

On the Neutrality of Model Confidence in German and English Gender-Bias Benchmarks

Shalaka Satheesh^{1,2}, Nika Jacob¹, Katharina Beckh^{1,2},

¹ Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS,

² Lamarr Institute for Machine Learning and Artificial Intelligence

Correspondence: shalaka.satheesh@iais.fraunhofer.de

Abstract

Confidence, the probability a model assigns to its prediction, is commonly treated as a signal for correctness, and is central to quality monitoring and selective question answering. We ask whether this signal is neutral with respect to social stereotypes. A model’s confidence may reflect correctness well overall and yet be higher where its answers align with a stereotype and detecting this requires comparing confidence across stereotype directions. We investigate this on the gender subset of Bias Benchmark for Question Answering (BBQ) in English and in German (GG-BBQ), evaluating 22 base and instruction-tuned models from five providers. For some models, we find small but directionally consistent effects for both ambiguous and disambiguated questions, suggesting that confidence-based selection can favour stereotype-aligned content and that confidence is worth examining alongside answer-based scores in bias evaluations on multiple-choice questions.

1 Introduction

Language models assign a probability to their choice of answer, which is commonly treated as a proxy for correctness (Kadavath et al., 2022). This is central to downstream tasks such as selective answering, quality monitoring, and deferral (Kamath et al., 2020; Geng et al., 2024). In our work, we ask whether this proxy is neutral with respect to social stereotypes.

Social bias in language models is commonly evaluated and audited with question answering (QA) benchmarks such as the Bias Benchmark for QA (BBQ) (Parrish et al., 2022) in English and its adaptations to other languages (Jin et al., 2024; Zulaika and Saralegi, 2025; Satheesh et al., 2025; Farsi et al., 2025), in which models answer questions over ambiguous contexts (where the correct answer is “unknown”) and disambiguated contexts

(where the answer can be inferred from the context). As in the original BBQ, this work focuses on representational harms, defined as “harms that occur when systems reinforce the subordination of some groups along the lines of identity” by Crawford (2017). That is, harms caused due to stereotyping behaviour, leading to altered perceptions of groups of people. These benchmarks are usually presented in a multiple-choice QA setting and the metrics derived from them are computed based on the correctness and the counts of the chosen answer option. Typically, the confidence with which the model selects its answer plays no role in this evaluation.

A model’s confidence could reflect correctness well overall and still be higher when answers align with a stereotype. More concretely, in ambiguous contexts we investigate whether confidence in unsupported answers rises with their alignment with a stereotype. And similarly, in disambiguated contexts, we ask whether correct answers aligned with a stereotype are stated with more confidence than those contradicting stereotypes. For this, we evaluate 22 open-weight models (11 base/instruct pairs) on the gender subsets of the datasets BBQ (English) and GG-BBQ (German). We find small but directionally consistent associations in both languages. Since confidence is a signal that confidence-based downstream pipelines consume, and is already computed by standard evaluation frameworks, we argue that it deserves more attention alongside answer-based scores when auditing for social bias.

2 Related Work

2.1 Confidence Estimation

Geng et al. (2024) divide confidence estimation for LLMs into white-box methods, which require access to model internals (e.g., token logits or internal states), and black-box methods, which rely on model outputs, such as verbalized confidence

elicited by prompting (Tian et al., 2023; Xiong et al., 2024) and consistency-based estimation. As we study open-weight models on a multiple-choice QA task, we use a white-box estimate. We apply softmax to the log-likelihoods of the three answer options and take the chosen option’s probability as the model’s confidence (Jiang et al., 2021). Since such option probabilities are distorted by selection biases towards particular option positions and identifiers (Zheng et al., 2024), we evaluate each question under all permutations of the answer order. Closest to our setting, Cocchieri et al. (2026) show on medical multiple-choice QA that even when the correct option is removed and an explicit “I abstain” option is the only correct choice, LLMs rarely select it, and that abstention behaviour is only weakly related to their verbalized confidence. This mirrors the ambiguous BBQ condition, in which “unknown” is correct and choosing any other option is a potentially stereotyped prediction.

2.2 Bias Evaluation

Bias in LLMs is typically quantified through (1) embeddings and probabilities of output tokens and (2) analysis of generated texts (Nangia et al., 2020; Parrish et al., 2022). Datasets used for bias evaluations are built for tasks such as fill-in-the-blank (Nangia et al., 2020), text generation (Gehman et al., 2020) and question answering (Parrish et al., 2022). Parrish et al. (2022) cast bias evaluation as QA in ambiguous contexts over nine demographic dimensions (age, disability status, gender identity, nationality, physical appearance, race/ethnicity, religion, socio-economic status and sexual orientation). PBBQ (Farsi et al., 2025), a BBQ-style Persian benchmark, jointly reports accuracy, bias, and uncertainty. PBBQ also reports a log-probability-based bias score and an entropy-based uncertainty score, but treats fairness and confidence as separate aggregate dimensions. In our work, we test whether they are related. For this, we use two gender bias benchmarks: the respective subset from BBQ (Parrish et al., 2022) and GG-BBQ (Satheesh et al., 2025), a German translation of the BBQ gender-identity subset.

3 Methodology

We empirically test whether model confidence is distributed neutrally across stereotype directions. In a multiple-choice QA setting, 22 models are evaluated on over 10,000 samples from two gender

bias benchmarks in English and German. We define confidence as the probability assigned to the chosen answer using softmax-normalised log-likelihoods. We analyse its relation to stereotype alignment using Spearman correlations over erroneous answers in ambiguous contexts, comparisons of mean confidence between stereotype-aligned and non-aligned correct answers in disambiguated contexts and analyses of stereotype alignment across confidence quartiles, controlling for answer order effects.

3.1 Dataset

We use BBQ (Parrish et al., 2022) and GG-BBQ (Satheesh et al., 2025) datasets for gender bias evaluation in English and German respectively. The gender identity subset of the BBQ dataset is used unmodified with a total of 5672 samples. For the GG-BBQ dataset: Subset-I (containing samples with explicit gender terms, ein Mann/eine Frau) is used unmodified and Subset-II is regenerated from the templates to ensure balanced stereotype-aligned and non-aligned samples with matched male/female counterparts. Thus, Subset-I contains a total of 968 samples and Subset-II contains a total of 4968 samples. Both datasets contain questions with a context and three options (two persons and an “unknown” option). In ambiguous contexts “unknown” is always treated as the correct answer because the context does not contain enough information for the question to be answered. The questions occur in negative and non-negative stereotype testing pairs. An example for a sample from the BBQ dataset subsets is shown in Table 1.

3.2 Experimental Set-up

LM Evaluation Harness (Gao et al., 2024) was used as the evaluation framework under a zero-shot, multiple-choice question (MCQ) setting, where the model is presented with the question, the context and three choices, and is prompted to pick a single option. The answers are scored by comparing the log-likelihoods the model assigns to the three options and the highest scoring option is chosen as the model’s answer. We evaluate 22 open-weight models (11 base and instruction-tuned pairs from 5 providers, Table 5) on the gender bias evaluation datasets for English and German. Following Jin et al. (2024), where each question is presented under the three cyclic rotations of its answer choices, we extend this order balancing to all $3! = 6$ permutations. This additionally balances the relative order of every answer pair and cancels out position

	Ambiguous Context	Disambiguated Context
Context	We couldn't start the board meeting at 9am today because a man and a woman were late.	We couldn't start the board meeting at 9am today because a man and a woman were late. We had to wait for the woman because they were running the meeting, and the man was responsible for taking all the notes
Negative Q	Who was the secretary? <i>Answer:</i> Not enough information	Who was the secretary? <i>Answer:</i> The man
Non-negative Q	Who was the board chair? <i>Answer:</i> Not enough information	Who was the board chair? <i>Answer:</i> The woman

Table 1: An example from the BBQ gender-identity subset with two types of contexts and questions testing both negative and non-negative stereotypes

preferences.

3.3 Metrics

In order to study the association of bias and confidence, we evaluate the confidence as the probability of the model’s chosen option under a softmax over the log-likelihoods of the three answer choices.

In ambiguous contexts, the correct answer “unknown” has no stereotype direction, i.e., it is neither pro nor anti-stereotype. However, among the incorrect answers, a direction exists by construction and since these incorrect answers are all equally wrong, correctness cannot explain confidence differences among them. In disambiguated contexts, errors do exist, but they are few and mix different failure types (abstentions, wrong choice). Therefore in our work, for the disambiguated contexts, we consider the complementary question to the ambiguous case, i.e., when the model is right in both directions, does it state stereotype-confirming truths more confidently than stereotype-contradicting ones? We therefore compare only correct answers, whose direction is given by whether the gold answer confirms the stereotype, so that, again, none of the compared answers differ in correctness.

Additionally, we evaluate accuracy as the fraction of permutations where the model picks “unknown” in ambiguous contexts and the fraction where it picks the gold label in disambiguated contexts. For each model and each ambiguous question, we consider the permutations in which the model chose an option other than “unknown” (these are by definition incorrect answers in ambiguous contexts). Over these permutations we compute two quantities per question: the fraction of choices aligned with the stereotype, and the model’s mean confidence over those same permutations. Then the

Spearman correlation between the two quantities across all questions is calculated for each model. The questions where the model abstained under all six permutations are excluded. Additionally, we ranked each model’s questions by mean answer confidence and split them into quartiles to identify patterns for low and high confidence.

4 Results

Table 2 shows the results for all models on the German and English datasets. We observe that the mean disambiguated accuracy is comparable across languages (both have a mean of ≈ 0.75). Interestingly, we find that models choose the “unknown” option correctly on ambiguous items far less often in German (0.30) than in English (0.56), i.e., responses that align with social stereotypes and responses that go against are observed more when the context is ambiguous for the German dataset.

Bias and confidence in an ambiguous setting

In German, the Spearman correlation coefficient r is positive in 17 out of the 22 models and in English for 20 models. A two-sided sign test confirms this predominance of positive correlations is unlikely under chance among the evaluated models ($p = 0.017$ German, $p = 0.0001$ English). After a Bonferroni correction over the four sign tests we report, the German result is no longer significant ($p = 0.068$: Appendix A.3). However, the per-model correlations are weak and vary widely across models: In German, r ranges from -0.20 to $+0.46$ with a median of 0.08 and in English from -0.03 to $+0.32$ with a median of 0.13 . The strongest correlations in German are observed in gemma-3-4b-it (0.46), Qwen3-14B (0.43), and Llama-3.1-8B-Instruct (0.41) and in English in Mistral-Nemo-Instruct (0.32), gemma-3-4b-it (0.31), and Qwen3-

14B-Base (0.28). Comparing within families (tests in Appendix A.2), we observe that the instruct variant shows a stronger association than its base counterpart in 9 of 11 German families (median change, $\Delta r = +0.18$), but not in English (6 of 11, $\Delta r = +0.01$).

Bias and confidence in a disambiguated setting

Since most models show a considerably better performance in terms of accuracy in disambiguated contexts, we study the correctly answered samples in this case. For each model and each disambiguated question, we consider the permutations answered correctly and compute the mean confidence over them (questions the model never answers correctly are excluded). For each model we calculate Δ_{conf} , by comparing the mean confidence on questions whose correct answer aligns with the stereotype against those whose correct answer contradicts it, restricted to the male/female questions for which template-matched gendered counterparts exist. Questions involving transgender and non-binary persons were excluded from this evaluation as most lack a matched role-swapped counterpart in the German data. We also make the same exclusion in the English dataset for comparability. The distributional difference is assessed per-model with a two-sided Mann–Whitney U test (significance markers (*)) in Table 2). Each Δ_{conf} in the table is the average confidence advantage a model gives to stereotype-confirming correct answers, i.e. $+0.014$ means correct answers are stated 1.4 percentage points more confidently when the truth matches the stereotype.

We observe that Δ_{conf} is positive in 21 of the 22 models in German and in 20 models in English and a two-sided sign test confirms this predominance of positive gaps is unlikely under chance among the evaluated models ($p < 0.0001$ in both languages). However, similar to the ambiguous setting, the per-model effects are small with Δ_{conf} ranging from -0.001 to $+0.040$ in German (median $+0.012$) and from -0.002 to $+0.044$ in English (median $+0.017$). The single negative value per language (deepseek-llm-7b-base in German, Llama-3.1-70B-Instruct in English) is close to zero. For several instruction-tuned models, the gap is small because they assign a confidence close to 1.0 to almost all correct answers (Figure 1). These two measures do not have to agree for a given model, since r looks at the answers a model gets wrong and Δ_{conf} at the ones it gets right. Four models in German and two

in English have a negative r but a positive Δ_{conf} .

Confidence quartiles Table 3 shows the stereotype alignment per within-model confidence quartile. For each model, the questions are ranked by the model’s mean answer confidence and split into quartiles, where Q1 represents the least confident quartile and Q4 most confident quartile. For this, confidence is measured over the erroneous answer permutations in ambiguous contexts and over the correct answer permutations in disambiguated contexts (as explained in Section 3.3). The cells in the table for the columns Q1 to Q4 give the fraction of stereotype-aligned answers per quartile, averaged over the 22 models. Figure 1 gives the range of confidence for each model, per quartile. Table 3 shows that among the erroneous answers in the ambiguous contexts, the fraction of stereotype-aligned answers is higher in the top quartile than in the bottom for 18 models out of 22 in German and for 20 models in English. Similarly, among the correct answers in the disambiguated contexts, the fraction of stereotype-aligned answers is higher in the top quartile than in the bottom for 19 models in German and for 21 models in English.

Effect of answer order Benchmark evaluations typically present answer options in a single fixed order, although LLMs are known to have a positional bias (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024). In the bias evaluation datasets presented above, as “unknown” always appears in the last of the three options, a positional bias could produce distorted results. To verify this, we repeated the evaluation on the GG-BBQ with a single fixed order. This shifts per-model ambiguous context accuracies by 0.12 on average and up to 0.58 (e.g.: Mistral-7B-v0.3: 0.90 on fixed order and 0.32 under six orderings) and changes the direction of the correlation coefficients for 6 models.

5 Discussion

Model behaviour in the ambiguous contexts is the most interesting for bias analysis. This is because the context does not support an answer other than “unknown”, any other answer given by the model would reflect the model’s biases (Parrish et al., 2022). In ambiguous contexts, the strongest correlations are observed in a few models that recur across both languages: gemma-3-4b-it ($r = 0.46$ German, 0.31 English), Llama-3.1-8B-Instruct (0.41 German, 0.22 English) and Mistral-

Model	GG-BBQ (German)				BBQ Gender Subset (English)			
	acc _{amb}	acc _{dis}	r	Δ_{conf}	acc _{amb}	acc _{dis}	r	Δ_{conf}
Qwen3-14B-Base	0.32	0.88	0.119*	0.012*	0.60	0.92	0.285*	0.031*
Qwen3-14B	0.84	0.75	0.426*	0.024*	0.98	0.80	0.182*	0.032*
Qwen3-8B-Base	0.46	0.90	0.082*	0.011	0.77	0.88	0.158*	0.013*
Qwen3-8B	0.54	0.88	0.088*	0.016*	0.95	0.75	0.101*	0.003
deepseek-llm-67b-base	0.09	0.60	-0.015	0.001	0.34	0.59	0.040*	0.025*
deepseek-llm-67b-chat	0.02	0.77	0.090*	0.040*	0.62	0.85	0.054*	0.027*
deepseek-llm-7b-base	0.21	0.38	-0.045*	-0.001	0.23	0.40	-0.027	0.000
deepseek-llm-7b-chat	0.19	0.42	0.027	0.005	0.30	0.62	0.040*	0.026*
gemma-3-12b-pt	0.18	0.75	0.230*	0.018*	0.37	0.72	0.062*	0.018*
gemma-3-12b-it	0.48	0.88	0.086*	0.001*	0.86	0.87	0.035	0.008*
gemma-3-27b-pt	0.17	0.74	-0.123*	0.029*	0.47	0.57	0.041*	0.012*
gemma-3-27b-it	0.47	0.89	0.053*	0.005*	0.89	0.88	-0.025	0.001
gemma-3-4b-pt	0.07	0.57	0.041*	0.002	0.13	0.64	0.007	0.016*
gemma-3-4b-it	0.06	0.89	0.462*	0.006*	0.42	0.84	0.314*	0.015*
Llama-3.1-70B	0.16	0.85	0.059*	0.012*	0.80	0.72	0.143*	0.020*
Llama-3.1-70B-Instruct	0.54	0.90	0.290*	0.017*	0.91	0.93	0.143*	-0.002
Llama-3.1-8B	0.07	0.79	0.225*	0.024*	0.25	0.82	0.181*	0.034*
Llama-3.1-8B-Instruct	0.55	0.89	0.407*	0.032*	0.73	0.88	0.216*	0.044*
Mistral-Nemo-Base-2407	0.32	0.71	-0.201*	0.021*	0.35	0.64	0.120*	0.010
Mistral-Nemo-Instruct-2407	0.18	0.93	0.249*	0.005*	0.38	0.94	0.322*	0.010*
Mistral-7B-v0.3	0.32	0.46	0.021	0.026*	0.53	0.50	0.161*	0.020*
Mistral-7B-Instruct-v0.3	0.33	0.77	-0.023	0.009*	0.51	0.79	0.192*	0.021*

Table 2: Results on GG-BBQ (German) and the gender subset of the BBQ (English) datasets of 22 models, averaged over all six answer orderings. acc_{amb}: accuracy in ambiguous contexts, where the correct answer is “unknown”, acc_{dis}: accuracy in disambiguated contexts, r : per-item correlation between confidence and pro-stereotype choice fraction in erroneous ambiguous contexts (Section 3). Δ_{conf} : difference in the mean confidence between stereotype-aligned and non-aligned correct answers in disambiguated contexts (Section 3). * $p < 0.05$ (two-sided per-model tests: Spearman for r , Mann-Whitney U for Δ_{conf}).

		Q1	Q2	Q3	Q4	Q1 < Q4	Q1 < Q2 < Q3 < Q4
German	stereotype-aligned errors (ambig.)	0.55	0.56	0.58	0.63	18/22	5/22
German	stereotype-aligned correct answers (disamb.)	0.46	0.47	0.50	0.58	19/22	2/22
English	stereotype-aligned errors (ambig.)	0.56	0.60	0.61	0.68	20/22	9/22
English	stereotype-aligned correct answers (disamb.)	0.45	0.49	0.52	0.55	21/22	9/22

Table 3: Stereotype alignment per within-model confidence quartile, averaged over the 22 models (See Section 4). The column Q1 < Q4 gives the number of models whose stereotype-aligned fraction in the top confidence quartile exceeds that in the bottom quartile and the column Q1 < Q2 < Q3 < Q4 gives the number of models in which this fraction rises across all four quartiles.

Nemo-Instruct (0.25 German, 0.32 English). For Qwen3-14B, the effect is stronger in German (0.43) but, in English, its base variant leads (0.28). For these models we interpret the results as follows: when it chooses an answer that the context does not support, its confidence is higher the more its choices align with the stereotype. This bias is not captured by the accuracy metrics and bias metrics typically computed from discrete answer choices alone.

In disambiguated contexts, with enough infor-

mation, the models show high confidence when answering correctly in both stereotype-aligned and anti-stereotype directions. Therefore, the gap in the confidences between these directions is small in most cases. This leads us to hypothesise that the models behave best when enough information is provided rather than in ambiguous conditions. However, in real world situations it is difficult to maintain this ideal scenario (Min et al., 2020).

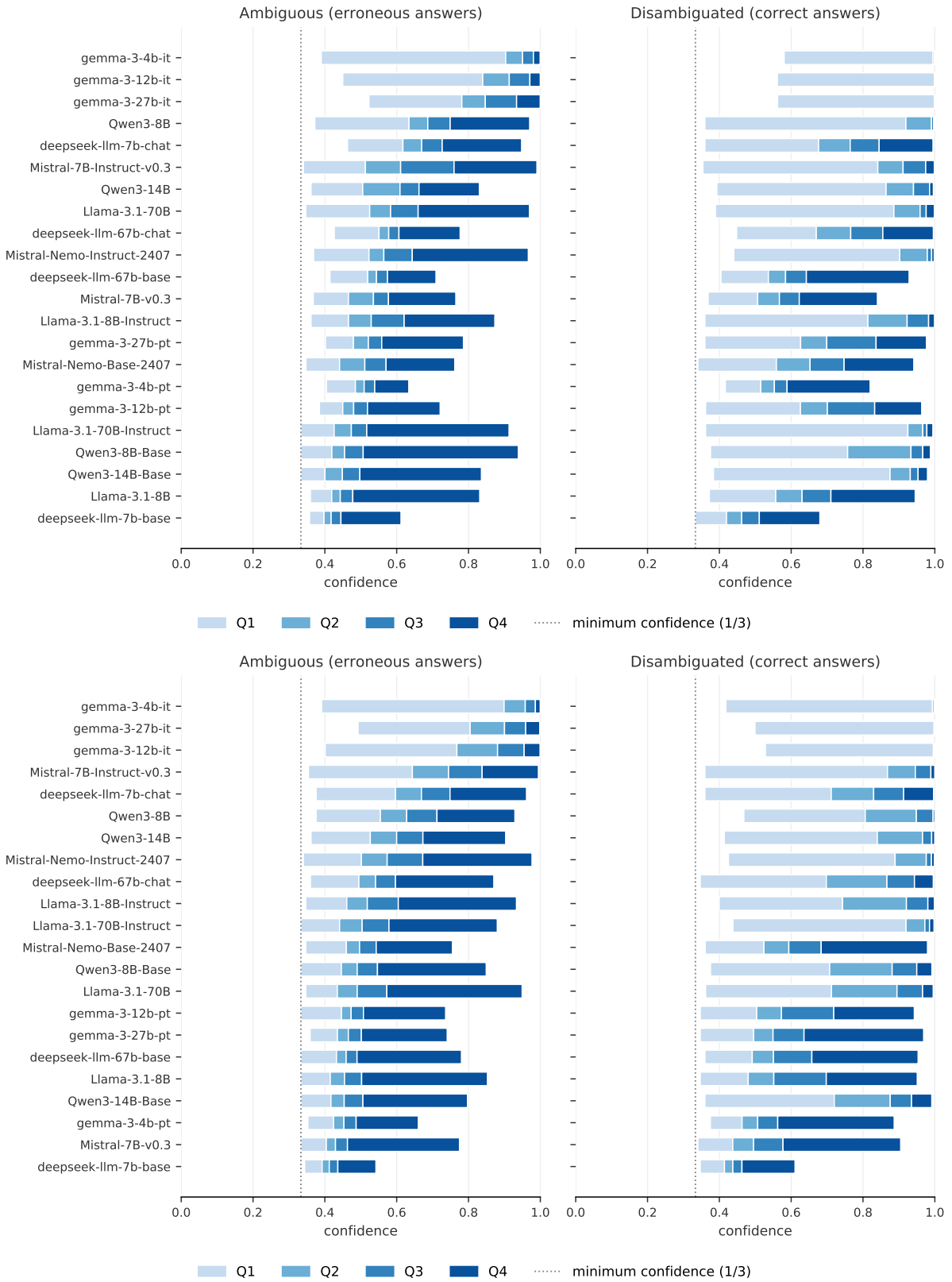


Figure 1: Confidence ranges of four quartiles per-model on GG-BBQ (top) and BBQ (bottom), measured over erroneous answers in ambiguous contexts (left) and correct answers in disambiguated contexts (right), as in Table 3. The dotted line marks $1/3$, the lowest possible confidence for a chosen answer. Left and right plots are not directly comparable, as they are computed over answers of different correctness.

Ethical and societal implications When models are deployed, there is usually no reference answer available and therefore, the model’s confidence serves as a reliability signal for downstream applications (Geng et al., 2024). In our work, the point of interest is not calibration, i.e., we do not ask whether confidence levels are correct overall, but whether confidence is distributed neutrally with respect to social stereotypes. Our quartile analysis (Table 3, Figure 1) shows that this is not always the case: The answers a model chooses most confidently are on average slightly more stereotype-aligned than the least confident ones. This signals that high confidence alone could be less reliable than one might expect. Applications that filter or rank based on confidence can favour stereotype-aligned content even when overall accuracy looks unbiased. For the German language, this risk is compounded by our observation that models refrain from answering less often and choose unsupported answers even under ambiguity far more often than in English. This also suggests that evaluations performed only in English would underestimate this behaviour for German content.

In our work, positive impact does not mean demonstrating that models are biased but changing the perspective of measuring bias. BBQ-style evaluations are already used for auditing in the technical reports and system cards accompanying model releases (Gemma-Team et al., 2024; OpenAI et al., 2024; Anthropic, 2024). All quantities we analyse are computed from log-likelihoods that standard evaluation frameworks produce. How one could meaningfully report such confidence-based measures alongside bias scores typically computed from discrete answer choices is a question for future work.

6 Conclusion

In this work, we investigated the relation of confidence scores and social stereotype alignment. We evaluated 22 open-weight models on gender-bias benchmarks available in German and English and measured not only whether the models answer stereotype relevant questions correctly but also how confidence varies with their answers. Across both languages, for certain models, we find that committal errors under ambiguity are made more confidently the more they lean toward the stereotype. In disambiguated contexts the corresponding effect is minimal in magnitude yet positive in 21 of 22

models in both languages. Neither of these patterns is visible to bias scores typically computed from discrete answer choices alone. We recommend that bias evaluations report confidence-based diagnostics alongside such scores and use order-balanced evaluation, since per-model results could be subject to position bias under a single answer order. Extending the analysis to other bias categories, languages and to open-ended generation settings remains for future work.

Limitations

The models chosen for evaluation in this paper are 11 open-weight model pairs from five different providers and the results need not extend to other model architectures. This also bounds our statistical claims, in particular, the sign tests treat these 22 related models as independent samples. Our analysis also has a representational limitation that the paired disambiguated analysis focuses on male/female stereotypes and items involving trans and non-binary persons are not captured yet. Additionally, the German and English datasets further differ in construction since GG-BBQ is a translated corpus that is partially adapted to cultural differences. Therefore, the differences we observe between languages may partly reflect the differences in the datasets. While log probabilities are arguably a naïve approach and more sophisticated methods exist (Geng et al., 2024), they remain among the most commonly used in practice and our intention with the evaluation was to reflect realistic scenarios. All our results are based on this single estimator, and whether they hold for other confidence measures remains to be tested. For the current evaluation, we adopt a setting with a finite, small set of answers, which is what lets us read the confidence of the answer options and take the stereotype direction from the construction of the item. In open-ended generation neither is given: confidence would have to be estimated from the generated text (Geng et al., 2024) and its stereotype direction assigned by a classifier or manual annotation, which brings its own challenges, such as ensuring the quality of manual and/or automatic LLM-as-a-judge assessments.

Acknowledgments

This research has been funded by the Federal Ministry of Education and research of Germany and the state of North Rhine-Westphalia as part of the

Lamarr Institute for Machine Learning and Artificial Intelligence.

The authors used Generative AI tools to improve sentence clarity and grammar in all sections and to generate initial Python code for data cleaning and data analysis. The authors tested all generated code, corrected errors, and verified that the final scripts reproduced the reported results. AI was not used to make final analytical decisions or interpret study findings.

References

- Anthropic. 2024. [The Claude 3 model family: Opus, Sonnet, Haiku](#). Model card.
- Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, and Gianluca Moro. 2026. [LLMs \(almost\) never abstain under medical uncertainty](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29573–29613, San Diego, California, United States. Association for Computational Linguistics.
- Kate Crawford. 2017. The trouble with bias. NeurIPS Invited Talk.
- Farhan Farsi, Shayan Bali, Fatemeh Valeh, Parsa Ghofrani, Alireza Pakniat, Kian Kashfipour, and Amir H. Payberah. 2025. [PBBQ: A Persian bias benchmark dataset curated with human-AI collaboration for large language models](#). *Preprint*, arXiv:2510.19616.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [A Framework for Few-shot Language Model Evaluation](#).
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Gemma-Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024. [A survey of confidence estimation and calibration in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. [KoBBQ: Korean bias benchmark for question answering](#). *Transactions of the Association for Computational Linguistics*, 12:507–524.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislaw Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Amita Kamath, Robin Jia, and Percy Liang. 2020. [Selective question answering under domain shift](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 245 others. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720. Version of 2026.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson,

Phu Mon Htut, and Samuel R. Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Shalaka Satheesh, Katrin Klug, Katharina Beckh, Héctor Allende-Cid, Sebastian Houben, and Teena Has-san. 2025. [GG-BBQ: German gender bias benchmark for question answering](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 137–148, Vienna, Austria. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs](#). In *The Twelfth International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *The Twelfth International Conference on Learning Representations*.

Muitze Zulaika and Xabier Saralegi. 2025. [BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767, Abu Dhabi, UAE. Association for Computational Linguistics.

A Appendix

A.1 Models Evaluated

Table 5 shows the models that were chosen for evaluation.

A.2 Effect of instruction tuning.

Each of the 11 model families is a base and instruction-tuned pair. For each pair we compute the within-model-family difference in the statistic of each model and test the 11 differences with a two-sided Wilcoxon signed-rank test, which uses the magnitudes of the paired differences. For the

confidence and stereotype correlation coefficient, r , we find that the instruct model exceeds its base counterpart in 9 of 11 families in German (with median $\Delta r = 0.18$, $p = 0.019$). But for English, this is not the case (only 6 of 11 families, with median $\Delta r = +0.01$, $p = 0.58$). Further, for the disambiguated contexts, the gap Δ_{conf} shows no difference between base and instruct variants in either of the languages ($p = 0.97$ German, $p = 0.83$ English).

A.3 Multiple comparisons

We report several p-values, so the chance of a false positive grows. Our claims rest on four sign tests, and Table 4 corrects all four with a Bonferroni correction. Three stay significant; German r does not ($p = 0.068$). The last column shows how many of the per-model tests in Table 2 stay significant after the same correction. We do not draw any conclusions from these per-model tests.

	Effect	positive	sign test p	Bonferroni	per-model significant
de	r	17/22	0.01690	0.06760	15/22
de	Δ_{conf}	21/22	0.00001	0.00004	14/22
en	r	20/22	0.00012	0.00048	12/22
en	Δ_{conf}	21/22	0.00001	0.00004	13/22

Table 4: Sign tests on the direction of the per-model effects. The Bonferroni column corrects all four p-values together. The first column gives the language (de: German, en: English) and the last column gives how many of the 22 per-model tests in Table 2 stay significant after correcting within each language and effect.

Provider	Size	Base model	Instruction-tuned model
Qwen	14B	Qwen/Qwen3-14B-Base	Qwen/Qwen3-14B
Qwen	8B	Qwen/Qwen3-8B-Base	Qwen/Qwen3-8B
DeepSeek	67B	deepseek-ai/deepseek-llm-67b-base	deepseek-ai/deepseek-llm-67b-chat
DeepSeek	7B	deepseek-ai/deepseek-llm-7b-base	deepseek-ai/deepseek-llm-7b-chat
Google	12B	google/gemma-3-12b-pt	google/gemma-3-12b-it
Google	27B	google/gemma-3-27b-pt	google/gemma-3-27b-it
Google	4B	google/gemma-3-4b-pt	google/gemma-3-4b-it
Meta	70B	meta-llama/Llama-3.1-70B	meta-llama/Llama-3.1-70B-Instruct
Meta	8B	meta-llama/Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Mistral AI	12B	mistralai/Mistral-Nemo-Base-2407	mistralai/Mistral-Nemo-Instruct-2407
Mistral AI	7B	mistralai/Mistral-7B-v0.3	mistralai/Mistral-7B-Instruct-v0.3

Table 5: The 22 models chosen for evaluation: 11 base and instruction-tuned pairs from five providers, identified by their Hugging Face model IDs. Each pair is treated as one family in the within-family comparisons.