the LLM to generate examples of specific form of sexism: subtle, edge, hostile vs. benevolent, ...	"... giving examples of everyday (e.g. workplace) as well as online sexism."
Reasoning	Forcing the LLM into a dialectic (or socratic) reasoning with a back and forth of multiple prompts	"What is a woman? ... What is an adult female human being? ... You define women as biologically female human adults. What is a biological female?"
Testing: side tasks	Asking the LLM to define other (similar) construct and tell the difference, or to classify comment and rewrite it in a non-sexist way	"Define sexism and misogyny. What are the differences?", "... how should the sentence be re-written to be non-sexist"
Enhancing	Asking the LLM to rewrite the definition to enhance quality and clarity	"Edit your definition for better flow and clarity"

Table 2: Qualitative analysis: a taxonomy of expert strategies to co-create a definition of sexism (Part III)

Performance of definition types on "sexist" class There is a bigger difference in GPT4o's performance when looking at the performance of the sexist vs non-sexist class. Co-created definitions have a small gain in performance over the LLM-definitions on the sexist class: LLM-generated definitions on average perform at $F1$ = .738, co-created definitions with a mean of $F1$ .740, and expert-written with a mean of $F1$ = .501).

4.5 Difference over experts

The difference between different experts is small: the expert with the lowest mean performing definitions is Expert 1 ($F1$ = .671, SD = .184), and the Expert with the highest mean over definitions is expert 9 ($F1$ = .702, SD = .166).

The upper half of Figure 3 shows that some experts are able to define more effective definitions of sexism than others, while others succeed better in the co-creation. Expert 5 shows comparatively higher scores for the co-creation with the LLM. This is notable because on average the LLM-generated definitions perform better.

The difference between experts becomes more pronounced when the results are disaggregated by dataset. The lower half of Figure 3 shows that CallMeSexist dataset performs > .80 in $F1$ for all experts, while the RedditGuest dataset shows great fluctuation depending on which expert has provided the definitions, with Expert 4 managing results around $F1$ = .60 while Expert 11 remains at $F1$ = .50.

Connection between the parts of the study This paper consists of four components: a survey, two interactive experiments, and zero-shot classifica-

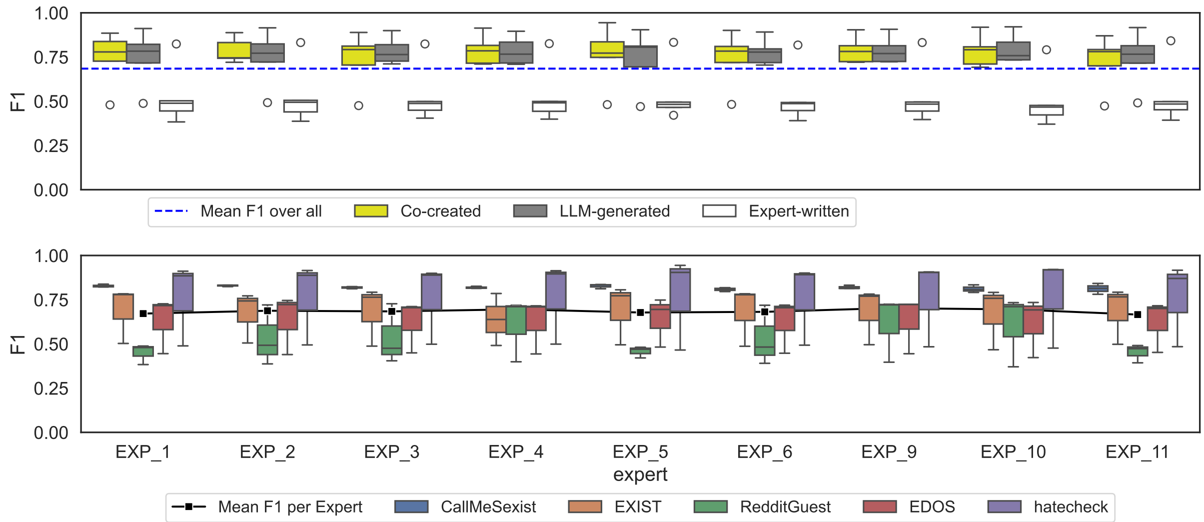

Figure 3: $F1$ (macro) performance of `GPT4o` per participant over the three definitions (upper plot) and over the five datasets (bottom row).

tion. When connecting results from all parts in the pipeline, we obtain additional insights. For instance, Expert 6 - less experienced with LLMs - was sometimes successful at co-creating definitions that perform better than the LLM-written definitions. In contrast, the co-created definition of experts confident in the use of LLMs (such as Expert 4) perform worse than the LLM-generated ones. This difference does not seem to be influenced by the difference in length of the definitions, or even to definition similarity to dataset definitions.

5 Conclusion

Experts have nuanced knowledge of complex constructs in their domain of expertise. Our aim was to connect this knowledge to zero-shot construct detection with LLMs, and test whether model and expert can collaborate for *hybrid intelligence*: model and expert complementing one-another. We addressed two high-level questions:
Q1 How do experts interact with instruction-tuned LLMs when assessing them for further use in research? **Q2** What do we learn from the expert / LLM interactions, and can we use these insights for improving computational detection?

Answering these questions led us to develop a four-part pipeline that connects LLM-expert interactions to computational detection of sexism. In **Part I**, nine sexism experts first answered questions on their use of LLMs and expertise in sexism research. **Part II** consisted of an interaction experiment, where experts assess and evaluate the knowledge of an LLM (`GPT3.5`) about their domain of expertise. These interactions also allowed us to observe experts' interactions with LLMs. **Part III** consisted of a second expert-LLM interaction, where each of the nine experts was tasked to create three definitions of sexism: an expert-written, an LLM-generated, and a co-created definition. We then evaluated zero-shot classification in **Part IV** by prompting the LLM `GPT4o` with each of these 27 created definitions on 2,500 texts from five sexism benchmark dataset. We release the interaction framework as well as the anonymous LLM-expert interactions and the definitions for future research.

Answering **Q1**, Part II and Part III found that sexism experts use different strategies for evaluating LLMs on their domain of expertise: content generation, asking questions, and labelling examples. Most experts were moderately satisfied with the LLM's knowledge of sexism.

On **Q2**, our modeling experiments in Part IV showed that LLM-written definitions help performance on benchmarks more than co-created definitions - which counters the hypothesis that co-creation is a fruitful manner to add expert knowledge into construct definitions for zero-shot classification. However, some experts do obtain higher zero-shot performance with co-created definitions, and confidence in LLM usage does not necessarily relate to more effective definitions: experts with low self-confidence in LLM expertise were often able to co-create more effective definitions than more LLM-savvy colleagues.

Acknowledgements

MR's contributions were funded by a GESIS Junior Visiting Researcher Grant for a fully funded research visit to GESIS in Cologne, Germany from March to April 2024. MR was until 4 December 2024 also funded by the Netherlands Organization for Scientific Research (NWO) through the *Rethinking News Algorithms* project, funded via the Open Competition Digitalization Humanities & Social Science grant (406.D1.19.073). We thank GESIS for covering the costs of proprietary models as well as payments to the participants.

We thank our colleagues for their comments and feedback on earlier versions of this paper, especially prof. dr. Suzan Verberne and prof. dr. Antske Fokkens. The feedback provided by this paper's ARR reviewers was also excellent and enhanced this paper. All remaining errors, flaws, or unclarities are ours.

Limitations

We identified at least three aspects that limit the generalizability of the findings in our study.

Scope We only test our framework on one social construct, sexism. The results - both on the interactions and on the classifications - may therefore not generalize to other complex constructs. Additionally, we tested our framework with only one LLM (`GPT4o`), and a limited set of nine experts. Future work may want to increase the scope of this work by adding more experts, constructs, or LLMs. We did complete additional zero-shot classification experiments with different temperature settings of `GPT4o`, as well as a different LLM (`LlaMa-3.1-70B-Instruct`), which showed lower performance and less variance over differences in prompt. These analyses can be found in the appendix.

Representativeness of Participants We use a limited sample of experts from WEIRD (Henrich et al., 2010) contexts: Western, Educated, Industrialized, Rich and Democratic. This is also relevant in a researcher context (e.g. researchers from the Global South), and limits generalizability to other research contexts.

Language Additionally, our datasets are only in the English language: both the benchmark datasets and the experiments were conducted in English, which also does not allow experts with different contexts to be included in these results.

Ethical Considerations

Harmful Content Sexism is a great harm to society and the world. We explicitly condemn sexism, and additionally want to warn any person using these models or datasets for potentially harmful utterances in them.

Experimental Safety of Participants Our participants were advised that they were free to leave the conversation at any time they felt the need. A large harm reduction factor was that our participants are experts already working on sexism, and therefore not as unprepared as crowd workers or other non-specialists to see sexist or hateful content. We caution researchers wanting to use a similar framework when participants are non-specialists, i.e. crowdworkers.

Writing Assistance Overleaf's integrated language model Writefull was used to assist in polishing and clarifying the language in this paper. This assistance, in accordance with the ACL Ethics Policy and Responsible Research Checklist, was solely used for improving the language in the paper rather than for producing new content or new ideas. All final writing is ultimately done by the authors, who are responsible for it.

Proprietary models Models of the GPT family are closed: developers have not released all information relating to the development and workings of these models. The openness of LLMs is not a binary variable: as Liesenfeld et al. (2023) indicates, there are several dimensions that make a model open, from shared code and training data, licensing, and whether payment is required for access. Despite this closedness, we still chose to use models of the GPT family because GPT models are especially popular with social science researchers researching constructs such as sexism. Additionally, the large-scale nature of our experiments and the lack of GPU access available to us made API requests more feasible for our experiments.

However, we are aware that open models are better for the scientific community due to their more reproducible nature and also the lack of payment required. We realize that our choice for proprietary models leads us to contribute to additional attention to these models, which is not ideal.

References

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *arXiv preprint arXiv:2406.11980*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149.

John Brewer. 2000. *Ethnography*. McGraw-Hill Education (UK).

Kathy Charmaz. 2015. Grounded theory. *Qualitative psychology: A practical guide to research methods*, 3:53–84.

Gabriele de Seta, Matti Pohjonen, and Aleksi Knuutila. 2024. Synthetic ethnography: Field devices for the qualitative study of generative models. *Big Data & Society*, 11(4):20539517241303126.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.

Alain Désilets, Louise Brunette, Christiane Melançon, and Geneviève Patenaude. 2008. Reliable innovation: A tecchie's travels in the land of translators. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Government and Commercial Uses of MT*, pages 339–345.

Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H. Andrew Schwartz. 2024. SOCIALITE-LLAMA: An instruction-tuned model for social scientific tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–468, St. Julian's, Malta. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bahareh Fatemi, Fazle Rabbi, and Andreas L. Opdahl. 2023. Evaluating the effectiveness of gpt large language model for news classification in the iptc news ontology. *IEEE Access*, 11:145386–145394.

Nathan Gagné and Léon Franzen. 2023. How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience. *Swiss Psychology Open: the official journal of the Swiss Psychological Society*, 3(1).

Mirta Galesic. 2006. Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of official statistics*, 22(2):313.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82.

Monique Hennink and Bonnie N Kaiser. 2022. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science & medicine*, 292:114523.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

John Hughes, Val King, Tom Rodden, and Hans Andersen. 1994. Moving out from the control room: ethnography in system design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, page 429–439, New York, NY, USA. Association for Computing Machinery.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2025. DefVerify: Do hate speech models reflect their dataset's definition? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4341–4358, Abu Dhabi, UAE. Association for Computational Linguistics.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. *arXiv preprint arXiv:2303.04222*.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. *Findings of the Association for Computational Linguistics: NAACL 2025*.

Kalle Kusk and Claus Bossen. 2022. Working with wolt: An ethnographic study of lenient algorithmic management on a food delivery platform. *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP).

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–6.

Leping Liu. 2023. Analyzing the Text Contents Produced by ChatGPT: Prompts, Feature-Components in Responses, and a Predictive Model. *Journal of Educational Technology Development and Exchange*, 16(1):49–70.

Riia Luhtanen and Jennifer Crocker. 1992. A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and social psychology bulletin*, 18(3):302–318.

Dhavalkumar Patel, Prem Timsina, Ganesh Raut, Robert Freeman, Matthew Levin, Girish Nadkarni, Benjamin S Glicksberg, and Eyal Klang. 2024. Exploring temperature effects on large language models across various clinical tasks. *medRxiv*, pages 2024–07.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions Matter: Guiding GPT for Multi-label Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021. No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 45–55, Online. Association for Computational Linguistics.

Francisco Rodriguez-Sanchez, Jorge Carrillo de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0).

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:573–584.

Anna Schjøtt Hansen and Jannie Møller Hartley. 2023. Designing what's news: An ethnography of a personalization algorithm and the data-driven (re) assembling of the news. *Digital Journalism*, 11(6):924–942.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People Make Better Edits: Measuring the Efficacy of LLM-Generated Counterfactually Augmented Data for Harmful Language Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Carlo Simon, Stefan Haag, and Lara Zakfeld. 2023. Experiments on gpt-3 assisted process model development. *ECMS*, 2023:37th.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Maartje Ter Hoeve, Julia Kiseleva, and Maarten Rijke. 2022. What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 46–75, Seattle, United States. Association for Computational Linguistics.

Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Ding Wang, Santosh D Kale, and Jacki O'Neill. 2020. Please call the specialism: Using wechat to support patient care in china. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Maximilian Weber and Merle Reichardt. 2023. Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models. ArXiv:2401.00284 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland. Association for Computational Linguistics.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Lining Zhang, Mengchen Wang, Liben Chen, and Wenxin Zhang. 2022. Probing gpt-3's linguistic knowledge on semantic tasks. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 297–304.

Appendix

A Precautions for a Responsible Participant Study

We completed an ethics form from the Social Science department of the Vrije Universiteit Amsterdam. This form is designed to decide whether a study requires additional ethical review. We received confirmation that indicated our methods were not harmful to participants, and could proceed without additional ethics review if taking into account responsible study design in the form of informed consent, participant payment, and warnings on the potential of harm.

A factor that mitigated the harm to our participants was that they were sexism research experts, who were duly informed about the purposes of this study. Participants were not directly exposed to hateful language and were informed that they could quit the experiment at any time. They were offered a payment for their time, at the minimum wage level in Germany.

A.1 Informed Consent Statement

Explainer This task is part of a scientific study about the use of AI in social science. The purpose of the task is to obtain data on how people - specifically social science experts - interact with so-called Large Language Models (LLMs). The task Experts are briefly asked to give information on their experience with LLMs. Then, experts are asked to interact with the model to determine whether this model is sufficient to detect the construct "sexism" in texts. Experts are then also asked to generate definitions with the model, which will be later used to improve LLM detection of sexism.

Compensation We estimate the survey takes 30 minutes. Our compensation for your time is 6 euros. This payment can be received by providing your email address after completing the survey. Your details will not be used for any other purpose, and not stored. This study is conducted by researchers connected to GESIS - Leibniz Institute for the Social Sciences in Cologne, the Vrije Universiteit Amsterdam, and the University of Konstanz. Researchers are reachable through email: myrthe.reuver@vu.nl. Save this email address for any questions or concerns about this study.

Outcome The generated texts and data gathered in the study may be published, or made available to other researchers. Your name or identifying information will not be stored, and will not be shared with anyone.

Note: Despite practices on model safety that went into model development, the LLM can potentially generate harmful output that may be upsetting or offensive. Such output is not condoned by the researchers or their institutes. You can stop interacting with the model at any time, for any reason.

Consent statement

I understand the purpose of this study, and participate in this task out of my own free choice.

I understand I can withdraw at any time, for any reason, without any consequences of any kind.

I understand that in any report on the results of this research my identity will remain anonymous.

I understand that I am free to contact the researchers through myrthe.reuver@vu.nl to seek further clarification and information.

B Survey and Experiment

B.1 Survey and Experimental Design

We use Qualtrics as a survey platform, and connect calls to the OpenAI API through a Web Service component in the workflow. This Web Service had as input a text field presented to the participant, and collected the output that the API provided in an object that was presented in a text presentation question that was presented to the participant. Our Qualtrics templates, both in pdf and in Qualtrics format, are released through our GitHub repository.

The 10-turn conversational loop was achieved by a chain of if-clauses in the survey flow: if a participant chose to end the conversation, the participant was re-directed outside of the Web Service. If the participant indicated she was not yet done with the experimental interaction, she was instead re-referred to another Web Service field, but one that received as input the entire previous conversation. This made the conversation continuous despite requiring multiple API calls due to Qualtrics not allowing one call loop towards the same Web Service.

B.2 Additional Participant Info

Seven of our participants were PhD candidates or equivalent, one was a postdoc, and three participants assistant professor. When asked for their main research discipline, four experts answered with communication science, three experts answered with (computational) social science. The remaining participants responded with Human-Computer Interaction or computer science. This broad range of answers assures a wide range of experiences with LLMs for sexism research in our study.

B.3 Additional Expert Survey results

LLM Experience Nine out of the 11 social science experts had used LLMs at least once for research purposes. Of the ones who use LLMs, one expert indicates using LLMs daily, 5 indicate a weekly use, one indicates monthly usage while two experts say they use LLMs less than monthly. All experts who used LLM were familiar with ChatGPT, with Claude and Mistral being the second-most popular LLM (both N = 4).

All experts who had used LLMs before were familiar with interacting with an LLM through a chat interface, and eight experts were familiar with using LLMs through API or code. Fine-tuning was used before by seven experts. A minority of experts (2 experts) indicated experience with designing LLM agents and prompt chains.

Experts indicated from a list which tasks they most used LLMs for: data annotation (N = 5), followed by assistance with text writing (N = 4), followed by summarization (N = 2) and information seeking (N = 2). One experts mentioned help with data visualization. Interestingly, a connection could be seen with background expertise: experts with a background in communication science all mentioned text writing, while experts mentioning computational social science all mentioned data annotation.

In an open question about LLM use in research, expert 3 indicated feeling uncomfortable with over-use of LLMs for data annotation and expert 11 indicated to find anthropomorphization of LLMS to be a potential danger for social scientists using LLMs. Expert 6 indicated they did not use LLMs herself, but relied on co-authors to use LLMs in their shared projects. Expert 9 shared that they not only use LLMs in sexism research, but researches the use of LLMs in social science.

B.4 Heatmaps of Questionnaire Likert Scales Results

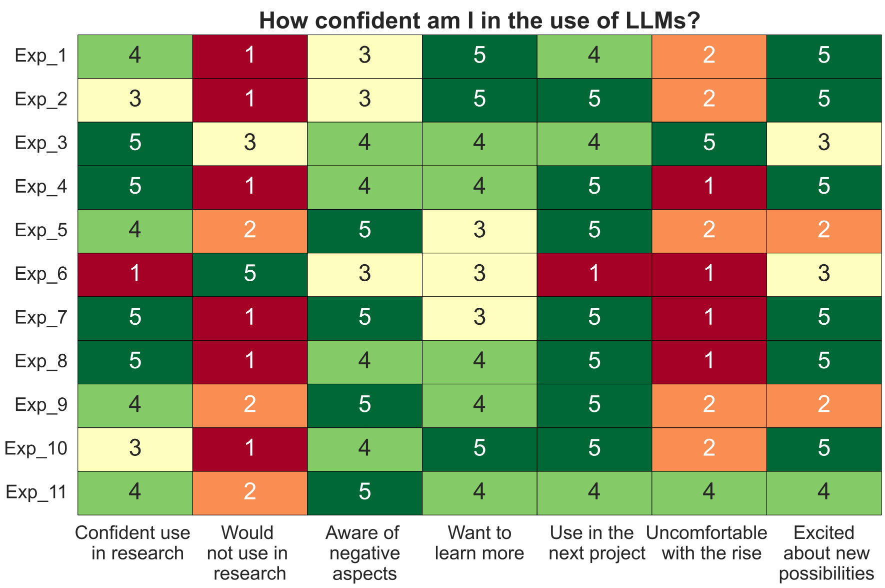

Figure 4: Heatmap of Likert scale on participants experience on LLMs.

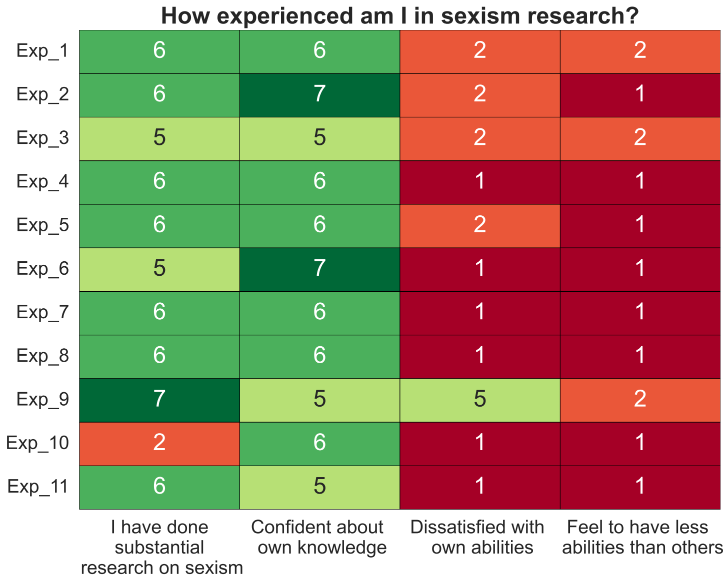

Figure 5: Heatmap of Likert scale auto-reported experience on sexism research.

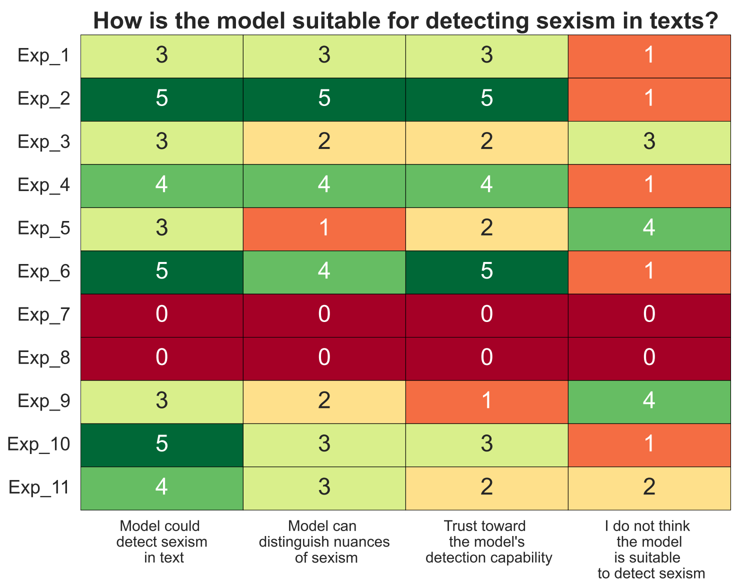

Figure 6: Heatmap of Likert scale on suitability of the model in detecting sexism.

C Modelling Specifications

Our experimental code is written in Python 3.9.5 and it uses the openai package version 1.51.2 for calling to the OpenAI API.

C.1 Model Hyperparameters and Compute Infrastructure

The generation parameters are set to: temperature = 0, max_tokens=200, top_p=1, frequency_penalty=0, presence_penalty=0. We then repeated this with another run with a set temperature of 0.7, and max_tokens to 20, to analyze robustness of this approach in different temperatures. Results of these experiments and an analysis of them are presented in Appendix G.

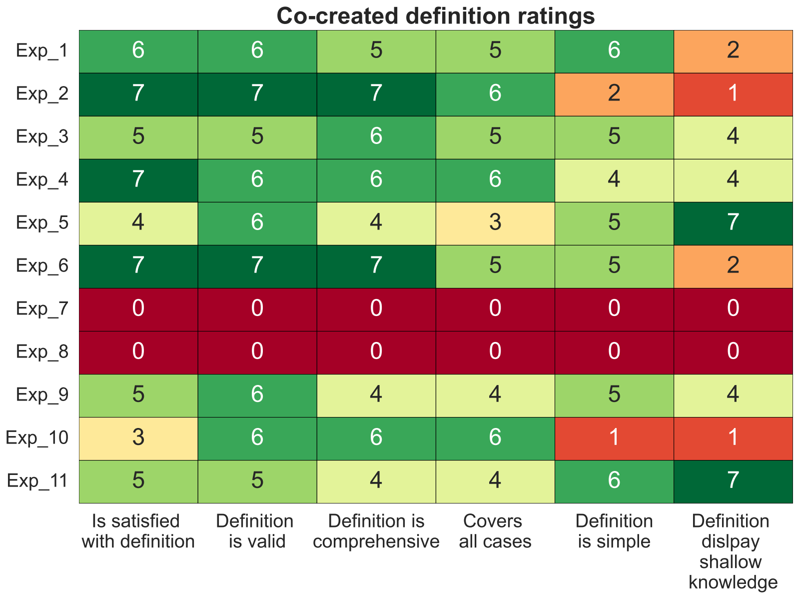

Figure 7: Heatmap of Likert scale on the suitability fo the co-created definition.

C.2 Costs for proprietary models

Our costs for using proprietary models were paid by GESIS.

Costs related to interaction experiments Participant interactions came at maximum 20 calls per participant, for 9 participants, at approximately 0.00317 dollar cents per call. This meant that the total costs for the experiments were around $0.50 cents.

Costs related to classification experiments For the `GPT4o` classification experiments, one total run equals 9 participants times 3 different definitions, which is 27 definition prompts. These prompts were used on 2,500 texts, which resulted in 67,500 prompts sent to `GPT4o` through the OpenAI API. These 67,500 classification decisions cost approximately 0.00317 dollars per call, or $213 in total for the entire dataset. We did these experiments twice to check for robustness, leading to a total cost of $426.

C.3 Prompt for classification

"Your task is to annotate if a piece of text is sexist or not. Later, you get a detailed definition of sexism.

Definition of sexism: **definition**

With this definition in mind and a piece of text, please reply with either of the three options and nothing else: 1) sexist, 2) non-sexist, 3) can't say. Please use the last option as little as possible and if you do, provide a one sentence explanation. Do not give an explanation if you pick 'sexist' or 'non-sexist'.

We are ready to begin now. Is this sentence sexist based on the past definition' : **text** - 'Please reply with either of the three options and nothing else: 1) sexist, 2) non-sexist, 3) can't say."

D Evaluation datasets: class balance

Relative	CMSB	EDOS	REDDIT	EXIST	HateCheck
Sexist	0.448	0.232	0.13	0.476	0.734
Non sexist	0.552	0.768	0.87	0.524	0.266

Table 3: Original class distribution in the five datasets - relative amount of positive and negative class

In raw counts	CMSB	EDOS	REDDIT	EXIST	HateCheck
Sexist	534	4854	699	1636	373
Non sexist	690	15146	5856	1800	136

Table 4: Original class distribution in the five datasets - raw counts of positive and negative class

Relative	CMSB	EDOS	REDDIT	EXIST	HateCheck
Sexist	0.436	0.243	0.107	0.476	0.739
Non sexist	0.564	0.757	0.893	0.524	0.267

Table 5: Class distribution in the dataset subsamples we used for evaluation - relative amount of positive and negative class

In raw counts	CMSB	EDOS	REDDIT	EXIST	HateCheck
Sexist	224	116	65	238	367
Non sexist	276	384	435	262	133

Table 6: Class distribution in the dataset subsamples we used for evaluation - raw counts of positive and negative class

E Definitions Analysis

Table 7 lists the definitions collected from our experts in the second part of our survey. For each expert, we report the definition provided in the beginning of the experiment (*Expert-written*), the definition that was chosen as the best provided by the model (*LLM-generated*), as well as the co-created definition that was submitted by the expert as a response in the end of the survey (*Co-created*). We also display the dataset-specific definitions employed for the annotation in the five datasets we conduct our modeling experiments on. Taken together, the definitions listed in Table 7 constitute the *prompt repository* we evaluate and release in this paper.

Source	Type	Definitions
Expert 1	Expert-written	Sexism is discrimination based on gender, including stereotypes and prejudice. It can take many forms, from offensive language to sexual violence.
	Co-created	Sexism refers to prejudice, discrimination, or stereotyping based on a person's sex or gender, typically with a bias against one gender over the other. It can manifest in various forms, including unequal treatment, unfair expectations, and harmful attitudes or beliefs about individuals based on their sex. Examples of everyday sexism in the workplace could include paying women less than men for the same job or assuming certain roles are only suitable for one gender. Online sexism could involve making derogatory comments based on gender or perpetuating harmful stereotypes in digital spaces.
	LLM-generated	Sexism refers to prejudice, discrimination, or stereotyping based on a person's sex or gender, typically with a bias against one gender over the other. It can manifest in various forms, including unequal treatment, unfair expectations, and harmful attitudes or beliefs about individuals based on their sex. Examples of everyday sexism in the workplace could include paying women less than men for the same job or assuming certain roles are only suitable for one gender. Online sexism could involve making derogatory comments based on gender or perpetuating harmful stereotypes in digital spaces. If you have any more questions or need further clarification, feel free to ask!

Source	Type	Definitions
Expert 2	Expert-written	I would define sexism as a way to imply stereotypical or unrespectful opinion on gender.
	Co-created	Sexism is a form of prejudice, stereotyping, or discrimination based on gender, often targeting women. This bias reflects the belief in one gender's superiority, resulting in unfair treatment, negative stereotypes, and limited opportunities solely due to gender. For example, a common illustration of sexism is the gender pay gap, where women are paid less than men for the same work. Sexism can manifest in different forms like unequal treatment, unfair stereotypes, or restrictions on opportunities, impacting both individuals and society. If you require more examples or details for each form of sexism, feel free to ask! Sexism's effects can be harmful, leading to inequalities, perpetuating stereotypes, and limiting opportunities for individuals. It primarily affects women but can also impact men in different contexts. Resolving sexism involves efforts at individual, societal, and institutional levels through education, promoting gender equality, implementing fair policies, and challenging discriminatory practices. If you need further clarification or details on this topic, please let me know!
	LLM-generated	Sexism is a complex issue that involves prejudice, stereotypes, or discrimination against individuals based on their gender. This kind of bias often targets women and embodies the belief that one gender is superior to another. It can take the form of unfair treatment, harmful stereotypes, or limitations on opportunities solely because of a person's gender. Sexism is not limited to individual interactions but can also be ingrained in societal structures, leading to significant adverse effects on both individuals and society.
Expert 3	Expert-written	Believing that one sex is superior to another. It includes different forms of harmful stereotyping including benevolent sexism, etc.
	Co-created	Sexism refers to discrimination, prejudice, or stereotyping based on a person's gender. It can manifest in various forms, such as treating individuals differently because of their gender, assuming certain characteristics or behaviors based on gender, or denying opportunities to someone because of their gender. Sexism can affect people of all genders, but it is often directed towards women and can have harmful effects on their lives and opportunities. There are different types of sexism, such as hostile sexism, benevolent sexism, and ambivalent sexism.
	LLM-generated	Sexism refers to discrimination, prejudice, or stereotyping based on a person's gender. It can manifest in various forms, such as treating individuals differently because of their gender, assuming certain characteristics or behaviors based on gender, or denying opportunities to someone because of their gender. Sexism can affect people of all genders, but it is often directed towards women and can have harmful effects on their lives and opportunities. There are different types of sexism, such as hostile sexism, benevolent sexism, and ambivalent sexism.
Expert 4	Expert-written	A prescriptive set of behaviors or qualities, that women (and men) are supposed to exhibit to conform to traditional gender roles.

Source	Type	Definitions
	Co-created	It’s fascinating to see your expertise in discerning subtle linguistic cues to identify sexism in texts. The definition of sexism centers on discrimination, prejudice, stereotyping, and biased behaviors toward individuals based on gender, leading to unequal treatment and marginalization. Evaluating texts for sexism entails examining portrayals of gender superiority or inferiority, reinforcement of stereotypes, and gender-based limitations, which involves analyzing the text’s structure, tone, and phrasing to ascertain if it perpetuates gender inequality.
	LLM-generated	Sexism is a form of discrimination, prejudice, or stereotyping based on a person’s gender. It involves attitudes, beliefs, and behaviors that differentiate, marginalize, or treat individuals differently based on their gender. Sexism can manifest in various ways, such as through language, actions, or systemic biases that perpetuate inequality between genders. When analyzing a text for sexism, it’s important to assess if it portrays one gender as superior or inferior to the other, reinforces harmful stereotypes, or limits an individual’s opportunities or worth based on their gender. This thorough evaluation considers the structure, tone, and phrasing of the text to determine if it perpetuates gender inequality. If you need further assistance or examples, feel free to ask.
Expert 5	Expert-written	Abusive language targeted to an individual or group because of their sex and/or gender.
	Co-created	Sexism specifically refers to discrimination, prejudice, or stereotyping directed towards individuals or groups based on their gender, where gender is the underlying cause of the abusive behavior. Mentioning the gender of the target alone may not be sufficient for a sentence to be considered sexist. Your task is to help identify and address harmful and unfair attitudes and behaviors towards people because of their gender, while also recognizing that counter speech to challenge these harmful attitudes is permissible.
	LLM-generated	Certainly! Here’s an updated prompt that includes the clarification about allowing counter speech: "Given a piece of text, please analyze and identify any instances of sexism present. Sexism specifically refers to discrimination, prejudice, or stereotyping directed towards individuals or groups based on their gender, where gender is the underlying cause of the abusive behavior. Mentioning the gender of the target alone may not be sufficient for a sentence to be considered sexist. Your task is to help identify and address harmful and unfair attitudes and behaviors towards people because of their gender, while also recognizing that counter speech to challenge these harmful attitudes is permissible."
Expert 6	Expert-written	Statements or behaviours that reinforce the idea that men are given more power and status than women, that men and women are (or should be) distinguished on the basis of their social roles and characteristics; and relationships between men and women are defined by sexual reproduction and intimacy. Sexism can take the form of hostile and benevolent sexism.

Source	Type	Definitions
	Co-created	Hostile sexism refers to overtly negative and hostile attitudes and behaviors towards women based on the belief that women are inferior to men. This can involve open discrimination, derogatory remarks, and devaluation of women. Benevolent sexism, on the other hand, involves attitudes and behaviors that may appear positive or chivalrous on the surface, but actually serve to reinforce traditional gender roles and keep women in subordinate positions. These attitudes often arise from the societal norms and expectations that distinguish men and women based on their perceived social roles and characteristics. Examples of benevolent sexism include patronizing behavior, protective gestures, or idealizing women as pure and delicate. To detect sexism in text, look for language or scenarios that suggest unequal treatment, devaluation, or reinforcement of traditional gender roles. Pay attention to how women are portrayed, whether they are being reduced to stereotypes, denied agency, or discriminated against. Both hostile and benevolent sexism can manifest in various ways in written content, so it's important to analyze the overall context and underlying messages being conveyed.
	LLM-generated	I'm glad you found the previous information helpful. I understand you would like a new definition that incorporates the idea that men and women are distinguished based on their social roles and characteristics. Here's an updated definition for you: Hostile sexism refers to overtly negative and hostile attitudes and behaviors towards women based on the belief that women are inferior to men. This can involve open discrimination, derogatory remarks, and devaluation of women. Benevolent sexism, on the other hand, involves attitudes and behaviors that may appear positive or chivalrous on the surface, but actually serve to reinforce traditional gender roles and keep women in subordinate positions. These attitudes often arise from the societal norms and expectations that distinguish men and women based on their perceived social roles and characteristics. Examples of benevolent sexism include patronizing behavior, protective gestures, or idealizing women as pure and delicate. To detect sexism in text, look for language or scenarios that suggest unequal treatment, devaluation, or reinforcement of traditional gender roles. Pay attention to how women are portrayed, whether they are being reduced to stereotypes, denied agency, or discriminated against. Both hostile and benevolent sexism can manifest in various ways in written content, so it's important to analyze the overall context and underlying messages being conveyed. If you have any more questions or need further clarification, feel free to ask!
Expert 7	Expert-written	No Answer
	Co-created	No Answer
	LLM-generated	No Answer
Expert 8	Expert-written	No Answer
	Co-created	No Answer
	LLM-generated	No Answer

Source	Type	Definitions
Expert 9	Expert-written	Sexism is linked to beliefs around the fundamental nature of women and men and the roles they should play in society. Sexist assumptions about women and men, which manifest themselves as gender stereotypes, can rank one gender as superior to another.
	Co-created	Sexism is a form of discrimination, prejudice, or stereotyping based on a person's sex or gender, typically towards women or girls. It involves treating people differently or unfairly because of their gender, and can be expressed in various forms, such as unequal treatment, stereotypes, or systemic sexism within societal structures
	LLM-generated	Sexism is a form of discrimination, prejudice, or stereotyping based on a person's sex or gender, typically towards women or girls. It involves treating individuals differently or unfairly because of their gender, and can be expressed in various forms, such as unequal treatment, stereotypes, or systemic sexism within societal structures. Let me know if you need more information or have any other questions!
Expert 10	Expert-written	Broadly, sexism is prejudice or discrimination based on one's sex or gender - a form of oppression that results primarily in disadvantages for women and queer people. A more technical definition of sexism is a system of norms, beliefs, and practices that normalise, perpetuate, and legitimise a binary, heteronormative, and/or deterministic framework of understanding sex and gender. Sexism can manifest in many different ways and it is deeply embedded in our institutions, practices, traditions, cultural norms and values. So detecting and understanding how sexism works necessarily involves studying how society is structured, organised, and governed.
	Co-created	Sexism is the manifestation of discriminatory, prejudiced, or stereotypical attitudes towards individuals based on their sex or gender. Understanding sexism includes recognizing and addressing the systemic, structural components deeply ingrained in societal norms, practices, and institutions that perpetuate unequal treatment and opportunities among genders. Understanding the power dynamics and societal frameworks that fuel gender inequality in diverse contexts is crucial for combating sexism effectively.
	LLM-generated	Certainly! Here's a refined version of the comprehensive definition of sexism that you requested: Sexism is the manifestation of discriminatory, prejudiced, or stereotypical attitudes towards individuals based on their sex or gender. This includes recognizing and addressing the systemic, structural components deeply ingrained in societal norms, practices, and institutions that perpetuate unequal treatment and opportunities among genders. Understanding the power dynamics and societal frameworks that fuel gender inequality in diverse contexts is crucial for combating sexism effectively.
Expert 11	Expert-written	A view of an identity that is defined by social beliefs but is not necessarily supported by a ground truth or doesn't reflect their actual abilities

Source	Type	Definitions
	Co-created	Sexism is a form of prejudice or discrimination based on a person's sex or gender, typically against women but also impacting men. Discrimination can manifest in various ways, including unequal treatment, stereotypes, or demeaning attitudes towards individuals based on their gender. Positive stereotypes, even if seemingly complimentary, can still contribute to sexism by reinforcing harmful beliefs about certain genders. An example of a positive stereotype that is harmful is the belief that women are naturally more nurturing and caring than men, which can limit opportunities for women by prioritizing caregiving roles over personal and professional aspirations. In response to your question about whether the statement that "women are stronger than men since they can sustain labor pains, whose intensity is known to be one of the worst kinds of pain known to humans" is sexist, the emphasis on labor pain as a defining strength can perpetuate stereotypes and reinforce traditional gender roles. It restricts the recognition of other diverse strengths individuals possess, which ultimately contributes to the reinforcement of gender norms and inequality. To make the sentence non-sexist, it could be rephrased to acknowledge various strengths and abilities without attributing them solely to gender, promoting inclusivity and respect for individual differences. If you would like more examples of cases that might not be overtly sexist but still have a sexist undertone, please let me know.
	LLM-generated	Sexism is a form of prejudice or discrimination based on a person's sex or gender, typically against women but also impacting men. Discrimination can manifest in various ways, including unequal treatment, stereotypes, or demeaning attitudes towards individuals based on their gender. Positive stereotypes, even if seemingly complimentary, can still contribute to sexism by reinforcing harmful beliefs about certain genders. An example of a positive stereotype that is harmful is the belief that women are naturally more nurturing and caring than men, which can limit opportunities for women by prioritizing caregiving roles over personal and professional aspirations. In response to your question about whether the statement that "women are stronger than men since they can sustain labor pains, whose intensity is known to be one of the worst kinds of pain known to humans" is sexist, the emphasis on labor pain as a defining strength can perpetuate stereotypes and reinforce traditional gender roles. It restricts the recognition of other diverse strengths individuals possess, which ultimately contributes to the reinforcement of gender norms and inequality. To make the sentence non-sexist, it could be rephrased to acknowledge various strengths and abilities without attributing them solely to gender, promoting inclusivity and respect for individual differences. If you would like more examples of cases that might not be overtly sexist but still have a sexist undertone, please let me know.

Source	Type	Definitions
Datasets	CMSB	Something can be sexist because of its content (what the speaker believes) or because of its phrasing (because of the speaker's choice of words). For example, the speaker may express sexist attitudes towards gender inequality, either endorsing it or disbelieving it's existence. The speaker may express stereotypes (how genders are traditionally seen and compared to each other) and behavioral expectations (how individuals of a gender should behave according to traditional views). On the other hand, a message may be sexist simply because of how the speaker phrases it–independently from what general beliefs or attitudes the speaker holds. A message is sexist, for example, when it contains attacks, foul language, or derogatory depictions directed towards individuals because of their gender.
	EDOS	We define sexist content as any abuse, implicit or explicit, that is directed towards women based on their gender, or on the combination of their gender with one or more other identity attributes (e.g. Black women or Muslim women).
	REDDIT	*For Misogynistic content, we defined four categories: (i) Misogynistic Pejoratives, (ii) descriptions of Misogynistic Treatment, (iii) acts of Misogynistic Derogation and (iv) Gendered Personal attacks against women.*
	EXIST	Sexism as "prejudice, stereotyping, or discrimination, typically against women, on the basis of sex."
	HateCheck	*Hate Speech as abuse that is targeted at a protected group or at its members for being a part of that group. We define protected groups based on age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation, which broadly reflects international legal consensus (particularly the UK's 2010 Equality Act, the US 1964 Civil Rights Act, and the EU's Charter of Fundamental Rights).*

Table 7: Definitions collected. Among the dataset definitions, those in *italic* refer to datasets having definitions of the related concepts of misogyny and hate speech instead of sexism.

Comparing LLM-generated and co-created definitions It is true for the large majority of the experts that the *co-created* definition is either identical to the *LLM-generated* one (experts 3 and 11) or just an edited, cleaned-up version of the *LLM-generated* one (experts 1, 5, 6, 9, 10). Such minimal edits involved the removal of the ChatGPT-specific jargon (i.e., "If you have any more questions or need further clarification, feel free to ask!") or minimal word changes.

From the perspective of the prompting experiments we conducted with these definitions, we can consider these cases as robustness tests: either the two definitions are identical, or their semantics is nearly identical. In the case of expert 2 and 4, the best *LLM-generated* definition and the *co-created* definition are very different (resulting in very different prompts for the model). In the case of Expert 2, the *co-created* definition is much longer and richer.

Comparing Expert-written and LLM-generated definitions Definitions provided by the Experts (M = 34.44 tokens, SD = 25.24) are generally quite short and thus, low in informativeness, especially when compared with the LLM-generated definitions (M = 119.89, SD = 58.29) and the co-created definitions (M = 110.55, SD = 56.44) which are generally much longer. Only Expert 10 provides a self-written definition that is slightly longer than the co-created one.

To unpack these differences, we employ SBERT (Reimers, 2019)(`all-mpnet-base-v2`) and TF-IDF (Sparck Jones, 1972) to encode the 27 definitions obtained in the previous steps as well as the five definitions from the sexism benchmarks. Once we obtained the embedding for each definition, we compute the cosine similarities by first normalizing the vectors and then computing their dot product.

Table 8 displays, per expert, the cosine similarities between expert-written and co-created, expert-written

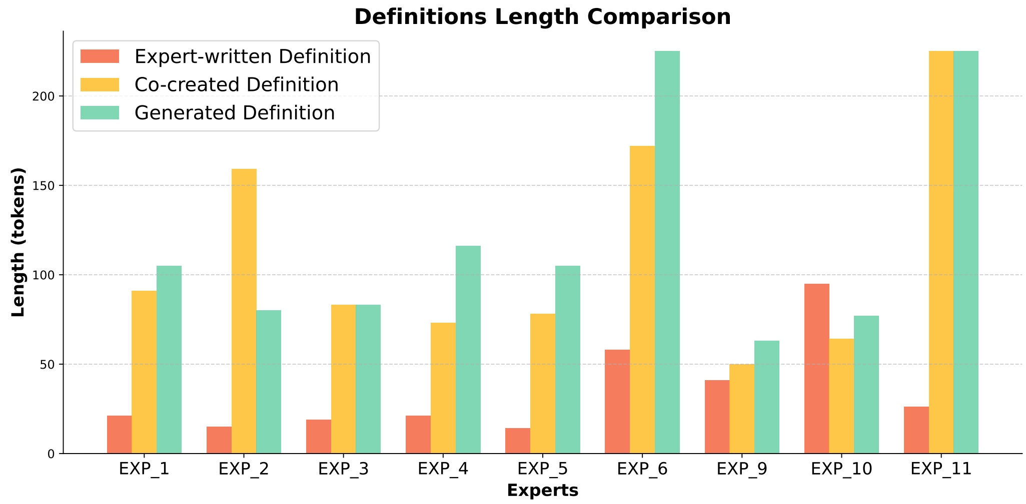

Figure 8: Length comparison, per expert: Expert(-written), (LLM-)generated, Co-created

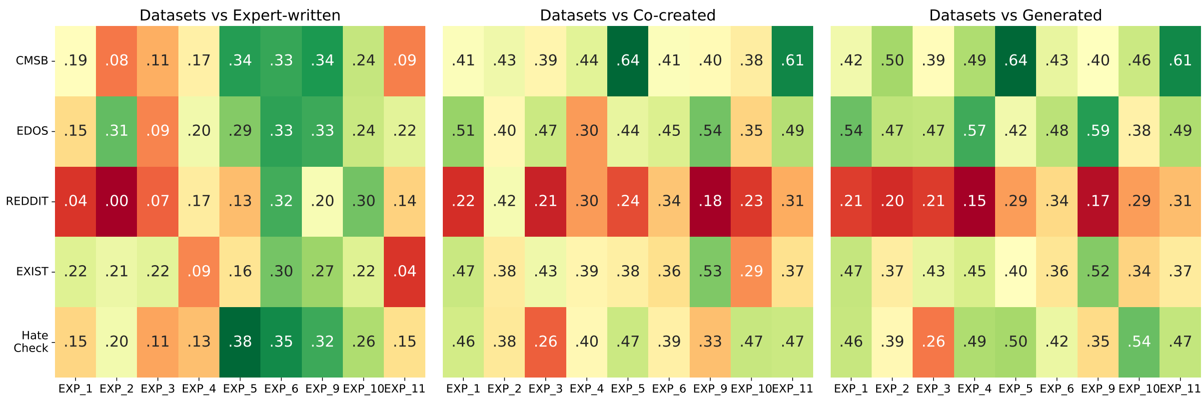

Figure 9: Cosine Similarity Heatmaps between Datasets definitions and Expert-written (left) Co-created (center) and Generated (right) definitions using TF-IDF.

and LLM-generated, co-created and LLM-generated.

Method	Definitions	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_9	Exp_10	Exp_11
TF-IDF	Expert-written vs Co-created	.37	.19	.31	.32	.36	.67	.26	.62	.24
	Expert-written vs LLM-generated	.36	.20	.31	.28	.36	.71	.24	.59	.24
	Co-created vs LLM-generated	.99	.72	1.00	.74	.96	.97	.96	.95	1.00
SBert	Expert-written vs Co-created	.94	.83	.71	.52	.65	.76	.91	.94	.27
	Expert-written vs LLM-generated	.95	.86	.71	.53	.47	.76	.90	.84	.27
	Co-created vs LLM-generated	.99	.96	1.00	.94	.62	.87	.99	.85	1.00

Table 8: Cosine similarity between Expert-written, Co-created and LLM-generated definitions.

Another interesting comparison is the one between the different types of definitions produced by our experts in the three scenarios and the definitions of datasets employed for our modeling experiments. The heatmaps in figures 9 and 10 display the cosine similarity between experts and dataset definitions.

E.1 Impact of definition variation and quality on performance

As a next step, we consider the relationship between the quantitative properties of the definitions discussed above and the performance of the corresponding models. In particular, in table 9 we report the correlation between model performance and:

- Quality of the co-created definition as rated by each expert in table 7 (*Quality*): does expert assessment

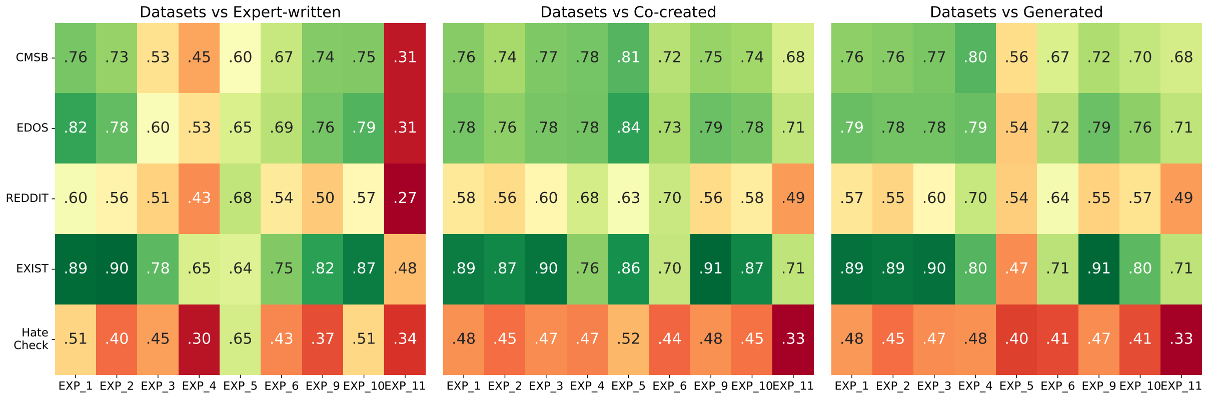

Figure 10: Cosine Similarity Heatmaps between Datasets definitions and Expert-written (left), Co-created (center) and Generated (right) definitions using an SBERT model (`all-mpnet-base-v2`).

reflect classification performance?

- Definition length (*Length*): does a richer definition lead to a better performance?
- Similarity between prompted definition and original definitions employed to in the annotation of the benchmark dataset (*Similarity*): does semantic overlap with the original definition guarantee better performance?

	CMSB	EDOS	REDDIT	EXIST	HateCheck
Quality	.28	.12	.20	-.35	-.33
Length - LLM-generated	-.67	-.38	-.15	.18	-.25
Length - Co-created	-.48	-.05	-.38	-.14	-.61
Length - Expert-written	-.85	-.62	-.65	-.85	-.24
Similarity (TF-IDF) - LLM-generated	-.09	-.03	-.17	-.62	33
Similarity (TF-IDF) - Co-created	-.14	.13	.12	-.05	.23
Similarity (TF-IDF) - Expert-written	-.24	.08	-.25	-.20	-.65
Similarity (SBert) - LLM-generated	.31	.66	.60	-.19	-.33
Similarity (SBert) - Co-created	.73	.46	-.03	-.27	.68
Similarity (SBert) - Expert-written	-.49	-.32	.10	-.09	-.71

Table 9: Correlations between experts' rating, length and cosine similarity with performance ($F1$) on each dataset

F Prompts beating the majority class

Since each of our five benchmarks have either balanced or imbalanced distributions, and these data sets sometimes have a relatively large majority class, classifying all data as the majority class could lead to an accuracy of 75% without any construct understanding. To analyze whether our zero-shot experiments actually improve over majority class, we do a brief analysis of this per prompt and participant. In the plot in the main paper, these analyses are done with majority $F1$.

Here, we also explore majority class accuracy as a baseline per dataset. For CallMeSexist all prompts are above majority class in accuracy, for the EXIST data the majority of prompts are, for the hatecheck dataset this is 75% of all prompts above the majority class, and for the EDOS dataset there are only a few prompts from a few experts that are accurate above majority class. For the RedditGuest this is basically none of the prompts.

There is a clear effect with expert, and especially definition type: on the EDOS dataset, a co-created definition can bring performance above majority class, as does it for some experts with Hatecheck. However, on the RedditGuest dataset, the GPT definition is more often successful.

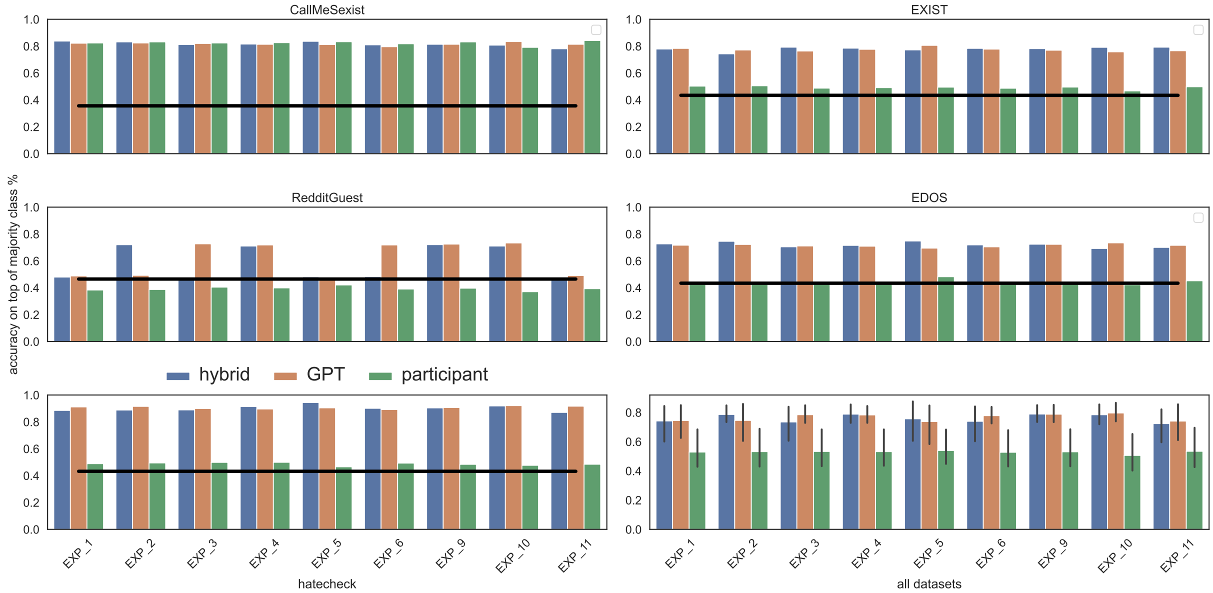

Figure 11: Showing per-definition macro $F1$ for each dataset, with a plotted line for the macro-$F1$ of the majority class in each dataset. For most, this is non-sexist, though for some it is sexist.

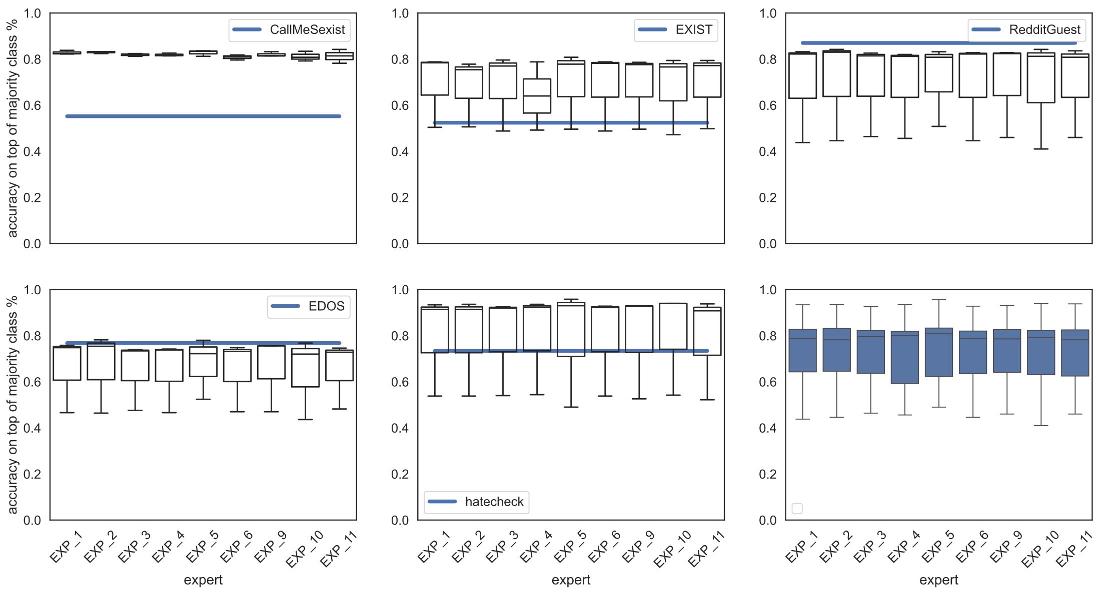

Figure 12: Showing per dataset **accuracy** with a plotted line for the majority class in each dataset. For most datasets, the majority class is non-sexist, though for some it is sexist.

G Effects of Temperature

The temperature hyperparameter in generative Large Language Models affects the softmax function and is related to randomness, or the probability of generating tokens (Renze, 2024). A lower temperature leads to more deterministic responses, always generating the most probable response, and therefore less diverse tokens in output. A lower setting is often chosen in high-stakes domains such as the medical domain (Patel et al., 2024), or settings in which being factually correct is important, while higher temperature settings are have been linked to creativity and unexpectedness (Peeperkorn et al., 2024). It is not fully clear how properietary models by OpenAI have implemented their temperature hyperparameter.

Several works identify no difference in LLM classification performance with different temperature settings (Yang et al., 2022; Zhang et al., 2022; Patel et al., 2024), while others optimize this hyperparameter and find different results with higher temperatures (Simon et al., 2023). Other work considers a temperature setting of 0 or close to it best for classification due to its more deterministic nature (Fatemi et al., 2023). Recent work analyzing the effects of different definitions in prompts also opts for a temperature of 0 (Korre et al., 2025). We want to determine whether temperature has an effect on our results.

G.1 Method

We ran GPT4o zero-shot classification pipeline as described in Section 3.4, with the only difference being a temperature set to 0.7. While an average higher performance and robustness to prompt versions are often desirable in models and model results, our intent was also to look at how individual definitions can affect performance and naturalistically replicate social scientists' prompting experiences. This makes a model showing less sensitivity to prompt variance less desirable for our study.

For zero-shot prompting, we had the exact same specifications as outlined in Section**??** of the main paper. We also used the exact same datasets as specified in the main paper, with five sexism datasets and three different definitions per participant.

G.2 Results

With the higher temperature, the results fluctuate more, with larger differences between different datasets. We do also see that the expert-written definition performs marginally worse (M $F1$ = .748) than the GPT definitions (M $F1$ = .760), and in fact the co-created definition shows on average the highest perfomance (M $F1$ = .762). See Figure13 for a figure comparable to Figure3

Overall, we see much smaller differences and much more randomness seemingly less related to different prompt versions. Average performance is overall a bit higher than with a lower temperature: M $F1$ = .760 vs M $F1$ = .68 for the temperature=0 run. This is mostly due to the expert-written definition performing much more comparable to the other two definitions than in the temperature=0 results.

In general, the results with a higher temperature seem to show more difference and sensitivity to datasets than to different prompt types.

The RedditGuest dataset shows the highest variance in results, meaning results within this dataset vary a lot with different participants and version definitions.

The higher temperature in general shows more variance *across* different datasets than *within* different datasets, e.g. less difference with different prompts or experts. The opposite is the case for the zero temperature setting, which shows more variation for different prompts and experts, and less for results.

G.3 Discussion

Unlike some previous work, we found that a higher temperature leads to a higher average performance in $F1$. However, a higher temperature also seems to show less variance to different prompt versions (Likely due to generating more randomness and less probable tokens in each response).

While an average higher performance and robustness to prompt versions are often desirable in models, our intent was also to look at how individual definitions can affect performance and naturalistically replicate social scientists' prompting experiences. This makes a model showing less sensitivity to prompt variance less desirable, though users requiring a higher overall performance and less prompt sensitivity should probably increase their temperature setting.

H LLaMa prompting

As a robustness check, we also prompted LLaMa3 70B (Dubey et al., 2024) with the same 27 definitions as GPT4o in the main paper.

H.1 Methods

We ran LLaMa on its default temperature setting of 0.5, and were run on 2 A100 GPUs. We used the Huggingface package for the modelling. Furthermore, we used default hyperparameters and a quantization of 0.5. For zero-shot prompting, we had the exact same specifications as outlined in Section 3.4 of the

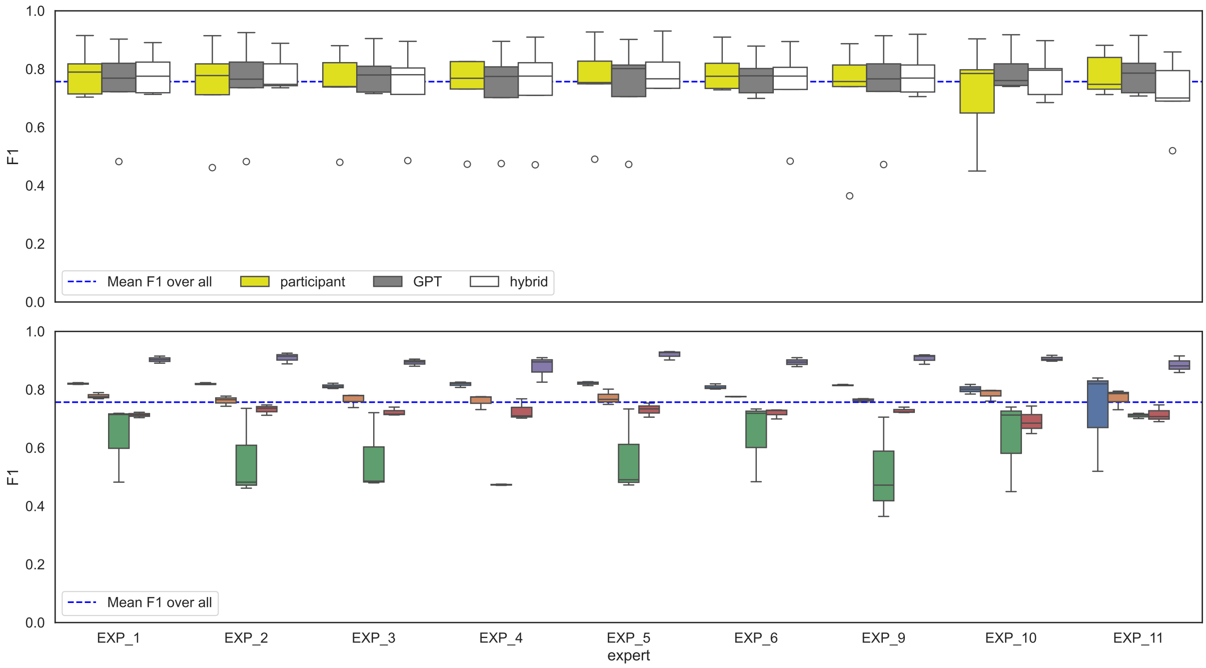

Figure 13: Difference per participant over definition types (above) and datasets (bottom) for the modelling experiments with a temperature of 0.7

main paper. We also used the exact same datasets as specified in the main paper, with five sexism datasets and three different definitions per participant.

H.2 Results

We find overall lower performance (F1 = .695) than the results of GPT4o, and a low variability over expert, dataset, or definition. This shows that GPT with temperature 0 seems indeed most susceptible to prompt changes. There is little performance difference between the three definition types: the LLM-generated definition performs slightly lower (F1 = .699) than the co-written definition (F1 = .584) or participant-written definitions (F1 = .703). Results over different datasets are spread from F1 = .671 (EDOS dataset) to .758 (hatecheck dataset). See Figure 14, representing the LLaMa results in a similar format as the GPT result in the main paper, showing much less variability in performance over different conditions.

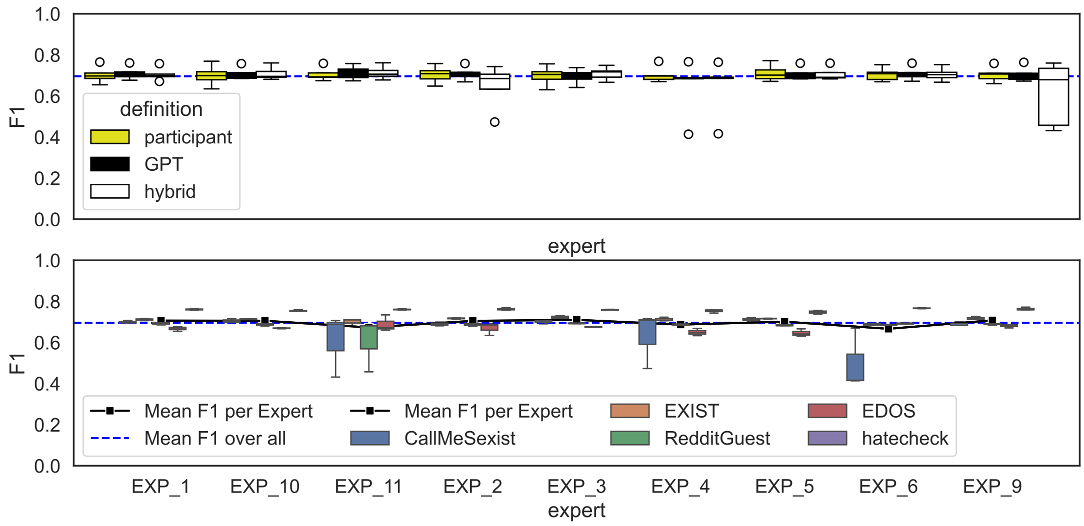

Figure 14: Difference per participant over definition types (above) and datasets (bottom) for the modelling experiments with LLaMa