# From Events to Impacts: Calibrated Decomposition for LLM-Based Geopolitical Forecasting

Anonymous Authors[1]

## Abstract

Predicting geopolitical risk often requires estimating how public events propagate into downstream exposures, a task poorly captured by occurrence-only benchmarks. Many operational decisions depend not merely on whether an event occurs, but on whether it transmits into a supplier, route, regulatory obligation, market-access constraint, or other exposure. We propose Decomposed Impact Forecasting (DIF), which factorizes an impact forecast into event occurrence, exposure transmission, and material impact conditional on transmission, followed by post-hoc calibration. We evaluate Seed-20, a public-source pilot set with frozen forecast dates, temporally constrained evidence, archived prompts and outputs, and controlled exposure labels. In held-out pilot results, validation-selected calibrated DIF improves Brier and log score relative to calibrated direct and retrieval-augmented baselines, while calibration error remains unstable at this sample size. The contribution is an auditable protocol and pilot evidence for impact forecasting, not a definitive model ranking.

## 1. Introduction

Recent work evaluates large language models (LLMs) on unresolved questions, dynamic benchmarks, forecasting tournaments, and live forecasting arenas (Halawi et al., 2024; Karger et al., 2025; Schoenegger and Park, 2023; Schoenegger et al., 2024; Yang et al., 2026). These efforts ask whether general-purpose AI systems can assign useful probabilities to future events. Yet many real decisions are not exhausted by asking whether a public event occurs. A sanctions announcement, port closure, export-control measure, or military escalation matters operationally only if it transmits into an exposure such as a supplier, shipping route, market-access condition, regulatory obligation, or production dependency.

We formulate this as impact forecasting. The target is not only whether an event occurs, but whether it propagates into a specified exposure and becomes materially relevant for a downstream decision. Directly asking an LLM for the probability of material impact can conflate three mechanisms: event occurrence, transmission to an exposure, and impact conditional on transmission. DIF makes these mechanisms explicit and evaluates them under leakage-audited evidence constraints.

Our claim is intentionally modest. Seed-20 is a pilot demonstrating that exposure-aware geopolitical impact forecasts can be specified, audited, calibrated, and scored with proper scoring rules (Brier, 1950; Gneiting and Raftery, 2007). It is not a final benchmark of geopolitical foresight, and its small size prevents stable conclusions about model rankings. The paper contributes: (i) a task formulation separating event, transmission, and impact probabilities; (ii) an auditable protocol for historical public-source impact forecasting; (iii) pilot evidence that decomposition plus calibration can improve held-out proper scores; and (iv) an error analysis motivating larger, live, temporally blocked benchmarks.

Relation to AI forecasting. DIF is complementary to recent work on LLM forecasting tournaments, benchmark construction, and prediction arenas. Those settings primarily test whether a model can produce calibrated probabilities for event questions. Our focus is the additional modeling layer needed when an event is only an intermediate variable for a downstream exposure. This connects the workshop themes of benchmark design, probabilistic reasoning, retrieval, and decision relevance: the same forecast can be correct at the event level and still unhelpful if it does not identify the path from event to material impact. We therefore treat exposure specification, temporal leakage control, and proper scoring as part of the task definition rather than as implementation details.

## 2. Event Forecasting Is Not Impact Forecasting

Most forecasting questions are framed as event occurrence: "Will country X announce measure Y by date Z?" This is useful, but operational decisions and strategic foresight often require a different target (Courtney, 2001; Tetlock and Gardner, 2015): "Will measure Y materially affect exposure K by date Z?" The difference matters because the mapping from event to impact is neither deterministic nor uniform across exposures. An event may occur without transmitting to a given exposure; alternatively, an event with moderate occurrence probability may have high conditional impact for a vulnerable exposure if it occurs and transmits.

This gap is especially important for geopolitical risk. First, geopolitical events are multi-hop: policy announcements, military movements, border frictions, or sanctions can propagate through logistics, finance, suppliers, legal obligations, and customer demand. Second, the relevant exposure is local to a decision maker. A port disruption is not simply "bad"; it matters differently for firms using that route, relying on substitute capacity, or holding inventory. Third, free-form LLM rationales can appear persuasive while hiding which part of the forecast is uncertain. By decomposing the target, DIF exposes whether uncertainty is driven by event occurrence, transmission, or impact severity.

The decomposition also changes evaluation. Binary event benchmarks can score a model correctly even when it reasons about the wrong exposure. Impact forecasting therefore needs exposure labels, temporal evidence controls, and scoring against a joint event-transmission-impact target. This makes the task narrower than open scenario generation but more decision-relevant than occurrence-only prediction.

## 3. Method

Each question has forecast date $d$, horizon $h$, information set $\mathcal{F}_d$, event family, region, binary outcome, and controlled exposure vocabulary. Evidence available to the forecaster must satisfy $published_date \leq d$. Let $E$ be event occurrence, $T_k$ transmission to exposure category $k$, and $I_k$ material impact on that exposure. The exposure-specific target is

$$P_d(I_k) = P(E \mid \mathcal{F}_d)P(T_k \mid E, \mathcal{F}_d)P(I_k \mid T_k, E, \mathcal{F}_d). \tag{1}$$

This is a chain-rule factorization of the target, not an independence assumption. Dependence is retained through conditioning: transmission is conditional on the event, and material impact is conditional on both event and transmission. For binary benchmark scoring, $Y = 1$ when the event, transmission, and material impact criteria are jointly satisfied for the primary target exposure specified by the question; additional predicted exposures are used only for exposure-label F1.

Following the forecasting literature, we require numeric probabilities rather than ordinal risk levels or unconstrained narratives (Tetlock and Gardner, 2015; Mellers et al., 2015). DIF prompts return $p_E$, $p_{T,k}$, $p_{I,k}$, predicted exposure categories, and a rationale. We compute $p_{\text{raw},k} = p_E\, p_{T,k}\, p_{I,k}$ for the primary target exposure. The calibration layer is an $L_2$-regularized logistic regression on features $\mathbf{x} = (\text{logit}(p_E), \text{logit}(p_{T,k}), \text{logit}(p_{I,k}), \text{logit}(p_{\text{raw},k}))$, with logits clipped to $[\text{logit}(\epsilon), \text{logit}(1-\epsilon)]$, $\epsilon = 10^{-6}$. The penalty $\lambda$ is selected from a small grid by validation Brier score against an identity-calibration baseline on the four validation questions; the selected calibrator and all hyperparameters are frozen before test-set scoring, so no test-split information enters fitting or model selection. Isotonic calibration is reserved for larger datasets because Seed-20 is small (Platt, 1999; Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005; Guo et al., 2017). Calibration is not an afterthought: decomposed probabilities can be miscalibrated even when the factorization is semantically useful, and this layer tests whether the structured signal becomes reliable under proper scoring rules.

Scoring target. The benchmark target is deliberately conservative. The positive class requires that the event occur, that the event transmit to the named exposure, and that the resulting effect satisfy the materiality rule in the question. This avoids awarding probability mass for broad geopolitical salience alone. In addition to Brier and log score, we record exposure-label F1 because a model that predicts the correct probability for the wrong exposure is not operationally interpretable. The exposure score is diagnostic rather than a primary ranking metric because Seed-20 is too small to support stable multi-label conclusions.

Figure 1 summarizes the pipeline. The design is intended to make failure modes inspectable. If $p_E$ is high but $p_{T,k}$ is low, the model believes the event is likely but not relevant to the exposure. If $p_{T,k}$ is high but $p_{I,k}$ is low, the event may reach the exposure without material impact. This structure supports error analysis that a single direct probability cannot provide.

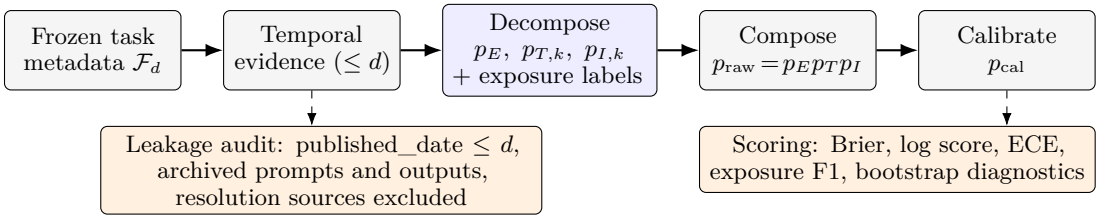


Figure 1. DIF pipeline. The model receives frozen task metadata $\mathcal{F}_d$ and temporally valid evidence (publication date $\leq d$), returns the decomposed probabilities $p_E$, $p_{T,k}$, $p_{I,k}$ and predicted exposure labels, composes a raw impact probability $p_{\text{raw}}$, and applies validation-selected calibration before final scoring. The leakage audit and scoring guards (orange) are part of the experimental design, not post-hoc filters.

## 4. Evaluation

Seed-20 pilot. Seed-20 is a manually constructed public-source pilot set containing 20 historical questions selected for clear forecast dates, horizons, binary resolution criteria, public resolution sources, and at least one temporally valid evidence item. It has 13 positive and 7 negative outcomes with train/validation/test splits of 8/4/8. Event families include shipping disruption, border or customs disruption, armed conflict escalation, energy supply disruption, and sanctions or export controls. Exposure labels use a controlled vocabulary; model-generated free-text labels are normalized before exposure F1 is computed.

Question construction. Each question is written from the perspective of an analyst observing the world at date $d$. The task statement contains the event family, region, horizon, primary exposure, and materiality criterion, but excludes resolution information. We prefer questions with public resolution sources and unambiguous negative cases, because geopolitical pilots otherwise overrepresent salient positive events. Evidence items are selected before scoring and frozen with timestamps so that the experiment can be replayed without changing the information set. These design choices trade breadth for auditability; the goal of Seed-20 is not coverage, but a reproducible test harness for the decomposition.

Leakage controls. Retrieval cutoff is necessary but not sufficient: modern LLMs may have parametric memory of historical outcomes (Paleka et al., 2026). We therefore disable web search, include only frozen task metadata and temporally valid evidence in prompts, exclude outcomes, resolution sources, and gold labels, archive raw prompts and outputs, and audit evidence timestamps. We still treat Seed-20 as a historical pilot rather than live forecasting; recent unresolved questions or older model checkpoints would be stronger tests.

Table 1. Seed-20 held-out test results ($n = 8$). Lower is better for Brier, log score, and ECE.

| Method | Brier ↓ | Log ↓ | ECE ↓ |
|---|---|---|---|
| Train prevalence | 0.2031 | 0.5977 | 0.1250 |
| Direct LLM | 0.4434 | 1.1759 | 0.5513 |
| Direct calibrated | 0.1371 | 0.4496 | 0.1785 |
| RAG LLM | 0.2894 | 2.2099 | 0.4625 |
| RAG calibrated | 0.1870 | 1.9608 | 0.3172 |
| Multi-agent direct | 0.3410 | 0.9123 | 0.4631 |
| DIF raw | 0.4049 | 1.0636 | 0.5279 |
| DIF calibrated | 0.0762 | 0.2675 | 0.2108 |

Baselines and metrics. We compare a train-prevalence base rate, Direct LLM, calibrated Direct, RAG LLM, calibrated RAG, a multi-agent direct aggregation baseline, DIF raw, and DIF calibrated. All LLM methods use the same frozen frontier checkpoint, deterministic decoding, and one temporally valid evidence item per question for RAG/DIF; prompts and raw outputs are archived and will be released as supplementary material. Primary metrics are Brier score (Brier, 1950), log score, and expected calibration error (ECE), following standard practice for probabilistic forecasting and calibration evaluation (Gneiting and Raftery, 2007; Guo et al., 2017); probabilities are clipped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-6}$ for log score. Exposure F1 is reported as a DIF diagnostic. Test metrics are reported separately from train and validation. Bootstrap intervals and leave-one-out analyses are diagnostic because $n$ is small.

On the held-out test split, calibrated DIF obtains the best Brier score (0.0762) and log score (0.2675), beating the train-prevalence baseline (Brier 0.2031). It improves Brier score relative to calibrated Direct LLM (0.1371) and calibrated RAG LLM (0.1870), while uncalibrated DIF raw is worse than RAG in Brier score but better in log score. This pattern suggests that decomposition alone does not consistently improve performance; its value in this pilot appears only after cal-

ibration. DIF is not a magic prompt template. It is a structured probabilistic representation whose outputs require calibration before being used as forecasts.

Train prevalence has the lowest ECE, and Direct calibrated has lower ECE than DIF calibrated, so we do not claim calibration dominance. Because calibration is fit on only eight training questions and selected on four validation questions, calibrated results are a stress test of the protocol rather than evidence of stable generalization. The paired bootstrap estimate for calibrated DIF versus calibrated Direct is $-0.0632$ Brier, with a 95% interval that narrowly overlaps zero $[-0.1300, 0.0020]$ over 500 resamples, reinforcing our interpretation of the result as diagnostic rather than conclusive. Controlled-vocabulary exposure F1 for DIF is 0.7143 on the test split.

## 5. Discussion and Error Analysis

Impact forecasting reframes geopolitical foresight as a resolvable probabilistic task rather than open-ended scenario generation, while retaining the decision orientation emphasized in strategic foresight and superforecasting work (Courtney, 2001; Tetlock and Gardner, 2015). The decomposition exposes where uncertainty enters: an event can be likely yet fail to transmit to an exposure, while a rare event can have high conditional impact. This is valuable for decision support because it allows users to inspect whether a forecast is driven by geopolitical uncertainty, exposure uncertainty, or impact uncertainty.

The pilot also surfaces operational issues. Exposure categories must be controlled; otherwise, models can evade evaluation through vague labels. Raw outputs must be archived; otherwise, prompt and parsing artifacts cannot be audited. Calibration must be separated from final test scoring; otherwise, the evaluation risks rewarding post-hoc fitting. Finally, retrieval controls must be treated as part of the experimental design, not merely as implementation details.

For the normalized calibrated DIF run, full-sample Brier is 0.0841. The largest leave-one-out shift occurs when excluding one negative Israel-Lebanon question, changing Brier by $-0.0172$; the same question is the worst individual calibrated-DIF error with per-question Brier 0.4111. This case illustrates a typical failure mode: when geopolitical tension is salient, a model can overestimate transmission to a specific exposure even when the event family is broadly plausible. Such errors motivate exposure-level rationales, counterfactual evidence checks, and larger negative sets.

## 6. Limitations and Future Work

The main limitations are scale and leakage. Seed-20 is author-constructed and underpowered; calibration uses only 8 train and 4 validation questions; and historical questions can still be contaminated through model pretraining. We therefore present the results as a reproducibility and methodology pilot. Scaling to GEF-Impact-100 should add independent annotation, inter-annotator checks for exposure labels, blocked temporal splits, live or recent unresolved questions, and factor-level diagnostics that distinguish errors in $p_E$, $p_T$, and $p_I$ under pre-registered validation protocols.

Three extensions stand out. First, exposure categories could be linked to explicit graphs of suppliers, routes, facilities, contracts, or regulatory obligations, turning the forecasting target into a query over a structured exposure graph while retaining resolvable binary outcomes. Second, larger benchmarks should report performance by event family, region, horizon, and exposure type under temporally blocked test scoring, since calibrators selected before regime shifts may fail after them. Third, future work should compare DIF with decision-theoretic baselines: a forecast that is only slightly better under Brier score may be much more valuable if it separates low-probability/high-impact from high-probability/low-impact cases, while a better aggregate score may not improve decisions if the wrong exposure is identified. We therefore view DIF as a bridge between AI forecasting benchmarks and decision support: it preserves resolvable outcomes while making the forecast target closer to the structure of real decisions.

## 7. Conclusion

We introduced DIF, an exposure-aware decomposition for LLM-based geopolitical impact forecasting, and implemented a leakage-audited Seed-20 pilot with archived prompts, raw outputs, calibration, scoring, and robustness artifacts. The results support the feasibility of evaluating downstream impact forecasting with proper scoring and exposure-level labels, while leaving definitive model comparisons to larger and more live evaluations. The central claim is not that Seed-20 proves DIF is superior, but that geopolitical forecasting should be evaluated as event-to-impact propagation rather than occurrence-only prediction.

## References


Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1):1–3, 1950.

Courtney, H. 20/20 Foresight: Crafting Strategy in an Uncertain World. Harvard Business School Press, 2001.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In International Conference on Machine Learning, 2017.

Halawi, D., Zhang, F., Chen, Y.-H., and Steinhardt, J. Approaching human-level forecasting with language models. In Advances in Neural Information Processing Systems, 2024.

Karger, E., Bastani, H., Chen, Y.-H., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In International Conference on Learning Representations, 2025.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. E. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science, 10(3):267–281, 2015.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the International Conference on Machine Learning, 2005.

Paleka, D., Goel, S., Geiping, J., and Tramèr, F. Pitfalls in evaluating language model forecasters. In International Conference on Learning Representations, 2026.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers, 1999.

Schoenegger, P. and Park, P. S. Large language model prediction capabilities: Evidence from a real-world forecasting tournament. arXiv:2310.13014, 2023.

Schoenegger, P., Tuminauskaite, I., Park, P. S., Valdece Sousa Bastos, R., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. Science Advances, 10(45):eadp1528, 2024. doi:10.1126/sciadv.adp1528.

Tetlock, P. E. and Gardner, D. Superforecasting: The Art and Science of Prediction. Crown, 2015.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. LLM-as-a-Prophet: Understanding predictive intelligence with Prophet Arena. In International Conference on Learning Representations, 2026.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.


## A. Seed-20 Question List

Table 2 lists all 20 Seed-20 questions with their forecast date, horizon, event family, region, primary exposure category, ground-truth outcome, and split assignment. Event families are abbreviated as SHP (shipping disruption), BRD (border or customs disruption), ACE (armed conflict escalation), ENG (energy supply disruption), and SXC (sanctions or export controls). Exposure categories are abbreviated as LRD (logistics route disruption), CED (commodity export disruption), BCD (border-customs disruption), MAE (market-access exposure), ESE (energy-supply exposure), RSE (regulatory-sanctions exposure), and SUP (supplier disruption). Resolution sources, full positive and negative resolution criteria, and gold exposure label sets are included in the artifact bundle.

## B. Resolution Examples and Evidence

To illustrate the Seed-20 question-construction protocol without relying on the supplementary bundle, we record three resolved questions in full—one clear positive case, one clear negative case, and one mixed/borderline case—each with its question text, an evidence item visible to the forecaster, full resolution criteria, and a separation between within-horizon evidence and the post-horizon confirmation that establishes the recorded outcome. For negative cases, we explicitly distinguish within-horizon reporting (which can document the trajectory but not the absence of a future event) from post-horizon evidence (which establishes the resolution). Every Seed-20 question follows the same template: forecast date $d$, horizon $h$, primary

Table 2. Seed-20 questions with forecast date $d$, horizon $h$ (days), event family, region, primary exposure, outcome $y$, and split. Outcome $y = 1$ indicates the joint event-transmission-impact criterion was met within the horizon.

| # | Topic | Family | Region | $h$ | Exposure | $y$ | Split |
|---|---|---|---|---|---|---|---|
| 01 | Red Sea / Houthi (2023-12-14) | SHP | MENA | 30 | LRD | 1 | train |
| 02 | Black Sea grain deal (2023-07-14) | SHP | E. Europe | 30 | LRD, CED | 1 | train |
| 03 | Niger ECOWAS sanctions (2023-08-08) | BRD | W. Africa | 180 | BCD, MAE | 1 | train |
| 04 | Panama Canal drought (2023-09-06) | SHP | C. America | 120 | LRD, CED | 1 | train |
| 05 | China Ga/Ge export controls (2023-07-10) | SXC | E. Asia | 60 | RSE, MAE, SUP | 1 | train |
| 06 | Venezuela-Guyana Essequibo (2023-12-06) | ACE | S. America | 90 | ESE, MAE | 0 | train |
| 07 | DPRK nuclear test (2024-01-15) | ACE | E. Asia | 180 | MAE | 0 | train |
| 08 | Taiwan Strait drills (2023-04-10) | SHP | E. Asia | 90 | LRD | 0 | train |
| 09 | Kosovo-Serbia border (2023-09-29) | BRD | Europe | 60 | BCD | 0 | val |
| 10 | Israel-Lebanon clashes (2023-10-22) | ACE | MENA | 60 | LRD | 0 | val |
| 11 | Poland-Ukraine truckers (2023-11-29) | BRD | Europe | 60 | BCD, MAE | 1 | val |
| 12 | Haiti Port-au-Prince (2024-03-05) | BRD | Caribbean | 60 | LRD, MAE | 1 | val |
| 13 | Sudan conflict exports (2023-04-16) | ACE | E. Africa | 60 | LRD, CED | 1 | test |
| 14 | Nagorno-Karabakh access (2023-09-20) | BRD | Caucasus | 30 | BCD, MAE | 1 | test |
| 15 | Rhine low water (2022-08-08) | SHP | Europe | 30 | LRD, CED | 1 | test |
| 16 | Strait of Hormuz closure (2024-04-09) | SHP | MENA | 45 | LRD, ESE | 0 | test |
| 17 | Taiwan post-election (2024-01-13) | SHP | E. Asia | 60 | LRD | 0 | test |
| 18 | Ukraine gas transit end (2024-06-04) | ENG | Europe | 240 | ESE, MAE | 1 | test |
| 19 | Texas-Mexico inspections (2022-04-09) | BRD | N. America | 30 | BCD, SUP | 1 | test |
| 20 | Canada Pacific port strike (2023-07-05) | SHP | N. America | 30 | LRD, CED | 1 | test |

exposure $k$, paired positive/negative resolution criteria, resolution date $d + h$, resolution summary, and at least two public references. Free-text exposure labels generated by models are normalized to the controlled vocabulary before exposure F1 is computed.

Positive: Red Sea / Houthi (#01). Question. "As of 2023-12-14, what is the probability that Houthi attacks will materially disrupt commercial container shipping through the Red Sea within the next 30 days?" Forecast date: 2023-12-14. Horizon: 30 days. Region: MENA. Primary exposure: logistics-route disruption (LRD). Evidence (visible at $d$). Reuters, "Yemen's Houthis claim responsibility for strike on Norwegian tanker," 2023-12-12, reporting the missile attack on the Strinda near the Bab al-Mandab Strait. Positive criterion. At least one major container carrier pauses Red Sea container traffic, reroutes around Africa, or imposes disruption-related surcharges within the horizon. Negative criterion. No major container carrier pauses or materially reroutes Red Sea services before the resolution date. Post-horizon resolution. Maersk paused Red Sea container traffic on 2023-12-15, and shipping firms subsequently imposed disruption-related fees within the horizon; outcome $y = 1$. Resolution sources. Reuters, "Maersk to pause all container ship traffic through the Red Sea," 2023-12-15; Reuters, "Shipping firms impose extra fees as Red Sea attacks hit trade," 2023-12-22.

Negative: DPRK nuclear test (#07). Question. "As of 2024-01-15, what is the probability that North Korea will conduct a nuclear test within the next 180 days?" Forecast date: 2024-01-15. Horizon: 180 days (resolution date 2024-07-13). Region: East Asia. Primary exposure: market-access exposure (MAE). Evidence (visible at $d$). Reuters, "North Korea fires missile, minister to visit Russia as tensions rise," 2024-01-14. Positive criterion. North Korea conducts a nuclear detonation during the horizon. Negative criterion. North Korea does not conduct a nuclear detonation during the horizon. Within-horizon trajectory. Rhetoric and conventional weapons tests continued, with no nuclear detonation reported through the horizon. Post-horizon resolution source. Reuters, "What to know about North Korea's nuclear weapons programme," 2024-09-13, which lists six confirmed DPRK nuclear tests in 2006, 2009, 2013, January 2016, September 2016, and September 2017, and no further detonations during the horizon; outcome $y = 0$. We mark this as a negative-by-absence resolution: the cited within-horizon source documents escalation patterns, while the post-horizon Reuters compilation establishes that no detonation occurred during the horizon.

Mixed: Israel-Lebanon clashes (#10). Question. "As of 2023-10-22, what is the probability that Israel-Lebanon border clashes will cause material port or maritime logistics closure within the next 60 days?" Forecast date: 2023-10-22. Horizon: 60 days (res-

olution date 2023-12-21). Region: MENA. Primary exposure: logistics-route disruption (LRD). Positive criterion. Clashes escalate into full-scale war causing material port or maritime logistics closure during the horizon. Negative criterion. Clashes do not become a full-scale war causing material port or maritime logistics closure during the horizon. Within-horizon trajectory. Reuters, "Israel expands evacuations as Lebanon border clashes escalate," 2023-10-22, and subsequent reporting through November and December 2023 documented sustained but contained cross-border exchanges, evacuations, and precautionary advisories without a declared full-scale war. Post-horizon resolution sources. Reuters, "Dangerous stasis on Israel's northern border leaves evacuees in limbo," 2024-01-11; Reuters, "Israeli soldiers on northern border dig in and wait as tensions mount," 2024-01-12. These sources describe daily skirmishes and a lower-intensity cross-border standoff without escalation into full-scale Israel–Hezbollah war during or immediately after the horizon, supporting outcome $y = 0$. This question is the largest leave-one-out shift for calibrated DIF on the test split and illustrates a typical failure mode in which geopolitical salience outweighs the strict materiality criterion.

Template summary and limitations. The remaining 17 Seed-20 questions follow the same fields. We note explicitly that several Seed-20 negative cases rely on absence-of-event resolutions, where within-horizon reporting documents trajectory and post-horizon reporting establishes the negative outcome; the supplementary artifact bundle (Appendix C) records both for each question. A larger benchmark should pre-register at least one post-horizon resolution source per negative case, ideally from an independent compilation rather than within-horizon news reporting alone.

## C. Reproducibility Artifacts

To preserve anonymity during review, we do not provide external artifact links. Instead, the paper itself includes the Seed-20 metadata (Appendix A) and worked resolution examples with sources (Appendix B) required to inspect the main reproducibility claims. The full artifact bundle—containing dataset files, prompt files, raw model outputs, parsed forecast logs, audit reports, calibration reports, reliability data, bootstrap intervals, error-analysis files, and table-ready CSV files—will be released with the camera-ready version. Proprietary internal tool code, private data, internal prompts, and API keys are excluded. For each question, the bundle records the forecast date, horizon, temporally valid evidence item, model output, parsed probabilities, predicted exposure labels, gold resolution, and scoring row. This structure is intended to make the pilot inspectable even when model rankings remain statistically fragile.

## D. Recommended Scaling Protocol

A larger GEF-Impact-100 benchmark should pre-register task construction rules, use independent annotators for event and exposure labels, block splits by time and event family, and reserve a live unresolved subset. Calibration models should be selected only on validation data and frozen before test scoring. We also recommend releasing factor-level diagnostics so that future work can distinguish event-forecast errors from transmission and exposure-impact errors.