Invisible Conflicts: Media Coverage Asymmetry and Categorical Failure in LLM Conflict Forecasting

Poli Nemkova [1]

Abstract

Media coverage of armed conflict is deeply asymmetric: we document a 224× gap between the most and least covered conflict zones in English-language media across 22 countries from 2020–2026, measured as distinct news articles per ACLED conflict event (GDELT, deduped URL-level). We investigate whether this asymmetry shapes LLM parametric knowledge by evaluating zero-shot conflict escalation forecasting across all 22 countries using two LLM backbones (Llama-3.3-70B, GPT-4o). Results reveal a more troubling pattern than a simple performance gradient: LLMs do not forecast conflict — they categorize it. Models apply near-universal escalation priors to under-covered active conflict zones (recall = 0.918) while exhibiting near-zero recall on over-covered zones (recall = 0.231), suggesting that parametric knowledge from media coverage produces conflict-zone labels rather than dynamic understanding of when escalation actually occurs. A system that predicts escalation every month for six years in DRC or Myanmar is not forecasting — it is reciting a category. Critically, this failure cannot be resolved at inference time: augmenting prompts with structured ACLED event count evidence degrades under-covered performance below the trivial always-predict-escalation baseline for both models (Llama F1: 0.329 vs. 0.344; GPT-4o F1: 0.211 vs. 0.344), confirming that the deficit is in parametric knowledge and not in evidence access. This categorical failure is arguably more dangerous than low F1 for humanitarian early warning systems, where missing a crisis costs lives. We argue that the NLP community must treat geographic conflict coverage asymmetry as a first-class fairness problem: under-covered populations receive not just less accurate AI, but qualitatively different AI that cannot distinguish stable from escalating periods within their conflicts. We call for coverage-stratified benchmarking, conflict NLP datasets for under-covered zones, and training data documentation standards for geographic conflict representation.

1. Introduction

Artificial intelligence systems for humanitarian decision support are proliferating rapidly. Conflict early warning platforms, displacement forecasting tools, and resource allocation systems increasingly rely on large language models (LLMs) as reasoning engines (Rost & Ronco, 2026; UNHCR, 2022) — yet the epistemic foundations of these systems remain poorly understood. LLMs acquire world knowledge through exposure to training corpora drawn predominantly from English-language web text, and that text does not represent the world's conflicts equally. This paper asks a precise question: does the asymmetry in how conflicts are covered by media translate into asymmetry in how LLMs reason about them — and if so, what are the consequences for humanitarian AI deployment where the stakes are highest?

The unequal coverage of global conflicts by Western media is a well-documented phenomenon. Galtung & Ruge (1965) established that proximity, cultural consonance, and elite-nation involvement systematically shape which events receive coverage. Entman (1993) showed that framing — what is selected and made salient — reflects the cultural and political context of the media producing it. Quantitative work has confirmed that these dynamics have real-world consequences: Eisensee & Strömberg (2007) demonstrated that US disaster relief is driven by news coverage to such a degree that a disaster in Africa requires roughly 40 times more casualties than one in Eastern Europe to receive equivalent television coverage. More recent computational work has extended these findings to the digital era, documenting

[1]Department of Computer Science, University of North Texas, Denton, Texas, USA. Correspondence to: Poli Nemkova <poli.nemkova@unt.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

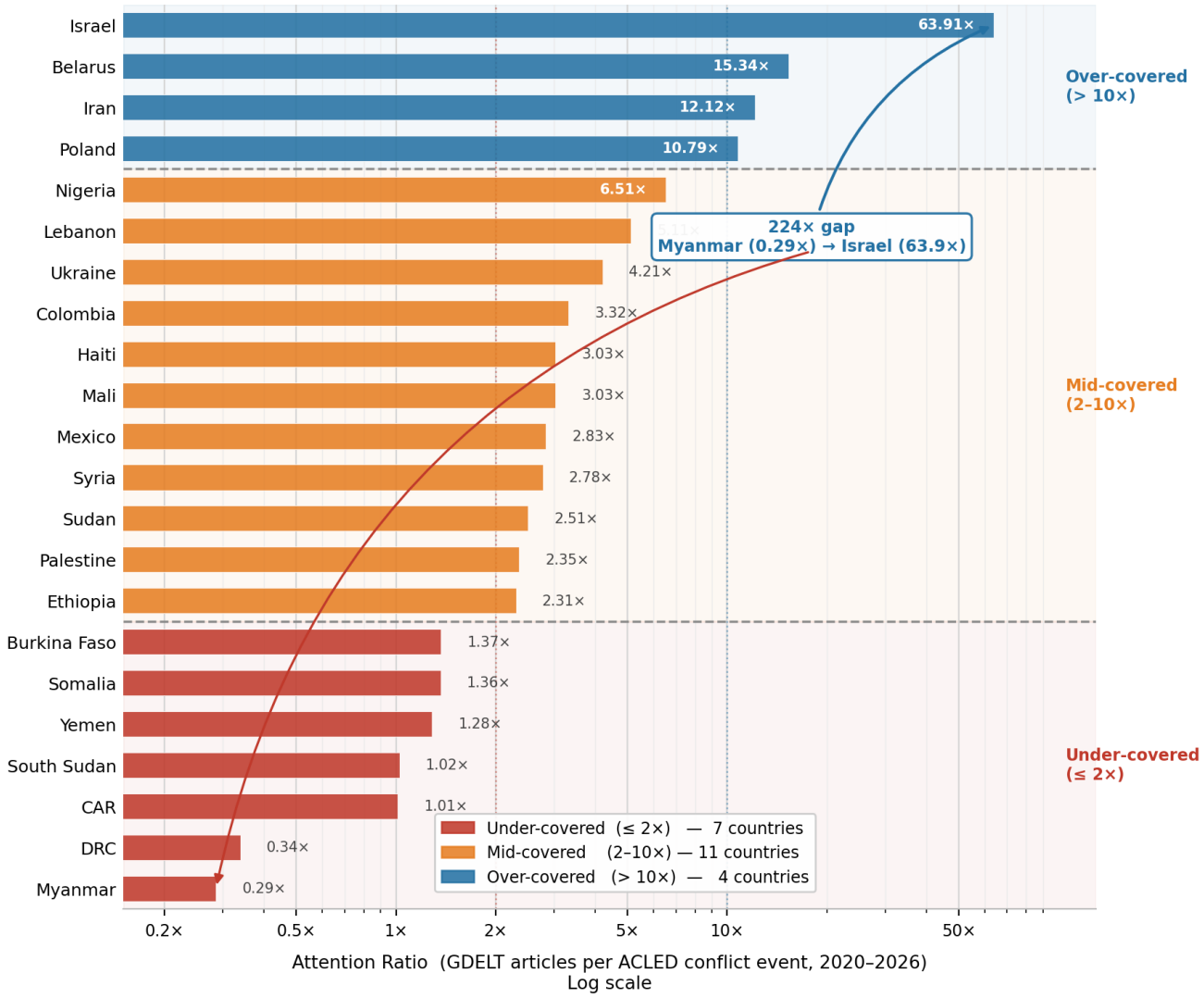

Figure 1. Media attention ratio across 22 conflict zones, measured as GDELT English-language news articles per ACLED conflict event, 2020–2026 (log scale). A 224× gap separates the most and least covered zones (Myanmar 0.29× → Israel 63.9×). Coverage tiers: Under (≤2×), Mid (2–10×), Over (>10×).

systematic under-coverage of conflict zones in Sub-Saharan Africa and Southeast Asia relative to Europe and the Middle East, even when controlling for conflict intensity (Croicu & von der Maase, 2024). What has not been examined is whether this asymmetry propagates into AI systems trained on media-derived corpora — and whether it produces qualitative differences in how those systems reason about conflict, not merely quantitative differences in accuracy.

Computational approaches to conflict forecasting have grown substantially, with machine learning models now achieving meaningful predictive accuracy on civil war onset and escalation tasks (Ward et al., 2013; Blair & Sambanis, 2020). What these approaches share, however, is reliance on structured datasets and hand-engineered features rather than the parametric knowledge encoded in LLMs — leaving open the question of whether LLMs can serve as general-purpose conflict reasoning engines, and for whom.

The connection between training data composition and model behavior is well established in the NLP literature. Bender et al. (2021) documented how LLMs trained on web text inherit the demographic and geographic biases of that text. Dodge et al. (2021) showed that C4 — a widely used pretraining corpus — substantially overrepresents English-language content from a small set of domains and geographies, and that blocklist filtering disproportionately removes text from and about minority communities. Gebru et al. (2021) argued for systematic documentation of training data provenance as a prerequisite for responsible deployment.

Beyond corpus composition, geographic bias manifests directly in model outputs: Manvi et al. (2024) demonstrated that LLMs exhibit consistent and measurable geographic bias across a range of prediction tasks, with performance degrading systematically for regions underrepresented in training data. Multilingual NLP research has further shown that model performance correlates strongly with resource availability in the training corpus (Wu & Dredze, 2020). Despite this growing awareness of training data bias, the specific question of geographic conflict coverage asymmetry as a source of qualitative reasoning failure has not been addressed (Nemkova et al., 2025a;b).

We present four contributions. First, we quantify the media attention gap across 22 active conflict zones using ACLED ground-truth event data (Raleigh et al., 2010) and GDELT media coverage (Leetaru & Schrodt, 2013) spanning 2020–2026, documenting a 224× gap between the most and least covered conflicts in English-language media (Myanmar 0.29× vs. Israel 63.9×). Second, we provide behavioral evidence that this asymmetry produces not a gradient of forecasting skill, but a qualitative difference in reasoning mode: models apply near-universal escalation priors to under-covered active conflict zones (Llama recall = 0.918) while failing categorically on over-covered zones where they apply equally static NO priors (GPT-4o recall = 0.000). This pattern reveals that LLMs have learned which countries are conflict zones from media exposure, but have not learned when escalation occurs within them — precisely because that fine-grained temporal signal is absent from training data for under-covered regions. Third, we demonstrate that this failure cannot be resolved at inference time: augmenting prompts with structured ACLED event counts degrades under-covered performance below the trivial Always-YES baseline for both models, confirming that the deficit is in parametric knowledge and not in evidence access. Fourth, we argue that this constitutes a compounding equity problem of a specific character: the communities most affected by armed conflict and least covered by media receive AI systems that cannot distinguish a stable month from an escalating one (Bender et al., 2021; Eisensee & Strömberg, 2007). For humanitarian early warning — where the cost of a missed crisis far exceeds the cost of a false alarm — a system that predicts escalation every month is not a safety net; it is noise. We call on the NLP community to treat coverage-stratified benchmarking and conflict NLP dataset creation for under-covered zones as priority research directions (Gebru et al., 2021), drawing an explicit parallel to the mobilization around low-resource machine translation in the preceding decade.

2. Measuring the Gap

2.1. Data Sources

We operationalize media attention using two independent data sources that together provide ground-truth conflict intensity and media coverage volume across 22 active conflict zones from January 2020 through February 2026.

Conflict ground truth: ACLED. The Armed Conflict Location and Event Dataset (Raleigh et al., 2010) provides event-level records of political violence and protest events, coded from news reports, NGO communications, and government sources. We use ACLED as our measure of actual conflict intensity, aggregating monthly event counts per country. ACLED's systematic global coverage and real-time coding make it the standard ground truth for conflict research (Muchlinski et al., 2016). Crucially, ACLED itself is subject to reporting bias — events in countries with less media access are systematically under-coded (Raleigh et al., 2010) — meaning our attention gap estimates are *conservative*: the true asymmetry between conflict intensity and media coverage is likely larger than we measure.

Media coverage: GDELT. The Global Database of Events, Language and Tone (Leetaru & Schrodt, 2013) indexes English-language news articles from thousands of outlets worldwide, coding events using the CAMEO conflict taxonomy. We query the GDELT BigQuery dataset for conflict-relevant CAMEO event codes (codes 14–20, 173, 180–196, 200–203) and aggregate monthly counts of *distinct source URLs* per country. URL-level deduplication is essential: each news event in GDELT generates multiple rows (one per actor pair, one per event code), so raw row counts massively overcount article volume. We use distinct URLs as the closest available proxy for distinct articles. We note that GDELT's coverage is substantially English-language biased, systematically under-indexing local-language reporting on conflicts in Sub-Saharan Africa and Southeast Asia. This is not a limitation we correct for — it is precisely the bias we seek to measure, since English-language web text constitutes a substantial portion of LLM pretraining corpora (Dodge et al., 2021).

Country selection. We select 22 countries spanning five regions: Sub-Saharan Africa (Sudan, Ethiopia, Somalia, DRC, South Sudan, Nigeria, Mali, Burkina Faso, Central African Republic), Southeast Asia (Myanmar), Eastern Europe (Ukraine, Belarus, Poland), Middle East and North Africa (Israel, Palestine, Lebanon, Iran, Yemen, Syria), and Latin America (Colombia, Haiti, Mexico). Countries were selected to provide structural diversity across conflict types — interstate war, post-coup civil conflict, protracted insurgency, multi-actor fragmented conflict, and political crisis — while ensuring sufficient ACLED event density for meaningful monthly statistics. Countries with near-zero ACLED

event counts (Belarus, Poland) are retained as implicit controls for the model behavior analysis in Section 4.

2.2. The Attention Ratio

We define the *media attention ratio* for country c as:

$$\text{AttentionRatio}(c) = \frac{\overline{A}_{\text{GDELT}}(c)}{\overline{E}_{\text{ACLED}}(c)} \quad (1)$$

where both quantities are averaged over the 74-month study period. This ratio captures media attention *per unit of conflict activity*: a high ratio indicates that each conflict event generates many articles (over-covered); a low ratio indicates that conflict events generate few articles (under-covered).

2.3. Results: A 224× Attention Gap

Figure 1 reports the attention ratio for all 22 countries, sorted from least to most covered. The gap between the most and least covered active conflict zones spans **224**×: Myanmar receives 0.29 articles per conflict event while Belarus receives 15.3× and Israel 63.9×. Restricting to countries with substantial active conflict (ACLED events $>$ 1,000 over the study period) narrows the comparison but preserves the essential finding: Myanmar (0.29×) and South Sudan (1.02×) receive approximately **15–220× less media attention per conflict event** than Ukraine (4.21×), Lebanon (5.11×), or Iran (12.1×).

Three structural patterns are visible. First, the seven most under-covered active conflict zones are without exception located in Sub-Saharan Africa or Southeast Asia: Myanmar, DRC, South Sudan, CAR, Yemen, Somalia, and Burkina Faso. Second, the most over-covered zones — Israel, Belarus, Iran, Poland — share a common feature: their media coverage reflects geopolitical salience rather than conflict intensity. Poland's coverage reflects spillover from the Ukraine refugee crisis rather than domestic armed conflict. Third, Ukraine's ratio (4.21×) falls in the mid-covered tier despite being the largest absolute recipient of media attention in our dataset (1.1 million articles), because its ACLED event count is also the largest (266,664 events). High absolute coverage does not necessarily imply high relative coverage.

2.4. Methodological Notes

Two limitations deserve explicit acknowledgment. First, GDELT indexes primarily English-language sources. Local-language reporting on conflicts in Bambara, Hausa, Tigrinya, or Burmese is substantially absent, meaning our attention ratios understate the true coverage available to locally-informed observers. Since LLM pretraining corpora also predominantly reflect English-language sources (Dodge et al., 2021), this shared bias is precisely what we seek to characterize rather than correct. Second, the attention ratio captures volume but not quality of coverage: a country may receive many articles but superficial treatment, or few articles but detailed analysis. We treat volume as a proxy for training signal density, acknowledging that this is an approximation. Future work employing document-level analysis of coverage depth could refine this measure.

3. Experimental Setup

We evaluate zero-shot conflict escalation forecasting across all 22 countries using two LLM backbones: Llama-3.3-70B (Grattafiori et al., 2024) served via the Groq inference API, and GPT-4o (OpenAI, 2023). The zero-shot parametric design is deliberate: we provide only the country name and prediction period, with no evidence bundle, no few-shot examples, and no contextual data. Any performance difference across coverage tiers must therefore reflect parametric knowledge — what each model learned during pretraining — rather than evidence quality or prompting strategy.

The zero-shot design follows the evaluation paradigm established in Brown et al. (2020): by withholding in-context examples, we isolate what the model has internalized during pretraining from what it can infer from prompt context. This is the appropriate design for our research question, which concerns parametric knowledge rather than in-context reasoning.

The task is binary escalation forecasting: given a country and a 30-day prediction window, predict whether armed conflict fatalities will exceed 130% of the recent baseline. Ground truth is derived from ACLED as described in Section 2. We evaluate across the full 1,628-case dataset (74 cases per country, 2020–2026), using temperature $T = 0$ for deterministic output and a single run per model.

We compare against two baselines. The **majority-class baseline** predicts the most frequent label per country; because all 22 countries have escalation rates below 0.5, this always predicts NO, yielding F1 = 0.000 everywhere and providing no discriminative signal. The **Always-YES baseline** predicts escalation on every case, maximizing recall at the cost of precision. Because the majority-class baseline is trivially uninformative, Always-YES serves as the operative comparison: it represents the simplest possible strategy for catching escalation events in chronically conflicted zones, and any claimed improvement should be measured against it.

To test whether structured evidence can close the performance gap on under-covered zones, we run a second condition (**evidence-augmented**) in which each prompt includes the three-month ACLED event count trend preceding the prediction window. This ablation directly addresses the hy-

pothesis that the parametric condition's failures stem from insufficient information rather than miscalibrated priors. We report evidence-augmented results for both models in Section 6.

4. LLM Behavior Under Attention Asymmetry

4.1. Results: Two Models, Two Failure Modes

Table 1 reports per-tier performance for both models alongside the Always-YES baseline. The headline finding is not a gradient of forecasting skill correlated with attention ratio — it is a qualitative divergence in how each model handles uncertainty about unfamiliar conflict zones, relative to a trivial predictor.

Finding 1: Both models fail to forecast — they categorize. Neither model exhibits the temporal discrimination required for forecasting: neither reliably distinguishes an escalating month from a stable month within a given conflict zone. Instead, both apply country-level priors that are largely stable across the 74-month observation window. Llama defaults to YES for active conflict zones (under-covered recall = 0.918), behavior that closely mirrors the Always-YES baseline (recall = 1.000) and suggests the model has learned a conflict-zone label rather than a dynamic signal. GPT-4o is more conservative overall but follows the same categorical logic.

Finding 2: The models exhibit opposite failure modes on under-covered zones. Llama achieves recall = 0.918 on under-covered countries by predicting YES on nearly every case — a blanket prior that catches actual escalation events at the cost of massive over-prediction (precision = 0.257). GPT-4o's recall on the same countries is 0.518, suggesting its richer parametric knowledge has introduced a conservative NO prior for less-familiar countries. Both are wrong in complementary ways: Llama's under-covered F1 (0.401) exceeds the Always-YES baseline (0.344) by only 0.057 points, gained entirely through marginal precision improvements rather than genuine temporal discrimination.

Finding 3: GPT-4o produces zero recall on the over-covered tier. GPT-4o predicts NO on every case in Belarus, Iran, Israel, and Poland (F1 = 0.000, recall = 0.000), falling below the Always-YES baseline (F1 = 0.084). This is the strongest evidence of categorical reasoning: GPT-4o has learned that these countries do not fit the armed conflict zone category in the traditional sense, and applies that label unconditionally across all 296 cases — missing the 13 actual escalation events that did occur. More capable parametric knowledge produces more confident miscategorization.

4.2. Two Failure Modes, One Equity Problem

The divergence between models is meaningful precisely because it cannot be resolved by model selection. Choosing Llama maximizes recall on under-covered zones but at a false positive rate that would overwhelm any operational humanitarian system. Choosing GPT-4o reduces false positives but at the cost of missing the majority of actual crises in both under-covered and over-covered zones. Neither model provides a deployable signal; the choice between them is a choice between failure modes, not between adequate and inadequate performance.

5. The Compounding Equity Problem

The behavioral findings of Section 4 describe a specific injury to specific populations. The seven most under-covered active conflict zones — Myanmar, DRC, South Sudan, CAR, Yemen, Somalia, and Burkina Faso — are simultaneously among the most conflict-affected populations on earth, the least covered by media (0.29–1.37 articles per ACLED event), and the contexts where AI-assisted early warning is most urgently needed and least institutionally resourced (Rost & Ronco, 2026). This conjunction is not coincidental: the same dynamics that make these conflicts less newsworthy to Western media (Galtung & Ruge, 1965; Eisensee & Strömberg, 2007) make them less represented in the corpora LLMs train on (Dodge et al., 2021). These populations are first ignored by the press, then ignored by AI systems trained on the press.

The deployment risks we identify are not hypothetical. Vinck et al. (2019) documented that humanitarian organizations increasingly rely on predictive tools without adequate understanding of their failure modes, and called for systematic evaluation of AI reliability across the geographic contexts where these tools are deployed. Our results provide precisely that evaluation — and the findings are alarming.

5.1. Two Channels, One Equity Problem

Channel 1: Training signal deficit. LLMs have fewer examples of escalation patterns, actor behavior, and structural drivers for under-covered countries. The result is visible in our data: neither model can distinguish a stable month from an escalating one in DRC, Myanmar, or South Sudan. They know these countries are conflict zones — but lack the temporal granularity to forecast when escalation occurs.

Channel 2: Calibration asymmetry. GPT-4o expresses uncertainty through conservative NO predictions — a pattern that works for countries it understands but produces silent failure elsewhere. For humanitarian practitioners, silence is indistinguishable from safety. GPT-4o catches only 2 of 20 escalation events in Somalia; the other 18 pass undetected.

Condition	Tier	F1	P	R	N+
Always-YES baseline	Under	0.344	0.208	1.000	135
	Mid	0.332	0.199	1.000	162
	Over	0.084	0.044	1.000	13
Llama-3.3-70B (parametric)	Under	0.401	0.257	0.918	135
	Mid	0.373	0.253	0.710	162
	Over	0.082	0.050	0.231	13
GPT-4o (parametric)	Under	0.340	0.253	0.518	135
	Mid	0.371	0.279	0.556	162
	Over	0.000	0.000	0.000	13

Table 1. Zero-shot escalation forecasting by coverage tier. Coverage tiers defined by attention ratio: Under ($\leq 2\times$), Mid (2–10×), Over ($>10\times$). N+ = positive cases (actual escalation events) per tier across 1,628 total cases.

These channels compound in the humanitarian context specifically because false negatives are not symmetric with false positives: a missed escalation event may mean delayed preposition of supplies or failure to evacuate civilians, while a false alarm costs only analyst time. And AI early warning systems are most likely to be deployed where analyst capacity is lowest — precisely the under-covered contexts where our results show they fail most severely. The populations most likely to experience AI failure are the populations for whom AI failure is most likely.

5.2. Scope of the Gap

Our 224× figure is a lower bound. GDELT's English-language bias means local-language journalism in Bambara, Hausa, Tigrinya, Burmese, and Somali — which documents conflict dynamics in the granular temporal detail LLMs need — is largely absent from both our measurement and from pretraining corpora. The structural parallel to low-resource machine translation is direct: the knowledge gap for DRC or Myanmar is not a consequence of conflict complexity but of training data scarcity rooted in coverage asymmetry. The solution is not a better model — it is better data, produced through the same deliberate community investment that drove progress in low-resource MT (Wu & Dredze, 2020).

6. Ablation: Does Evidence Help?

The parametric results in Section 4 leave open a natural objection: perhaps the failure on under-covered zones reflects not a fundamental knowledge deficit but simply an absence of real-time signal. If so, providing structured conflict evidence at inference time should close the gap. We test this directly by augmenting each prompt with the three-month ACLED event count trend preceding the prediction window (see Section 3), and re-evaluating both models across all 1,628 cases.

6.1. Results

Table 2 reports the full five-condition comparison. Two findings stand out.

Finding 4: Adding evidence degrades under-covered performance below the trivial baseline. Reading the under-covered recall column from top to bottom: 1.000 → 0.918 → 0.518 → 0.378 → 0.178. Every step toward more sophisticated reasoning reduces the ability to catch actual crises in under-covered zones. Most critically, both evidence conditions fall below the Always-YES baseline on under-covered F1 (Llama: 0.329 vs. 0.344; GPT-4o: 0.211 vs. 0.344). A practitioner who replaced the trivial baseline with either evidence-augmented model would catch fewer escalation events in the world's most neglected conflict zones.

The mechanism is interpretable. When presented with event counts, both models attempt to reason about whether the observed trend constitutes meaningful escalation — but they lack the country-specific calibration to do so. A three-month count of 200 ACLED events means something very different in Myanmar (chronically high-intensity) than in South Sudan (episodic). Without knowing what "normal" looks like for a given country, the models treat stable high-count series as evidence against escalation and treat small absolute increases as noise. The parametric prior — blanket YES for conflict zones — is disrupted, but nothing useful replaces it.

Finding 5: Evidence helps where parametric knowledge already exists. The over-covered tier shows the opposite pattern: both evidence conditions improve on their parametric counterparts (Llama: 0.111 vs. 0.082; GPT-4o: 0.119 vs. 0.000). For Iran, the evidence condition correctly identifies several escalation periods marked by sharp event count spikes; for Israel, GPT-4o's evidence condition catches the single escalation event in the dataset (October 2023) where the parametric condition predicts NO on every case. The structured signal is interpretable precisely because the models have sufficient background knowledge to contextualize it.

Condition	Tier	F1	P	R	N+
Always-YES baseline	Under	0.344	0.208	1.000	135
	Mid	0.332	0.199	1.000	162
	Over	0.084	0.044	1.000	13
Llama-3.3-70B (parametric)	Under	0.401	0.257	0.918	135
	Mid	0.373	0.253	0.710	162
	Over	0.082	0.050	0.231	13
GPT-4o (parametric)	Under	0.340	0.253	0.518	135
	Mid	0.371	0.279	0.556	162
	Over	0.000	0.000	0.000	13
Llama-3.3-70B (evidence)	Under	0.329	0.291	0.378	135
	Mid	0.212	0.176	0.265	162
	Over	0.111	0.065	0.385	13
GPT-4o (evidence)	Under	0.211	0.261	0.178	135
	Mid	0.161	0.151	0.173	162
	Over	0.119	0.074	0.308	13

Table 2. Five-condition comparison across coverage tiers. Evidence condition augments each prompt with the three-month ACLED event count trend preceding the prediction window. All conditions evaluated on the same 1,628-case dataset. N+ = positive cases per tier.

This pattern is consistent with findings on LLM calibration more broadly: Kadavath et al. (2022) showed that larger models are better calibrated about their own knowledge boundaries, expressing higher confidence on questions within their training distribution. Our results suggest this calibration advantage extends to geographic conflict knowledge — but only where that knowledge exists.

This asymmetry confirms that the failure on under-covered zones is a knowledge deficit, not a prompting problem. Evidence augmentation is not a viable fix: it improves performance only where improvement is least needed, and worsens it where the humanitarian stakes are highest.

6.2. Implications for System Design

These results carry a direct implication for practitioners building LLM-assisted humanitarian early warning systems. The standard design assumption — that richer prompts produce better forecasts — does not hold across coverage tiers. On under-covered zones, the parametric Llama model with no evidence outperforms both evidence-augmented conditions, and even it barely exceeds the Always-YES trivial baseline. The appropriate benchmark for any proposed LLM-based early warning system is not zero-shot performance on well-covered zones; it is whether the system can outperform predicting escalation every month in the countries it is most likely to be deployed for.

None of the five conditions tested here clear that bar reliably.

7. Recommendations

Our findings motivate three concrete recommendations, framed as actionable research directions with clear precedent.

7.1. Coverage-Stratified Benchmarking

Conflict NLP evaluations should report performance stratified by media attention ratio alongside aggregate metrics. A system that performs well on Ukraine and Palestine while failing on DRC and Myanmar is not a general-purpose conflict AI system — it is calibrated to well-covered conflicts. We propose a simple standard: report F1 separately for under-covered ($\leq 2\times$), mid-covered (2–$10\times$), and over-covered ($>10\times$) tiers using the attention ratios we provide. This requires no additional data collection — only a change in reporting practice, analogous to the per-language evaluation norms that exposed systematic disparities in multilingual NLP (Wu & Dredze, 2020).

7.2. Dataset Creation for Under-Covered Conflict Zones

The training signal deficit for DRC, Myanmar, South Sudan, and similar contexts cannot be addressed by better architectures or prompting alone — our evidence condition results confirm this directly. We call on the NLP community to treat annotated conflict datasets for under-covered zones as priority infrastructure: creating NLP-ready resources from ACLED and UCDP (Sundberg & Melander, 2013) for under-covered regions, supporting annotation projects involving local researchers, and funding shared tasks targeting these contexts — the same deliberate investment that drove progress in low-resource MT (Wu & Dredze, 2020). The data infrastructure exists. What is missing is the commu-

nity's attention to it.

7.3. Training Data Documentation

LLM developers should document geographic conflict coverage statistics — specifically, what fraction of training data discusses each conflict zone and how this correlates with task performance. This instantiates the agenda of Gebru et al. (2021) and Dodge et al. (2021) in a domain where underdocumented bias has measurable humanitarian consequences. GPT-4o's recall of 0.100 on Somalia is not discoverable from any model card; it requires empirical evaluation most humanitarian organizations cannot conduct. We do not claim developers must equalize coverage — only that asymmetry, where it exists, should be documented so deployers can make informed decisions.

8. Conclusion

We set out to ask whether media attention asymmetry shapes LLM behavior on conflict forecasting tasks. The answer is yes — but the mechanism is more fundamental than a performance gradient. LLMs do not forecast conflict; they categorize it. Both models apply stable country-level priors across six years of observation windows, and the 224× attention gap translates directly into a 224× gap in the training signal available to calibrate those priors.

The evidence condition results sharpen this conclusion. Adding structured ACLED event counts does not fix the failure — it worsens it, because models lack the country-specific calibration to interpret what trends mean for countries they know poorly. The failure is not a prompting problem. It is a training data problem, and it cannot be patched at inference time.

The communities most affected by conflict and least covered by media receive AI that cannot distinguish a stable month from an escalating one. Closing this gap requires coverage-stratified benchmarks, training datasets for under-covered zones, and documentation standards that tell deployers where a model can and cannot be trusted. The data infrastructure exists. What has been missing is the NLP community's recognition that general-purpose conflict AI is not theirs until it works for the conflicts the world is not watching.

Limitations

We identify six limitations of the present study.

GDELT as an attention proxy. GDELT indexes primarily English-language outlets, underrepresenting local-language reporting. A conflict zone with rich Arabic, French, or Swahili coverage will appear under-covered in our measure even if it is not under-covered globally. Our attention ratios are therefore conservative estimates: the true asymmetry in LLM training data is likely larger than we report. Additionally, GDELT aggregates across all news types, not conflict-specific reporting, meaning politically salient countries receive inflated ratios relative to their actual conflict coverage.

ACLED coverage and conflict definition. ACLED's coverage is not uniform — under-covered countries may have thinner event records because ACLED's own documentation infrastructure is less developed there, not because conflict is less frequent. Our escalation threshold (130% of a 3-month rolling baseline) is operationally motivated but arbitrary; absolute F1 values are threshold-dependent even if the qualitative tier-level pattern is not.

Two LLMs may not generalize. We evaluate two frontier model families. The categorical reasoning pattern we document may not hold for smaller models, models with more recent training cutoffs, or models trained on conflict-specific corpora. Our claim is that the pattern exists in two widely-deployed systems that are plausible candidates for humanitarian AI deployment.

Zero-shot evaluation only. Few-shot prompting with conflict-specific examples or retrieval-augmented generation may improve performance on under-covered zones. We regard this as an open question: the evidence condition results suggest that adding structured data does not fix the calibration problem, but more sophisticated inference-time strategies remain untested.

Training cutoff confounds. Our dataset spans 2020–2026 and both models have training cutoffs within this window. We do not stratify by pre- and post-cutoff periods. However, the stable YES or NO priors we observe across the full 74-month window are inconsistent with a model making genuine use of temporal information, suggesting the structural explanation dominates.

Causal interpretation. We document a correlation between media attention ratio and LLM behavior and interpret it as a training data exposure effect. We cannot establish this causally — we have no access to either model's training data distribution. The alternative explanation (that under-covered countries are harder to forecast due to ACLED sparsity rather than LLM pretraining gaps) is argued against by the evidence condition results: if ACLED sparsity were the cause, providing ACLED evidence at inference time should help. It does not.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings*

of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623, 2021. doi: 10.1145/3442188.3445922.

Blair, R. A. and Sambanis, N. Forecasting civil wars: Theory and structure in an age of "big data" and machine learning. *Journal of Conflict Resolution*, 64(10):1885–1915, 2020.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Croicu, M. and von der Maase, H. Combining transformer-based actor embeddings with Gaussian processes for conflict forecasting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, 2021. doi: 10.18653/v1/2021.emnlp-main.98.

Eisensee, T. and Strömberg, D. News droughts, news floods, and U.S. disaster relief. *The Quarterly Journal of Economics*, 122(2):693–728, 2007. doi: 10.1162/qjec.122.2.693.

Entman, R. M. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993. doi: 10.1111/j.1460-2466.1993.tb01304.x.

Galtung, J. and Ruge, M. H. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1):64–90, 1965. doi: 10.1177/002234336500200104.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

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

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jackson, J., Jones, A., Tran-Johnson, J., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Leetaru, K. and Schrodt, P. A. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pp. 1–49. Citeseer, 2013.

Manvi, R., Khanna, S., Burke, M., Lobell, D., and Ermon, S. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, 2024. URL https://arxiv.org/abs/2402.02680.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103, 2016. doi: 10.1093/pan/mpv024.

Nemkova, P., Adhikari, A., Pearson, M., Sadu, V. K., and Albert, M. V. Cross-lingual stability and bias in instruction-tuned language models for humanitarian nlp. *arXiv preprint arXiv:2510.22823*, 2025a.

Nemkova, P., Ubani, S., Polat, S. O., Kim, N., and Nielsen, R. D. Detecting human rights violations on social media during russia-ukraine war. *AAAI 2025*, 2025b.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. Introducing ACLED: An armed conflict location and event dataset. volume 47, pp. 651–660, 2010. doi: 10.1177/0022343310378914.

Rost, N. and Ronco, M. Anticipating humanitarian emergencies with a high risk of conflict-induced displacement. *International Journal of Forecasting*, 42(1):138–157, 2026. doi: 10.1016/j.ijforecast.2025.04.006.

Sundberg, R. and Melander, E. Introducing the UCDP georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532, 2013.

UNHCR. Project jetson: Predictive analytics for forced displacement in Somalia. Technical report, United Nations High Commissioner for Refugees, 2022. URL https://www.unhcr.org/innovation/jetson.

Vinck, P., Pham, P., Kreutzer, T., and Kluth, W. Responsible use of artificial intelligence in the humanitarian sector. *Harvard Humanitarian Initiative*, 2019.

Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., and Weschle, S. Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review*, 15(4):473–490, 2013.

Wu, S. and Dredze, M. Are all languages created (morally) equal? Probing NLP for dialect and language differences.

In *Proceedings of the 4th Workshop on Ethics in Natural Language Processing*, 2020. URL https://arxiv.org/abs/2010.12435.