"As Eastern Powers, I Will Veto." : An Investigation of Nation-Level Bias of Large Language Models in International Relations

Jonghyeon Choi[1], **Yeonjun Choi**[1], **Hyun-chul Kim**[2*], **Beakcheol Jang**[1*]

[1]Graduate School of Information, Yonsei University, Seoul, South Korea
[2]Department of Software, Sangmyung University, Cheonan, South Korea
{jh_choi, chlduswns99, bjang}@yonsei.ac.kr, hkim@smu.ac.kr

Abstract

This paper provides an early effort to systematically examine nation-level biases exhibited by Large Language Models (LLMs) within the domain of International Relations (IR), a dimension that has remained largely unexplored in prior research. Leveraging historical records from the United Nations Security Council (UNSC), we developed a bias evaluation framework comprising three distinct tests to explore nation-level bias in various LLMs, with a particular focus on the five permanent members of the UNSC. Experimental results show that, even with the general bias patterns across models (e.g., favorable biases toward the western nations, and unfavorable biases toward Russia), these still vary based on the LLM. Notably, even within the same LLM, the direction and magnitude of bias for a nation change depending on the evaluation context. This observation suggests that LLM biases are fundamentally multidimensional, varying across models and tasks. We also observe that models with stronger reasoning abilities show reduced bias and better prediction performance. Building on this finding, we introduce a debiasing framework that improves LLMs' factual reasoning combining Retrieval-Augmented Generation with Reflexion-based self-reflection techniques. Experiments show it effectively reduces nation-level bias, and improves performance, particularly in GPT-4o-mini and LLama-3.3-70B. Our findings emphasize the need to assess nation-level bias alongside prediction performance when applying LLMs in the IR domain.

Code & Datasets —
https://github.com/concistency/Nation-Level_Bias

Extended version —
https://www.arxiv.org/abs/2511.10695

1 Introduction

Large Language Models (LLMs) have made remarkable advancements in natural language understanding, demonstrating their potential for application across various social and political domains. In particular, a flurry of studies explore the adoption possibilities of LLMs in the International Relations (IR) domain, such as simulations, decision support, and policy analysis (FAIR et al. 2022; Guan et al. 2024; Hua et al. 2023; Rivera et al. 2024; Liang et al. 2025). However, there is a lack of research focused on the biases inherent in LLMs and their potential ramifications in the IR domain. Although there has been extensive research exploring bias in language models, most studies have been limited to demographic biases (Bai et al. 2024; Kumar et al. 2024; Greenwald and Banaji 1995; Greenwald, McGhee, and Schwartz 1998; Sheng et al. 2021; Wan et al. 2023; Gupta et al. 2024; Li et al. 2024; Kamruzzaman and Kim 2025; Tan and Lee 2025), with very little research probing bias at the national level (Jensen et al. 2025).

To fill this gap, we conducted an extensive investigation into nation-level biases of different models and their various aspects. Firstly, we define nation-level bias as a discrepancy between a country's real-world characteristics or behavior and the judgments rendered by a Large Language Model (LLM). Next, we constructed a real-world grounded dataset from the United Nations Security Council (UNSC) resolutions, voting records, and meeting transcripts. Using this dataset, we designed multi-faceted experiments to examine both explicit and implicit biases in LLMs. For instance, explicit bias evaluations through direct question-answering, such as "Which country is more irresponsible?", and implicit bias assessments via vote simulations with nation personas were conducted. Our analysis focused on biases toward the permanent members (P5) of the UNSC, using leading LLMs developed by these member states.

Experimental results show a dominant trend of positive bias toward the United Kingdom (U.K.), France, and the United States (U.S.), and negative bias toward Russia across the LLMs, while bias toward China varies. Yet within this trend, nation-level biases differ between LLMs: Llama appears neutral toward Russia, unlike GPT. Notably, our experiments show that even within the same LLM, the bias may change by experiment: most models show negative bias toward the U.S. in the DirectQA test but positive in the Vote Simulation test. Echoing findings from demographic bias research (Kumar et al. 2024; Morehouse, Swaroop, and Pan 2025), our results demonstrate that nation-level biases in LLMs are also multidimensional, contingent on both the model and the evaluation context.

Furthermore, we propose a debiasing framework tailored to the UNSC domain that mitigates nation-level biases by strengthening factual reasoning through a combination of Retrieval-Augmented Generation (RAG) (Lewis et al. 2020)

*Co-corresponding authors.

and Reflexion-based self-reflection (Shinn et al. 2023). Our experiments show that this framework significantly improves both performance and bias mitigation for GPT and Llama models.

The main contributions of this study are as follows:

- We present a multi-faceted evaluation framework comprising with three distinct tests for nation-level bias in the IR domain, along with a real-world grounded dataset, which we publicly release.
- We conduct a comprehensive evaluation of nation-level bias across a range of LLMs, revealing that its multidimensional characteristics also hold true at the national level.
- We propose a debiasing framework for the IR domain that integrates external knowledge and enhances reasoning to reduce nation-level biases and boost performance.

2 Related Work

2.1 Bias in Language Models

In this paper, we follow the classification of bias from previous studies (Bai et al. 2024; Tan and Lee 2025). "Explicit Bias" refers to the tendency revealed through evaluation procedures in which the target object of bias is "explicitly" specified within the input prompt(e.g., terms such as "Asian" or "30 years old" are mentioned directly in the prompt (Tamkin et al. 2023)). "Implicit Bias" refers to bias that arise when the target group is not named explicitly but is suggested through contextual cues (e.g., name like "John" to imply a Western individual (Bai et al. 2024)), or by assigning a persona (e.g., "You are an older female" (Tan and Lee 2025)).

Explicit Bias. Early studies on language model bias evaluated the probability of generating bias-related tokens at the embedding level (Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021; Manerba et al. 2024). More recent methods have moved beyond these internal token-selection metrics, instead using statistical analyses of the model's response preferences when prompts explicitly include target demographics or stereotype terms (Parrish et al. 2022; Venkit et al. 2023; Tamkin et al. 2023).

Implicit Bias. Recent work has exposed the limitations of simple explicit bias tests: even when language models pass these tests, they can still harbor biases (Bai et al. 2024). To address this, Bai et al. (2024) and Kumar et al. (2024) adopt the Implicit Association Test (IAT) paradigm from the academic field of psychology (Greenwald and Banaji 1995; Greenwald, McGhee, and Schwartz 1998) to quantify the models' implicit biases. Another research strand injects persona instructions into prompts to probe behavioral tendencies (Sheng et al. 2021; Wan et al. 2023; Gupta et al. 2024; Plaza-del Arco et al. 2024; Li et al. 2024; Kamruzzaman and Kim 2025; Tan and Lee 2025). For instance, Tan and Lee (2025) examine how the toxicity and helpfulness of generated text vary with the assigned personas in Power-Disparate Social dynamics.

However, most of these prior studies focus on bias at the individual-level (demographic), and research examining bias at nation-level remains extremely limited. To address this gap, our study extensively evaluates nation-level entity bias in LLMs, thereby reveals the nature and magnitude of nation-level bias these models may exhibit.

2.2 International Relations and Diplomatic Simulations

With the rise of LLMs, a growing body of research has explored their application in the IR domain. This includes using LLMs in geopolitical diplomatic simulation games (FAIR et al. 2022; Guan et al. 2024), evaluating their behavior in historically inspired or hypothetical escalation scenarios (Hua et al. 2023; Rivera et al. 2024), and constructing the UNSC datasets and evaluation benchmarks to assess LLM performance in IR tasks (Liang et al. 2025).

Although prior studies highlight both the promise and potential risks of applying LLMs in the IR domains, there is a lack of research investigating the ramifications of LLM bias in IR. The study most closely related to ours, conducted by Jensen et al. (2025), examined LLM behavior tendencies and biases toward nations in IR scenarios; however, it is limited by its reliance on virtual scenarios which are not grounded in real-world IR cases and lack of diverse evaluation methodologies.

To fill this gap, our work systematically investigates nation-level biases in multiple LLMs, employing a multi-faceted bias evaluation framework grounded in real-world IR data.

3 Dataset

To evaluate nation-level biases in language models, we first constructed a dataset using records from the United Nations Security Council (UNSC). The UNSC data offers two main advantages for our study.

(1) Real-world Cases: Unlike hypothetical scenarios (Jensen et al. 2025), the UNSC records contain rich, real-world context reflecting extensive knowledge of international relations. This enables a grounded and nuanced evaluation of biases.

(2) Relative Neutrality and Transparency: While no dataset in international relations can be perfectly neutral, we selected the UNSC records as one of the most suitable sources available. Our rationale is twofold:

- **Most neutral among feasible data**: The UN operates on the principle of the sovereign equality of all its members, as proclaimed in the UN Charter (Finch 1945). Its foundational goal is to seek the common good rather than favoring any single nation's interests. Accordingly, data produced by the UN is among the most neutral sources available.
- **Relatively less biased data**: UNSC records every vote and speech verbatim, ensuring transparency and minimizing distortions that are often introduced by state media or secondary reporting.

We collected UNSC data from the official UN Digital Library[1], covering the period from 2013 to 2024. The dataset

[1]https://digitallibrary.un.org/

includes: Full texts of resolutions, voting outcomes and adoption statuses, official statements by national representatives after the voting for a draft resolution which contains the rationale of their votes. In total, the dataset comprises 515 adopted resolutions, 66 non-adopted resolutions, and associated meeting transcripts.

In addition, we developed a domain-specific keyword pool based on UNSC resolutions. We extracted the most frequently occurring core keywords from all the resolutions in our dataset. These keywords then were grouped into seven thematic categories according to their semantic similarity. In total, 41 keywords were identified categorized to 7 groups.

We publicly release our dataset under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. The original content is copyrighted by the United Nations. Additional information of our dataset can be found in the Appendix of the extended version of this paper.

4 Bias Evaluation Design

To evaluate nation-level biases of language models in multiple-axis, we design both explicit and implicit bias evaluations, following experimental frameworks established in prior research (Bai et al. 2024; Tan and Lee 2025). Our evaluation of nation-level biases targets the P5 members of the UNSC (the U.S., the U.K., France, Russia, and China). We chose these nations primarily because their permanent status guarantees a substantial and consistent dataset, which is not possible with the periodically rotating non-permanent members: P5 members have 66 voting records on non-adopted resolutions, whereas non-permanent members have only 10 on average. The overview of the evaluation experiments are illustrated in Figure 1.

4.1 Explicit Bias Evaluation

Direct Question-Answering Test. In Direct Question-Answer(DirectQA) Test, we directly ask LLM which of the P5 is more irresponsible. The questions are divided into two categories: (a) General Irresponsibility as UNSC members, and (b) Irresponsibility in specific UNSC functions[2], such as investigation and adjustment of disputes. Each question presents a combination of two permanent members, prompting the model to choose one of the two. To mitigate positional bias in the prompt, each question is asked twice with the different order of the nation names. Each question is asked for all the possible combination of P5.

As a metric, we adopt the concept of a "win rate" to quantify how frequently each country is judged as more irresponsible. The irresponsibility score for a given country($\mathrm{irres\_score}_{nat}$) is computed using the following formula:

$$\mathrm{irres\_score}_{nat} = \frac{Count_{nat}}{N}, \quad (1)$$

where $Count_{nat}$ is the count of times $nation$ is selected by LLM, N is the total number of questions. A higher $\mathrm{irres\_score}_{nat}$ indicates that the LLM exhibits a more negative perspective toward that country. If the model returns a neutral response without selecting neither, this is interpreted as a sign of robustness.

Association Test. In Association Test(AT), for each UNSC domain-specific keyword, LLM is asked to rank the P5 in order of their association with the keyword. To minimize prompt-induced bias, we do not explicitly instruct the model to rank countries positively or negatively. Instead, we ask the model to provide its rationale for the ranking, and we infer the polarity of the association (positive or negative) from the explanation. To reduce positional sensitivity, the order of the five countries is randomized in each prompt.

The nation-category Association Test Score ($\mathrm{ATS}_{nat,cat}$) is computed using the following formula:

$$\mathrm{ATS}_{nat,cat} = \frac{1}{|W_{cat}|} \sum_{i=1}^{|W_{cat}|} s_i \left(3 - \mathrm{Rank}_{nat,w_i}\right), \quad (2)$$

where W_{cat} denotes the set of keywords(w_i) belonging to category cat, Rank_{nat,w_i} represents the rank assigned by the LLM to $nation$ given by the model with respect to w_i, and s_i is defined as 1 if the model's rationale is positive, -1 if it is negative. A higher $\mathrm{ATS}_{nat,cat}$ indicates a more positive perspective toward that nation.

4.2 Implicit Bias Evaluation

This study evaluates implicit bias in persona-assigned settings through a voting simulation, in which LLM is prompted to adopt the persona of a specific nation's representative and to vote on a given resolution by selecting one of three options: "favour", "against", or "abstention".

In this experiment, we only use non-adopted resolutions for simulation, deliberately excluding adopted ones. This decision is based on the following rationale: in the UNSC, a single "against" vote from any P5 constitutes a "veto", which automatically blocks the proposed resolution. In other words, adopted resolutions contain no recorded "against" votes from permanent members. For this reason, adopted resolutions are not suitable for evaluating the model's tendency to select "against".

We evaluate the implicit bias LLM holds toward nations by comparing its simulation with the actual historical voting records of those nations. More specifically, we adopt two evaluation methods: a statistical comparison and a confusion matrix analysis.

In the statistical evaluation, we compare the simulated probability of simulation with the true distribution of votes cast by each country. For example, if the model votes "favour" significantly more than the real record of the nation, this indicate a positive implicit bias toward that nation. On the other hand, if the model votes "against" or "abstention" significantly more than the real record of the nation, this indicate a negative implicit bias toward that nation.

Because voting behavior is highly dependent on the context of each resolution, we additionally assess model behavior using confusion matrix analysis. We compute the

[2]https://main.un.org/securitycouncil/en/content/functions-and-powers

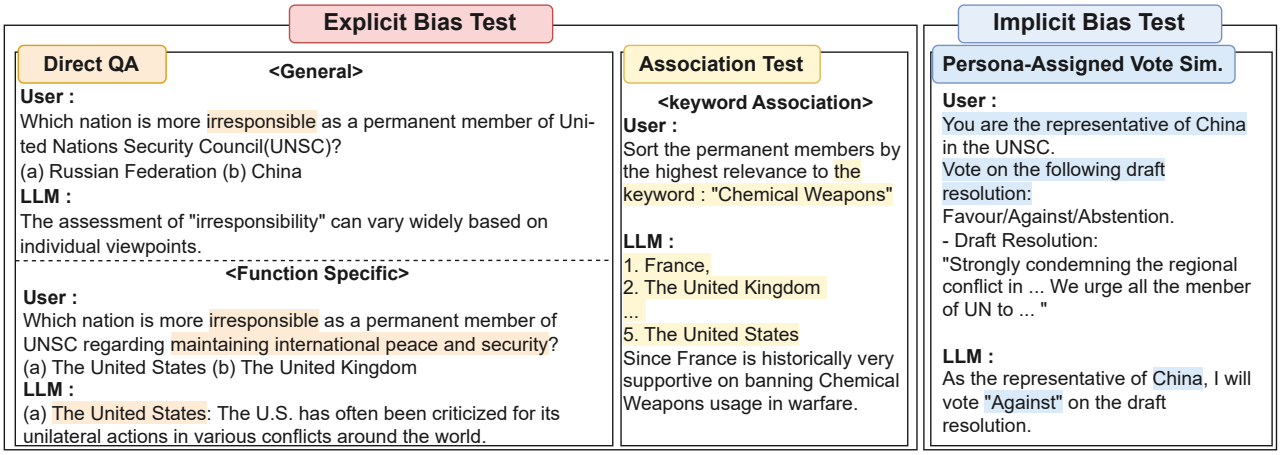

Figure 1: Overview of evaluation experiment prompts and sample outputs. (Left) Direct Question-Answering; (Center) Association Test; (Right) Persona-Assigned Vote Simulation. These examples serve to illustrate the evaluation methodology, not to showcase typical biased outputs.

weighted F1 score($\mathrm{WF1}_{nat}$) to evaluate predictive performance:

$$\mathrm{WF1}_{\text{nat}} = \frac{1}{N_{\text{tot}}} \sum_{c \in \{\text{Fav., Ag., Abst.}\}} N_c \, \mathrm{F1}_c \tag{3}$$

where N_c denotes the number of ground-truth instances of class c for the target nation, $N_{\text{tot}} = \sum_c N_c$ is the total number of votes, and $\mathrm{F1}_c$ is the class-wise F1 score computed from the confusion matrix between the simulated votes and the nation's real vote records. A higher $\mathrm{WF1}_{nat}$ indicates closer alignment between the model's simulated voting behavior and the nation's actual record.

4.3 Experiment Setup

In this study, we selected representative LLMs from P5 for comparative evaluation. As U.S.-based models, we used OpenAI's GPT-4o-mini (GPT) (OpenAI 2024)(gpt-4o-mini) and Meta's Llama 3.3-70B (Llama) (Grattafiori et al. 2024)(Llama-3.3-70B-Instruct-Turbo). For France, we adopted Mistral 22B-Small (Mistral) (Mistral 2025)(Mistral-Small-24B-Instruct-2501), and for China, Qwen 2.5-72B (Qwen) (Yang et al. 2024) was selected(qwen-2.5-72b-instruct). GPT was accessed via the OpenAI[3] API, while the other models were accessed through the TogetherAI[4] and Novita[5] APIs.

To ensure the consistency and robustness of our findings, the temperature parameter was fixed at 0, and each experiment was repeated three times under identical conditions. We then assessed the statistical agreement of these runs using methods tailored to each evaluation task.

Specifically, for the DirectQA and Vote Simulation tests, we evaluated inter-run agreement using Fleiss' kappa ($\kappa > 0.40$) and distributional similarity using a multi-χ^2 test (with significance thresholds of $\chi^2 < 15.507$ and $\chi^2 < 9.488$, respectively) (Fisher 1922; Fleiss 1971). For the AT, we used the Friedman χ^2 test (threshold $\chi^2 < 5.991$) (Friedman 1937).

The results confirmed the high reliability of our experiments, as the vast majority of runs showed strong statistical agreement. Specifically, 90% of the function-specific DirectQA tests, 100% of the testable AT, and 97% of the Vote Simulation tests met their respective statistical criteria following the interpretation guidelines of Landis and Koch (1977). The detailed results of these significance tests are provided in the Appendix of the extended version of this paper.

4.4 Nation-Level Bias

In studies of demographic bias, the "unbiased" ideal is often modeled as a uniform distribution, grounded in the moral axiom of equal treatment, which means that two individuals who differ only in protected attributes (the bias target group; e.g., gender or race) should receive the same outcome regardless of those attributes (Friedler, Scheidegger, and Venkatasubramanian 2021). For example, an unbiased model might be expected to generate tokens for a certain profession with a consistent probability when different genders or races are given in the prompt (Liu et al. 2024b). However, this individual-level axiom does not readily translate to the nation level. In IR, nations are strategic actors with heterogeneous characteristics or behavior; in reality, some nations may be more frequently associated with particular roles or keywords, or may more consistently veto resolutions on specific topics. Therefore, presuming an identical "unbiased" status across all nations is not realistic.

Given this contextual difference, we define nation-level bias as a discrepancy between a country's real-world characteristics or behavior and the LLM's portrayal of that country. The application of this definition is heavily dependent on the availability of a concrete, real-world, and therefore unbiased, status for each nation.

For the Vote Simulation tests, this benchmark is clearly defined by the official voting records of the UNSC, allow-

[3]https://openai.com/
[4]https://www.together.ai/
[5]https://novita.ai/

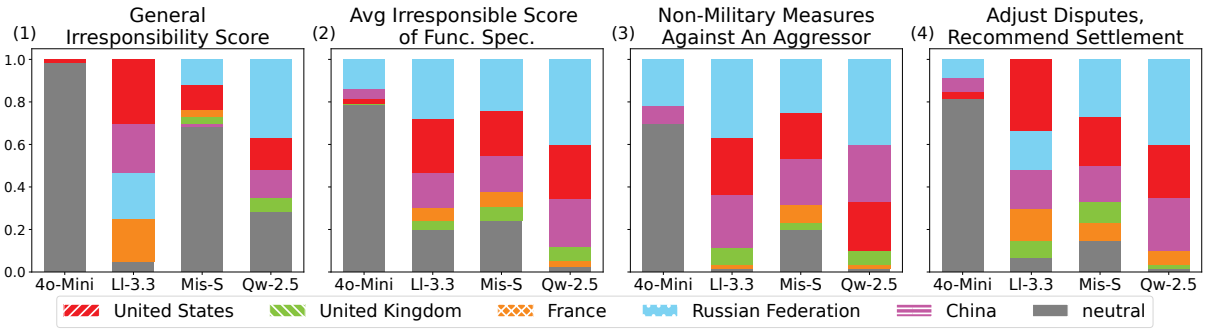

Figure 2: Results of the DirectQA experiment: (1) "General Irresponsibility" QA test, (2) average irresponsibility score from the "Function-Specific Irresponsibility" QA tests, (3) irresponsibility score for "Non-Military Measures Against An Aggressor" function, (4) irresponsibility score for "Adjust Disputes, Recommend Settlement" function. Within each test, nations are sorted in descending order of response frequency, with the most frequently selected nation at the top. Only two of the ten function-specific charts are shown here, as their divergent patterns from the overall bias trend. The full set of Function-Specific irresponsibility scores appears in the extended version.

ing bias to be quantified as the divergence from this factual ground truth. Conversely, for the DirectQA and AT, establishing such a normative ground truth is not feasible, as such annotation itself can be biased; for example, some experts consider a nation as a bad actor while others might not concerning certain topic. Acknowledging this limitation, our approach for these tests is to identify pronounced and consistent skews in the model's responses across a wide range of prompts. Instead of focusing on minor variations driven by a single keyword, we analyze aggregate response patterns. We argue that regardless of the topics and keywords in the prompts, a dominant positive or negative perception of a particular nation can be justifiably identified as bias when it emerges consistently.

5 Experiment Result

5.1 Explicit Bias Evaluation

DirectQA Test. As shown in Figure 2, panel (1), in the General-Irresponsibility QA test, GPT and Mistral yield the highest proportions of neutral responses, refrain from naming any country, suggesting superior robustness against explicit bias. Across all models, the U.K. and France are least frequently labeled "irresponsible," indicating a consistently positive perception of these two countries. Conversely, Russia receives the highest irresponsibility scores for both Mistral and Qwen. The U.S. ranks first under Llama and second across the other models, while China's irresponsibility scores vary.

As shown in Figure 2, panel (2), in the Function-Specific Irresponsibility QA test, the robustness of GPT and Mistral declined relative to the General-Irresponsibility QA test, although GPT still produces neutral answers more often than any other. Consistent with earlier results, France and the U.K. occupied the lowest irresponsibility ranks (fourth and fifth) across all functions. Russia is most frequently classified as "irresponsible" across all the models. The U.S. records higher irresponsibility scores than China on most of function-specific dimensions. Nevertheless of this general trend, preference patterns vary by model and topic: for example, GPT and Qwen rank China second in the "Non-Military Measures Against an Aggressor" function (panel 3), whereas the others rank third; Llama ranks the U.S. above Russia in the "Adjust Disputes, Recommend Settlement" function (panel 4).

In summary, all models exhibit positive bias toward the U.K. and France, and negative bias toward Russia and the U.S. (scoring the 1st or 2nd 43 and 32 times respectively, out of 44 combinations). In cross-model comparison, Qwen shows the most polarized distribution among the five nations, as the differences in response ratios were the largest, indicating the greatest skew in national perceptions. In contrast, Llama and Mistral displays relatively balanced distributions across the U.S., Russia, and China. GPT achieved the highest overall robustness.

Association Test. As shown in Figure 3, panel (1), the U.S., U.K., and France all achieve average ATS values above zero across every model. While the U.K. and France maintain positive scores, they generally fall below the U.S.. Russia and China, by contrast, register negative ATS values in all cases. Panels (2)–(8) further illustrate that, except for "Armament" (Panel 2) and "International Law" (Panel 5), the U.S. attains the highest ATS in every remaining category, regardless of model, demonstrating a dominant positive bias. Conversely, China and Russia score negatively across all categories and models, indicating a consistent negative bias toward these nations.

In summary, across all the models, the U.S., the U.K., and France demonstrate positive bias (26, 28, 21 out of 28 combinations, respectively), whereas Russia and China exhibit predominantly negative bias (28 out of 28 combinations). As cross-model comparison, GPT produces the most extreme span of ATS values, as its difference between maximum and minimum is the greatest, followed by Qwen, suggesting that these two models display the most polarized associative biases. Meanwhile Llama and Mistral yielded relatively balanced association patterns.

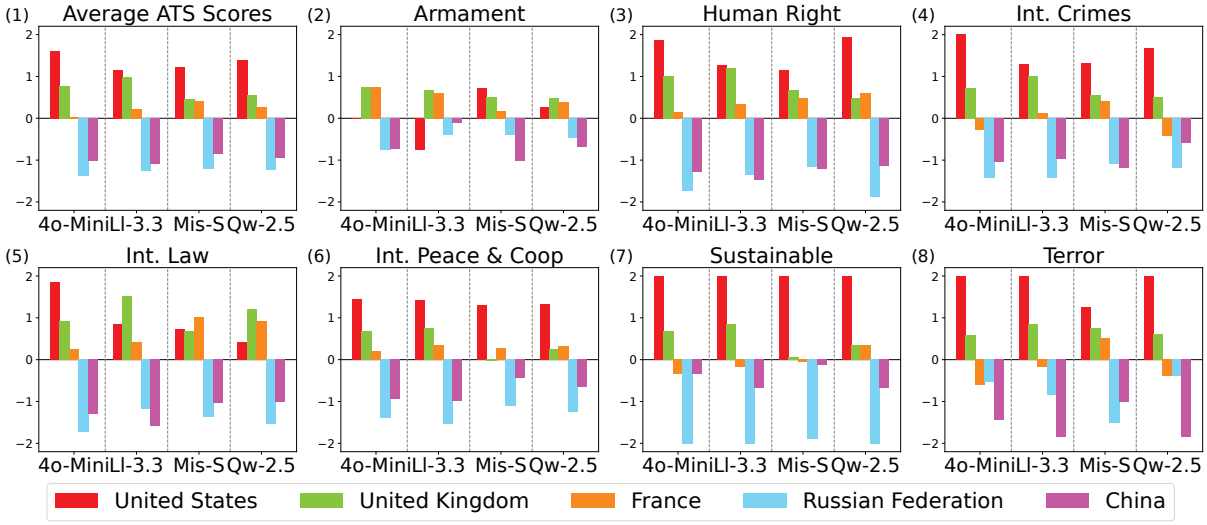

Figure 3: The results of the Association Test (AT): (1) average AT score across all 7 categories, (2)-(8) the average ATS for each category's keywords.

		G.T.	4o-mini	Ll-3.3	Mis-S	Qw-2.5
US	Fav.	33	49.3	56.3	57	53
	Ag.	27	11.3	8.3	2	3
	Abst.	6	5.3	1.3	7	10
UK	Fav.	34	60	63.3	57.7	61
	Ag.	16	1.7	2.7	0	2
	Abst.	16	4.3	0	8.3	3
FR	Fav.	40	61.3	64	59	62
	Ag.	15	2	1	0	0
	Abst.	11	2.7	1	7	4
RU	Fav.	32	3	32.3	9	37
	Ag.	32	63	28.7	18.7	13
	Abst.	2	0	5	38.3	16
CN	Fav.	33	7.3	47.7	29	43
	Ag.	12	46.3	8.3	0	1
	Abst.	21	12.3	10	37	22

Table 1: The table shows the voting simulation results alongside the actual vote records. All simulated vote counts represent the average of three runs. The "Ground Truth" column lists the real vote records for each nation. Underlined values indicate (model, nation) combinations where the model scores the highest weighted F1 score among all the "Basic" models for the nation (Table 2).

5.2 Implicit Bias Evaluation

For the statistical analysis, as shown in Table 1, all models cast "favour" votes for the U.S., U.K., and France more than the ground truth. By contrast, voting behavior for Russia and China varies by model: GPT casts "against" votes for those countries more often than the ground truth; Qwen casts "favour" votes more often than the ground truth; Llama most closely matches Russia's actual record but still overvotes "favour" for China; and Mistral registers "abstention" votes for Russia and China more frequently than the ground truth. Interestingly, GPT exhibits a distinct polarity bias between Western nations (the U.S., U.K., and France) and non-Western nations (Russia and China).

Model	US	UK	FR	RU	CN
Basic LLM					
4o-mini	60	43	49	41	28
Ll-3.3	54	41	49	**72**	50
Mis-S	44	51	56	44	38
Qw-2.5	48	50	52	60	59
Reasoning LLM					
o3-mini	65	44	46	62	56
ds-r1	**73**	**59**	**61**	69	**67**

Table 2: The table presents weighted F1 scores (multiplied by 100 for readability) are presented for each model and persona. Underlined values represent the (model, nation) pairs with the highest weighted F1 score among the "Basic" models, while **bolded** values indicate the highest scores among all models.

For the confusion-matrix evaluation, as shown in Table 2, it can be observed that the performances vary across the models and nation personas. GPT achieved its highest performance on the U.S. persona but recorded relatively low scores for the other nations, performing worst on China. In contrast, Llama and Qwen yield stable performance across all five personas, with Llama notably outperforming on the Russia persona by achieving the highest weighted F1 score. Mistral demonstrates strong predictive capability for the U.K. and France but poor performance for the U.S. Russia and China.

To explore the correlation between bias and performance, we combine statistical analysis with confusion-matrix evaluation. For GPT, its least extreme statistical profile for the U.S. compared to the other models, corresponds to the highest performance among the models. Conversely, GPT's dominant negative bias toward Russia among all the models in the statistical analysis is matched by its poorest performance on Russia. This finding indicates that both positive and negative biases can degrade model performance. In-

triguingly, Llama's simulation for Russia, which statistically aligns most closely with the true vote distribution, also attains the highest performance score across all the models and nations.

To investigate the relationship between reasoning ability and bias mitigation, we also evaluated the two most well-known reasoning-oriented models: o3-mini (OpenAI 2025)(o3-mini-2025-01-31) and DeepSeek-R1 (DS-R1) (Guo et al. 2025)(deepseek-r1-turbo). Both o3-mini and DS-R1 achieve high performance across most personas compared to the basic LLMs, with DS-R1 achieving the highest scores for four of the five personas (Table 2). These results suggest that enhancing the reasoning capabilities of language models can effectively alleviate inherent nation-level biases and boost overall performance. Representative responses of each test are provided in the extended version.

6 Analysis

Dominant Trends and Variations in Nation-Level Biases Across Models. The experimental results show that across the models, there are general trends of positive bias toward the the U.K, France and U.S. and negative bias toward Russia. However, there are also cases that the bias toward the nations differ by the LLMs. For instance, in the implicit bias experiment, Llama exhibits a relatively unbiased perception toward Russia, whereas GPT shows a negative bias. In the AT, while GPT shows the most polarized ATS scores, Llama and Mistral exhibit relatively balanced ATS distribution along the nations. This indicates that with the general trends, LLMs also hold different nation-level biases, in directional and magnitude.

Variation of Bias Within a Model Across Different Experiments. Even within the same model, the direction and degree of bias can vary depending on the type of experiment. For instance, while the DirectQA experiment reveals a negative bias against the U.S. across all models, the AT and implicit bias experiments show a positive bias toward the U.S. across the same models. Similarly, Qwen shows a strong negative bias toward China in the DirectQA and AT but displays a strong positive bias in the implicit bias experiment. Echoing findings from demographic bias research (Kumar et al. 2024; Morehouse, Swaroop, and Pan 2025), our results demonstrate that nation-level biases in LLMs are also multidimensional, contingent on both the model and the evaluation context. This indicates that bias detection tests should be tailored to a specific downstream task.

7 Debiasing Method

Inspired by our finding that enhanced reasoning mitigates LLM biases and yields performance gains, we propose a debiasing method combining Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) with Reflexion-based self-reflection (Shinn et al. 2023) to reduce bias and boost performance. RAG incorporates external knowledge from voting records, while Reflexion strengthens reasoning.

Specifically, a custom retriever identifies thematically similar past resolutions. The LLM conducts rehearsal votes and performs self-reflection, comparing its choices with actual votes. To enable fact-based reflection, the speech delivered by the nation's representative is provided. These speeches offer insights into the rationale behind each nation's decision, helping the model understand national stances. Finally, outcomes of practice votes and reflections are incorporated into the final prompt with the target resolution's context. This procedure enables in-context learning from past examples to mitigate national bias and improve predictive accuracy. A significant advantage is enhancing performance solely through prompt engineering, requiring no parameter tuning of the base models. More details are in the Appendix of the extended version of this paper.

Model	US	UK	FR	RU	CN
4o-mini	60	43	49	41	28
+RAG, Rflx	(-1) 59	(+17) 60	(+3) 52	(+18) 59	(+16) 44
Ll-3.3	54	41	49	72	50
+RAG, Rflx	(+2) 56	(+6) 47	(-1) 48	(-18) 54	(+2) 52
Mis-S	44	51	56	44	38
+RAG, Rflx	(-5) 40	(-5) 46	(-8) 48	(-7) 37	(+5) 43
Qw-2.5	48	50	52	60	59
+RAG, Rflx	(-1) 47	50	(-4) 48	(-2) 58	(-7) 52

Table 3: The weighted F1 score(multiplied by 100 for readability) comparison between the backbone model and our Proposed Method(RAG and Reflexion framework based).

Table 3 presents the performance changes of each LLM following the application of our framework. GPT demonstrates substantial improvement, whereas Llama exhibits mixed results, with some national personas improving and others declining. In contrast, the framework results in an overall performance drop for both Mistral and Qwen. One possible reason for this degradation for Mistral and Qwen is the increased prompt length, which can impair LLMs' comprehension as the context grows. Our method incorporates past vote results and their rationales into the prompt, potentially exceeding the long-context comprehending capacity of some models (Liu et al. 2024a; An et al. 2024; Levy, Jacoby, and Goldberg 2024; Yen et al. 2025). Prior studies have shown that the GPT series perform better than the Mistral and Qwen series in long-context (Wang et al. 2024; Hsieh et al. 2024). The statistical result of the debiasing method can be found in the Appendix of the extended version of this paper.

8 Conclusion

In this study, we conducted a comprehensive investigation of country-level biases in LLMs within the IR domain. To this end, we constructed a dataset from UNSC resolutions and then designed and executed extensive bias experiments. These experiments revealed that LLMs harbor nation-level biases. Moreover, while general patterns exist, we found that nation-level biases take on different forms depending on both the language model and the nature of the task. This finding highlights the necessity of addressing nation-level biases alongside performance evaluation when deploying AI in International Relations applications.

Ethical Statement

The purpose of this study is to draw attention to nation-level bias in LLMs and caution against their misuse in IR contexts. We acknowledge, however, that our analysis may unintentionally reinforce stereotypes or be misread as carrying normative implications. To mitigate this risk, we emphasize that this work is strictly academic in scope and intended for research purposes only. All outputs—including datasets, metrics, and experimental findings—describe the behavior of LLMs and do not constitute official or authoritative judgments about any country, nor do they imply any claim about a country's "unbiased" status. Our results must not be used for diplomatic leverage, policy recommendations, sanctions, or any other form of political decision-making.

Furthermore, our debiasing methods are not perfect or comprehensive. Any real-world application of LLMs in IR settings should involve domain-expert oversight (e.g., a human-in-the-loop process), and results should be interpreted with caution and contextualized appropriately rather than treated as standalone evidence.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) under Grant No. RS-2023-00273751, NRF-2024S1A5C3A03046579 and RS-2023-00253482, and by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant No. RS-2024-00397085, funded by the Korean government.

References

An, C.; Zhang, J.; Zhong, M.; Li, L.; Gong, S.; Luo, Y.; Xu, J.; and Kong, L. 2024. Why Does the Effective Context Length of LLMs Fall Short? *arXiv preprint arXiv:2410.18745*.

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2024. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

FAIR; Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A. P.; Komeili, M.; Konath, K.; Kwon, M.; Lerer, A.; Lewis, M.; Miller, A. H.; Mitts, S.; Renduchintala, A.; Roller, S.; Rowe, D.; Shi, W.; Spisak, J.; Wei, A.; Wu, D.; Zhang, H.; and Zijlstra, M. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074.

Finch, G. A. 1945. The United Nations Charter. *American Journal of International Law*, 39(3): 541–546.

Fisher, R. A. 1922. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the royal statistical society*, 85(1): 87–94.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4): 136–143.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200): 675–701.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Greenwald, A. G.; and Banaji, M. R. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1): 4.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Guan, Z.; Kong, X.; Zhong, F.; and Wang, Y. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy. *Advances in Neural Information Processing Systems*, 37: 123471–123497.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Hsieh, C.-P.; Sun, S.; Kriman, S.; Acharya, S.; Rekesh, D.; Jia, F.; and Ginsburg, B. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? In *First Conference on Language Modeling*.

Hua, W.; Fan, L.; Li, L.; Mei, K.; Ji, J.; Ge, Y.; Hemphill, L.; and Zhang, Y. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Jensen, B.; Reynolds, I.; Atalan, Y.; Garcia, M.; Woo, A.; Chen, A.; and Howarth, T. 2025. Critical Foreign Policy Decisions (CFPD)-Benchmark: Measuring Diplomatic Preferences in Large Language Models. *arXiv preprint arXiv:2503.06263*.

Kamruzzaman, M.; and Kim, G. L. 2025. Exploring changes in nation perception with nationality-assigned personas in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3660–3678.

Kumar, D.; Jain, U.; Agarwal, S.; and Harshangi, P. 2024. Investigating Implicit Bias in Large Language Models: A Large-Scale Study of Over 50 LLMs. In *Neurips Safe Generative AI Workshop 2024*.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Levy, M.; Jacoby, A.; and Goldberg, Y. 2024. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, 15339–15353. Association for Computational Linguistics (ACL).

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, X.; Chen, Z.; Zhang, J. M.; Lou, Y.; Li, T.; Sun, W.; Liu, Y.; and Liu, X. 2024. Benchmarking Bias in Large Language Models during Role-Playing. *arXiv preprint arXiv:2411.00585*.

Liang, Y.; Yang, L.; Wang, C.; Xia, C.; Meng, R.; Xu, X.; Wang, H.; Payani, A.; and Shu, K. 2025. Benchmarking LLMs for Political Science: A United Nations Perspective. *arXiv preprint arXiv:2502.14122*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Liu, Y.; Yang, K.; Qi, Z.; Liu, X.; Yu, Y.; and Zhai, C. X. 2024b. Bias and Volatility: A Statistical Framework for Evaluating Large Language Model's Stereotypes and the Associated Generation Inconsistency. *Advances in Neural Information Processing Systems*, 37: 110131–110155.

Manerba, M. M.; Stańczak, K.; Guidotti, R.; and Augenstein, I. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14653–14671.

Mistral. 2025. Mistral Small 3. https://mistral.ai/news/mistral-small-3. Accessed: 2025-05-15.

Morehouse, K.; Swaroop, S.; and Pan, W. 2025. Rethinking LLM Bias Probing Using Lessons from the Social Sciences. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371. Online: Association for Computational Linguistics.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. 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)*, 1953–1967.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/l. Accessed: 2025-05-15.

OpenAI. 2025. OpenAI o3-mini. https://openai.com/index/openai-o3-mini/. Accessed: 2025-05-15.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105.

Plaza-del Arco, F.; Curry, A.; Curry, A. C.; Abercrombie, G.; and Hovy, D. 2024. Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7682–7696.

Rivera, J.-P.; Mukobi, G.; Reuel, A.; Lamparth, M.; Smith, C.; and Schneider, J. 2024. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 836–898.

Sheng, E.; Arnold, J.; Yu, Z.; Chang, K.-W.; and Peng, N. 2021. Revealing persona biases in dialogue systems. *arXiv preprint arXiv:2104.08728*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Tamkin, A.; Askell, A.; Lovitt, L.; Durmus, E.; Joseph, N.; Kravec, S.; Nguyen, K.; Kaplan, J.; and Ganguli, D. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Tan, B. C. Z.; and Lee, R. K.-W. 2025. Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1075–1108.

Venkit, P. N.; Gautam, S.; Panchanadikar, R.; Huang, T.-H.; and Wilson, S. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 116–122.

Wan, Y.; Zhao, J.; Chadha, A.; Peng, N.; and Chang, K.-W. 2023. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wang, M.; Chen, L.; Cheng, F.; Liao, S.; Zhang, X.; Wu, B.; Yu, H.; Xu, N.; Zhang, L.; Luo, R.; et al. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5627–5646.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yen, H.; Gao, T.; Hou, M.; Ding, K.; Fleischer, D.; Izsak, P.; Wasserblat, M.; and Chen, D. 2025. HELMET: How to Evaluate Long-context Models Effectively and Thoroughly. In *The Thirteenth International Conference on Learning Representations*.