Multi-View Behavioral Probing for Political Bias in Arabic and Multilingual Transformers Before and After Domain Adaptation

Ahmad Abdelhameed[1], Ensaf Hussein Mohamed[2], Walaa Medhat[3]
ITCS School, Nile University, Giza, Egypt[1]
Computers & AI, Helwan University, Cairo, Egypt[2]
Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt[3]

***Abstract*—Political bias in transformer-based language models poses a critical challenge for applications involving politically sensitive Arabic news, yet systematic evaluation remains limited. This paper presents a multi-view behavioral framework to detect political bias in four pre-trained transformer models: AraBERTv2, CAMeLBERT, mBERT, and XLM-R. The framework integrates four complementary probes: sentiment drift, emotion drift, counterfactual actor-swapping for identity sensitivity, and masked language model probing to detect lexical preference shifts. Each model is evaluated before and after domain-adaptive fine-tuning on the FigNews Arabic political news dataset to analyze how politically sensitive training data influences representational bias. To synthesize signals from these probes, a Decision and Bias Reporting Agent (DBRA) aggregates the evidence using a structured hierarchy that prioritizes implicit bias indicators. Results show that bias is already present in base checkpoints and can significantly shift after adaptation. For example, mBERT's masked preference for SideA drops from 40.7% to 0.0%, indicating complete directional collapse, while XLM-R shows a large increase in masked preference toward SideA (ΔPR = +32.8%).**

Keywords*—NLP; political bias; Arabic transformers; domain-adaptive pretraining; masked probing; actor-swapping; bias detection; behavioral evaluation

I. Introduction

Bias in large language models has become a critical concern as these systems are increasingly used to analyze politically sensitive news and public discourse. In such settings, even subtle shifts in sentiment, emotion, or entity framing can shape public perception, reinforce polarization, and lead to high-stakes misinterpretations. This challenge is particularly relevant for Arabic NLP, where complex morphology, dialect variation, and context-dependent political cues make model behavior highly sensitive to lexical choices. As a result, Arabic transformer models are frequently deployed and adapted for news analysis, despite limited systematic understanding of how political bias manifests in their predictions.

Importantly, political bias is often treated as an inherent property of a pre-trained checkpoint. However, in real deployments, models rarely remain frozen: they are continuously adapted through continued pre-training or domain-adaptive pre-training (DAPT) to better match target distributions. This adaptation step can reshape bias-related behavior, sometimes stabilizing outputs and sometimes amplifying implicit preferences.

Most prior work evaluates adaptation primarily through task-level accuracy improvements. In contrast, fewer studies examine how adaptation reshapes model behavior under counterfactual edits, affective prediction tasks, or likelihood-based probing. This motivates the need for a unified evaluation framework that can capture bias-related drift beyond standard performance metrics. Given the societal impact of politically biased model behavior, this study is structured around the following research questions:

RQ1: Do Arabic and multilingual transformer models exhibit measurable political bias?

RQ2: What is the impact of fine-tuning on the Arabic political dataset FIGNEWS on the bias behavior of Arabic and multilingual transformers?

RQ3: How can political bias in Arabic and multilingual transformer models be systematically evaluated?

To answer these questions, this paper introduces a unified behavioral evaluation framework and applies it to four widely used models: AraBERTv2 [28], CAMeLBERT [29], mBERT [30], and XLM-R [31] before and after domain-adaptive fine-tuning on the FIGNEWS [27] dataset. The proposed framework combines four complementary diagnostics: (1) zero-shot sentiment drift, (2) zero-shot emotion drift, (3) actor-swapping for identity sensitivity, and (4) masked word prediction probing for lexical preference drift. Cross-seed analysis is conducted to assess whether observed bias effects are consistent across random initializations.

The contributions of this work can be summarized as follows:

- Introducing a comprehensive multi-view behavioral framework for political bias evaluation, through which bias is systematically examined across complementary diagnostic dimensions before and after domain adaptation, enabling structured and holistic assessment.
- Presenting an evidence-aggregation and bias reporting mechanism that consolidates multi-probe behavioral signals into a unified and interpretable bias

characterization, facilitating principled model comparison and structured bias categorization.

- Providing empirical evidence that political bias is inherently measurable in pretrained representations, demonstrating that bias may originate from foundational model pretraining rather than emerging solely during task-specific adaptation.
- Demonstrating that domain adaptation in politically sensitive contexts can induce divergent bias trajectories across architectures, and revealing that apparent robustness at the surface level may coexist with substantial latent representational distortion.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the methodology and experimental setup, Section 4 presents the results, and Section 5 concludes with key findings and future directions.

II. Related Work

Political bias in language models has become a central concern as transformer-based systems are increasingly deployed in high-stakes settings such as news analysis, public discourse monitoring, and conflict-related content moderation. While existing studies have examined bias across multiple languages, Arabic political discourse remains substantially underexplored despite its unique challenges: complex morphology, dialect variation, and culturally embedded political framing create bias patterns that standard multilingual benchmarks fail to capture. This work focuses specifically on Arabic political news covering the Palestinian-Israeli conflict, a domain where actor framing, emotional salience, and lexical sensitivity make bias evaluation particularly critical. Bias in such settings can emerge not only in explicit model outputs but also in affective predictions and internal representational preferences [1], and is further amplified when models undergo domain-adaptive pre-training on politically sensitive data, which may reshape bias signals in unpredictable ways [22].

A. Political Bias in Large Language Models

A growing body of work has examined political bias as a measurable and socially consequential phenomenon in modern LLMs. For example, recent studies propose systematic evaluation approaches that quantify political leaning in model outputs and demonstrate that political bias can be expressed through both *what* the model states and *how* it frames responses [16]. Other work evaluates political bias in real-world interactive settings, showing that political preferences can be observed through downstream decision-making and user-facing recommendation behavior [12]. Beyond surface-level outputs, political bias has also been analyzed through target-oriented sentiment paradigms, which measure how sentiment shifts when the target entity changes while the surrounding context remains fixed [10]. Recent work also highlights that political bias assessments may diverge between humans and LLM-based evaluators, emphasizing the need for careful behavioral diagnostics in real news settings [24]. These approaches align closely with our motivation, as political bias in news discourse often manifests as subtle differences in framing and affect rather than overt ideological statements.

B. Cultural and Arabic-Centered Political Evaluation

Bias evaluation becomes especially important in Arabic contexts, where political and cultural cues are deeply tied to lexical choice, identity references, and sociopolitical framing. Recent benchmarks such as PALMX 2025 highlight the need for culturally grounded benchmarks for Arabic and Islamic culture, demonstrating that standard multilingual evaluation does not capture culturally sensitive failure modes [2]. More broadly, recent Arabic-focused surveys emphasize persistent gaps in resources, evaluation coverage, and robust model assessment, particularly for socially sensitive applications [7], [8]. These findings support the need for bias evaluation frameworks that explicitly probe political framing and actor sensitivity in Arabic news settings, rather than relying solely on accuracy-based task performance. Political framing bias has also been studied in multilingual contexts beyond Arabic, suggesting that framing effects can generalize across sociopolitical settings and languages [25].

C. Domain Adaptation and Fine-Tuning Effects on Bias

A key challenge is that domain adaptation may not simply improve task performance but can also reshape bias-related behavior. Recent work on Arabic domain-adaptive pre-training for aspect-based sentiment analysis shows that continued pre-training can significantly affect downstream sentiment behavior and domain robustness [3]. In parallel, A-MASA provides multi-domain Arabic sentiment resources that further highlight how domain shifts and topic variation can influence affective prediction patterns [4]. In high-stakes political contexts, this implies that fine-tuning on politically oriented data may alter not only accuracy but also the stability of sentiment and emotion predictions, potentially amplifying bias even when headline metrics appear improved. Related studies on temporal drift in transformer sentiment systems similarly show that model behavior can shift substantially over time or adaptation, even without explicit changes in evaluation metrics [23].

D. Counterfactual Evaluation and Actor Sensitivity

Bias in political discourse is often triggered not only by polarity but also by *who* is being discussed. For this reason, counterfactual evaluation has become a widely used approach for measuring identity sensitivity, where protected or salient entities are swapped within otherwise identical inputs. FairFlow introduces a model-based counterfactual augmentation framework that demonstrates how controlled substitutions can expose systematic bias and improve robustness in NLP pipelines [5]. However, recent work cautions that template-based counterfactual probes may introduce artifacts and should be interpreted carefully when concluding bias [21]. In politically sensitive settings, actor-swapping provides a particularly direct diagnostic of identity-dependent behavior, as it isolates whether models treat different political actors symmetrically under matched contexts. This line of work motivates our actor-swapping experiment, where sentiment and emotion predictions are tested for invariance under politically salient entity substitution.

E. Masked Language Model Probing and Implicit Bias

While behavioral metrics such as sentiment and emotion flips capture overt instability, masked language modeling

(MLM) probes can reveal deeper representational preferences that may remain hidden under stable categorical outputs. Recent work by Sweeney proposes measuring social biases in masked language models through prediction-based proxies, showing that bias can persist in the likelihood landscape even when downstream behavior appears stable [6]. This perspective is critical for political bias evaluation; as representational drift may lead to systematic lexical preferences toward specific actors or narratives. Recent probing studies further demonstrate that political ideology can be encoded in latent representations and generalize across tasks, motivating representational analysis beyond surface-level predictions [18]. Similarly, studies on implicit political stereotypes show that bias may remain hidden even when explicit outputs appear neutral, reinforcing the need for implicit probing signals such as masked likelihood preferences [19].

F. Bias in Emotion Inference and High-Stakes Affective Prediction

Emotion inference models are increasingly used in news analysis, moderation, and public sentiment tracking. However, recent evidence suggests that such systems may exhibit a high risk of political bias, where affective predictions shift systematically across political groups or contexts [9]. This is particularly relevant for our setting, since our results show that emotion predictions are substantially more volatile than sentiment and can exhibit large flip rates under actor substitution. These findings reinforce the need to treat affective analysis as a bias-sensitive task rather than a neutral diagnostic.

G. Broader Bias Foundations and Mitigation Directions

Several works provide broader conceptual grounding for understanding bias sources and mitigation strategies. Prior analyses identify multiple sources of bias in NLP pipelines, including dataset construction, representation imbalance, and annotation artifacts, which can interact with model training and adaptation [11]. In addition, work on quantifying and alleviating political bias proposes structured measurement and mitigation techniques, highlighting that political bias is not merely an ethical issue but also a measurable modeling phenomenon [13]. More recently, mitigation research has expanded toward multi-objective and multi-agent frameworks that attempt to reduce social bias while preserving utility, emphasizing that bias mitigation remains an open challenge even for modern large models [17]. Complementary frameworks such as Nbias provide bias identification tools for text, reinforcing the importance of systematic detection pipelines [14]. Finally, recent reviews of transformer-based sentiment analysis further confirm that transformer models can achieve strong performance while still exhibiting sensitivity to domain and bias-related factors [15].

Overall, prior work establishes that political and social bias can be measured across multiple layers of model behavior, from explicit outputs to affective predictions and masked likelihood preferences. However, existing studies often focus on either high-level political bias in generation [12], [16] or general social bias in MLMs [6], without systematically connecting these perspectives under a unified before/after domain adaptation setting. Beyond interpretability concerns, political bias has been shown to directly affect downstream task performance, such as stance classification, reinforcing that bias is also a reliability issue [26]. This work bridges this gap by introducing a multi-view behavioral probing framework that evaluates Arabic and multilingual encoders before and after politically oriented domain-adaptive fine-tuning. By combining sentiment and emotion drift, counterfactual actor-swapping, and masked lexical probing, the proposed framework captures both surface-level robustness and deeper representational bias amplification, including cases where improved stability masks severe implicit drift. This study also extends a prior study on detecting bias in Arabic political news [20] by shifting from supervised detection pipelines to a probing-based framework that analyzes model behavior before and after domain-adaptive fine-tuning. Furthermore, this study introduces a Decision and Bias Reporting Agent (DBRA) that synthesizes outputs from all four behavioral experiments into a unified, evidence-driven bias assessment and ranked model classification, directly addressing the need for systematic evaluation frameworks in politically sensitive Arabic NLP.

III. Methodology

This section presents the proposed framework for analyzing behavioral drift in Arabic and multilingual transformer models before and after domain-adaptive pre-training (DAPT) using masked language modeling (MLM) on politically sensitive Arabic news text as shown in (Fig. 1). The framework is designed to identify potential political bias through four complementary behavioral probes and to synthesize their signals into a unified bias assessment using a Decision and Bias Reporting Agent (DBRA). The objective of this framework is not to optimize downstream task accuracy, but rather to systematically diagnose representational and behavioral bias in the models. To ensure the validity of the evaluation, all datasets used for behavioral probing are strictly held out from the adaptation training process, preventing any form of data leakage between training and evaluation stages.

A. Models

Four widely used transformer-based encoder models are examined, covering both Arabic-specialized and multilingual settings:

- AraBERTv2, an Arabic-focused BERT model pretrained on large-scale Arabic corpora [28].
- CAMeLBERT, a transformer pretrained on a mixture of Modern Standard Arabic and dialectal data [29].
- mBERT, a multilingual BERT model trained on more than 100 languages [30].
- XLM-R, a multilingual RoBERTa-based encoder with strong cross-lingual representations [31].

These models enable comparison of whether behavioral drift under DAPT differs between Arabic-specialized and multilingual pre-training systems.

B. Domain-Adaptive Pretraining (DAPT)

For each model, domain-adaptive pre-training is performed using the masked language modeling (MLM) objective on Arabic news text FIGNEWS [27]. DAPT aligns the model with the target domain while preserving the original architecture.

Formally, given an input sequence $x = (x_1 \ldots x_n)$,

The model is optimized to maximize the conditional likelihood:

$$\mathcal{L}_{MLM} = -\sum_{i \epsilon M} \log P(x_i | x_{-M}) \quad (1)$$

MLM training masks a subset of token positions M, and the model predicts each masked token x_i conditioned on the remaining visible tokens x_{-M}.

DAPT is applied for two epochs using a masking probability of 0.15 using Adam optimizer and a learning rate of $2e$-5.

To ensure reproducibility and measure stability, three adapted variants per model are trained using different random seeds:

$$S = \{13,42,100\} \quad (2)$$

To distinguish systematic bias drift from random training variability. Results are reported as mean and standard deviation across seeds.

C. Data Splits and Data Hygiene

Data Splits. An unlabeled subset from FIGNEWS is used [27] for domain-adaptive pre-training (DAPT) containing 13,320 training instances. All behavioral evaluations are conducted on strictly held-out data, including a general test set of 1100 instances, an actor-swapping evaluation set of 100 paired samples, and a masked-probing set of 140 samples. Table I summarizes the dataset splits and their purposes. No near-duplicate overlap is verified using exact string matching.

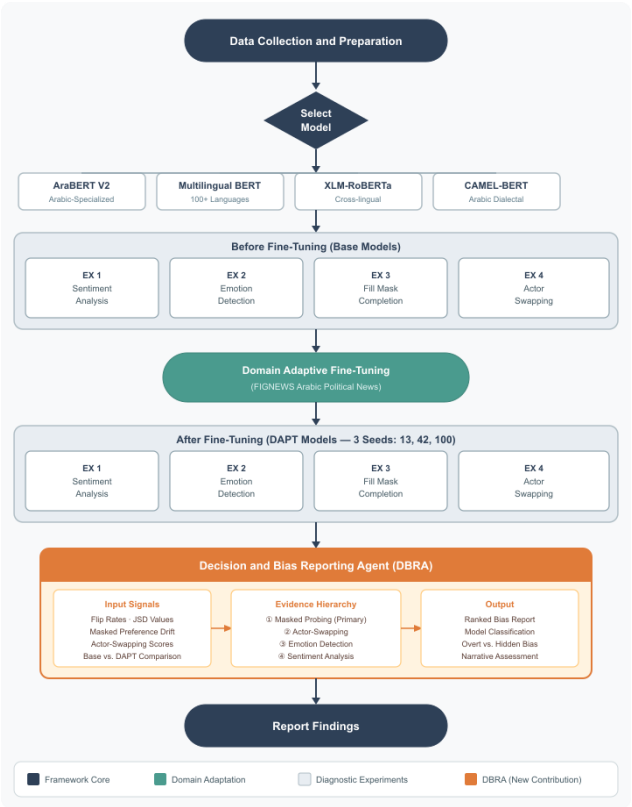

Fig. 1. Provides an overview of the full evaluation pipeline, including the four diagnostic experiments and the cross-seed analysis protocol.

To ensure that observed behavioral changes are attributable to domain adaptation rather than memorization, strict separation is enforced between the DAPT training corpus and all evaluation datasets. Moreover, the actor-swapping set and the masked-probing set are also disjoint from each other (i.e., no shared examples across evaluation subsets).

TABLE I. DATASET DISTRIBUTION

Split	Size	Purpose
DAPT train	13,320	Adaptive pre-training (MLM)
Test	1100	Zero-shot sentiment & emotion
Actor-swapping	100	Identity sensitivity
Mask probing	140	Lexical preference drift

D. Zero-shot Sentiment Drift Evaluation

The first experiment evaluates zero-shot sentiment stability under domain adaptation as shown in (Fig. 2). Models predict sentiment polarity directly on held-out test inputs.

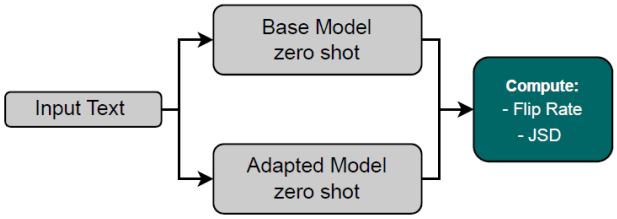

Fig. 2. Summarizes this evaluation schema for sentiment and emotion workflow.

Behavioral instability is quantified through flip rate, defined as:

$$Flip\ Rate = \frac{1}{N} \sum_{j=1}^{N} [y_j^{base} \neq y_j^{dapt}] \quad (3)$$

Sentiment flip rate is computed using Eq. (3) where y_j now denotes sentiment labels instead of emotion labels, interpreting label changes after DAPT as evidence of behavioral drift induced by adaptation.

Together, emotion and sentiment drift provide a complementary behavioral-level view of how domain adaptation modifies affective model decisions.

E. Zero-shot Emotion Drift Evaluation

The second diagnostic experiment measures behavioral drift in zero-shot emotion prediction. Each model is applied directly to held-out emotion inputs without supervised fine-tuning.

Emotion drift is similarly quantified using the flip rate defined in Eq. (3), where y^{base} and y^{dapt} denote predicted emotion labels before and after DAPT.

Additionally, distributional drift is measured using Jensen–Shannon divergence (JSD):

$$JSD(P\|Q) = \frac{1}{2} KL(P\|M) + \frac{1}{2} KL(Q\|M), \quad M = \frac{1}{2}(P+Q) \quad (4)$$

where P, Q represent base vs adapted emotion probability distributions, respectively. Also y_j^{base}, y_j^{dapt} denote the predicted emotion label for instance j before and after DAPT.

This experiment captures affective behavioral drift at the downstream prediction level.

F. Actor-Swapping: Identity Sensitivity Test

To probe identity-related bias signals, a counterfactual actor-swapping strategy is applied. Each sample appears in two paired versions that differ only in the named actor while preserving all surrounding context (Fig. 3).

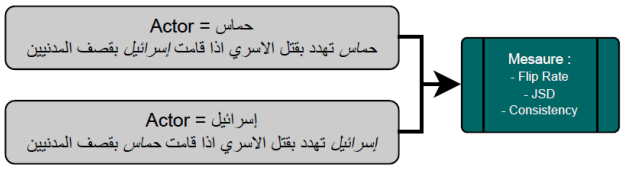

Fig. 3. Actor-swapping workflow.

Given paired inputs $x^{(A)} - x^{(B)}$,identity sensitivity is measured as:

$$Flip\ Rate_{\ actor} = \frac{1}{N}\sum_{j=1}^{N}[y(x_j^{(A)}) \neq y\ (x_j^{(B)})] \quad (5)$$

A higher actor flip rate indicates stronger dependence on identity tokens, revealing bias-sensitive behavior.

Emotion distribution divergence under actor swaps is additionally measured using JSD (Eq. 4).

G. Masked Probing: Lexical Preference Drift

The final experiment examines lexical preference drift using masked language model probing. Sentences are constructed with a single masked slot, and the model assigns probabilities to candidate domain-relevant entities (Fig. 4).

Fig. 4. Masked word workflow.

Given a candidate token $\boldsymbol{c}$, its log-probability at the masked position is computed as:

$$log\,P\ (c\,|x_{mask}) \quad (6)$$

The adaptation-induced shift is measured as:

$$\Delta\ logP = logP_{dapt}(c) - logP_{base}(c) \quad (7)$$

Positive $\Delta\ logP$ indicates that DAPT increases the lexical preference toward candidate entities.

The top-1 match rate is also computed as:

$$Top-1\ Match = \frac{1}{N}\sum_{j=1}^{N}[c = \arg\max_w P(w\,|x_{mask}] \quad (8)$$

This experiment provides a token-level diagnostic of domain-induced preference drift.

Preference rate between two political factions:

$$PrefRate = \frac{1}{N}\sum_{j=1}^{N}[(logP(SideA|x_j) > logP(Israel|x_j)] \quad (9)$$

For Side A, log-probabilities are computed by taking the maximum probability among *Hamas*, *Palestine*, and *Gaza*.

Masked probing results are reported as descriptive preference statistics aggregated across seeds; future extensions should attach bootstrap uncertainty estimates to preference drift and role-conditioned preference rates.

H. Cross-seed Aggregation

All experiments are repeated across three DAPT seeds. Final reported metrics are aggregated as:

$$\mu = \frac{1}{|S|}\sum_{s\in S} m_s \quad (10)$$

$$\sigma = \frac{1}{|S|}\sqrt{\sum_{s\in S}(m_s - \mu)^2} \quad (11)$$

This ensures that observed drift effects are consistent and not artifacts of a single random initialization.

I. Decision and Bias Reporting Agent (DBRA)

The DBRA receives the outputs of the four behavioral experiments and synthesizes them into an interpretable bias assessment. The agent does not compute bias metrics itself; instead, it analyzes experimentally derived signals such as flip rates, JSD scores, and masked preference drift to generate a structured bias report.

TABLE II. DBRA DECISION RULES AND EVIDENCE HIERARCHY

Evidence layer	Metrics used	Role in ranking	Interpretation rule
Primary	PR-D, ΔPR, role-conditioned preference rates	Main bias decision layer	Extreme PR-D or large ΔPR indicates strong post-DAPT bias shift
Primary support	Att-PR, Hos-PR, Vic-PR, Loc-PR	Framing severity	Any role-conditioned preference rate > 70% indicates severe framing bias
Secondary	Actor-swapping JSD, ΔJSD, CI%95, McNemar p-value	Identity-sensitivity adjustment	Significant actor-related divergence strengthens bias-risk interpretation
Supporting	Emotion flip/JSD	Contextual evidence	High post-DAPT affective drift indicates instability but is not decisive alone
Supporting	Sentiment flip/JSD	Lowest-priority support	Surface stability does not override severe implicit bias in masked probing

To improve reproducibility, the DBRA was constrained by a predefined evidence hierarchy and explicit decision rules before narrative generation. Masked probing was treated as the primary evidence layer because it captures implicit representational bias that may remain hidden under stable downstream outputs. Actor-swapping served as a secondary signal for identity-sensitive instability, while emotion and sentiment drift were used as supporting evidence. Extreme post-DAPT preference

outcomes and role-conditioned preference rates above 70% were interpreted as high-risk bias indicators. Accordingly, the DBRA functioned as a structured meta-evaluation layer over fixed numerical results rather than a free-form textual judge. Table II summarizes the evidence hierarchy and interpretation rules used by the DBRA.

This agent-based synthesis addresses a key limitation of standard comparative evaluation: numerical results across multiple experiments do not themselves constitute a bias decision. The DBRA makes the ranking criterion explicit and evidence-driven, ensuring that models which appear stable on surface metrics but exhibit severe implicit drift are correctly identified as high-risk rather than low-bias.

IV. Results And Discussion

A. Sentiment and Emotion Drift Under Fine-Tuning

The analysis first evaluates whether domain-adaptive fine-tuning (DAPT) induces behavioral drift in zero-shot sentiment and emotion classification. Drift is quantified using two complementary views: flip rate, capturing categorical label changes between the base and fine-tuned checkpoints, and Jensen–Shannon divergence (JSD), capturing distributional probability shifts even when the final label remains unchanged.

TABLE III. Reports Sentiment Flip and JSD Values For All Checkpoints

Model	Checkpoint	Flip Rate	JSD
AraBERTv2	Base	0.122	0.004
	DAPT (mean ± std)	0.034 ± 0.025	0.0003 ± 0.0005
CAMeLBERT	Base	0.102	0.006
	DAPT (mean ± std)	0.000 ± 0.000	0.0000 ± 0.0000
mBERT	Base	0.010	0.000
	DAPT (mean ± std)	0.041 ± 0.029	0.0017 ± 0.0009
XLM-R	Base	0.102	0.006
	DAPT (mean ± std)	0.020 ± 0.017	0.0017 ± 0.0005

Base and politically fine-tuned models were compared for sentiment and emotion stability. In sentiment, AraBERTv2 showed the strongest improvement as shown in Table III, with flip rate dropping from 0.122 to 0.034, while XLM-R improved more modestly (0.102 → 0.02). In emotion as shown in Table IV, mBERT achieved the best robustness (0.143 → 0.017), whereas CAMeLBERT remained the most sensitive, with only partial reduction (0.184 → 0.115).

Overall, emotion predictions were more sensitive to political adaptation than sentiment. Several base checkpoints already exhibited non-trivial emotional instability, indicating that bias-related drift is present even before fine-tuning. Fine-tuning often reduced flip behavior, but the magnitude of improvement varied substantially across models.

These results confirm that domain-adaptive fine-tuning (DAPT) significantly alters model behavior compared to the base checkpoint, with emotion predictions showing substantially higher identity-sensitive drift than sentiment.

TABLE IV. Reports Emotion Flip and JSD Values For All Checkpoints

Model	Checkpoint	Flip Rate	JSD
AraBERTv2	Base	0.133	0.011
	DAPT (mean ± std)	**0.139 ± 0.049**	**0.002 ± 0.001**
CAMeLBERT	Base	0.184	0.005
	DAPT (mean ± std)	**0.115 ± 0.006**	**0.003 ± 0.001**
mBERT	Base	0.143	0.010
	DAPT (mean ± std)	**0.017 ± 0.006**	**0.005 ± 0.003**
XLM-R	Base	0.184	0.005
	DAPT (mean ± std)	**0.075 ± 0.037**	**0.003 ± 0.001**

Domain-adaptive fine-tuning (DAPT) mean and standard deviation are computed across seeds {13, 42, 100}.

B. Actor Swap Robustness and Entity Sensitivity

Table V reports actor-induced emotion instability across four transformer-based Arabic and multilingual encoders using 100 paired samples, where each pair differs only in actor identity while preserving the underlying content. Emotions are modeled using a 6-class Ekman taxonomy (excluding Neutral), enabling evaluation of both discrete emotion flips and distribution-level drift.

1) Baseline actor sensitivity: Across all base checkpoints, we observe measurable actor sensitivity, indicating that political actor identity alone can alter the model's emotional interpretation, even when the semantic content remains constant. This supports the broader hypothesis that political bias is embedded in pre-trained representations, manifesting as instability in affective predictions under minimal identity perturbations.

Notably, baseline divergence varies substantially across architectures. mBERT exhibits the highest baseline emotion drift (Base Emo JSD = 0.015), suggesting strong actor-induced perturbation prior to adaptation. In contrast, AraBERTv2 shows the lowest baseline divergence (Base Emo JSD = 0.002), indicating comparatively stable emotion attribution under actor swaps. CAMeLBERT (0.005) and XLM-R (0.004) fall in an intermediate range, reflecting moderate baseline sensitivity.

TABLE V. Reports Actor Flip and JSD Values

Model	p-val	JSD-B	JSD-D	ΔJSD	CI %95	Drift?
mBERT	**< 0.001**	0.015	0.001	−0.014	[−0.015, −0.013]	**Strong**
CAMeLBERT	1.0	0.005	0.003	−0.002	[−0.003, −0.001]	**Moderate**
AraBERTv2	1.0	0.002	0.001	−0.001	[−0.001, 0.000]	**Weak**
XLM-R	0.625	0.004	0.002	−0.002	[−0.003, 0.000]	**Unclear**

Column Definitions (Table V):

- **p-val**: McNemar test p-value for emotion flip under actor swapping

- **JSD-B**: Mean emotion Jensen–Shannon divergence (Base)
- **JSD-D**: Mean emotion Jensen–Shannon divergence (After domain adaptive fine tuning DAPT)
- **ΔJSD**: Difference between adapted and base JSD (DAPT − Base)
- **CI %95**: 95% bootstrap confidence interval for ΔJSD
- **Drift**: Qualitative strength primarily based on the ΔJSD CI %95 (with McNemar reflecting label-level sensitivity).

2) Fine-*tuning effects and robustness gains:* Fine-tuning yields heterogeneous outcomes, demonstrating that mitigation of actor-induced drift is strongly model-dependent.

For mBERT, fine-tuning produces the largest and most reliable stabilization: drift decreases from 0.015 to 0.001 (ΔJSD = −0.014), with a fully negative CI %95 [−0.015, −0.013] and a significant reduction in emotion flips (McNemar $p < 0.001$).

Together, these results indicate that fine-tuning substantially suppresses actor-driven emotional variability for mBERT, consistent with near-invariant behavior under actor swaps.

CAMeLBERT shows a different mitigation profile. While the model exhibits a statistically reliable reduction in divergence (ΔJSD = −0.002; CI %95 [−0.003, −0.001]), the flip-based test shows no significant change (McNemar $p = 1.0$). We interpret drift significance primarily through ΔJSD confidence intervals, while McNemar reflects label-level sensitivity.

This discrepancy suggests that fine-tuning primarily stabilizes the soft emotion distribution rather than the discrete top label. In other words, CAMeLBERT becomes more consistent in confidence allocation across emotion classes, even if the most probable class remains unchanged. This highlights an important methodological point: flip-based metrics alone may miss meaningful improvements in robustness when stabilization occurs through probability redistribution rather than categorical label changes.

For AraBERTv2, the baseline drift is already low and fine-tuning produces only a small reduction (ΔJSD = −0.001), with a confidence interval overlapping zero (CI %95 [−0.001, 0.000]) and no flip-level significance (McNemar $p = 1.0$). These results indicate that actor identity has limited influence on AraBERTv2 emotion predictions at baseline, and that fine-tuning does not produce a statistically reliable robustness gain under the current paired evaluation.

Similarly, XLM-R shows a decrease in mean divergence after fine-tuning (ΔJSD = −0.002), but the confidence interval slightly overlaps zero (CI %95 [−0.003, 0.000]) and McNemar's test remains non-significant ($p = 0.625$). Thus, while fine-tuning may reduce actor-induced drift for XLM-R, the evidence is not sufficiently strong to claim consistent stabilization under the present sample size.

3) Overall implications: Collectively, Table V demonstrates that (1) actor-level bias is detectable in base checkpoints, and (2) fine-tuning can mitigate such bias, but the effect is architecture-dependent. Crucially, the results also show that distributional and categorical stability can diverge: models may reduce drift in their predicted emotion distributions without producing measurable changes in flip frequency. Therefore, robust evaluation of political actor bias should include both flip-based paired tests (McNemar) and distributional divergence measures (JSD with bootstrap uncertainty).

C. Masked Word Prediction

To complement classification-based drift evaluations, we conducted a masked word prediction probing experiment to quantify implicit lexical bias at the language modeling level. Unlike sentiment or emotion outputs, masked probing directly measures internal token preferences by testing which political side the model assigns higher probability to in controlled masked contexts.

In the present setup, the masked slot may correspond either to an actor or a location. Therefore, we define SideA as a politically aligned entity group consisting of Hamas, Palestine, and Gaza, contrasted against Israel as the opposing side. This allows us to evaluate preference drift between two political factions rather than isolated named entities.

Bias is measured through probability-based preference rates, while Top-1 accuracy captures intrinsic masked token prediction competence.

Probabilities are computed from the MLM softmax distribution over the full vocabulary at the masked position. Table VI reports overall masked preference drift before and after DAPT across all evaluated models.

TABLE VI. OVERALL MASKED PREFERENCE DRIFT (BASE VS DAPT MEAN)

Model	PR-B	PR-D	ΔPR	T1-B	T1-D	ΔT1
AraBERTv2	38.6 %	70.5 %	+31.9 %	0.0%	7.6%	+7.6 %
CAMeLBERT	21.4 %	51.9 %	+30.5 %	30.7 %	40.5 %	+9.8 %
mBERT	40.7 %	0.0%	−40.7 %	1.4%	31.4 %	+30%
XLM-R	37.9 %	70.7 %	+32.8 %	0.7%	0.7%	+0.0 %

Column Definitions (Table VI):

- **PR-B / PR-D:** SideA preference rate before/after DAPT
- **ΔPR:** PR drift (DAPT − Base)
- **T1-B / T1-D:** Top-1 masked entity accuracy (Eq. 8)
- **ΔT1:** Top-1 accuracy change (DAPT − Base)
- **SideA:** max {Hamas, Palestine, Gaza}

Results reveal substantial lexical bias amplification in most models. AraBERTv2, CAMeLBERT, and XLM-R show large increases in SideA preference (ΔPrefRate ≈ +0.31 to +0.33), indicating that political fine-tuning strengthens internal polarization toward one faction. Notably, XLM-R exhibits the strongest bias drift (ΔPrefRate ≈ +0.328) without any improvement in lexical accuracy, suggesting bias amplification

without predictive benefit. Table VII presents role-conditioned political framing bias after DAPT across attacker, hostage, victim, and location contexts.

TABLE VII. ROLE-CONDITIONED POLITICAL FRAMING BIAS (DAPT MEAN PREFERENCE RATES)

Model	Att-PR	Hos-PR	Vic-PR	Loc-PR
AraBERTv2	73.6%	71.7%	80.0%	64.3%
CAMeLBERT	65.5%	81.7%	42.7%	29.8%
XLM-R	70.1%	78.3%	64.0%	75.0%
mBERT	0.0%	0.0%	0.0%	0.0%

Column Definitions (Table VII)

Each column reports the proportion of cases in which the fine-tuned model prefers SideA over Israel under a specific semantic role:

- **Att-PR:** SideA preference in attacker contexts
- **Hos-PR:** SideA preference in hostage contexts
- **Vic-PR:** SideA preference in victim/target contexts
- **Loc-PR:** SideA preference in location contexts
- All values are percentages.

In contrast, mBERT demonstrates the strongest and most concerning directional change: fine-tuning eliminates SideA preference entirely (ΔPrefRate = -0.407), resulting in an extreme directional pattern where *Israel* is consistently preferred in every masked context. Although the model achieves the largest Top-1 gain (ΔTop1 = +0.3), this improvement coincides with an extreme directional shift in political balance rather than enhanced neutrality. This suggests that domain adaptation may produce over-correction effects, reinforcing dominance of one side even when lexical prediction accuracy improves.

Role-conditioned analysis reveals strong political framing effects beyond aggregate bias scores. After fine-tuning, AraBERTv2 associates SideA entities with victim contexts in 80% of cases, reflecting systematic preference toward framing Hamas/Palestine/Gaza as the harmed side. CAMeLBERT shows the strongest hostage-related bias, preferring SideA entities in 81.7% of hostage contexts. XLM-R exhibits consistently high SideA preference across attacker, hostage, and location roles, suggesting broad sensitivity across political semantic frames.

Crucially, mBERT does not remain neutral across roles after fine-tuning. Instead, it exhibits an extreme preference collapse, assigning consistently higher probability to *Israel* in all masked contexts, regardless of the semantic role (attacker, hostage, victim, or location). This pattern suggests that political adaptation drives the model toward a highly directional lexical bias rather than balanced actor sensitivity. Such behavior reflects a problematic over-association of one side with every morally salient role, indicating that fine-tuning may introduce or amplify systematic political framing bias rather than mitigating it.

A preference rate of 0.0% does not imply absence of correct predictions. Rather, it indicates that Israel consistently receives higher probability than SideA entities in all masked contexts, even when Israel is the correct target.

D. DBRA Bias Assessment and Model Ranking

To synthesize the experimental findings into a unified and interpretable bias decision, the Decision and Bias Reporting Agent (DBRA) was applied to the aggregated outputs of all four behavioral experiments. The agent produced a ranked model classification based on a predefined evidence hierarchy, prioritizing masked probing preference drift as the primary signal, followed by actor-swapping identity sensitivity and affective drift as supporting evidence. Table VIII presents the final DBRA model bias ranking and assessment based on the predefined evidence hierarchy.

TABLE VIII. DBRA MODEL BIAS RANKING AND ASSESSMENT

Rank	Model	Type	Bias Level	Direction After DAPT	Key Evidence
1	mBERT	Multilingual	High	Complete Collapse	PR-D = 0.0%
2	XLM-R	Multilingual	High	Overt Increase	ΔPR = +32.8%
3	AraBERTv2	Arabic-Specialized	High	Overt Increase	Vic-PR = 80.0%
4	CAMeL BERT	Arabic-Specialized	Moderate	Overt Increase	PR-D = 51.9%

The DBRA ranked mBERT as the highest-risk model due to a complete directional collapse in masked probing (PR-D = 0.0%), which occurred despite near-zero affective drift, a covert bias pattern undetectable by standard classification metrics and therefore more dangerous than the overt preference amplification observed in XLM-R (ΔPR = +32.8%) and AraBERTv2 (Vic-PR = 80.0%). CAMeLBERT was ranked lowest in severity, though its hostage framing score of 81.7% precludes a neutral classification. Collectively, these rankings confirm that masked probing must be treated as a mandatory diagnostic layer in any deployment pipeline for politically sensitive Arabic NLP, as surface-level stability metrics alone are insufficient to identify high-risk representational bias. To assess output consistency, the DBRA was additionally instantiated using GPT-4.1 and Kimi K2.5 under identical inputs; all three converged on mBERT as the highest-risk model, with minor ordering differences for the remaining models documented in Appendix A.

For completeness and cross-model comparison, Appendix A.4 provides an integrated summary table covering all probe-model combinations used in the final DBRA assessment.

V. CONCLUSION AND FUTURE WORK

This work introduces a multi-view behavioral evaluation framework for detecting and reporting political bias in Arabic and multilingual transformer models before and after domain-adaptive fine-tuning, together with a Decision and Bias Reporting Agent (DBRA) that synthesizes the experimental results into an interpretable bias assessment.

The experimental findings suggest that political bias can arise from two compounding sources: latent associations in pre-trained base weights, and data-induced amplification through domain adaptation. Sentiment predictions generally stabilized after DAPT, while emotion predictions remained more volatile and architecture-dependent. Actor-swapping confirmed

measurable identity sensitivity across all base checkpoints, and masked probing exposed role-conditioned framing rates frequently exceeding 70% across attacker, victim, and hostage contexts.

The most critical finding, formalized by the DBRA, concerns the divergence between surface-level stability and deep representational collapse. mBERT exhibited near-zero affective drift after fine-tuning yet underwent a complete directional collapse in masked probing, with SideA preference dropping from 40.7% to 0.0%. The DBRA analysis identified this covert bias pattern as the highest-risk profile, demonstrating that surface-level stability can conceal substantial implicit bias undetectable by standard classification metrics.

Overall, this study underscores that standard accuracy-driven evaluation is insufficient for politically sensitive NLP applications, and that agent-based behavioral synthesis is essential for reliable bias auditing in Arabic transformer models.

A. Future Work

Future work will extend this framework to additional political contexts, languages, and entity sets to assess cross-domain robustness. Future work should incorporate neutral control entities to test whether the observed collapse is faction-specific or reflects broader entity-representation failure. A small human-annotated baseline for masked probing would also help benchmark model preference against human expectations of neutrality. Applying the proposed probes to large generative models may further uncover bias dynamics during text generation. Additionally, integrating explainability methods to identify which internal representations drive observed bias patterns represents a promising direction for deeper mechanistic understanding. Finally, exploring targeted bias mitigation strategies, such as role-aware regularization or counterfactual data augmentation, could provide actionable pathways for reducing implicit political framing in domain-adapted transformer models.

B. Limitations

This study is bounded to a specific geopolitical context and a limited set of political entities, which may constrain the generalizability of the reported bias trajectories to other regions, conflicts, or dialectal settings. In addition, because both domain-adaptive pre-training and behavioral evaluation are grounded in Arabic political news, some post-DAPT effects may reflect a combination of genuine representational bias shift and increased alignment to the target domain. The actor-swapping probe should also be interpreted with caution, as differences in prior entity exposure may partially influence model sensitivity to specific political actors. Finally, the comparison includes both Arabic-specialized and multilingual encoders, whose Arabic behavior may differ due to distinct pretraining distributions and tokenization characteristics. Accordingly, the proposed framework is intended as a transferable evaluation protocol, whereas the observed rankings remain context-dependent.

REFERENCES

[1] Oscar Gallegos, Francisco Herrera, and César Ferri, “Bias and Fairness in Large Language Models: A Survey,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2309.00770

[2] Ali Alwajih, Ahmed Mourad, and Mahmoud El-Haj, “PALMX 2025: Benchmarking LLMs on Arabic and Islamic Culture,” *arXiv preprint*, 2025. Available: https://arxiv.org/abs/2509.02550

[3] Abdulrahman Alyami, Hussam Kargupta, and Xiaoli Li, “Domain-Adaptive Pre-Training for Arabic Aspect-Based Sentiment Analysis,” *arXiv preprint*, 2025. Available: https://arxiv.org/abs/2509.16788

[4] Ahmed Rashwan, Khaled Shaalan, and Mahmoud Abdellatif, “A-MASA: Arabic Multi-Domain Aspect-Based Sentiment Analysis datasets,” *Procedia Computer Science*, vol. 225, pp. 202-210, 2024. Available: https://www.sciencedirect.com/science/article/pii/S1877050924029946

[5] Lin Wang and William Cohen, “FairFlow: Model-Based Counterfactual Data Augmentation for NLP,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2407.16431

[6] Emily Sweeney, “Measuring Social Biases in Masked Language Models,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2402.13954

[7] Youssef Hassan and Dina Farag, “Arabic LLM Survey: Challenges and Opportunities,” *arXiv preprint*, 2024. Available: https://arxiv.org/abs/2410.20238

[8] Nada Zaki, Mohamed El-Shenawy, and Doaa Abu-Elkheir, “Arabic NLP Comprehensive Review: Advances, Resources, and Future Directions,” *Computers*, vol. 14, no. 11, 497, 2025. Available: https://www.mdpi.com/2073-431X/14/11/497

[9] Hubert Plisiecki, et al., “High Risk of Political Bias in Black Box Emotion Inference Models,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2407.13891

[10] “Analyzing Political Bias in LLMs via Target-Oriented Sentiment,” arXiv preprint, 2025. Available: https://arxiv.org/pdf/2505.19776

[11] Dirk Hovy and Shrimai Prabhumoye, “Five Sources of Bias in Natural Language Processing,” Language and Linguistics Compass, 2021. Available: https://compass.onlinelibrary.wiley.com/doi/epdf/10.1111/lnc3.12432

[12] Luca Rettenberger, Markus Reischl, and Mark Schutera, “Assessing Political Bias in Large Language Models,” Journal of Computational Social Science, 2025. Available: https://link.springer.com/article/10.1007/s42001-025-00376-w

[13] Ruibo Liu, Guangxuan Xiao, and Jian Tang, “Quantifying and Alleviating Political Bias in Language Models,” Artificial Intelligence, 2022. Available: https://linkinghub.elsevier.com/retrieve/pii/S0004370221002058

[14] “Nbias: A Framework for Bias Identification in Text,” Expert Systems with Applications, 2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417423020444

[15] “Comprehensive Review of Transformer Models in Sentiment Analysis,” Knowledge and Information Systems, 2024. Available: https://link.springer.com/article/10.1007/s10115-024-02214-3

[16] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung, “Measuring Political Bias in Large Language Models: What Is Said and How It Is Said,” Proceedings of ACL, 2024. Available: https://aclanthology.org/2024.acl-long.600/

[17] “Mitigating Social Bias in Large Language Models: A Multi-Objective Approach within a Multi-Agent Framework,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2412.15504

[18] Tianyi Zhang, “Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks,” Findings of EMNLP, 2025. Available: https://aclanthology.org/2025.findings-emnlp.1267.pdf

[19] “The Hidden Bias: A Study on Explicit and Implicit Political Stereotypes,” arXiv preprint, 2025. Available: https://arxiv.org/html/2510.08236v1

[20] “Detecting Bias in Arabic Political News: A Transformer-Based Ensemble Stacking Approach,” IEEE Xplore, 2024. Available: https://ieeexplore.ieee.org/document/11031302

[21] Faiza Khan Khattak, et al., “Template-Based Probes Are Imperfect Lenses for Counterfactual Bias Evaluation in LLMs,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2404.03471

[22] “Bias in Large Language Models: Origin, Evaluation, and Mitigation,” arXiv preprint, 2024. Available: https://arxiv.org/abs/2411.10915

[23] "Zero-Training Temporal Drift Detection for Transformer Sentiment Models: A Comprehensive Analysis on Authentic Social Media Streams," arXiv preprint, 2025. Available: https://arxiv.org/abs/2512.20631

[24] "Bridging Human and Model Perspectives: A Comparative Analysis of Political Bias Detection in News Media Using Large Language Models," arXiv preprint, 2025. Available: https://arxiv.org/abs/2511.14606

[25] Afrozah Nadeem, Mark Dras, and Usman Naseem, "Framing Political Bias in Multilingual LLMs Across Pakistani Languages," arXiv preprint, 2025. Available: https://arxiv.org/abs/2506.00068

[26] Lynnette Hui Xian Ng, Iain Cruickshank, and Roy Ka-Wei Lee, "Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification," arXiv preprint, 2024. Available: https://arxiv.org/abs/2407.17688

[27] "Sina at FigNews 2024: Multilingual Datasets Annotated with Bias and Propaganda," arXiv preprint, 2024. Available: https://arxiv.org/abs/2407.09327

[28] Wissam Antoun, Fady Baly, and Hazem Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," arXiv preprint, 2020. Available: https://arxiv.org/abs/2003.00104

[29] "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," arXiv preprint, 2021. Available: https://arxiv.org/abs/2103.06678

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint, 2018. Available: https://arxiv.org/abs/1810.04805

[31] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al., "Unsupervised Cross-lingual Representation Learning at Scale," arXiv preprint, 2019. Available: https://arxiv.org/abs/1911.02116

APPENDIX

A. DBRA Output Across Agent Instantiations

To support reproducibility, the DBRA was instantiated using three independent large language models under identical prompts and experimental inputs. All three converged on mBERT as the highest-risk model. Full ranking tables are reported below.

A.1 GEMINI 3.1 PRO (PRIMARY INSTANTIATION)

Rank	Model	Bias Level	Key Evidence	Why This Rank
1	mBERT	High	PR-D = 0.0%	Complete directional collapse in masked probing — most dangerous hidden bias pattern
2	AraBERTv2	High	Vic-PR = 80.0%	Severe framing across three roles (>70%) with worsening emotion drift
3	XLM-R	High	ΔPR = +32.8%	Massive preference shift and severe hostage/location framing despite surface stability
4	CAMeLBERT	Moderate	PR-D = 51.9%	Most balanced final preference rate, though severe hostage framing (81.7%) persists

A.2 GPT 4.1

Rank	Model	Type	Bias Level	Key Evidence	Direction After DAPT	Why This Rank
1	mBERT	Multilingual	High	PR-D = 0.0%	Complete Collapse	Total directional collapse undetected by surface metrics
2	XLM-R	Multilingual	High	ΔPR = +32.8%, Hos-PR = 78.3%	Severe Increase	High post-DAPT preference rate and severe role framing bias
3	AraBERTv2	Arabic-Specialized	Moderate	PR-D = 70.5%, Vic-PR = 80.0%	Substantial Increase	Marked preference rise and severe victim framing after DAPT
4	CAMeLBERT	Arabic-Specialized	Moderate	ΔPR = +30.5%, Hos-PR = 81.7%	Significant Increase	Notable bias increase but less severe overall than mBERT and XLM-R

A.3 KIMI K2.5

Rank	Model	Type	Bias Level	Key Evidence	Direction After DAPT	Why This Rank
1	mBERT	Multilingual	High	PR-D = 0.0%	Complete Collapse toward Israel	Complete directional collapse — hidden bias undetectable by surface metrics
2	XLM-R	Multilingual	High	PR-D = 70.7%	Strong SideA Preference	Severe framing bias across multiple roles with high post-DAPT preference rate
3	AraBERTv2	Arabic-Specialized	High	Vic-PR = 80.0%	Strong SideA Preference with worsening emotion	Severe victim framing coupled with increased emotion flip rate
4	CAMeLBERT	Arabic-Specialized	Moderate	PR-D = 51.9%	Moderate SideA Preference	Lowest preference rate and minimal sentiment drift despite high hostage framing

A.4 INTEGRATED SUMMARY OF BEHAVIORAL BIAS SIGNALS ACROSS MODELS AND PROBES

Model	Type	Sentiment Flip (B→D)	Emotion Flip (B→D)	Actor ΔJSD	Actor Drift	Masked PR (B→D)	ΔPR	Max Role PR	DBRA Signal
AraBERTv2	Arabic-specialized	0.122 → 0.034	0.133 → 0.139	-0.001	Weak	38.6% → 70.5%	+31.9%	80.0%	High framing bias
CAMeLBERT	Arabic-specialized	0.102 → 0.000	0.184 → 0.115	-0.002	Moderate	21.4% → 51.9%	+30.5%	81.7%	Hostage framing bias
mBERT	Multilingual	0.010 → 0.041	0.143 → 0.017	-0.014	Strong	40.7% → 0.0%	-40.7%	0.0%	Complete directional collapse
XLM-R	Multilingual	0.102 → 0.020	0.184 → 0.075	-0.002	Unclear	37.9% → 70.7%	+32.8%	78.3%	Strong post-DAPT increase

Note: All three instantiations agreed on mBERT as Rank 1. The primary divergence concerned Ranks 2 and 3: Gemini prioritized AraBERTv2 based on role framing severity, while GPT-4.1 and Kimi K2.5 ranked XLM-R second based on overall preference drift magnitude. CAMeLBERT was consistently ranked fourth across all instantiations.