The Effects of Demographic Instructions on LLM Personas

Angel Felipe Magnossão de Paula
Universitat Politècnica de València
València, Spain
adepau@doctor.upv.es

J. Shane Culpepper
University of Queensland
Brisbane, Australia
s.culpepper@uq.edu.au

Alistair Moffat
The University of Melbourne
Melbourne, Australia
ammoffat@unimelb.edu.au

Sachin Pathiyan Cherumanal
RMIT University
Melbourne, Australia
s3874326@student.rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Johanne Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Abstract

Social media platforms must filter sexist content in compliance with governmental regulations. Current machine learning approaches can reliably detect sexism based on standardized definitions, but often neglect the subjective nature of sexist language and fail to consider individual users' perspectives. To address this gap, we adopt a perspectivist approach, retaining diverse annotations rather than enforcing gold-standard labels or their aggregations, allowing models to account for personal or group-specific views of sexism. Using demographic data from Twitter, we employ large language models (LLMs) to personalize the identification of sexism.

Our empirical results show that OpenAI's LLMs (GPT-3.5, GPT-4, and GPT-4o) and two open-source LLMs (Mistral and Qwen) exhibit higher Krippendorff's alpha label agreement with female annotators than with male annotators. As well, each LLM presents higher Krippendorff's alpha agreement with a specific annotator age group. We then sought to counter these trends by providing "persona" instructions as part of the LLM prompt, with somewhat surprising outcomes, highlighting the potential of user-centered perspectivist methods to improve content moderation systems.

CCS Concepts

• **Information systems** → **Retrieval effectiveness**; *Task models*; Sentiment analysis.

Keywords

Evaluation, sexism detection, perspectivism, large language models, bias, unbiased methods

ACM Reference Format:
Angel Felipe Magnossão de Paula, J. Shane Culpepper, Alistair Moffat, Sachin Pathiyan Cherumanal, Falk Scholer, and Johanne Trippas. 2025. The Effects of Demographic Instructions on LLM Personas. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730255

SIGIR '25, Padua, Italy.

ACM ISBN 979-8-4007-1592-1/25/07
https://doi.org/10.1145/3726302.3730255

1 Introduction and Motivation

Relevance judgments are usually undertaken to develop a set of gold-standard labels, seeking relevance consensus; with those labels then used to support measurement of the extent to which the documents retrieved by some information retrieval (IR) system align with users' information needs, thereby allowing benchmarking of systems. One major challenge in relevance judgments is the high annotator costs for the labeling that is required. To reduce that cost, researchers have proposed the use of LLMs, which have been shown to usefully supplement human labeling and at scale [26].

Bias and fairness in IR systems have also received attention [6, 17]. Nor are LLMs immune – they too can be affected by bias, stereotypical associations [2, 15], and adverse sentiments towards specific groups [12]. For example, gender bias evaluation in natural language processing is a topic that has received much attention [5], with de-biasing techniques having been also been proposed [4].

Faggioli et al. [9] propose a human-machine collaboration spectrum that categorizes judgment strategies based on how much humans rely on machines, and suggest that "AI Assistance" is a likely path for employment of LLMs. Faggioli et al.'s pilot study finds a reasonable correlation between highly-trained human assessors and a fully automated LLM, concluding that while the technology is promising, it requires further study. Use of LLMs for relevance judgment is thus an emerging area of interest [20, 24].

Here, we explore if LLMs can be directed to adopt a *persona* when conducting relevance judgments. In particular, different annotators might react differently when asked "is this tweet sexist", with their answers coming from subjective viewpoints influenced by, amongst other factors, gender and age. That is, the "is it sexist" question may not always have a single "right" answer. Hasler et al. [11] suggest the use of augmented test collections that include user-centric evaluation and anonymized demographic information such as age, gender, and education level, plus task-specific details such as the assessor's expertise, interest, motivation, confidence, and degree of document relevance. Given that range of influencing factors, we are interested in whether an LLM – a "stochastic parrot" [3] – is capable of reflecting different demographic responses. For example, can an LLM be (reliably) instructed to "be a male over 45"?

To investigate if LLMs can mimic subjectivity, we use the Social neTworks (EXIST) Shared Task at CLEF 2023 [18] test collection for sEXism Identification. This dataset contains labels (opinions) from

diverse human judges for a set of tweets[1], labels which do indeed exhibit demographic patterns. By comparing the human labels with LLM-generated labels, we are thus able to infer a "persona" for several commercial and open-source LLM models. The specific research questions we consider are:

(i) Do LLMs exhibit bias toward certain demographics when classifying text as sexist or non-sexist; and

(ii) Can adopting a demographic-based persona mitigate bias in LLMs when classifying text as sexist or non-sexist.

Krieg et al. [14] suggest that female stereotypes are influential in relevance judgments. Our experiments show a similar outcome, with all five tested LLMs correlating more closely with female opinions than with male. The second part of our work here then asks the same set of LLMs to adopt a range of specific demographic personas as they respond to the "is this tweet sexist" question. Surprisingly, all of the LLMs tested seemed to be incapable of doing so. That is, the LLMs were unable to "empathize" and take on different patterns of opinion; and their "personalities" seem to be relatively inflexible.

2 Background

Perspectivism in Text Classification. In traditional text classification annotation, disagreements are resolved into a single "gold standard" label via aggregation [10]. This approach has recently been challenged, especially in subjective tasks like hate speech and sexism detection [18], as it risks enforcing a single ground truth. Perspectivism is a machine learning approach that takes data annotated by different individuals and models the varied perspectives that influence their opinions and world view [10]. We adopt the perspectivist approach to examine biases in sexism detection; and then consider the specific question to whether LLMs can simulate personas based on demographic attributes.

Sexism Detection. Sexism detection is the task of deciding if text contains sexist content. Traditional sexism detection systems have relied on predefined labels and fixed perspectives, overlooking the nuanced and subjective nature of sexist statements; moreover, as social media have expanded their influence, researchers have focused on developing scalable approaches to sexism detection.

A significant advancement towards addressing this issue is the EXIST (sEXism Identification in Social neTworks) Lab at CLEF [18], designed for perspectivist learning by highlighting annotation disagreements rather than imposing gold-standard labels. In particular, different annotators might react differently when asked if a given tweet "is sexist". The EXIST initiative acknowledges the inherent subjectivity in sexism classification, and aims to improve model robustness by incorporating diverse perspectives in the annotation process. Various approaches have been proposed, ranging from rule-based methods [23] to machine learning techniques [7, 8, 21]. In this study we investigate the role of LLMs.

LLM Personas. Recent research has explored how LLMs can mimic personas based on prompts that describe the demographics of a target user group using *persona prompting* [27]. However, the use of persona prompting is still poorly understood, and has led to somewhat inconsistent outcomes, in differences that might be attributed to the lack of clarity on how small edits to the input prompt can produce unexpected changes in the generated text [25, 27].

To mitigate this issue, Aguda et al. [1] proposed a "reliability index" called LLM-Relindex, which can be used to identify input prompts that may require a domain expert to review the output results. Aguda et al. found that LLM-Reindex was most reliable when prompts were customized by persona. Furthermore, LLM-generated personas have also shown to exhibit demographic biases. Salminen et al. [22] identified biases in age and occupation, and a strong tendency towards personas from the United States. Furthermore, Zheng et al. [28] demonstrate that such personas do not consistently improve performance; but that gender, and domain-based personas do sometimes lead to improved performance.

[1]User-provided ground truth labels are not publicly available, meaning that this dataset could not have been used as training by LLM, an important and necessary assurance.

Table 1: EXIST 2023 label distribution over 7,958 tweets.

	Gender		Age		
	Female	**Male**	**18–22**	**23–45**	**46+**
Sexist	10,961	10,790	6,933	7,422	7,396
Non-Sexist	12,913	13,084	8,983	8,494	8,520
Total	23,874	23,874	15,916	15,916	15,916

3 Methodology

Data. We use the EXIST 2023 shared task dataset [18, 19], designed for perspectivist learning by capturing annotation disagreements rather than assigning single labels. The collection focuses on sexism in tweets, and includes demographic data about the annotators, enabling further analysis of bias and subjectivity in the classification process. The primary task involves distinguishing sexist from non-sexist content, rated using a binary scale. Each tweet is annotated by six individuals, stratified across two factors: gender (male and female) and age group (18–22, 23–45, and 46+), ensuring diversity of perspective. There are 7,958 tweets in total, divided into training, development, and test sets. We merged the training and development subsets into a single file to increase the number of samples, and streamline analysis.

Instead of a rigid binary classification, a soft-labeling framework based on the proportion of human annotators that selected each category is used, capturing annotator disagreement by providing probabilistic distributions over the two categories, which sum to 1.0. The organizers of EXIST 2023 provided the annotations for the training and development sets. Table 1 presents the distribution of annotations (sexist or non-sexist) across gender and age groups. The dataset maintains a balanced composition of male and female annotators, as well as across different age groups.

We analyze label distributions by demographic groups, and test for statistically significant differences between group means using a t-test for gender (two levels), and a one-way ANOVA for age-group (three levels), against an alpha of 0.05. A post hoc Tukey's HSD test was performed to determine the specific groups contributing to these differences. Differences in the gender factor were not significant ($p = 0.237$). The ANOVA indicated a significant difference between the mean values of different age groups, with the follow-up Tukey's HSD test showing significant differences between the

18–22 and the 23–45 groups, and between the 18–22 and the 46+ groups, reinforcing the presence of annotation differences: the 18–22 age group identified sexist content less frequently (and non-sexist content more frequently) than the other groups, while the 23–45 and 46+ groups exhibited more similar annotation patterns. These results highlight the role of the perspectivist approach.

Large Language Models. Three LLMs from OpenAI were used to evaluate model performance in detecting sexism: GPT-3.5 (gpt-3.5-turbo-0125), GPT-4 (gpt-4-turbo-2024-04-09), and GPT-4o (gpt-4o-2024-08-06). Additionally, we included two open-source LLMs, Mistral (Mistral-Small-Instruct-2409 22B) and Qwen (Qwen2.5-14B), to provide a comparative analysis of sexism detection capabilities across these five model architectures.

Prompt Creation. Three prompt candidates were developed based on the EXIST 2023 task guidelines,[2] focusing on specificity, grammar, and clarity. These three prompts were used to annotate twenty randomly selected tweets from the EXIST 2023 dataset. The prompt with the highest output consistency across three LLMs was chosen, where consistency was measured as the percentage of cases in which all models produced the same classification.[3] The prompt with the highest output consistency achieved a 75% success rate, followed by the other two prompts, which attained 70% and 55%.

Following the lead of others [13, 29], we optimized the prompt using an LLM (o1-preview, the most advanced OpenAI model at the time of experimentation) to obtain the version shown in Figure 1. The refined prompt incorporates placeholders for demographic information to be inserted (gender and age group), allowing exploration of how LLMs judge sexism in tweets when instructed to do so from different perspectives.

Bias Analysis. To study whether LLM predictions are subject to bias, a series of evaluations were conducted. Model predictions were generated first using the baseline prompt without any demographic cues (that is, without the bold text shown in Figure 1). The two demographic factors – gender and age – were then incorporated into the prompt, to explore their impact on sexism classification. Agreement between model outputs and human annotations was quantified using Krippendorff's α. To establish the robustness of the analysis, we conducted confidence intervals using bootstrap resampling (10,000 iterations). All of the measured Krippendorff's α coefficients had confidence intervals smaller than 0.001.

```
You are an expert linguist specializing in detecting
sexism in social media texts. Your demographic information
is as follows:
-Sex: [insert sex: male/female]
-Age Group: [insert age group: 18-22/23-45/46+]

Task: Analyze the given tweet and determine whether it
contains sexist expressions or behaviors. This includes
tweets that:
-Are sexist themselves (contain sexist language or
promote sexist ideas).
-Describe a sexist situation (narrate or depict scenarios
where sexism occurs).
-Criticize sexist behavior (call out or oppose sexism).

Instructions:
1. Read the Tweet Carefully: Examine the content of the
tweet thoroughly.
2. Make a Determination: Decide if the tweet fits any of
the above categories related to sexism.
3. Classification:
-Assign "YES" if the tweet contains any form of sexist
content as defined.
-Assign "NO" if the tweet does not contain sexist content.
4. Output Format: Provide the assigned Category in plain
text.
5. Constraint: You must not retrieve any text apart from
the two possible categories, YES and NO.

TWEET: [insert tweet]
```

Figure 1: Prompt structure. Bold text indicates the parts that were added/varied in the various experiments, as extensions beyond the "baseline" prompt.

4 Results

Do LLMs exhibit bias toward certain demographics when classifying text as sexist or non-sexist? We evaluate Krippendorff's α between each LLM and human annotators across two demographics – gender and age.

First, we compare Krippendorff's α of LLMs against female and male annotators. Table 2 shows that LLMs consistently show higher agreement with female annotators than with male annotators when classifying text as sexist, highlighting an inherent gender-based bias, with GPT-3.5 demonstrating the highest agreement with female annotators, and GPT-4o the lowest. Additionally, as we progress from GPT-3.5 to GPT-4 and subsequently to GPT-4o, we observe a decrease in agreement with both male and female annotators, indicating a decline in alignment as the models have evolved.

On the other hand, age-related patterns do not follow a consistent trend of bias across LLMs. Table 3 shows that GPT-3.5 and Mistral align more closely with annotators aged 46+, while GPT-4 and Qwen show higher agreement with the 23–45 age group. Meanwhile, GPT-4o aligns most with annotators aged 18–22.

Overall, our results demonstrate that LLMs exhibit bias towards certain demographics when classifying text as sexist, with higher agreement with female annotators than with male annotators. Note that throughout these results the observed differences were greater than the computed confidence intervals of the measured values.

Can adopting a demographic-based persona mitigate bias in LLMs when classifying text as sexist or non-sexist? Demographic factors – specifically gender and age – were incorporated into the LLM prompt to mitigate biases by adopting a demographic-based persona, indicated by the bold text sections in Figure 1.

Table 2 shows that only GPT-4_F, GPT-$4o_F$, and Mistral$_F$ exhibited increased agreement with female annotators compared to their base models, where the subscripts indicate the instruction added to the LLM prompts. However, GPT-3.5_F demonstrated a decrease in agreement with female annotators. Similarly, only GPT-4_M and Qwen$_M$ demonstrated improved agreement with male annotators, while remaining LLMs showed a decline. These findings indicate that incorporating gender-based personas in prompts should not be assumed to mitigate gender bias in LLMs for this task.

[2] https://nlp.uned.es/exist2023/

[3] With small optimization changes required to employ the prompts against the two open source models, Mistral, and Qwen, both of which rank highly on leaderboards for their model size.

Table 2: Krippendorff's α scores comparing human annotators by gender with LLMs. Bold text indicates the model's highest agreement across all gender groups. Subscripts F and M denote gender-based personas through persona prompting in LLMs.

Model	F (Female)	M (Male)
Human Annotators (F)	**1.000**	0.477
Human Annotators (M)	0.477	**1.000**
GPT-3.5	**0.415**	0.371
GPT-3.5$_F$	**0.398**	0.358
GPT-3.5$_M$	**0.404**	0.360
GPT-4	**0.365**	0.325
GPT-4$_F$	**0.401**	0.360
GPT-4$_M$	**0.372**	0.336
GPT-4o	**0.228**	0.191
GPT-4o$_F$	**0.234**	0.198
GPT-4o$_M$	**0.213**	0.172
Mistral	**0.353**	0.310
Mistral$_F$	**0.363**	0.326
Mistral$_M$	**0.330**	0.293
Qwen	**0.378**	0.345
Qwen$_F$	**0.372**	0.337
Qwen$_M$	**0.382**	0.347

Table 3: Krippendorff's α scores comparing human annotators by age group with LLMs. Bold text indicates the model's highest agreement across all age groups. Subscripts denote age-based personas introduced through prompting in LLMs.

Model	18–22	23–45	46+
Human Annotators (18–22)	**1.000**	0.445	0.436
Human Annotators (23–45)	0.445	**1.000**	0.463
Human Annotators (46+)	0.436	0.463	**1.000**
GPT-3.5	0.382	0.408	**0.413**
GPT-3.5$_{18-22}$	0.372	0.399	**0.409**
GPT-3.5$_{23-45}$	0.365	0.398	**0.402**
GPT-3.5$_{46+}$	0.383	0.407	**0.419**
GPT-4	**0.421**	**0.421**	0.404
GPT-4$_{18-22}$	0.455	**0.462**	0.452
GPT-4$_{23-45}$	0.446	**0.484**	0.430
GPT-4$_{46+}$	0.463	**0.474**	0.457
GPT-4o	**0.316**	0.290	0.278
GPT-4o$_{18-22}$	**0.286**	0.261	0.247
GPT-4o$_{23-45}$	**0.302**	0.272	0.265
GPT-4o$_{46+}$	**0.302**	0.271	0.262
Mistral	0.368	0.384	**0.392**
Mistral$_{18-22}$	0.372	0.389	**0.392**
Mistral$_{23-45}$	0.378	0.392	**0.398**
Mistral$_{46+}$	0.360	0.377	**0.383**
Qwen	0.406	**0.418**	0.404
Qwen$_{18-22}$	0.421	**0.432**	0.424
Qwen$_{23-45}$	0.423	**0.437**	0.427
Qwen$_{46+}$	0.412	**0.419**	0.411

For age-based personas, Table 3 shows that GPT-4, Mistral, and Qwen consistently exhibited higher agreement with the prompt-included age grouping than the base models, whereas GPT-3.5 demonstrated increased agreement only for the 46+ age group. Meanwhile GPT-4o did not indicate any improvement in agreement along any age groups.

Overall, the findings indicate that instructing LLMs to adopt demographic-based personas has inconsistent and unpredictable effects. While persona prompting improves alignment with certain demographic groups in some models, it decreases alignment in others, meaning that it cannot be relied upon, and that it may not be presumed to provide a reliable way of mitigating bias.

5 Discussion and Conclusion

In this study, we investigated demographic biases in LLMs when classifying text as sexist or non-sexist. By analyzing Krippendorff's α agreement between LLM-based annotations and human annotations across gender and age breakdowns, we found that LLMs consistently align more closely with female annotators than male annotators, indicating a gender-based bias. However, age-based breakdowns exhibited no clear pattern of bias, with different models aligning with different age groups.

Adopting a structure that might be thought of as addressing these biases, we also explored the effectiveness of persona prompting, incorporating gender and age information into the LLM prompts. Our results indicate that this approach yields inconsistent and unpredictable outcomes. Some models improved alignment with specific demographic groups, but others showed decreased agreement. We thus cannot (yet) regard demographic-based prompting to be an effective way of mitigating LLM bias, an outcome that we expect will be of interest (and concern) to other researchers.

Future Work. We plan to extend our study beyond binary classification to explore whether incorporating more granular categories, such as those defined by Plaza et al. [19], can provide insights into the inconsistencies observed in model agreement. Moreover, humans possess multi-faceted personas that encompass the intersectionality of various demographics [16]. While the current study addresses one-dimensional personas, we aim to explore demographic intersectionality in future work by examining how multiple factors interact and whether this results in more consistent alignment across LLMs. Beyond age and binary gender, other characteristics may also shape annotators' perspectives on sexism, and we plan to assess their influence on model behavior. Lastly, we plan to extend our study to a wider range of LLMs, to identify what common factors may exist, and to assess the generalizability of our findings.

Acknowledgment. This work was supported in part by the Australian Research Council (projects DP190101113, DE200100064, and CE200100005) and was undertaken with the assistance of computing resources from RACE (RMIT AWS Cloud Supercomputing).

References

[1] T. D. Aguda, S. Siddagangappa, E. Kochkina, S. Kaur, D. Wang, and C. Smiley. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 10124–10145, 2024. URL https://aclanthology.org/2024.lrec-main.885/.

[2] C. Basta, M. R. Costa-jussà, and N. Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, 2019. doi: 10.18653/v1/W19-3805.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, 2021. doi: 10.1145/3442188.3445922.

[4] A. Bigdeli, N. Arabzadeh, S. Seyedsalehi, B. Mitra, M. Zihayat, and E. Bagheri. De-biasing relevance judgements for fair ranking. In *Advances in Information Retrieval*, pages 350–358. Springer Nature Switzerland, 2023.

[5] M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, editors. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 2019.

[6] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 6437–6447, 2024. doi: 10.1145/3637528.3671458.

[7] A. F. M. de Paula and R. F. da Silva. Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models. In *IberLEF@ SEPLN*, 2022.

[8] A. F. M. de Paula, R. F. da Silva, and I. B. Schlicht. Sexism prediction in Spanish and English tweets using monolingual and multilingual BERT and ensemble models. *arXiv preprint arXiv:2111.04551*, 2021.

[9] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, page 39–50, 2023. doi: 10.1145/3578337.3605136.

[10] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, and D. Bernardi. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*, pages 1–28, 2024.

[11] L. Hasler, M. Halvey, and R. Villa. Augmented test collections: A step in the right direction. *arXiv preprint arXiv:1501.06370*, 2015.

[12] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, 2020. doi: 10.18653/v1/2020.acl-main.487.

[13] D. Kepel and K. Valogianni. Autonomous prompt engineering in large language models. *arXiv preprint arXiv:2407.11000*, 2024.

[14] K. Krieg, E. Parada-Cabaleiro, M. Schedl, and N. Rekabsaz. Do perceived gender biases in retrieval results affect relevance judgements? In *Advances in Bias and Fairness in Information Retrieval*, pages 104–116. Springer International Publishing, 2022.

[15] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, 2019. doi: 10.18653/v1/W19-3823.

[16] A. Liu, M. Diab, and D. Fried. Evaluating large language model biases in persona-steered generation. *arXiv preprint arXiv:2405.20253*, 2024.

[17] R. Lunardi, D. La Barbera, and K. Roitero. The elusiveness of detecting political bias in language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, page 3922–3926, 2024. doi: 10.1145/3627673.3680002.

[18] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. Overview of EXIST 2023 – Learning with disagreement for sexism identification and characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 316–342. Springer Nature Switzerland, 2023.

[19] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. Overview of EXIST 2023: sEXism Identification in Social NeTworks. In *Advances in Information Retrieval*, pages 593–599. Springer Nature Switzerland, 2023.

[20] H. A. Rahmani, C. Siro, M. Aliannejadi, N. Craswell, C. L. A. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz. LLM4Eval: Large language model for evaluation in IR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3040–3043, 2024. doi: 10.1145/3626772.3657992.

[21] F. Rodríguez-Sánchez, J. Carrillo de Albornoz, and L. Plaza. Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 8:219563–219576, 2020. doi: 10.1109/ACCESS.2020.3042604.

[22] J. Salminen, C. Liu, W. Pian, J. Chi, E. Häyhänen, and B. J. Jansen. Deus ex machina and personas from large language models: Investigating the composition of AI-generated persona descriptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024. doi: 10.1145/3613904.3642036.

[23] M. Samory, I. Sen, J. Kohne, F. Flöck, and C. Wagner. "Call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584, 2021.

[24] J. A. Schnabel, J. Trippas, F. Scholer, and D. Hettiachchi. Multi-stage large language model pipelines can outperform GPT-4o in relevance assessment. In *Proceedings of the ACM Web Conference*, 2025. doi: 10.1145/3701716.3715488.

[25] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

[26] P. Thomas, S. Spielman, N. Craswell, and B. Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1930–1940, 2024. doi: 10.1145/3626772.3657707.

[27] P. Zhan, Z. Xu, Q. Tan, J. Song, and R. Xie. Unveiling the lexical sensitivity of LLMs: Combinatorial optimization for prompt enhancement. *arXiv preprint arXiv:2405.20701*, 2024.

[28] M. Zheng, J. Pei, L. Logeswaran, M. Lee, and D. Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing*, pages 15126–15154, 2024. doi: 10.18653/v1/2024.findings-emnlp.888.

[29] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/10850.