State Actor Bias in Non-Western Conflicts: Evidence from Large Language Models

Olukunle O. Owolabi *
Independent Researcher
olukunle.owolabi@alumni.tufts.edu

Abstract

Large Language Models (LLMs) encode and operationalize biases shaped by the socio-political contexts embedded in their training corpora and post-training alignment procedures. This study investigates state actor bias in conflict interpretation within a non-Western context, using ACLED event data from Nigerian conflicts (2014-2024). State actors (Police and Military) play a central role in the region's conflicts, where a rise in armed group activity necessitates legitimate state engagements ($\mathbf{B}$, Battles) even as violence against civilians ($\mathbf{V}$, illegitimate) remains a growing concern.

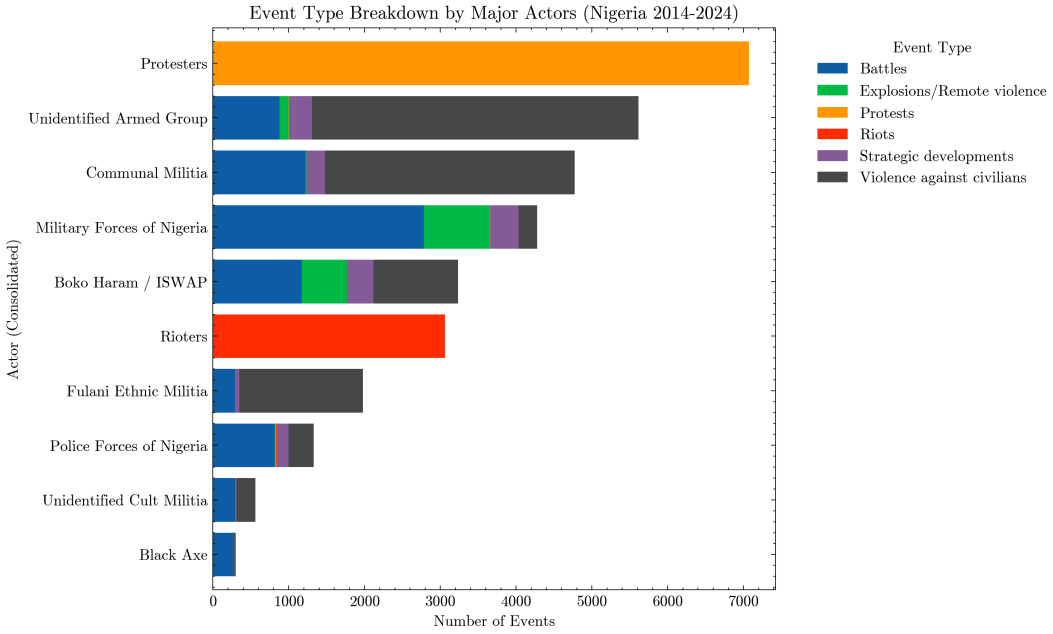
We examine the ability of three open-source models (Mistral7b, Llama3.2, and Qwen2.5) to distinguish between legitimate and illegitimate state actions. Bias polarization is defined using two principal error modes: False Legitimacy (FL), where $\mathbf{V} \rightarrow \mathbf{B}$ (illegitimate action is excused), and False Illegitimacy (FI), where $\mathbf{B} \rightarrow \mathbf{V}$ (legitimate action is condemned). The $\mathbf{FL}/\mathbf{FI}$ ratio serves as a measure of the direction and strength of this polarization.

The results reveal a bimodal, actor-specific bias in how LLMs interpret state violence. For military actions, **Mistral** shows a legitimation tendency ($\mathbf{FL}/\mathbf{FI} = \mathbf{5.57}$), **Llama** exhibits a pronounced illegitimation bias ($\mathbf{0.07}$), and **Qwen** remains comparatively balanced ($\mathbf{0.58}$). These differences are likely shaped by variations in training data and safety alignment that influence how models frame state legitimacy. Police actions display a shared homogeneous bias, with near-zero $\mathbf{FL}$ rates and high $\mathbf{FI}$ rates (Llama3.2: 76.20%, Qwen2.5: 20.61%)suggesting a consistent tendency to associate policing with civilian violence, likely reflecting global narratives embedded in pre-training data. Together, these findings suggest that LLMs may internalize politically conditioned reasoning, highlighting the need for more context- sensitive alignment strategies when applying such models to conflict analysis, especially in non-western settings.

1 Introduction

Understanding how Large Language Models (LLMs) interpret conflict dynamics in non-Western contexts is crucial as these models increasingly inform research and policy tools. Figure 1 shows that state actors (specifically the Police and Military) rank among the top ten conflict actors in Nigeria between 2014 and 2024, underscoring the duality of their role as both protectors in security operations and aggressors in civilian confrontations. As armed group activity has intensified, the state's dual role in legitimate engagements and potential civilian harm becomes more prominent in the region as the data shows.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.



Note: Military Forces of Nigeria, Police Forces of Nigeria are among the top 10 most active state actors.

Figure 1: Top ten conflict actors in Nigeria (2014-2024), highlighting the prominence of state actors such as the Police and Military.

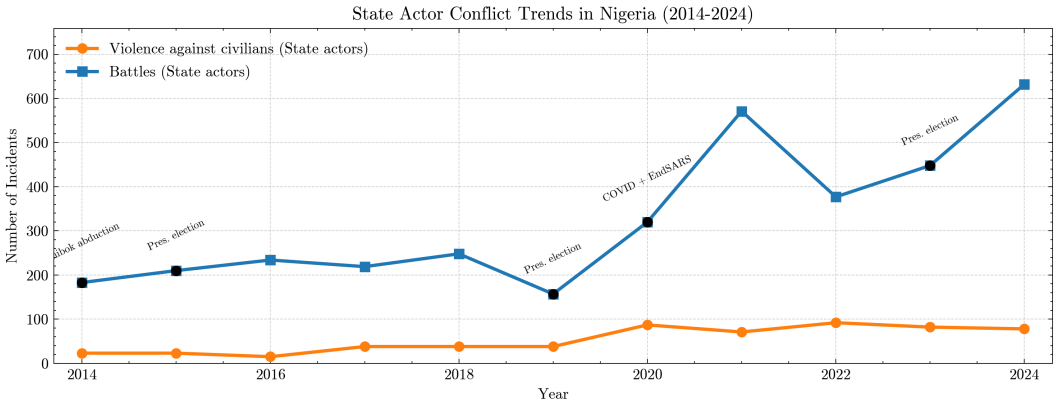


Figure 2: Trend of state actor engagements (2014-2024) showing an increase in both legitimate (Battles) and illegitimate (Violence against Civilians) actions over time.

Figure 2 further illustrates an upward trend in both legitimate military engagements (Battles) and illegitimate actions (Violence against Civilians) by state actors. This pattern aligns with key national moments such as increased in armed conflict such Boko Haram insurgency escalation (2014-), post-2019 presidential elections (2019-), and the 2020 EndSARS protests. These trends motivate the need to examine how LLMs, which draw heavily from global narratives and media data, interpret such state-led actions in non-Western conflict settings.

2 Data

This study uses Armed Conflict Location and Event Data (ACLED) covering Nigeria from 2014 to 2024. The dataset captures spatial and temporal details of conflict events, including actor type, event category, and fatalities. We focus on events attributed to state actors (Military and Police), comparing two primary forms of engagement: **Battles (B)** representing legitimate state action, and

Violence against Civilians (V) representing illegitimate force. These categories form the foundation for assessing LLM bias in distinguishing between lawful and unlawful state actions.

3 Results

Table 1: Polarity and Symmetry of Misclassification Bias ($\mathbf{V} \leftrightarrow \mathbf{B}$)

Model	Actor	True V	True B	FL Count	FL %	FI Count	FI %	FL/FI Ratio
Bias towards Police (V=335, B=815)								
mistral7b	Police	335	815	3	0.90%	87	10.67%	**0.08**
llama3.2	Police	335	815	0	0.00%	621	**76.20%**	**0.00**
qwen2.5	Police	335	815	0	0.00%	168	**20.61%**	**0.00**
Bias towards Military (V=250, B=2784)								
mistral7b	Military	250	2784	33	**13.20%**	66	2.37%	**5.57**
qwen2.5	Military	250	2784	10	4.00%	193	6.93%	**0.58**
llama3.2	Military	250	2784	7	2.80%	1163	**41.77%**	**0.07**

- **Illegitimation Bias (Critic)**: $\mathbf{FL/FI} \ll \mathbf{1}$ (model tends to condemn legitimate actions)
- **Legitimation Bias (Apologist)**: $\mathbf{FL/FI} \gg \mathbf{1}$ (model tends to excuse illegitimate actions)
- **Balanced Bias (Neutral)**: $\mathbf{FL/FI} \approx \mathbf{1}$ (model shows minimal directional bias)

The results reveal two distinct but consistent patterns of bias across state actor classifications. For the Military, models show differing directional tendencies. **Mistral** exhibits a *Legitimation Bias* with an $\mathbf{FL/FI\ Ratio}$ of **5.57**, indicating a tendency to excuse illegitimate state actions. In contrast, **Llama3.2** shows an *Illegitimation Bias* with a ratio of **0.07**, often misclassifying legitimate engagements as violence against civilians. **Qwen2.5** remains comparatively balanced with a ratio of **0.58**. These differences may reflect variations in training data composition and alignment procedures that influence how models interpret state legitimacy.

For the Police, all models display a *Homogeneous Bias*, characterized by near-zero **FL** rates and elevated **FI** rates (**Llama3.2**: 76.20%, **Qwen2.5**: 20.61%, **Mistral**: 10.67%). This shared pattern suggests that models are more likely to associate policing with civilian harm, possibly reflecting globally prevalent narratives embedded in pre-training data.