The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

The Confidence Trap: Gender Bias and Predictive Certainty in LLMs

Ahmed Sabir[1], **Markus Kängsepp**[1], **Rajesh Sharma**[1,2]

[1]Institute of Computer Science, University of Tartu, Estonia
[2]School of AI and CS, Plaksha University, India

Abstract

The increased use of Large Language Models (LLMs) in sensitive domains leads to growing interest in how their confidence scores correspond to fairness and bias. This study examines the alignment between LLM-predicted confidence and human-annotated bias judgments. Focusing on gender bias, the research investigates probability confidence calibration in contexts involving gendered pronoun resolution. The goal is to evaluate if calibration metrics based on predicted confidence scores effectively capture fairness-related disparities in LLMs. The results show that, among the six state-of-the-art models, Gemma-2 demonstrates the worst calibration according to the gender bias benchmark. The primary contribution of this work is a fairness-aware evaluation of LLMs' confidence calibration, offering guidance for ethical deployment. In addition, we introduce a new calibration metric, Gender-ECE, designed to measure gender disparities in resolution tasks.

Introduction

Large pretrained LLMs have demonstrated exceptional performance across a variety of tasks, yet their trustworthiness remains a pressing concern, especially regarding gender bias. These models not only inherit biases from their training data but can also amplify stereotypes, leading to unfair or inaccurate outputs in critical applications (Zhao et al. 2018; Rudinger et al. 2018; Nangia et al. 2020; Felkner et al. 2023; Gallegos et al. 2024; Sabir and Sharma 2025). The challenge lies not only in detecting bias but also in ensuring that users can reliably interpret and trust model predictions, particularly in high-stakes domains such as hiring, healthcare, and legal decision-making.

Calibration, which ensures that an LLM's confidence in its predictions accurately reflects the likelihood of being correct (Niculescu-Mizil and Caruana 2005; Si et al. 2022), is a crucial aspect of model reliability. Without proper calibration, an LLM might provide overconfident yet incorrect outputs, misleading users into trusting biased or unreliable responses. This issue becomes even more critical in cases where models exhibit skewed confidence distributions across different demographic groups, further exacerbating fairness concerns. Therefore, in this work, we aim to answer the question regarding calibration in a gender bias scenario. Here, calibration refers to how well a model's probability estimates match reality. For instance, when a model assigns an 80% probability to its prediction (*e.g.* for the pronoun 'he'), then, ideally, for those 80% confidence predictions, the model should be correct approximately 80% of the time.

The importance of calibration is particularly relevant in gender bias tasks. In many real-world settings, the objective is not only to measure which pronoun best matches human bias but also to assess the model's confidence in its predictions. Poor calibration means that the model might generate seemingly "confident" predictions even when it is frequently incorrect, posing a significant risk when deploying these models in production or high-stakes scenarios. In addition, advancing the understanding of fairness in the context of misrepresented groups contributes to a more comprehensive evaluation of model behavior (Ding et al. 2024).

While many studies have explored the biases and stereotypes propagated by LLMs (Gallegos et al. 2024), little to no research has examined how well these models are calibrated in their predictions or how closely they align with human biases. In this work, we investigate these gaps by analyzing the confidence calibration of LLMs in gendered and gender-neutral contexts. Specifically, inspired by model calibration evaluation methods such as Expected Calibration Error (ECE) (Guo et al. 2017), which help evaluate the model's predicted confidence with its actual outcomes, we seek to answer the following question: To what extent is the predictive confidence of LLMs calibrated in gendered pronoun resolution tasks?

To summarize our contributions: (1) we conduct a comprehensive analysis of widely used open-weight LLMs, offering practical guidance for their ethical deployment, and (2) we propose Gender-ECE, a novel adaptation of ECE designed to quantify gender-related calibration disparities in coreference resolution tasks.

Related Work

Gender Bias in LLM. Large language models have demonstrated remarkable capabilities in natural language understanding and generation. However, they also inherit and amplify societal biases, including gender bias. To measure Gender bias in downstream tasks in NLP applications, a variety of research efforts have investigated gender bias using tem-

---

plates based on structured sentences such as "He/She is a/an [occupation]," in which blanks are filled with occupations that convey either positive or negative stereotypes or associations (Stańczak and Augenstein 2021).

Additional research has been conducted using the format of Winograd Schemas (Levesque, Davis, and Morgenstern 2012). This method has been applied to various datasets such as WinoBias (Zhao et al. 2018), Winogender (Rudinger et al. 2018), and WinoMT (Stanovsky, Smith, and Zettlemoyer 2019). The Winograd Schema Challenge focuses on the task of coreference resolution, which requires commonsense reasoning. This challenge has been used to examine whether the interpretation of pronoun references in sentences is influenced by gender. This method has also been applied to evaluate gender-based stereotypes and neutral associations across various professions (Zhao et al. 2018; Rudinger et al. 2018; Sabir and Sharma 2025). More recently, most research has shifted beyond template-based approaches to address gender bias by evaluating the outputs of LLMs, which can encode and propagate societal stereotypes across various contexts, including personality traits, social roles, and cultural representations. For example, the work of Cheng *et al.* (Cheng, Durmus, and Jurafsky 2023) employs a method to assess stereotypes in LLMs by generating persona descriptions for various demographic groups. Extending this direction, prompt-driven frameworks such as RUTEd (Lum et al. 2025) measure bias by instructing LLMs to perform in-context tasks *e.g.* writing stories or adopting personas, shifting evaluation from short template-based sentences to longer-form generations that better surface bias.

However, the most dominant method to measure bias in LLMs is the use of template-based methods, where controlled sentence structures allow for systematic evaluation of gender bias in model predictions. These templates help isolate bias by ensuring that the only difference between sentence variations is the gender-related attribute while keeping all other linguistic components identical. In this study, we adopt a template-based approach.

Calibration in LLMs. Large language models are stated to be rather well-calibrated (Kadavath et al. 2022). Mostly, the studies involve using question-answer datasets with multiple answers (Krause et al. 2023; Kapoor et al. 2024; Yoon et al. 2025). There are many ways to check how well the model is calibrated, for example, using (1) prompting, (2) model token probabilities, and (3) training a model on top of the model's internal state or output. Firstly, many works on prompting (Kadavath et al. 2022; Xiong et al. 2023; Tian et al. 2023; Yoon et al. 2025) have relied solely on the model's ability, whether a LLM or a reasoning model, to convey confidence based on their answers or chain of thought. For example, Kadavath *et al.* (Kadavath et al. 2022) query a model about its answer confidence $P(true)$ after the model has answered a question in a question-answering setting. Secondly, token probabilities are a straightforward way to determine the model's uncertainty for each token. However, these probabilities might not be well-calibrated, for example, for low-resource languages (Krause et al. 2023) in the case of multilingual models. Also, having bigger or pretrained and fine-tuned models does not grant better calibrated models (Chen et al. 2022). Lastly, training an extra model on top of the LLM to assess the model's confidence more precisely. For example, (Kadavath et al. 2022) proposes to train models with an extra head to predict the probability that a model knows $P(IK)$ the answer. More recently, the work of (Kapoor et al. 2024) proposes a fine-tuning protocol to calibrate model confidences in question-answer tasks.

In this work, we are interested in token probabilities, more precisely, how well a model's probabilities for gender bias align with human bias. Thus, extracting the probability for pronouns and using simple post-hoc calibration methods suffices for our analysis.

Datasets

WinoBias (Zhao et al. 2018). WinoBias is a benchmark dataset for evaluating gender bias in LLMs. The dataset contains 3,160 sentences using templates inspired by Winograd-schema style sentences (Levesque, Davis, and Morgenstern 2012) with entities corresponding to people referred by their occupation (*e.g.* doctor). For each sentence, there are three variables person, occupation, and pronoun. The dataset is structured to test whether LLMs make gendered assumptions about professions by evaluating coreference resolution performance, *e.g.* "The developer argued with the designer and slapped her in the face".

Winogender (Rudinger et al. 2018). Winogender is a dataset similar to WinoBias, which also includes template sentences that refer to occupation and person. The dataset contains 720 template-based sentences. Unlike WinoBias, Winogender introduced a second entity to the gendered occupation, using gender neutral `someone` to avoid gender stereotype association. Each sentence in Winogender is carefully designed to evaluate coreference resolution accuracy in a way that minimizes unintended gender stereotypes.

GenderLex (Sabir and Sharma 2025). A similar template-based dataset to WinoBias and Winogender contains pronouns and biased occupations but also includes nouns and verbs directly related to the pronouns. This dataset uses ChatGPT 3.5-turbo (OpenAI 2022) to generate 1,676 sentence pairs (838 unique sentences), followed by human annotation and correction. However, unlike previous datasets, this template relies on the `last cloze` structure, *e.g.* "The lawyer stated that the legal case was won by him". This approach, adding the pronoun at the end, mimics how humans and LLMs generate text, postpones gender clues, and therefore provides a strong method to probe gender bias in LLMs.

Methods

The main idea of this work is to understand whether the model's LLM confidence (probability estimates) aligns with model prediction and its human alignment (*i.e.* human bias). In other words, we aim to evaluate the extent to which the model's self-assessed confidence (*i.e.* token probability) reflects both the correctness of predictions and the degree of correlation between these confidence levels and human judgment or biases, within the context of gender bias research. By examining this alignment, we can assess the extent to which the model's probability estimates serve as reliable indicators of actual performance in detecting, mitigating, or perpetuating gender bias.

To do this, for a given sentence with a pronoun, we extract the probability confidence of the model assigning to different pronouns. For example, in "The surgeon called the nurse and instructed [her/him] to prepare for surgery.", the "her" pronoun is more commonly expected here, which reflects occupational gender bias. However, we also evaluate the model's confidence when replacing the pronoun with "him" and check how confident the model is in its prediction. The idea is to evaluate the model's confidence by examining predictions involving a given pronoun and a stereotypically biased occupation (*e.g.* nurse). Therefore, the aim is to investigate the model's certainty in making biased predictions.

Let a sentence S contain T words, $S = (w_1, w_2, ..., w_T)$, If a pronoun w_p in the given sentence occurs at position k (*i.e.* $w_k = w_p$), the probability is computed as follows:

$$P(w_p \mid w_1, ..., w_{k-1}) = \frac{e^{z_{k-1,w_p}}}{\sum_{j=1}^{V} e^{z_{k-1,j}}}, \quad (1)$$

where z_{k-1,w_p} is the model logit score for the pronoun w_p at the position k. After extracting the probability, we investigate model's confidence using different calibration metrics:

Expected Calibration Error (ECE). ECE (Naeini, Cooper, and Hauskrecht 2015) measures how well the model-predicted confidence matches with the actual outcome and accuracy of the model. The model predictions are divided into M bins B_m. Bins can be divided into equal-width bins (each bin takes an equal amount of probability space) or equal-size bins (each bin has an equal number of instances). For each bin, the average confidence $\text{conf}(B_m)$ and average accuracy $\text{acc}(B_m)$ are calculated, and ECE is the average absolute bin-wise difference between average confidence and average accuracy. More precisely,

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|, \quad (2)$$

where n is the total number of instances. The smaller the ECE score, the more trustworthy the model's confidence (*i.e.* confidences reflect reality).

Reliability Diagrams. Reliability diagrams (Murphy and Winkler 1977) are a visualization method based on ECE (*e.g.* Figure 1). Each bin B_m is plotted on the y-axis based on the confidence. The blue bar shows the average accuracy of a bin $\text{acc}(B_m)$, visible on the x-axis. The red gap shows the difference between the average accuracy $\text{acc}(B_m)$ and average confidence $\text{conf}(B_m)$. Indicating the expected calibration error in the bin B_m. Ideally, the blue bars should be aligned with the dashed diagonal line to have a well-calibrated model (perfect calibration).

Instance Calibration Error (ICE). ICE (Si et al. 2022) measures the mean absolute difference between model confidence $\hat{p}_i$ and true label y_i for each instance i.

$$\text{ICE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{p}_i \right|. \quad (3)$$

Macro Calibration Error (MacroCE). MacroCE (Si et al. 2022) is similar to ICE. Except that it divides the instances into two subgroups, correct and incorrect predictions, and ICE is separately calculated for these subgroups. Finally, the average of ICE_{pos} and ICE_{neg} is taken. Formally:

$$\begin{aligned} \text{MacroCE} &= \frac{1}{2}(\text{ICE}_{pos} + \text{ICE}_{neg}), \\ \text{ICE}_{pos} &= \frac{1}{n_p} \sum_{i=1}^{n_p} (1 - \hat{p}_i), \forall \hat{y}_i = y_i, \\ \text{ICE}_{neg} &= \frac{1}{n_n} \sum_{i=1}^{n_n} (\hat{p}_i - 0), \forall \hat{y}_i \neq y_i, \end{aligned} \quad (4)$$

where n_p and n_n are number of positive and negative instances, $\hat{p}_i$ and $\hat{y}_i$ are predicted probability and predicted class-label of ith instance.

Brier Score. Brier score (Brier 1950) is a proper scoring rule that measures the mean squared difference between model predicted confidence $\hat{p}_i$ and true label y_i for each instance i. Brier score is commonly used as a loss measure for training machine learning models. The Brier score is defined as:

$$\text{Brier Score} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i)^2. \quad (5)$$

Gender-Aware ECE. All previous metrics for evaluating calibration in LLMs focus on comparing two gendered sentence completions, one with a male pronoun and one with a female pronoun. However, overall calibration, such as ECE, does not reveal how the model behaves differently with respect to male versus female pronouns. To address this gap, we propose Gender-Aware Group-ECE, a metric designed to capture calibration disparities across gendered pronouns.

$$\text{Gender-ECE} = \frac{1}{2} \left(\text{ECE}_{\text{male}} + \text{ECE}_{\text{female}} \right) \quad (6)$$

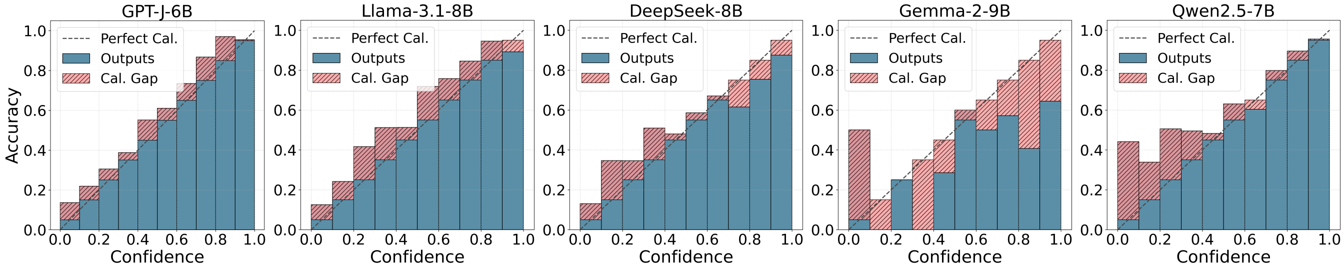
$$\text{ECE}_{\text{male}} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|, \quad \forall \hat{y}_i = 1,$$

$$\text{ECE}_{\text{female}} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|, \quad \forall \hat{y}_i = 0.$$

where $\hat{y}_i$ is predicted class, and [1/0] correspond to male/female label. Gender-ECE is conceptually inspired by MacroCE, however, in MacroCE the subdivision into groups (positive and negative) is based on whether prediction was correct or incorrect. Therefore, in the positive group are all the instances where model prediction is aligned with human bias (true label). On the other hand, Gender-ECE divides data into two groups (*i.e.* male and female) based on the predicted label, not relying on the true label for the division. By doing this, it shows how well-calibrated the model's scores are compared to the human bias for gender pronouns. In addition, MacroCE is calculated instance-wise, rather than bin-wise, like ECE and our proposed Gender-ECE. Instance-wise calibration error measuring leads to less reliable results, as it introduces more noise and instability compared to bin-wise measuring, which is more interpretable and provides an understanding of model confidence and calibration behavior.

Model	Standard Calibration Metrics				Gender-ECE			Human
	ECE	MacroCE	ICE	Brier Score	Group	M	F	
GPT-J-6B (Wang and Komatsuzaki 2021)	0.076	0.453	0.374	0.432	0.076	0.085	0.066	0.715
Llama-3.1-8B (AI@Meta 2024)	0.111	0.466	0.371	0.446	0.111	0.112	0.109	**0.727**
Gemma-2-9B (Team et al. 2024)	**0.327**	**0.493**	0.390	**0.559**	**0.267**	**0.330**	0.204	0.617
Qwen2.5-7B (Yang et al. 2024)	0.106	0.476	0.422	0.385	0.107	0.052	0.162	0.637
Falcon-3-7B (Team 2024)	0.161	0.491	**0.449**	0.356	0.149	0.081	**0.217**	0.605
DeepSeek-8B (DeepSeek-AI 2025)	0.085	0.461	0.369	0.470	0.090	0.074	0.106	0.686

Table 1: Comparison of calibration metrics across models on the GenderLex dataset. GPT-J-6B shows the best calibration (lowest ECE and balanced Gender-ECE). Gemma-2-9B performs the worst overall. **Bold** = highest (worst calibration and best human alignment), Underlined = lowest (best calibration and worst human alignment).



(a) ECE: 0.076 (b) ECE: 0.111 (c) ECE: 0.085 (d) ECE: **0.327** (e) ECE: 0.106

Figure 1: Reliability diagrams for the GenderLex dataset with the gender appearing at the end of the sentence having all context `last cloze` structure. Gemma-2-9B exhibits the **highest** ECE while GPT-J-6B is the most calibrated model.

Experimental Result

To answer our research questions and validate our findings, we experiment with a variety of gender bias benchmark datasets: WinoBias, Winogender, and GenderLex, In each dataset, there are two sentence pairs with a different pronoun. For example, in the WinoBias dataset: "The developer argued with the designer and slapped [him/her] in the face". Each sentence in the dataset is used directly as the input, and we perform a deterministic forward pass (w/o decoding parameters involved) and utilize offset mappings for precise token alignment of the pronoun, ensuring reproducibility.

Human Alignment. Each sentence is assigned a human-labeled bias score, indicating which sentence is more biased: "1" for male bias and "0" for female bias. Human alignment was conducted by three human subjects and one expert annotator, with final decisions made through majority voting. The expert also assessed the quality of the annotations. For example, if a human subject was unfamiliar with a particular occupation (*e.g.* demographer), they would consult the expert annotator to avoid making a random annotation. Notably, the Genderlex dataset showed the most annotation disagreement, likely due to many rare occupation terms, yielding moderate inter-annotator agreement (avg pairwise Cohen's $\kappa = 0.51$) (Cohen 1960). Overall, this process helps reduce noise from guessing or confusion in the final labels.

Model. We evaluate six state-of-the-art open-weight LLMs: GPT-J-6B (Wang and Komatsuzaki 2021), Falcon-3-7B-base (Team 2024), Llama-3.1-8B (AI@Meta 2024), Gemma-2-9B (Team et al. 2024), Qwen2.5-7B (Yang et al. 2024) and Llama-based DeepSeek-R1-8B or DeepSeek-8B (Distill from Llama-3.1-8B) (DeepSeek-AI 2025).

RQ1: How well do LLMs calibrate their confidence in predicting pronouns at the end of the sequence?

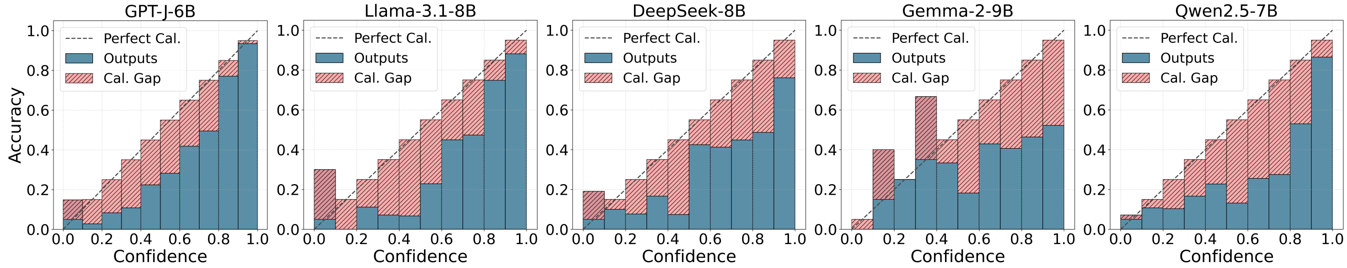
In this research question, the primary focus of the analysis is to assess the models' bias and confidence in predicting pronouns at the end of sentences, GenderLex dataset (`last cloze`), meaning that the model has access to the full context of the sentence before scoring a biased pronoun.

Table 1 compares gender bias calibration metrics across language models. Gemma-2-9B exhibits the highest overall miscalibration and a notable disparity between female- and male-associated pronouns. GPT-J-6B demonstrates the strongest calibration and fairness, with balanced performance across genders. Llama-3.1-8B achieves the best alignment with human bias judgments, showing nearly identical calibration across genders. Falcon-3-7B provides the most accurate probabilistic predictions (lowest Brier Score) but shows a large female–male gap. Qwen2.5-7B is well-calibrated for male pronouns, while DeepSeek-8B maintains a lower calibration error compared to Llama-3.1-8B.

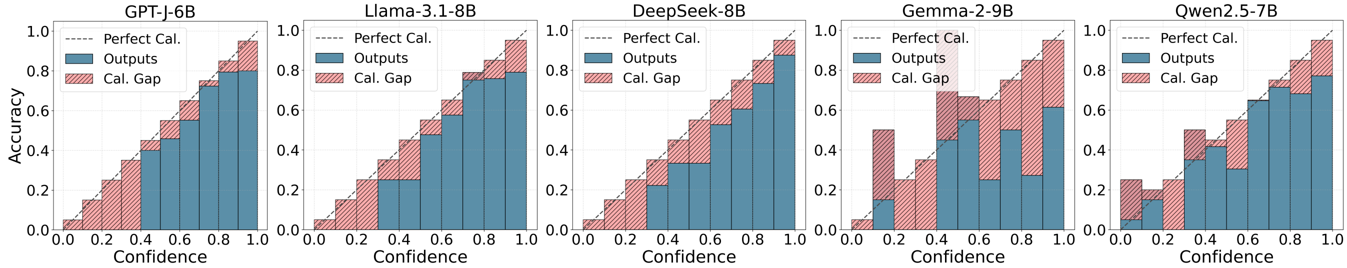
Finding 1 As shown in reliability diagrams (Figure 1), GPT-J-6B shows the best calibration (lowest ECE), with relatively small Gender-ECE gap differences. Gemma-2-9B performs worst overall, being consistently incorrect and high risk for deployment. Llama-3.1-8B is the most consistently human-aligned, while Falcon-3-7B, though under-calibrated, achieves the most accurate probabilities (lowest Brier) but exhibits a wider gender gap.

Model	WinoBias								Winogender							
	Standard Metrics				Gender-ECE			Human	Standard Metrics				Gender-ECE			Human
	ECE	MacroCE	ICE	Brier	Group	M	F		ECE	MacroCE	ICE	Brier	Group	M	F	
GPT-J-6B	0.157	0.444	0.356	0.481	0.164	0.150	0.179	**0.686**	0.086	0.473	0.400	0.406	0.118	0.066	0.170	0.685
Llama-3.1-8B	0.193	0.460	0.377	0.460	0.214	0.179	**0.249**	0.662	0.099	0.475	0.387	0.428	0.138	0.076	0.200	**0.707**
Gemma-2-9B	**0.429**	**0.490**	**0.482**	0.467	**0.297**	**0.438**	0.156	0.509	**0.373**	**0.486**	**0.422**	**0.533**	**0.396**	**0.372**	**0.421**	0.573
Qwen2.5-7B	0.234	0.442	0.362	**0.510**	0.190	0.259	0.121	0.630	0.136	0.461	0.379	0.463	0.129	0.139	0.119	0.657
Falcon3-7B	0.154	0.452	0.357	0.487	0.149	0.160	0.138	0.684	0.112	0.474	0.404	0.392	0.176	0.079	0.273	0.696
DeepSeek-8B	0.240	0.478	0.382	0.496	0.218	0.255	0.182	0.648	0.131	0.470	0.380	0.453	0.135	0.129	0.141	0.679

Table 2: WinoBias and Winogender benchmark datasets calibration evaluation using different metrics across different models. Gemma-2-9B shows the **worst** calibration overall, and GPT-J-6B is the more calibrated model.



(a) ECE: 0.157 (b) ECE: 0.193 (c) ECE: 0.240 (d) ECE: **0.429** (e) ECE: 0.234



(f) ECE: 0.086 (g) ECE: 0.099 (h) ECE: 0.131 (i) ECE: **0.373** (j) ECE: 0.136

Figure 2: (Top) Reliability diagrams for WinoBias. (Bottom) Reliability diagrams for Winogender. GPT-J-6B shows the most calibrated model performance, while Gemma-2-9B demonstrates the **worst** human-alignment and calibration.

RQ2: How well are LLMs calibrated when resolving pronouns in gender-biased coreference tasks?

In this experiment, we rely on two benchmark datasets designed to evaluate gender bias in coreference resolution.

WinoBias Dataset. This dataset is designed to evaluate gender bias in language models, particularly in coreference resolution tasks in which the model predicts pronouns in the middle of sentences rather than at the end. We employ the WinoBias-syntax (type-2) dataset, which consists of sentence pairs where answers can be resolved using syntactic information alone (without grounding or world knowledge), comprising 1,542 sentence pairs (771 unique sentences, each containing a single pronoun). We replace occupational titles with a gender-neutral term `person` (*e.g.* "The ~~developer~~ person argued with the [designer] and slapped [her] in the face") to reduce ambiguity in pronoun prediction, when two occupational entities co-occur in a sentence.

Table 2 (Left) shows that Gemma-2-9B has the worst calibration. In contrast, Falcon-3-7B and GPT-J-6B achieve the best calibration among the models evaluated. For gender calibration, Gemma-2-9B shows the largest gap, with much worse calibration for male pronouns. Llama-3.1-8B exhibits the highest miscalibration for female pronouns. Falcon-3-7B and GPT-J-6B again show the lowest Gender-ECE scores, reflecting a better balance across genders. Human alignment scores diverge notably for Gemma-2-9B, indicating weaker alignment with human bias and a higher risk of biased behavior toward the male group.

Finding 2 Figure 2 (`a` to `e`) show reliability diagrams for WinoBias dataset. Gemma-2-9B exhibits the worst calibration overall and a pronounced bias toward male pronouns, as indicated by its high Gender-ECE (male) score. These patterns indicate that Gemma-2-9B is both overconfident and inconsistent in gender-sensitive contexts. GPT-J-6B and Falcon3-7B are the fairest models, with the lowest Gender-ECE and minimal differences between pronouns.

Model	WinoBias		GenderLex	
	Male	Female	Male	Female
GPT-J-6B	0.206	0.508	0.373	0.377
Llama-3.1-8B	0.197	0.559	0.396	0.333
Gemma-2-9B	0.067	0.895	0.056	0.901
Qwen2.5-7B	0.130	0.596	0.426	0.416
Falcon3-7B	0.215	0.502	0.505	0.363
DeepSeek-8B	0.158	0.606	0.303	0.469

Table 3: ECE for male and female pronouns in the WinoBias and GenderLex datasets. Most models exhibit worse calibration for female pronouns, with Gemma-2-9B showing an extreme disparity, having the lowest ECE for male and the highest for female pronouns.

Model	WinoQueer			
	Gay	Lesbian	Trans	Queer
GPT-J-6B	0.121	0.790	0.816	0.700
Llama-3.1-8B	0.083	0.730	0.760	0.657
Gemma-2-9B	0.026	0.221	0.586	0.182
Qwen2.5-7B	0.189	0.898	0.919	0.788
Falcon3-7B	0.059	0.715	0.686	0.432
DeepSeek-8B	0.277	0.838	0.258	0.910

Table 4: ECE scores on WinoQueer dataset show that Qwen2.5-7B has the highest calibration errors in most categories, while Gemma-2-9B is the most calibrated overall.

Model	Metric	Someone	Person	Occ
GPT-J-6B	ECE	0.144	0.063	0.076
	GECE	0.138	0.077	0.076
Llama-3.1-8B	ECE	0.134	0.138	0.111
	GECE	0.132	0.130	0.111
Gemma-2-9B	ECE	0.364	0.367	0.327
	GECE	0.450	0.351	0.267
DeepSeek-8B	ECE	0.139	0.130	0.085
	GECE	0.138	0.137	0.090

Table 5: Gender-ECE (Group) and ECE scores on the GenderLex dataset show that models exhibit high calibration error with gender-neutral terms: `Someone` and `Person`.

Winogender Dataset. Unlike WinoBias dataset, which tests stereotypical occupational roles, Winogender evaluates gender preference in pronoun resolution without an inherent gender stereotype. *e.g.* "The paramedic helped the passenger because [he/she] was injured."

Table 2 (Right) shows that Gemma-2-9B has the weakest calibration, with the highest Gender-ECE and the lowest human alignment. GPT-J-6B achieves the lowest overall ECE but shows higher calibration error for female pronouns. Llama-3.1-8B shows moderate calibration with a higher Gender-ECE female error, yet the highest agreement with human judgments. Qwen2.5-7B exhibits balanced calibration with moderate gender gaps, while Falcon-3-7B shows moderate ECE but a pronounced female–male gap.

Finding 3 Figure 2 (`f–j`) shows reliability diagrams for the Winogender dataset. Gemma-2-9B exhibits the worst calibration, with the highest ECE and Gender-ECE, indicating poor handling of both male and female pronouns. GPT-J-6B achieves the lowest overall ECE and Gender-ECE but shows substantially higher error for female pronouns, indicating a notable gender imbalance in calibration.

RQ3: How well-calibrated are LLMs in predicting male and female gender-specific pronouns?

In this experiment, we examine model confidence in pronoun prediction for gender-biased occupations. We use human-labeled data to split the WinoBias and GenderLex datasets into male and female subsets.

Table 3 shows that most models have higher ECE (*i.e.* worse calibration) for female pronouns, suggesting that LLMs are generally less reliable when handling female pronouns. In particular, Gemma-2-9B achieves low ECE for male pronouns 7% in the WinoBias dataset, but extremely high ECE for female pronouns 93%, indicating a significant calibration gap between male and female references. A similar pattern is observed in GenderLex, where the male ECE is 6% and the female ECE is 94%. Moreover, GenderLex dataset exhibits higher ECE scores for male pronouns compared to the WinoBias dataset, demonstrating substantial miscalibration when pronouns are cloze last.

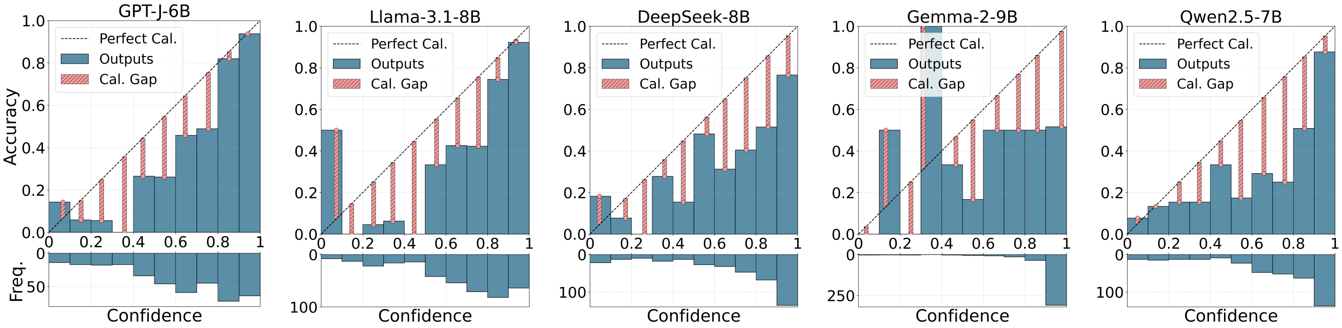
Finding 4 Most LLMs are less calibrated on female pronouns, revealing bias toward male references. Gemma-2-9B is well-calibrated for male pronouns but poorly calibrated for female, exhibiting a clear gender-specific bias.

RQ4: How well are LLMs calibrated on predictions about marginalized groups (*i.e.* LGBTQ community)?

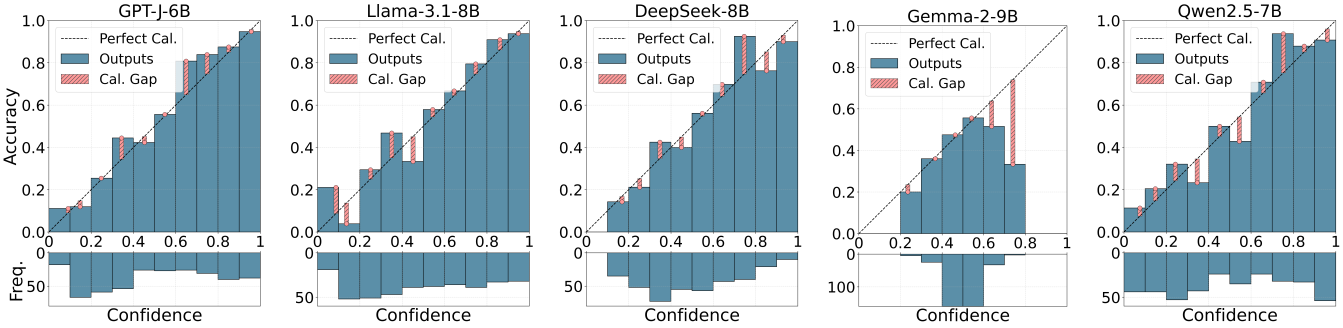
In this experiment, we use the WinoQueer dataset (Felkner et al. 2023). A Winograd-inspired benchmark designed to measure anti-LGBTQ+ biases in LLMs. From the 45,540 available sentences, we extract only 1,488 sentence pairs that, similar to GenderLex, end with a `last cloze` (*e.g.* "everyone hates Wren because they are [LGBTQ]").

Table 4 presents ECE scores for the WinoQueer dataset. Overall, the majority of models exhibit significantly greater miscalibration for Trans and Lesbian identity mentions, indicating increased bias and lower reliability when handling these identity groups. Interestingly, Qwen2.5-7B exhibits the highest ECE overall, particularly for Trans and Lesbian identities, revealing significant calibration issues. DeepSeek-8B also shows poor calibration for Queer identities. In contrast, Gemma-2-9B stands out as the most well-calibrated model in this setting, achieving the lowest ECE for Gay and Queer identities and relatively low scores overall. This suggests that Gemma-2-9B is better at maintaining confidence consistency across queer identity mentions. Meanwhile, GPT-J-6B and Falcon3-7B also perform poorly, especially on Lesbian and Trans instances, further highlighting the challenges these models face in equitably handling LGBTQ+ representations.

Finding 5 All Models exhibit the highest miscalibration for Lesbian and Trans identities in the WinoQueer dataset, with Qwen2.5-7B showing the worst overall.



(a) ECE: 0.148 (b) ECE: 0.196 (c) ECE: 0.235 (d) ECE: 0.429 (e) ECE: 0.234



(f) ECE: 0.045 (g) ECE: 0.067 (h) ECE: 0.062 (i) ECE: 0.025 (j) ECE: 0.076

Figure 3: Comparison of uncalibrated (Top) and Beta calibrated (Bottom) gender bias probabilities for WinoBias dataset. Instances are divided into 10 equal-width bins. The red bar shows the distance to perfectly calibrated probabilities in the bin, and the blue bar shows the accuracy. The histogram below the reliability diagram shows how many instances are in each bin.

Model	**ECE**	**Gender-ECE**		
		Group	**M**	**F**
Llama-3.1-8B	0.193	0.214	0.179	0.249
Llama-3.1-70B	0.216$_{+11.9\%\uparrow}$	0.230$_{+7.5\%\uparrow}$	0.207$_{+15.6\%\uparrow}$	0.253$_{+1.6\%\uparrow}$
Gemma-2-9B	0.429	0.297	0.438	0.156
Gemma-2-27B	0.366$_{-14.7\%\downarrow}$	0.361$_{+21.5\%\uparrow}$	0.341$_{-22.1\%\downarrow}$	0.381$_{+144.2\%\uparrow}$

Table 6: ECE and Gender-ECE on the WinoBias dataset. Red values indicate the % change in ECE score when increasing the model size within the same family.

Discussion

The main finding of this study is that Gemma-2 is the least calibrated model across genders, with larger calibration errors for female entities and worse human alignment. Results from all datasets indicate the model is biased toward male pronouns. Increasing model size improves confidence calibration for male-associated predictions but increases miscalibration for female pronouns, as shown in Table 6.

We selected commonly used models in the community, with DeepSeek chosen to analyze bias propagation and confidence during distillation. GPT-J-6B was included due to its minimally filtered training data, enabling analysis of how gender balanced data augmentation affects model confidence. Results indicate that less filtered models, such as GPT-J-6B, exhibit better calibration, suggesting that data filtering or augmentation, *e.g.* gender swapping (Zhao et al. 2018) or using gender neutral, influence model confidence. For example, using gender-neutral terms results in a less calibrated model, as shown in Table 5 with the GenderLex dataset. Therefore, providing explicit role context, such as the occupational title, reduces uncertainty and improves calibration. In contrast, gender-neutral *i.e.* `someone` and `person` result in the highest calibration error due to their inherent semantic ambiguity. DeepSeek-8B, distilled from Llama-3.1-8B, exhibits an overall higher gender calibration error in coreference tasks and less human alignment compared to the base model overall. This indicates that the distillation process negatively impacts model calibration.[1]

These findings highlight a gap between model confidence and its actual accuracy (*i.e.* human bias) that varies by predicted gender. This confidence–accuracy alignment serves as an indicator of model calibration, reflecting the extent to which the model's confidence corresponds to its true likelihood of being correct. Accordingly, in our analysis, we complement accuracy-based evaluation by examining confidence calibration disaggregated by gender, which reveals confidence-related bias (*e.g.* overconfidence in predictions for one gender despite comparable correctness). This is particularly important in trust-sensitive decision-making applications, *e.g.* AI-assisted resume screening.

---

[1] This effect diminishes in larger models *e.g.* DeepSeek-70B.

Calibration

As shown in the previous experiments, model output probabilities are often not well-calibrated, especially in our case, a deep neural network (Guo et al. 2017). Expected calibration error (ECE) and reliability diagrams can be used to check how well models are calibrated. Therefore, post-hoc calibration methods (*e.g.* Beta calibration (Kull, Silva Filho, and Flach 2017)) are needed, which are used to tune model predicted probabilities to match actual accuracies. For example, let's have 100 predictions with a probability of 80 per cent; in order to have well-calibrated probabilities, about 80 of 100 predictions should be correct. This kind of calibration helps to make the model more reliable and trustworthy.

Analysis of Table 2 and Figure 3 shows that the ECE is relatively high, suggesting poor calibration performance on the WinoBias dataset. In order to make the model predictions about its gender bias more reliable and trustworthy, we employ a simple post-hoc calibration method, Beta calibration (Kull, Silva Filho, and Flach 2017) on top of the data. The data is split into two: 385 validation and 386 test instances. As a result, Beta calibration method is able to alleviate calibration issues with all the models, resulting in about three times lower calibration error (ECE) (Figure 3).

Also, data is more evenly spread across confidences, visible from the frequency histogram below the reliability diagrams. Interestingly, Gemma-2-9B, the model before calibration, is predicting with high confidence towards one or another gender; however, based on human bias, the model should be much less confident about its predictions. As the model tunes its confidence predictions to be around 50 per cent. The model accuracies increase after calibration as follows: GPT-J-6B 69.2% to 76.9%, Llama-3.1-8B 65.8% to 74.9%, Deepseek-8B 63.5% to 69.9%, Gemma-2-9B 51.6% to 54.7%, and Qwen2.5-7B 61.1% to 76.4%. This indicates that post-hoc calibration can also increase the model's accuracy (*i.e.* reflecting human bias) for all models. This is expected, as the post-hoc calibrator is fitted on extra validation data, aligning the model predictions better with human bias.

Although calibration improves the trustworthiness of a model's confidence estimates, particularly their reliability and interpretability, techniques such as Beta calibration do not serve as bias mitigation strategies, as these procedures enhance confidence calibration without addressing underlying sources of bias.

Ablation Study

The effect of sample size on ECE has been studied in prior work on classifier uncertainty calibration (Minderer et al. 2021; Roelofs et al. 2022). To illustrate this effect in our setting, we conduct a small-scale experiment using a dataset of varying sizes and compute the corresponding ECE.

Table 7 reports the mean and standard deviation of ECE for five different sample sizes: 50, 100, 150, 250, 500. For each sample size, we draw 100 subsets without replacement from the full WinoBias dataset, which contains 771 instances. As a result, the standard deviation is substantially greater for smaller sample sizes, indicating increased instability when fewer instances are available.

	Sample size N				
	50	100	150	250	500
ECE (mean)	0.2630	0.2522	0.2429	0.2421	0.2401
ECE (std)	0.0381	0.0256	0.0182	0.0154	0.0076

Table 7: Comparison of ECE means and standard deviations across different sample sizes. Each subset of size N is sampled 100 times without replacement. Results are reported for the DeepSeek-8B model on the WinoBias dataset.

Thus, it is important to notice that with a small number of instances, the bins have less samples and the calibration error estimates are less reliable. As a consequence, both confidence estimates and derived measures, such as gender bias in our case, become unreliable under conditions of limited sample size. Therefore, it is essential to check the combination of complementary measures and ensure that each bin contains a sufficient number of instances to produce stable and meaningful estimates.

Conclusion

In this work, we investigate the alignment between predicted confidence scores and gender bias in large language models. Our results show that calibration error does not always correlate with human bias, and the degree of misalignment varies across models. We also propose a new metric, Gender-ECE, that can serve as an additional tool alongside existing calibration methods to better evaluate gender disparities in pronoun resolution tasks.

Ethical Statement

ECE and other calibration metrics measure how well a model's predicted confidence aligns with actual outcomes: how accurately the predicted probabilities reflect real-world (*i.e.* human bias). These metrics are primarily designed to assess trustworthiness, focusing on the relationship between confidence and accuracy. Gender-ECE extends this by evaluating calibration separately for subgroups, allowing comparison between, for example, male and female samples. However, these metrics do not fully capture other forms of bias, such as differences in prediction frequency, stereotypical language patterns, or unequal distribution of outcomes. Therefore, complementary fairness-oriented measures are required to comprehensively evaluate whether the model's behavior is fair across gender groups.

Acknowledgments

This work was supported by the Estonian Research Council grant "Developing human-centric digital solutions" (TEM-TA120) and by the Estonian Centre of Excellence in Artificial Intelligence (EXAI), funded by the Estonian Ministry of Education and Research and co-funded by the European Union and the Estonian Research Council via project TEM-TA119. It was funded by the EU H2020 program under the SoBigData++ project (grant agreement No. 871042) and partially funded by the HAMISON project.

References

AI@Meta. 2024. Llama 3 Model Card. *GitHub Repository*.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Chen, Y.; Yuan, L.; Cui, G.; Liu, Z.; and Ji, H. 2022. A close look into the calibration of pre-trained language models. *arXiv preprint arXiv:2211.00151*.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *ACL*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Ding, Z.; Liu, K. Z.; Peetathawatchai, P.; Isik, B.; and Koyejo, S. 2024. On fairness of low-rank adaptation of large models. *arXiv preprint arXiv:2405.17512*.

Felkner, V. K.; Chang, H.-C. H.; Jang, E.; and May, J. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *ACL*.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kapoor, S.; Gruver, N.; Roberts, M.; Pal, A.; Dooley, S.; Goldblum, M.; and Wilson, A. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *UncertaiNLP 2024*.

Krause, L.; Tufa, W.; Santamaría, S. B.; Daza, A.; Khurana, U.; and Vossen, P. 2023. Confidently wrong: exploring the calibration and expression of (Un) certainty of large language models in a multilingual setting. In *MM-NLG 2023*.

Kull, M.; Silva Filho, T.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, 623–631. PMLR.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Lum, K.; Anthis, J. R.; Robinson, K.; Nagpal, C.; and D'Amour, A. N. 2025. Bias in Language Models: Beyond Trick Tests and Towards RUTEd Evaluation. In *ACL*.

Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; and Lucic, M. 2021. Revisiting the calibration of modern neural networks. *NeurIPS*.

Murphy, A. H.; and Winkler, R. L. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1): 41–47.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *EMNLP*.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *ICML*.

OpenAI. 2022. Chatgpt: Optimizing LMs for dialogue.

Roelofs, R.; Cain, N.; Shlens, J.; and Mozer, M. C. 2022. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, 4036–4054. PMLR.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *NAACL*.

Sabir, A.; and Sharma, R. 2025. Exploring Gender Bias Beyond Occupational Titles. *arXiv preprint arXiv:2507.02679*.

Si, C.; Zhao, C.; Min, S.; and Boyd-Graber, J. 2022. Re-Examining Calibration: The Case of Question Answering. In *Findings EMNLP*.

Stanovsky, G.; Smith, N. A.; and Zettlemoyer, L. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Stańczak, K.; and Augenstein, I. 2021. A Survey on Gender Bias in Natural Language Processing. *ArXiv*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Team, T. 2024. Falcon 3 family of Open Foundation Models.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Yang, A.; et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yoon, D.; Kim, S.; Yang, S.; Kim, S.; Kim, S.; Kim, Y.; Choi, E.; Kim, Y.; and Seo, M. 2025. Reasoning models better express their confidence. *arXiv preprint arXiv:2505.14489*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.