UNIVERSITY OF TARTU
Institute of Computer Science

The Confidence Trap: Gender Bias and Predictive Certainty in LLMs

Ahmed Sabir, Markus Kängsepp, Rajesh Sharma
Institute of Computer Science, University of Tartu, Estonia

ahmed.sabir, markus.kangsepp, rajesh.sharma@ut.ee

Motivation

Calibration in LLMs: The task of ensuring that an LLM's expressed confidence levels accurately reflect the true likelihood of its responses being correct. Ideally, a well–calibrated model provides confidence scores that function as a reliable "trust meter" for its outputs.

Core Challenge of Miscalibration: LLMs often exhibit a discrepancy between confidence and accuracy. They may be *confidently wrong* or *uncertain but correct*, creating an unreliable foundation for decision-making.

Bias in LLMs: Systematic skew in model predictions arising from training data, model architecture, or alignment processes. Bias may manifest as stereotypes, unfair associations, or disparate treatment across demographic or thematic groups.

Why Calibration and Bias Matter: In high-stakes settings, miscalibration can amplify harmful biases. For example, a model might express high confidence in an answer shaped by biased data, leading to unjust outcomes in fields such as hiring or loan approvals.

Methods

Objective: Investigate the extent to which LLMs' predictive confidence is calibrated in gendered pronoun-resolution tasks.

1. Data curation

Input:

The **chef** mentioned that the recipe was crafted by [him/her]

2. Pronoun probability

Model token probabilities

$P(\text{him} \mid \text{context}) = 0.85$
$P(\text{her} \mid \text{context}) = 0.15$

3. Calibration & bias

Compare predicted confidence with

– ECE, Brier Score, ICE, MacroCE
– Human bias

Experiment

- **Diverse LLMs:** We evaluate six open-weight LLMs: GPT-J-6B (raw model), LLAMA-3.1-8B, Gemma-2-9B, Qwen2.5-7B, Falcon-3-7B, DeepSeek-R1-8B.
- **Benchmark:** Pronoun-resolution datasets: *WinoBias* and *Winogender*, *GenderLex* (last cloze pronoun), and *WinoQueer*. **Human alignment:** Each sentence pair is assigned a human-labeled bias score, indicating which sentence is more biased: "1" for male bias and "0" for female bias.
- **Evaluation metric:** Calibration metrics: ECE, MacroCE, ICE, Brier Score. Standard ECE does not reveal how the model behaves differently for male vs. female pronouns. To address this, we propose **Gender-Aware Group ECE**, a metric that captures calibration disparities across gendered pronouns:

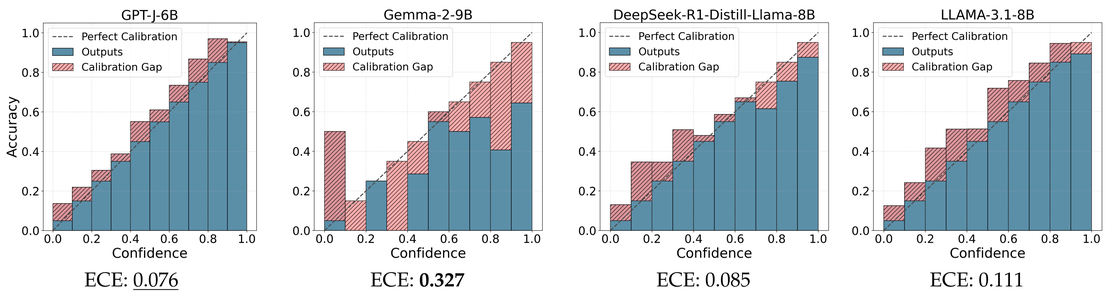
$$\text{Gender-ECE} = \frac{1}{2}\left(\text{ECE}_{\text{male}} + \text{ECE}_{\text{female}}\right)$$

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

Evaluation: Last Cloze Task

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 sentence context before scoring a biased pronoun.

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

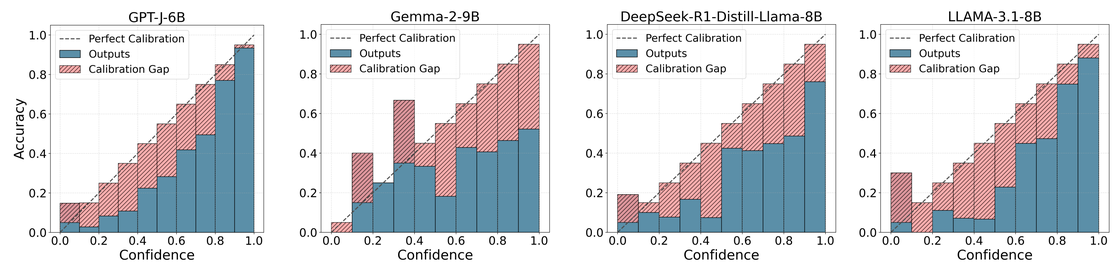


ECE: 0.076 | ECE: **0.327** | ECE: 0.085 | ECE: 0.111

Coreference Resolution Task

The analysis evaluates the model's confidence in coreference resolution by measuring the model's preferences when resolving pronouns.

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
Falcon-3-7B	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

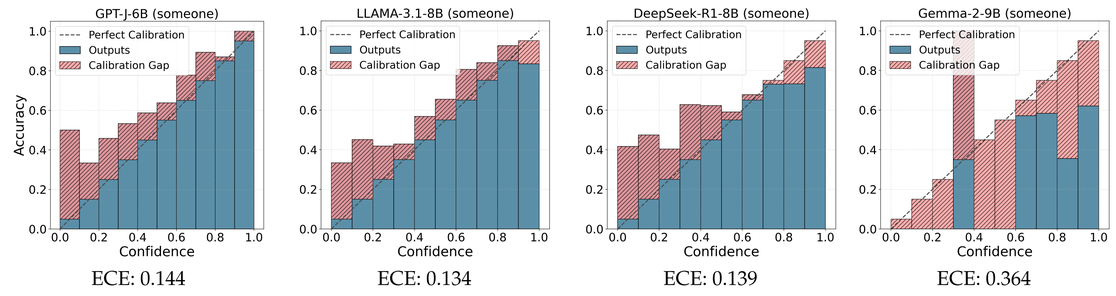


ECE: 0.157 | ECE: **0.429** | ECE: 0.240 | ECE: 0.193

Gender Neutral

We investigate model behavior under gender-neutral conditions by replacing occupation terms in the GenderLex dataset with gender-neutral expressions such as `person` and `someone`.

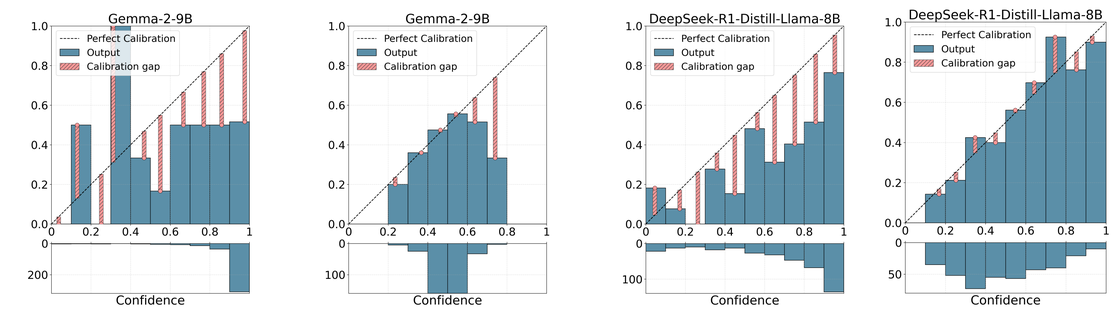
Metric	GPT-J-6B			LLAMA-3.1-8B			DeepSeek-R1-8B			Gemma-2-9B		
	Someone	Person	Occ	Someone	Person	Occ	Someone	Person	Occ	Someone	Person	Occ
ECE	0.144	0.063	0.076	0.134	0.138	0.111	0.139	0.130	0.085	0.364	0.367	0.327
MacroCE	0.484	0.476	0.453	0.483	0.478	0.466	0.482	0.481	0.461	0.493	0.493	0.494
ICE	0.454	0.442	0.374	0.436	0.445	0.371	0.452	0.443	0.369	0.397	0.393	0.390
Brier Score	0.331	0.341	0.432	0.358	0.348	0.446	0.352	0.361	0.470	0.560	0.581	0.574
Group-ECE	0.138	0.077	0.076	0.132	0.130	0.111	0.138	0.137	0.090	0.450	0.351	0.267
+ Male	0.106	0.031	0.085	0.115	0.071	0.112	0.048	0.065	0.074	0.363	0.367	0.330
+ Female	0.170	0.122	0.066	0.148	0.190	0.109	0.228	0.210	0.106	0.536	0.335	0.204
Human alignment	0.598	0.616	0.715	0.638	0.598	0.727	0.578	0.596	0.687	0.603	0.606	0.618



ECE: 0.144 | ECE: 0.134 | ECE: 0.139 | ECE: 0.364

Calibration

- Previous findings show that most models are poorly calibrated.
- We apply post-hoc **Beta calibration** (Kull, 2017) to adjust the model's confidence scores using a 50/50 validation–test split. This improves ECE, bringing predictions closer to the diagonal (perfect calibration).



ECE: **0.429** | ECE: 0.025 | ECE: **0.235** | ECE: 0.062

Conclusion

- We investigate how predicted confidence scores align with gender bias in large language models.
- We propose **Gender-ECE**, a complementary metric for evaluating gender disparities in pronoun-resolution calibration.

- Gemma-2 is the least well-calibrated model across all genders.
- The least filtered raw model (GPT-J-6B) is the most calibrated.
- LLAMA-3.1 shows the most balanced calibration and human alignment.
- Gender-neutral entities (*e.g.* someone) lead to poorer calibration.
- Distillation (DeepSeek-8B) amplifies gender-related calibration errors and degrades overall model calibration.

Code: `https://github.com/ahmedssabir/GECE`