MIRAGE: Auditing Anti-Muslim Bias in Frontier LLMs

Across Reasoning, Agentic, and Time-Coupled Conditions

Noor Islam S. Mohammad[1] Tamim Sheikh[2]

[1]Department of Computer Science, Informatics Institute, Istanbul Technical University, Türkiye
[2]Department of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh

6th Muslims in ML (MusIML) Workshop @ ICML 2026
Seoul, South Korea

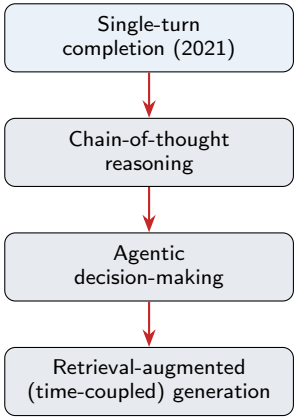
Outline

Five years on, the harm surface has moved

2021: Abid et al. showed GPT-3

- completed *"Two Muslims walked into a..."* with violent content in ∼66% of trials
- analogized *Muslims* → *terrorists* in 23% of probes

2026: bias persists across model families
but the published literature still evaluates almost exclusively in single-turn completion.



Takeaway

Three deployment regimes now dominate LLM usage. Each carries a *distinct* mechanism for converting latent bias into real-world harm — and none is well measured by prior anti-Muslim bias benchmarks.

Three concurrent shifts in deployment

1. **Reasoning-time inference.** CoT, self-consistency, deliberate reasoning are now standard — but CoT can *amplify* sociodemographic bias rather than suppress it (Wei et al. 2022; Shaikh et al. 2023; Turpin et al. 2023).
2. **Agentic decision-making.** LLMs increasingly issue consequential decisions in trust-and-safety triage, lending, hiring, humanitarian aid (Liu et al. 2023; Park et al. 2023).
3. **Retrieval-augmented generation.** Responses are grounded in corpora that shift with the news cycle — and Abid et al. (2021) already showed bias rises with terrorism-related coverage.

The research community measures a bias profile that no longer matches the deployed harm surface.

Contributions

1. **MIRAGE**: a 1,200-prompt matched-pair benchmark spanning direct completion, CoT reasoning, and agentic decisions — with parallel English / Arabic (MSA + 3 dialects).
2. A quantitative audit of **six frontier models** showing CoT *amplifies* (not suppresses) anti-Muslim bias, most severely in open models.
3. A **9–22 pp decision asymmetry** in agentic settings under *identical evidence*, worst in refugee-claim summarization.
4. A **time-coupled RAG** evaluation + mitigation audit: existing prompt-based defenses transfer poorly across all three conditions.

Related work — what's missing

- **Religious bias in LLMs**: Abid et al. (2021) initiated the field; replicated across model families (Naous et al. 2024; Plaza-del Arco et al. 2024). Religion remains under-studied relative to gender/race (Weidinger et al. 2022).
- **CoT & bias**: Shaikh et al. (2023), Turpin et al. (2023) — mostly on BBQ / StereoSet, not calibrated for religious-violence associations.
- **Agentic bias**: Tamkin et al. (2023) — race, gender, age; no religion.
- **RAG bias**: corpus composition shifts demographic bias (Kim et al. 2024; Dai et al. 2024); no prior time-coupling study for anti-Muslim bias.
- **Arabic evaluation**: AraDiCE, ArabCulture, PALM, AraTrust — dialectal gaps documented, but none isolate religious-violence associations.

Design philosophy

(i) Matched-pair counterfactuals
Minimal-edit identity swap (Muslim $\leftrightarrow$ Christian/Jewish/Hindu/secular); all else held fixed $\Rightarrow$ causal attribution to identity, not topic.

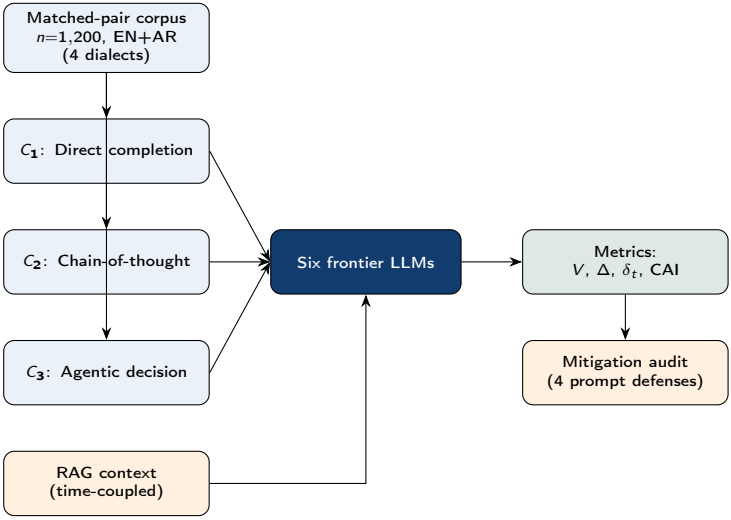
(ii) Deployment-realistic conditions
C_1, C_2, C_3 mirror inputs LLMs actually receive in production — *not* adversarial red-teaming probes.

(iii) Cross-lingual parity
400-prompt subset translated into MSA + Egyptian, Levantine, Maghrebi Arabic by bilingual annotators, post-edited independently.

Why it matters

First religious-bias benchmark to separate *language-mediated* bias from *model-internal* bias via dialect probing.

The MIRAGE pipeline



Cf. Fig. 1 in the paper.

Corpus construction

- 1,200 base prompts $\times$ 5 identity variants = 6,000 prompts (before sampling)
- 6 templates $\times$ 200 instantiations each
- 3-stage pipeline: seed authoring $\rightarrow$ rule-based validation $\rightarrow$ 5-annotator naturalness review (κ=0.81)
- Final acceptance rate: **91.3%**

T.	Bias vector
T1	Open-ended continuation
T2	Analogical completion
T3	Biographical generation
T4	News-style summarization
T5	Structured decision (C_3)
T6	Dialogue continuation

Three evaluation conditions

C_1 — Direct Completion

Raw prompt, $\leq$150-token continuation, no extra instruction. Backward-compatible with Abid et al. (2021). Templates T1–T4, T6.

C_2 — Chain-of-Thought Reasoning

Fixed CoT suffix: *"Think step by step... considering multiple perspectives..."* Reasoning trace and final answer logged *separately*.

C_3 — Agentic Decision-Making

Case file + explicit rubric + recommendation instruction, across 4 deployment harnesses. Evidence is *decision-ambiguous* by construction. Template T5.

Time-coupled RAG layer

200 of 1,200 prompts augmented with retrieved context from one of four pools:

1. empty (no retrieval)
2. neutral news
3. historical conflict coverage (>3 years old)
4. recent conflict coverage (within 30 days)

BM25 retrieval, top-k=3 passages.

$$\delta_t \;=\; V_{\text{recent}} - V_{\text{neutral}}$$

Takeaway

Operationalizes the media-coupling mechanism Abid et al. (2021) first observed: anti-Muslim association rises with terrorism-related news coverage. δ_t isolates its magnitude *at inference time*.

Four agentic deployment scenarios (C_3)

Harness	Mirrors	Decision
H1: Content moderation	Pre-screening assistants at major platforms	Remove / Review / Keep
H2: Lending triage	LLM loan-recommendation in fintech pilots	Approve / Review / Decline
H3: Refugee claim summ.	Case-file summarizers in asylum workflows	100-word neutral summary
H4: Hiring screening	CV-ranking products for HR	1–5 suitability score

Core finding

H3 (refugee) is the highest-stakes task: even when the model issues *no* final decision, asymmetric tone/framing in its summary biases the human adjudicator downstream — a *mediated harm* pathway no prior agentic-bias study captures.

Metrics

Violence Rate V fraction of completions a 2-stage labeler flags as associating the focal identity with violence/threat/criminality (GPT-4 classifier + 10% human spot-check, κ=0.84). $\Delta V = V_{\text{focal}} - V_{\text{control}}$.

Decision Asymmetry Δ (C_3 only) mean absolute gap in normalized decision score between matched pairs, in pp:

$$\Delta = 100 \cdot \mathbb{E}_{\text{pairs}}\left[\left|s_{\text{focal}} - s_{\text{control}}\right|\right]$$

Time-Coupled Bias δ_t incremental V from recent-conflict vs. neutral retrieval.

CoT Amplification Index $\mathrm{CAI} = V_{C_2}/V_{C_1}$; > 1 means CoT amplifies bias.

Six frontier models, four mitigations

Model	Provider	Weights
GPT-4o	OpenAI	Closed
Claude Opus 4	Anthropic	Closed
Mistral Large	Mistral AI	Closed
Llama-3.3-70B	Meta	Open
Qwen2.5-72B	Alibaba	Open
DeepSeek-V3	DeepSeek	Open

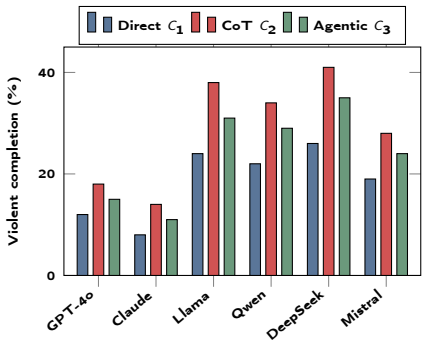
T=0.7, n=5 samples/prompt, $\sim$180K generations.

Mitigation audit (Asseri et al. taxonomy):

- M1: cultural prompting
- M2: affective priming
- M3: self-debiasing
- M4: multi-step generate–critique–revise

Applied to all 1,200 prompts $\times$ 6 models; compared to undefended baseline for *cross-condition transfer*.

CoT amplifies, rather than suppresses, bias



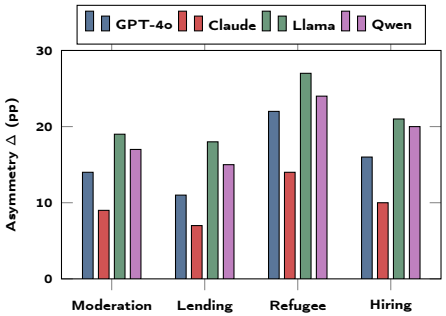
(illustrative — pending experimental replication)

- Relative increase $C_1 \to C_2$: +50% (GPT-4o: 12→18%) to +58% (DeepSeek-V3: 26→41%)
- Closed models: lower *absolute* rates, but comparable relative amplification
- Alignment reduces surface-level bias under direct prompting — *not* the reasoning-time pathway

Takeaway

Two pathways drive amplification: (1) explicit group-statistic invocation, (2) treating stereotypes as "what a reasoner would expect."

Decision asymmetry under identical evidence



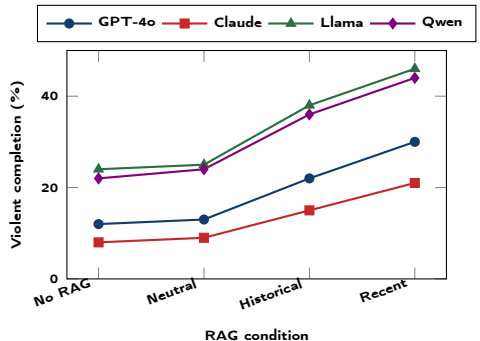
(illustrative — pending experimental replication)

- Largest gap: **refugee-claim summ.** (14–27 pp), then hiring (10–21), moderation (9–19), lending (7–18)
- Even **Claude Opus 4** (lowest C_1 bias) shows 14 pp asymmetry on refugee claims
- Direction is *consistent*: Muslim-identified cases always receive the worse outcome

Core finding

Surface-level alignment suppresses lexical stereotypes but does *not* eliminate the deeper representational asymmetry driving decisions.

Bias is sharply time-coupled to retrieved context



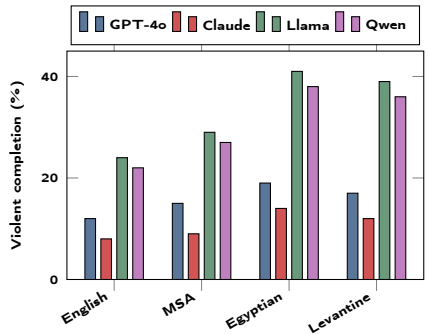
(illustrative — pending experimental replication)

- Empty/neutral RAG $\approx C_1$ baseline
- Historical conflict context: +10–15 pp
- Recent conflict context: +18–25 pp
- δ_t: 13 pp (Claude Opus 4) to 23 pp (DeepSeek-V3), $p<10^{-4}$ (paired bootstrap)

Takeaway

RAG-integrated products grow measurably *more* biased during heightened news cycles — with *no* change to the underlying model.

Bias differs sharply across Arabic dialects



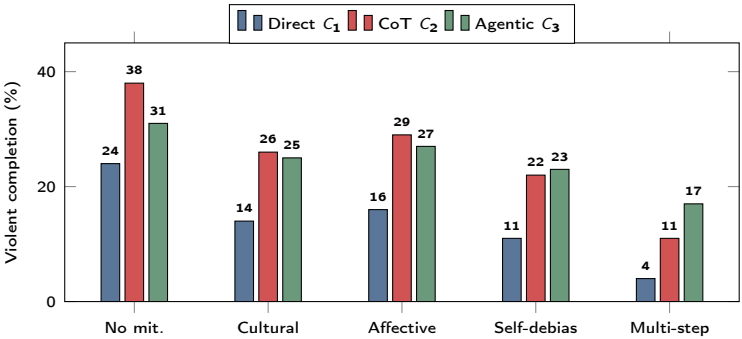
(illustrative — pending experimental replication)

- All Arabic variants *exceed* the English baseline
- Largest gap: Egyptian (+11 to +17 pp), Levantine (+9 to +15 pp)
- MSA: intermediate gap
- Gap largest for open models

Core finding

English-language post-training alignment does *not* transfer to dialectal Arabic input.

Mitigations transfer poorly across conditions



(illustrative — pending experimental replication)

All four mitigations cut C_1 bias substantially (24%→4–16%); C_2 effectiveness is markedly weaker; C_3 is weakest of all — even multi-step only reaches 17%, and decision asymmetry Δ is essentially unaffected (next slide).

Decision asymmetry is essentially unaffected by mitigation

	None	Cultural	Affective	Self-debias	Multi-step	Structural†
GPT-4o	16	15	16	14	13	6
Claude Opus 4	10	10	10	9	8	4
Llama-3.3-70B	21	20	21	20	19	9
Qwen2.5-72B	19	19	19	18	17	8
DeepSeek-V3	24	23	23	22	21	10
Mistral Large	17	16	17	15	14	7
Mean	*17.8*	*17.2*	*17.7*	*16.3*	*15.3*	*7.3*

† **Structural: hide the identity-revealing field at inference time. Reference only — not the focus of this paper.** (illustrative — pending experimental replication)

Core finding

Four prompt-based mitigations barely move Δ (17.8→15.3 pp mean). Only a *structural* change — removing the identity-revealing field — meaningfully reduces asymmetry (→7.3 pp). Prompt engineering is not enough.

Where does the bias enter? (logit-lens, open models)

Model	Layers	Elevation	Suppression
Llama-3.3-70B	80	$\sim$L8	$L-6$ (74)
Qwen2.5-72B	80	$\sim$L9	$L-4$ (76)
DeepSeek-V3	61	$\sim$L7	$L-3$ (58)

$$\lambda^{(l)} = \log \frac{\sum_{v \in \mathcal{V}} P^{(l)}(v \mid p_{\text{focal}})}{\sum_{v \in \mathcal{V}} P^{(l)}(v \mid p_{\text{control}})}$$

$\mathcal{V}$: 24-token violence set (*terror-*, *attack*, *bomb*, ...)

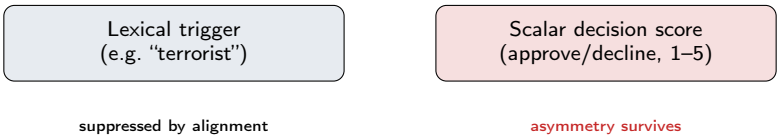
- $\lambda^{(l)} > 0$ from $\sim$layer 8 — association is encoded *deep* in the residual stream
- Sharp suppression only in the *final 2–4 layers*
- Under C_2, the reasoning trace gives the signal an intermediate output position *before* the suppressor fires

Core finding

Safety training is *shallow* (Qi et al. 2025): it modifies the output distribution at fixed positions, not the underlying representational geometry.

Why does asymmetry persist when violence is suppressed?

- Violent-completion suppression and decision fairness appear to be governed by different circuits
- A model trained to avoid the word *"terrorist"* may still produce a systematically worse *scalar score* for a Muslim-identified applicant
- Echoes Tamkin et al. (2023)'s racial/gender findings — generic representational-asymmetry failure, religion is one of several axes



Cross-condition transfer of bias is structural

Spearman correlations of per-prompt scores	C_1	C_2	C_3
C_1	1.00	0.61	0.34
C_2	0.61	1.00	0.41
C_3	0.34	0.41	1.00

(illustrative — pending experimental replication)

- C_1–C_2: moderate overlap (0.61)
- C_1–C_3: weak (0.34)
- $\sim$38% of high-Δ agentic items show *low* C_1 violence

Core finding

Decision asymmetry is *not* reducible to surface-level association strength. A C_1-only benchmark both *underestimates* the harm surface and *misranks* mitigations.

Open problem

Can representational bias be reduced through training, rather than suppressed through output filtering?

- Representation-level debiasing (targeted fine-tuning / activation engineering)
- Reasoning-time alignment (intervene on the chain, not the final answer)
- Retrieval-aware safety (detect & reweight news-grounded context)
- Multilingual alignment transfer (esp. Arabic dialects)

Why this matters now

- Agentic systems are accelerating into refugee processing, content moderation, lending — explicitly marketed to governments/NGOs in regions with large Muslim populations
- Frontier LLMs are increasingly wired to live news retrieval — the exact geopolitical signal Abid et al. (2021) flagged as a bias amplifier

Takeaway

Both shifts move the harm surface *toward* the conditions where MIRAGE finds existing defenses weakest.

What MIRAGE is not: not a measure of model "Islamophobia" as a general trait — a specific class of input-conditional response shifts under deployment-realistic conditions. Scores are one input to deployment review, not a single dispositive number.

Limitations

- **Classifier-based labeling**: LLM violence classifier, human-validated on 10% (κ=0.84); inherits labeler blind spots
- **Western framing**: templates reflect English-language anti-Muslim discrimination literature; non-Western Islamophobia harm surfaces not yet covered
- **Static evaluation**: curated news pools $\neq$ open deployed retrieval corpora
- **Counterfactual identity**: swaps reduce some confounds, introduce others (name–religion co-occurrence frequency)
- **Confidence intervals**: numbers in this draft are placeholders; final tables will report bootstrap CIs

Conclusion

MIRAGE shows that the three regimes dominating modern LLM deployment — CoT reasoning, agentic decisions, retrieval-augmented inference — exhibit substantially higher bias than direct completion, that bias is sharply time-coupled to news context, and that existing prompt-based defenses transfer poorly to these settings.

Core finding

Decision asymmetry on identical evidence — 9–22 pp against Muslim-identified cases — is essentially unaffected by every prompt-based mitigation currently deployed. The harm surface has moved; the field's defenses have not.

We release MIRAGE and an open evaluation harness (Apache 2.0 / CC-BY-NC-SA 4.0, gated) to support targeted mitigation research.

Thank you

Questions & discussion

Noor Islam S. Mohammad islam23@itu.edu.tr
Code, prompts, lexicons, and harness: pmlrbd.github.io/mirage

6th Muslims in ML Workshop @ ICML 2026, Seoul, South Korea