Belief Propagation in LLM World Models: Measuring Strategic Information Bias with Prediction Markets

Mykola Khandoga[1], **Yevhen Kostiuk**[1,2], **Anton Polishko**[1], **Yurii Filipchuk**[1], **Kostiantyn Kozlov**[1], **Dmytro Zamriy**[1], **Artur Kiulian**[1]
[1]Future Principle [2]Aarhus University

Abstract

Every information ecosystem produces beliefs that shape strategic decisions. Both human analysts and AI systems inherit the blind spots of their information sources. We show that LLMs, combined with prediction markets, function as a calibrated instrument for measuring how far ecosystem-induced beliefs deviate from an external reference: LLMs extract the beliefs a text corpus implies, and prediction market price trajectories – anchored at resolution by realised outcomes – provide the calibration reference against which to quantify the deviation.

We isolate the bias contribution of specific text through ablation: varying information context while holding the model fixed, with a contaminated model that knows actual outcomes as control. Applied to 111 Ukraine-related prediction markets (~93,000 predictions, four models), we find that English news context systematically biases territorial predictions, wrong 64–72% of the time ($p < 10^{-6}$). A contaminated model that knows actual outcomes shows the same error rate, indicating the bias originates primarily in the text. Supplementing with Ukrainian military-analytical sources reduces the bias for all clean models; absolute-error gains are partial and model-dependent.

We show that the distortion originates primarily in the sources, not the models. Consistent across four architectures, it will persist in any system that processes them and propagate into downstream decisions.

1 Introduction

The beliefs propagated by news coverage about ongoing events have major consequences for policy, public opinion, and resource allocation. Yet there exists no method to quantify how close they are to the public consensus. Existing approaches either detect framing properties of text without measuring their downstream cost (Ali and Hassan, 2022; Otmakhova et al., 2024), or evaluate LLM forecasting accuracy without analyzing the information diet that drives it (Karger et al., 2025; Halawi et al., 2024).

Closing this gap requires solving two problems. First, we need a model that internalizes discourse framing – not classifying frames from the outside, but absorbing them so its output reflects the belief the text induces. Second, we need a grounded scale against which to measure that belief – an external reference anchored by realised outcomes, not another model's opinion.

LLMs solve the first problem. In-context learning operates as implicit Bayesian inference over latent concepts in the input (Xie et al., 2022), mechanistically equivalent to gradient descent on internal representations (von Oswald et al., 2023). The model doesn't just read the text – it updates its beliefs toward what the text implies. The output probability is the induced belief.

Prediction markets solve the second. A belief without an external reference is just an opinion. Prediction markets (Wolfers and Zitzewitz, 2004, 2006) provide continuous, financially incentivized probability estimates that are eventually anchored by realised outcomes. We use them in two distinct roles: the binary resolution gives us a low-power but genuine ground truth (§4.1), while the continuous price trajectory serves as the calibration reference for all model comparisons. The delta between the LLM-induced belief and market price, measured in percentage points (pp), is our calibrated measure of framing cost.

Our goal is to measure the bias that a specific text corpus induces – how many pp closer to or further from the market estimate does this text push the model's prediction? An LLM prediction reflects both parametric knowledge from pretraining and in-context beliefs induced by the prompt. To isolate each source's contribution, we construct an ablation ladder:

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 108–120

- **A** provides market overview and current price data – the minimal context from which the model reasons using only parametric knowledge.
- **B** adds a price chart: structured numerical signal without narrative framing.
- **C** adds English news articles: narrative without the rest of the context.
- **D** combines all English-language sources – chart, news, war map, trader comments – representing the full English information ecosystem.
- $\mathbf{D_{UA}}$ supplements D with Ukrainian military sources: General Staff reports, frontline bloggers, defense media.

Each transition is a measurement in pp: A→B measures the value of enriched price history signal, A→C the cost of English news narrative alone, A→D the cost of the full English ecosystem, D→D_{UA} the value of Ukrainian source diversification. A formal framework is given in Appendix A; Appendix H reports an exploratory decomposition of parametric and context-induced bias components.

Our contributions are:

1. A method for measuring the bias a text corpus induces via belief propagation in LLMs, in calibrated probability units, validated against prediction markets;
2. An ablation structure that isolates parametric from context-induced bias, revealing that the English information ecosystem systematically distorts LLM predictions on territorial questions – a finding confirmed by a contaminated model control and by linguistic analysis of reasoning traces (offense-dominant verb framing, asymmetric counterfactual reasoning) – and that supplementing with Ukrainian military-analytical sources partially counterbalances the bias;
3. A dataset of ∼93,000 predictions with full reasoning traces.[1]

2 Related Work

LLMs as forecasters. ForecastBench (Karger et al., 2025) shows LLMs now outperform non-expert crowds; agentic systems have reached superforecaster-level performance (Alur et al., 2025). We depart from this line entirely: rather than benchmarking accuracy, we exploit LLMs' sensitivity to linguistic framing as a measurement instrument. Our models need not be good forecasters – they need to faithfully reflect what the text implies.

Computational framing analysis. Media framing shapes public understanding of conflict (Entman, 2004), and NLP work on factuality and bias of news media is surveyed in Nakov et al. (2024). Methods range from codebook annotation (Card et al., 2015) to LLM-based classification of political bias in conflict coverage (Baly et al., 2020; Chandra et al., 2026). Offense-dominant framing in Western coverage of Ukraine has been documented qualitatively (Ojala et al., 2024) and through topic modeling (Ptaszek et al., 2024). These approaches classify frames – detecting what framing exists. Our method measures what that framing costs, in calibrated probability units.

LLM bias. Biases in LLMs are documented across demographic dimensions (Gallegos et al., 2024; Feng et al., 2023). These treat the model as the object of study. Our method measures how biased text propagates *through* models into calibrated beliefs – the model is the instrument, not the subject.

3 Data and Methods

3.1 Dataset

We collected 111 Ukraine-related prediction markets from Polymarket[2] (January 2025 – January 2026): 65 territorial ("Will side X control [city] by [date]?") and 46 diplomatic (negotiations, sanctions, Zelenskyy suit). For each market, we constructed rolling prediction points at ∼8 cutoff dates with a median gap of ∼16 days, creating instances where we know the current price, the actual price at each horizon (6h, 12h, 1d, 2d, 3d, 5d and 7d), and all information available up to the cutoff.

Each prediction is made under five information conditions: **A** (market overview + current price data only), **B** (A + price chart image), **C** (A + English news blocks), **D** (A + chart + English news + war map + Polymarket trader comments), and $\mathbf{D_{UA}}$ (D supplemented with Ukrainian-language military sources: General Staff casualty reports, Telegram military bloggers, Militarnyi). Conditions are identical across models; the contaminated model (Gemini 3.1 Pro Preview) runs all conditions except D_{UA}.

[1] https://huggingface.co/datasets/OpenBabylon/unlp-ukraine-forecasting

[2] https://polymarket.com

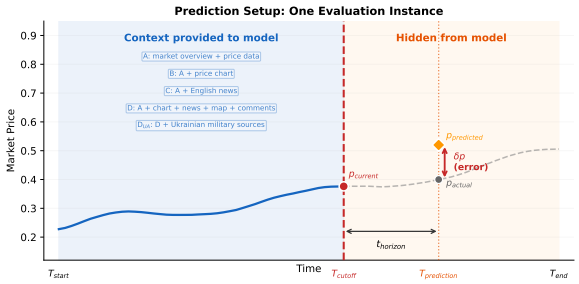

Figure 1: One evaluation instance. The model receives context up to T_{cutoff} under varying information conditions (A–D_{UA}); we compare its prediction at $T_{\text{prediction}}$ against the actual market price.

The English news corpus linked to our 111 benchmark markets comprises 16,457 articles from 2,217 domains (subset of a 122,290-article GDELT collection; Appendix F): 89% Western media, 3.4% think tanks (ISW), 2.1% Ukrainian English-language outlets, 2.8% Russian sources, and zero Ukrainian-language analytical sources. The D_{UA} corpus draws from DeepState frontline maps, General Staff daily loss reports, Ukrainian military bloggers, Militarnyi, Defense Express, and Ukrainian OSINT channels (Table 4).

3.2 Models

We evaluate three clean models – Gemini 2.5 Flash, Gemini 2.5 Pro, and GPT-5-mini – plus one contaminated control: Gemini 3.1 Pro Preview, whose training data extends past market resolution dates. The contaminated model's blind predictions show near-zero bias (+0.35 pp vs +2.0 pp clean average) and beat the no-change baseline by 10.4%, confirming knowledge of actual outcomes.

Flash and Pro share training data (Gemini 2.5 family) but differ in reasoning depth, enabling controlled comparison of processing effects. All models output structured predictions with full reasoning traces.

3.3 Statistical Framework

The market is the unit of independence throughout (N=65 territorial, N=46 diplomatic). Adjacent cutoffs have overlapping prediction windows (~44% overlap), so all tests use market-level aggregates with cluster-robust inference. We apply Bonferroni correction across 8 primary tests. With 65 clusters, we detect Cohen's $d \geq 0.35$ at 80% power. Our accuracy baseline is *no-change*: predict that the price at the horizon equals the current price. We use two distinct references: realised binary outcomes as the ground truth anchor in §4.1 (low-power but genuine), and the continuous market price trajectory as the calibration reference for all model comparisons in §4.2–4.4. Bias and MAE are therefore deviations from the market reference, not claims about reality. Full statistical details in Appendix B.

4 Results

4.1 The Benchmark Is Biased – But Useful

Polymarket overestimates Russian territorial capture by +3.5 pp relative to actual outcomes ($t(381) = 5.58$, $p < 10^{-7}$, $d = 0.29$; this test is point-level since it characterises a property of the benchmark itself, distinct from the market-clustered tests used for model comparisons in §4.2–4.4). Despite this bias, the market correlates strongly with outcomes ($r = 0.83$, $R^2 = 0.69$), and binary resolution provides insufficient power (3/65 territorial markets resolved YES). Note that Polymarket resolution for territorial markets relies on ISW and DeepState frontline updates, themselves expert assessments rather than direct observation; residual disagreements between these sources and on-the-ground reality are absorbed into the +3.5 pp figure. We use the continuous price trajectory as the calibration reference for all downstream comparisons.

4.2 English Context Destroys Signal

Blind models (condition A) show +2.0 pp average pro-capture bias – already present in training data, but *less* than Polymarket's +3.5 pp. Adding English news (condition D) pushes models to +3.4 pp, nearly matching the market's overestimation. English news does not add information – it replaces the model's more accurate prior with the discourse's less accurate framing. The shift is significant for Pro 2.5 (+2.4 pp, $p = 0.0001$) and GPT-5-mini (+1.4 pp, $p = 0.002$), both surviving Bonferroni correction. Flash shows a consistent but smaller effect (+0.5 pp, $p = 0.066$). All bias measurements below are relative to the market price trajectory; §4.1 establishes that this proxy is biased (+3.5 pp) but strongly correlated with outcomes ($r = 0.83$), meaning our estimates are conservative – the true distortion relative to reality is likely larger.

When context pushes predictions toward capture, those pushes are wrong 64–72% of the time across all clean models (binomial test vs 50%, $p < 10^{-6}$;

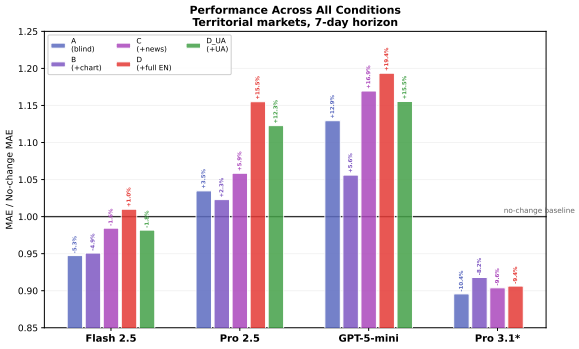

Figure 2: MAE vs no-change baseline across all information conditions. Adding English context (A→D) degrades predictions for all clean models; chart-only (B) outperforms full context (D). D_{UA} partially recovers accuracy. The contaminated model (Pro 3.1*) beats baseline under all conditions; D_{UA} was not run on it. Territorial markets, 7-day horizon.

Table 1). The bias is systematic and traceable: linguistic analysis of reasoning traces (Appendix L) shows Russia receives 2.3–3.7× more offensive verbs, "advances into" territory 77 times while Ukraine never does, and models reason about "if Russia succeeds" 60 times but "if Ukraine succeeds" zero times. Bias^2 accounts for only 2–9% of total error (Appendix G); we study it because it is directional, not because it dominates accuracy.

The full condition ladder (Figure 2; Appendix I) reveals that chart-only predictions (B) outperform full English context (D) for all clean models. Diplomatic markets serve as placebo: context does not damage predictions ($p = 0.03$ improvement for Flash), confirming a domain-specific mechanism (Appendix K). The bias is directionally asymmetric: downward predictions carry genuine signal, while upward predictions carry the discourse's offense-dominant distortion. Filtering out upward predictions converts all three models from losing to beating the no-change baseline (Appendix D).

4.3 Contaminated Model Ablation

The contaminated model (Pro 3.1*) knows actual outcomes – it beats no-change by 10.4%. Yet when English context pushes it toward capture, it is wrong at the same rate as clean models (Table 1).

The pro-capture push error rate is a property of the English news corpus, not of model ignorance (Appendix J). An inverted-question robustness check confirms this is not an artifact of question framing (Appendix C).

Model	Knows?	Push↑ acc.
Pro 3.1* (contam.)	Yes	27.9%
Flash 2.5	No	28.1%
GPT-5-mini	No	29.1%
Pro 2.5	No	36.3%

Table 1: When English context pushes predictions toward capture, accuracy is 28–36% across four models with different knowledge levels ($p < 10^{-6}$ for all, binomial test vs 50%). The error rate is a property of the corpus, not model ignorance.

4.4 Ukrainian Sources Reduce Bias; Accuracy Effects Are Model-Dependent

Supplementing English news with Ukrainian military sources (D → D_{UA}) reduces pro-capture bias across all three clean models (Figure 3). We use "correction" here in the sense of reducing deviation from the market reference; D_{UA} is source diversification, not fact-checking against a ground truth.

For Flash, D_{UA} eliminates the context-induced directional shift: bias drops from +0.4 pp ($p = 0.066$) to +0.1 pp ($p = 0.378$), indistinguishable from blind prediction. For Pro and GPT, significant pro-capture bias persists ($p < 0.01$). Ukrainian sources provide a comparable absolute correction across models, but the outcome differs because models accumulate different amounts of context damage: Flash takes little damage from English text, so the correction is sufficient; Pro amplifies the offense-dominant signal far beyond what source supplementation can repair (Appendices K, E, M).

Accuracy effects are more mixed. D_{UA} improves MAE on average across clean models, but not uniformly: the blind condition A remains competitive for some configurations, and D_{UA} hurts diplomatic predictions (Appendix K). The robust finding is bias reduction; absolute error improvements are partial and model-dependent.

5 Discussion and Conclusion

Our results indicate a measurable cost of information ecosystem misalignment with the market reference: English-language context induces a systematic pro-capture bias through both biased framing and source exclusion, and Ukrainian military sources partially counterbalance it. The bias is robust ($p < 10^{-6}$ for push accuracy), domain-specific (territorial but not diplomatic), and invariant to model knowledge (contaminated model shows same push error rate).

The practical implications are immediate. Sup-

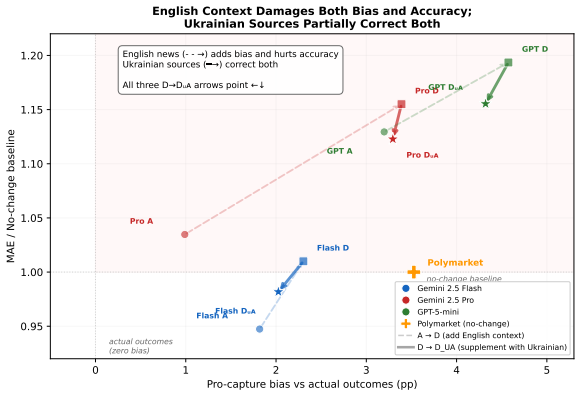

Figure 3: Bias vs accuracy for each model under conditions A (blind), D (English context), and D_{UA} (D supplemented with Ukrainian sources). Dashed arrows: A→D (English context damages). Solid arrows: D→D_{UA} (Ukrainian sources correct). All solid arrows point left and down. Territorial markets, 7-day horizon.

plementing English retrieval with Ukrainian sources through multilingual RAG reduces directional bias across all clean models; absolute-error improvements follow on average but are partial and model-dependent. Conservative reasoning (Flash) benefits most, while deeper reasoning (Pro) amplifies the offense-dominant signal beyond what source supplementation can repair. This makes model selection a critical lever: choosing conservative over deep reasoning can matter more than improving the information itself. The parametric bias already present in condition A reflects an English-centric information ecosystem and requires broadening source inclusion.

A case study illustrates the thesis in miniature: all clean models confidently predicted Zelenskyy would wear a suit to a papal funeral – logically sound but culturally blind (Appendix N). The method we introduce turns ecosystem misalignment from a qualitative concern into a grounded quantity.

Limitations

With 65 territorial market clusters, we achieve 80% power to detect Cohen's $d \geq 0.35$. Our MAE effects ($d = 0.25$–0.31) fall below this threshold, explaining non-significance after Bonferroni correction. The D_{UA} vs D improvement does not reach significance at the market-cluster level ($p = 0.10$–0.37). The Flash/Pro processing divergence, while controlled (same training data), is a single comparison. Polymarket's Ukraine markets may not generalize to other conflicts. While the case study uses Ukrainian sources, the methodology generalises to any conflict where one information ecosystem is suspected of systematic framing distortion relative to another.

Ethical Considerations

Our dataset spans 111 prediction markets about Ukraine – from whether specific cities will be captured and communities displaced, to diplomatic negotiations, international sanctions, and aid decisions. The territorial markets reduce the fate of real communities to price movements on a trading platform. We find this commodification of human catastrophe deeply troubling.

We use this data not to legitimize prediction markets, but because their structure inadvertently exposes something important: a measurable gap between what the English-language information ecosystem implies about events in Ukraine and what is actually happening. This gap – driven by both offense-dominant framing within included sources and exclusion of Ukrainian analytical ones – has consequences beyond prediction accuracy. Distorted understanding shapes international policy, humanitarian response, and the political will to support Ukraine.

The induced bias is not merely dovish or negotiation-oriented. It systematically pushes toward a low-agency, offense-dominant view of Ukraine in which Ukrainian leverage is discounted and territorial loss is treated as more inevitable than reality later shows. The worldview induced by the Anglo-American source ecosystem qualitatively resembles later concessionist policy rhetoric that downplays Ukrainian leverage. We leave this striking alignment for future study.

The bias we document is not abstract. When English-language AI systems systematically overestimate Russian territorial success, they reinforce a narrative of inevitable Ukrainian loss that Ukrainian soldiers, analysts, and journalists work daily to counter – through sources that English-language pipelines do not include.

This work used AI-based writing assistance tools for editing and formatting.

Acknowledgments

We gratefully acknowledge Amazon Web Services and DigitalOcean for the cloud compute credits and infrastructure that supported this work. Model inference, training, and evaluation pipelines were

run on AWS (EC2 GPU instances, SageMaker, S3); data collection and experiment tracking were hosted on DigitalOcean managed database and compute instances.

References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348.

Rohan Alur, Bradly C Stadie, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, Hayley Williamson, Nina Lozinski, Aaron Linsky, and Jasjeet S Sekhon. 2025. AIA forecaster: Technical report. *arXiv preprint arXiv:2511.07678*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4982–4991.

Dallas Card, Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 438–444.

Rohitash Chandra, Haoyan Chen, Yaqing Zhang, Jiacheng Chen, and Yuting Wu. 2026. An evaluation of LLMs for political bias in Western media: Israel-Hamas and Ukraine-Russia wars. *arXiv preprint arXiv:2601.06132*.

Robert M Entman. 2004. *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. University of Chicago Press.

Shangbin Feng, Chan Young Park, Yohan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3).

Danny Halawi, Fred Shi, Sebastian Borgeaud, Adam Lerer, Pieter Abbeel, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. 2025. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Mansurov, and Momin Mansurov. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Markus Ojala, Mervi Pantti, and Jenni Kangas. 2024. Framing the war in Ukraine: A comparative study of news coverage. *Journalism Studies*.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Grzegorz Ptaszek, Bogdan Yuskiv, and Serhii Khomych. 2024. War on frames: Text mining of conflict in Russian and Ukrainian news agency coverage on Telegram. *Media, War & Conflict*, 17(1):41–61.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174.

Justin Wolfers and Eric Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126.

Justin Wolfers and Eric Zitzewitz. 2006. Interpreting prediction market prices as probabilities. NBER Working Paper 12200, National Bureau of Economic Research.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*.

A Theoretical Framework

We formalize the mechanism studied in this paper. The goal is not a general theory of conflict forecasting, but to make precise what we mean by *offense-dominant* beliefs under source exclusion.

Latent battlefield state. Let i index a market-event pair. Let $y_i \in \{0, 1\}$ denote the realized outcome, where $y_i = 1$ corresponds to offensive success on the target event. Each event has an unobserved latent state $z_i \in \mathbb{R}$ capturing the true degree of offensive feasibility.

Information ecosystems as biased signals. We consider two information ecosystems: E (English-language) and U (Ukrainian military-analytical). Each provides a noisy signal about the same latent state:

$$x_i^E = z_i + \beta_E + \varepsilon_i^E, \qquad x_i^U = z_i + \beta_U + \varepsilon_i^U,$$

where $\varepsilon_i^E, \varepsilon_i^U$ are zero-mean noise and $\beta_E, \beta_U \in \mathbb{R}$ are systematic shifts. An ecosystem is *more offense-dominant* when it shifts beliefs further toward offensive success. The English ecosystem is more offense-dominant than the Ukrainian one whenever $\beta_E > \beta_U$.

LLMs as belief elicitation instruments. For model m, let μ_m denote the model's effective baseline belief under condition A (minimal context: market overview and current price data), absorbing both pretraining priors and the model's processing of the price signal. Given a source set S, the model produces

$$\hat{p}_i^{(m)}(S) = \sigma\big(g_i^{(m)}(S)\big),$$

where $\sigma(\cdot)$ is the logistic sigmoid and $g_i^{(m)}(S)$ is a latent score formed by combining the model prior with available signals:

$$g_i^{(m)}(S) = \frac{\lambda_{m,0}\mu_m + \sum_{e\in S} \lambda_{m,e}\, x_i^e}{\lambda_{m,0} + \sum_{e\in S} \lambda_{m,e}}.$$

Here $\lambda_{m,0} \geq 0$ weights the prior and $\lambda_{m,e} \geq 0$ weights ecosystem e. We treat these as effective parameters that absorb presentation-order effects and model-specific context processing. This maps onto our three experimental conditions:

$$\begin{aligned}\hat{p}_i^{(m)}(A) &= \sigma(\mu_m),\\ \hat{p}_i^{(m)}(D) &= \hat{p}_i^{(m)}(\{E\}),\\ \hat{p}_i^{(m)}(D_{UA}) &= \hat{p}_i^{(m)}(\{E, U\}).\end{aligned}$$

Proposition 1 (Source exclusion shifts latent scores toward offense)**.** *Assume ecosystem signals satisfy $x_i^e = z_i + \beta_e + \varepsilon_i^e$ with $\mathbb{E}[\varepsilon_i^e] = 0$, and that the latent score $g_i^{(m)}(S)$ is a positively weighted average of the model prior and available signals. If $\beta_E > \beta_U$ and $\lambda_{m,U} > 0$, then for every event i:*

$$\mathbb{E}\Big[g_i^{(m)}(D)\Big] > \mathbb{E}\Big[g_i^{(m)}(D_{UA})\Big].$$

Excluding Ukrainian sources systematically increases the expected latent score toward offensive success.

Proof. Substituting $\mathbb{E}[x_i^E] = z_i + \beta_E$ and $\mathbb{E}[x_i^U] = z_i + \beta_U$:

$$\mathbb{E}\Big[g_i^{(m)}(D)\Big] = \frac{\lambda_{m,0}\mu_m + \lambda_{m,E}(z_i + \beta_E)}{\lambda_{m,0} + \lambda_{m,E}},$$

$$\begin{aligned}&\mathbb{E}\Big[g_i^{(m)}(D_{UA})\Big]\\ &= \frac{\lambda_{m,0}\mu_m + \lambda_{m,E}(z_i + \beta_E) + \lambda_{m,U}(z_i + \beta_U)}{\lambda_{m,0} + \lambda_{m,E} + \lambda_{m,U}}.\end{aligned}$$

Since $\lambda_{m,U} > 0$ and $\beta_U < \beta_E$, the additional term pulls the weighted average below the English-only value. □

From scores to probabilities. Since $\sigma(\cdot)$ is monotone increasing, the score ordering implies $\hat{p}_i^{(m)}(D) > \hat{p}_i^{(m)}(D_{UA})$ pointwise for each realization of the noise. The ordering $\mathbb{E}[\hat{p}_i^{(m)}(D)] > \mathbb{E}[\hat{p}_i^{(m)}(D_{UA})]$ then follows by taking expectations. However, the *magnitude* of the probability gap depends on the curvature of σ and the noise distribution, so the proposition is stated at the level of latent scores where the result is exact.

Operational quantities. The framework yields three empirically measurable quantities. The *harm of English context*: $H_m = \mathbb{E}_i[\hat{p}_i^{(m)}(D) - \hat{p}_i^{(m)}(A)]$. The *source exclusion cost*: $X_m = \mathbb{E}_i[\hat{p}_i^{(m)}(D) - \hat{p}_i^{(m)}(D_{UA})]$. The *directional bias*: $B_m(S) = \mathbb{E}_i[\hat{p}_i^{(m)}(S) - y_i]$. On territorial markets, positive $B_m(S)$ indicates systematic overprediction of offensive success. Our main empirical finding is $B_m(D) > B_m(D_{UA})$, with $H_m > 0$ and $X_m > 0$.

Limitations of the formalization. This framework is deliberately modest. It does not claim that LLM outputs recover latent beliefs perfectly, nor that ecosystems differ only along a single dimension. The weighted-average assumption is an idealization: actual LLMs process context sequentially, and effective weights may depend on presentation order, prompt structure, and reasoning depth. The formalization makes explicit the mechanism tested

– if one ecosystem is more offense-dominant, excluding the other should produce more offense-dominant predictions – without claiming more than the experimental design supports.

B Statistical Methodology

Clustering. Adjacent cutoffs within a market have overlapping 7-day prediction windows (~44% overlap at the median 16-day gap). Intra-cluster correlation ranges from 0.008 (Flash) to 0.141 (GPT-5-mini). All tests use market-level aggregates.

Multiple testing. We apply Bonferroni correction across 8 primary tests. Market bias (territorial and diplomatic), directional bias shift (3 models), and MAE damage (3 models). After correction: market bias ($p < 10^{-4}$), Pro 2.5 bias shift ($p = 6.8 \times 10^{-4}$), and GPT bias shift ($p = 0.018$) survive. MAE tests and Flash bias shift do not survive. Push accuracy tests ($p < 10^{-6}$) are excluded from this family.

Power. With N=65 clusters, 80% power at $\alpha = 0.05$ requires $d \geq 0.35$. Observed: Pro 2.5 $d = 0.50$ (powered), GPT $d = 0.36$ (marginal), Flash $d = 0.19$ (underpowered). MAE effects $d = 0.25$–0.31 (underpowered). Flash-scale detection requires ~215 clusters.

Permutation tests. Sign-randomization tests (10,000 iterations) confirm all parametric results with concordant p-values.

C Question Inversion (Anti-X)

We re-run territorial predictions with inverted questions: "Will Russia *fail to capture* X?" with inverted prices ($1-p$). This is analogous to semantic entropy (Kuhn et al., 2023) but diagnoses *context quality* rather than model confidence.

When contradictions occur under condition D, they are directionally asymmetric. For Pro 3.1* (contaminated), the asymmetry is 47:1 – original-down/anti-up vastly dominates. English military reporting is stronger evidence *against failure* than *for capture*.

D Selective Trust

A zero-parameter rule – trust the model only when it predicts downward movement – converts losing strategies to winning ones. Flash selective: −3.4% vs no-change ($p < 10^{-21}$). Pro 2.5 selective: −3.3%. This reveals that downward predictions carry genuine signal while upward (pro-capture) predictions carry systematic noise. D_{UA} + selective is the best combination for GPT-5-mini (−0.8%), finally converting it to a winning strategy.

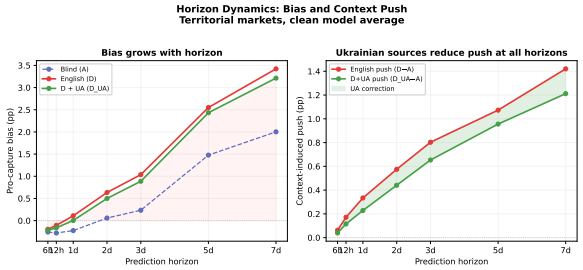

Figure 4: Left: pro-capture bias grows with prediction horizon for all conditions. Right: context-induced push (D−A) and Ukrainian-corrected push (D_{UA}−A) across horizons. D_{UA} reduces push at every horizon. Territorial markets, clean model average.

Hz	A	D	D_{UA}	D push	D_{UA} push
6h	−0.3	−0.2	−0.2	+0.1	+0.0
1d	−0.2	+0.1	+0.0	+0.3	+0.2
3d	+0.2	+1.0	+0.9	+0.8	+0.7
7d	+2.0	+3.4	+3.2	+1.4	+1.2

Table 2: Bias by horizon (pp, clean model average, territorial). Bias grows with horizon; D_{UA} reduces push at all horizons.

E Per-Horizon Analysis

Pro-capture bias grows with prediction horizon (Table 2, Figure 4). D_{UA} consistently produces less push than D at every horizon.

F Source Ecosystem and Utility

F.1 English Information Diet

The full GDELT collection spans all topics and comprises 122,290 articles from 6,232 domains. Of these, 16,457 articles from 2,217 domains are linked to our 111 Ukraine benchmark markets (§3). The composition statistics below describe the benchmark subset. The corpus is US-centric (63% of articles) and Western-oriented (Table 3).

The top domains by volume are Yahoo (5,864), marketscreener (1,142), freerepublic (781), Daily Mail (736), globalsecurity.org (659), ZeroHedge (595), ABC News (560), ISW (534), Independent (464), and Newsweek (462). The corpus reflects what GDELT surfaces as the English-language information diet about Ukraine – mainstream wire content, finance aggregators, and partisan outlets.

Category	Examples	%
Wire / mainstream	AP, Reuters, CNN, Guardian	36
Aggregators	Yahoo, AOL, bignewsnetwork	12
International	jpost, economictimes (India)	12
Conservative US	Fox, Breitbart, Epoch Times	5
Finance	marketscreener, fxstreet, Forbes	5
UK press	Daily Mail, Independent	5
Specialist / OSINT	ISW, globalsecurity.org	1
Other / regional	various	24

Table 3: English source composition by category. US sources account for 63% of articles. Ukrainian-language analytical sources: zero.

Zero Ukrainian-language analytical sources appear.

F.2 Ukrainian Information Diet

The D_{UA} condition supplements condition D with three source types (Table 4).

Source	Count	Type
Telegram bloggers	47,009 posts	16 channels
Militarnyi.com	2,352 articles	Defense news
General Staff losses	1,468 records	Daily casualty data

Table 4: Ukrainian sources added in D_{UA}.

The 16 Telegram channels represent the core of the Ukrainian military information ecosystem: 5 active/reserve military (including brigade commanders Magyar, Zhorin, Shtefan, Fedorenko, Krotevych), 4 journalists (Tsaplienko, Butusov, Kazanskyi), 4 analysts (Berezovets, Kovalenko, Mashovets, Zhdanov), 1 official channel (General Staff ZSU), 1 tech specialist (Beskrestnov), and 1 commentator. All channels represent Ukrainian perspectives – no Russian, neutral, or Western-OSINT channels are included. This is by design: D_{UA} tests whether adding the Ukrainian military-analytical ecosystem improves prediction accuracy, not whether balanced sourcing does.

F.3 Source Utility Analysis

We analyze which sources models actually cite in their reasoning traces (condition D, 7d horizon, clean models, $N = 2{,}130$ predictions) and whether citing a source correlates with better or worse predictions. The patterns below are correlational: ISW-cited predictions performing worse does not establish that citing ISW *causes* worse predictions; both may reflect harder markets or more contested events. Relative performance compares the model's error when citing a source against the no-change baseline for those same markets.

Source	Freq	Rel. Perf	Dir% (cited)	Dir% (not)
Russian officials	43.0%	−1.4%	**61.9**	51.9
Ukrainian officials	36.9%	−5.4%	60.0	54.0
Price action	76.9%	−8.6%	56.7	54.5
ISW	51.5%	−11.3%	50.9	**61.8**
Russian milbloggers	6.7%	−11.9%	54.8	56.3
Polymarket traders	17.7%	−6.5%	55.6	56.3
DeepState	3.8%	−1.0%	56.1	56.2
UA General Staff	5.4%	−15.1%	41.5	57.1
OSINT	1.9%	−14.6%	41.5	56.5

Table 5: Source utility: frequency of citation in reasoning traces and correlation with prediction quality. Relative performance = (NC_error − model_error) / NC_error; negative = worse than no-change. ISW – the most-cited analytical source (51.5%) – correlates with *worse* directional accuracy (50.9%) than predictions not citing it (61.8%).

The ISW paradox. ISW is the dominant analytical source (51.5% citation rate). Yet predictions citing ISW show 50.9% directional accuracy – a coin flip – compared to 61.8% when ISW is not cited. This is consistent with ISW's offense-dominant analytical framing: detailed, authoritative coverage of offensive operations that models internalize as evidence for territorial change. The authority of the source amplifies the framing effect.

Russian officials as accidental signal. Russian officials are cited at 43.0% with the best directional accuracy (61.9%) among high-frequency sources. This is counterintuitive until one considers the epistemic framing from Appendix L: models treat Russian claims with skepticism ("Russia *claims*..."), and this discounting accidentally produces better-calibrated predictions than absorbing ISW's authoritative framing uncritically.

Contaminated model source shift. The contaminated model (Pro 3.1*), which knows outcomes, shifts its citation patterns: ISW drops from 51.5% to 35.4%, Ukrainian officials from 36.9% to 23.1%, while price action rises from 76.9% to 81.0%. A model with outcome knowledge relies less on narrative sources and more on the price signal – further evidence that narrative sources add framing, not information.

G Bias-Variance Decomposition

Bias^2 accounts for 2–9% of model MSE; variance dominates at 91–98%. Polymarket: 8%

Model	Param.	Context	UA corr.
Flash 2.5	+1.8 pp	+0.5 pp	−0.3 pp
Pro 2.5	+1.0 pp	+2.4 pp	−0.1 pp
GPT-5-mini	+3.2 pp	+1.4 pp	−0.3 pp

Table 6: Three-layer bias decomposition. Parametric bias (A) is in the weights. Context-induced bias (D−A) comes from English news. Ukrainian correction ($D-D_{UA}$) is the bias reduction from supplementing with Ukrainian sources – roughly constant across models, but against vastly different context damage.

Model	Bias recov.	Accuracy recov.
Flash 2.5	57%	45%
Pro 2.5	4%	27%
GPT-5-mini	18%	60%

Table 7: D_{UA} recovery of context-induced damage. All models improve on both dimensions.

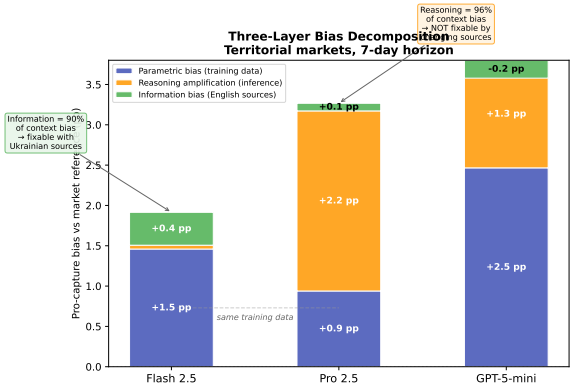

Figure 5: Bias decomposition. Flash and Pro share training data but differ 5× in context-induced bias. Ukrainian sources provide a similar absolute correction for all models, but it is swamped by Pro's context damage.

bias[2]. English context increases both components; Ukrainian sources reduce both. Variance reduction is the larger contributor to MAE improvement for bold models.

H Three-Layer Bias Decomposition

We present this decomposition as exploratory and correlational. With 65 territorial markets, per-model component estimates are noisy and we do not draw causal source-level conclusions from them. Our experimental conditions separate three components of pro-capture bias (Figure 5), measured on territorial markets at 7-day horizon.

Parametric bias (training data). Measured by condition A – no context, just the model's priors. Flash: +1.8 pp. Pro: +1.0 pp. GPT: +3.2 pp.

Context-induced bias (English news). Measured by D−A: the additional pro-capture shift from English context. This is where models diverge dramatically. Flash: +0.5 pp. Pro: +2.4 pp. GPT: +1.4 pp. Pro accumulates 5× more context damage than Flash despite sharing training data – deeper reasoning amplifies the offense-dominant signal in English text.

Ukrainian source correction. Supplementing with Ukrainian sources provides a roughly constant absolute correction across models: Flash −0.3 pp, Pro −0.1 pp, GPT −0.3 pp (Table 6). What differs is not the correction but the denominator – how much context damage each model accumulates. For Flash, 0.3 pp corrects most of the 0.5 pp context damage (57%). For Pro, a similar correction is negligible against 2.4 pp of damage (4%).

This is a model selection result: Ukrainian sources help all models roughly equally in absolute terms, but their impact depends on how aggressively the model amplifies English context. Conservative reasoning (Flash) keeps context damage small, making the correction sufficient. Deep reasoning (Pro) amplifies English bias beyond what source supplementation can repair.

I Full Condition Comparison

Table 8 reports MAE relative to the no-change baseline (in %) for every model under each of the five information conditions on territorial markets at the 7-day horizon. Negative values mean the model beats no-change; positive values mean it loses to it. Two patterns stand out. First, chart-only predictions (B) outperform full English context (D) for all three clean models, indicating that the bulk of the damage from condition D comes from narrative content rather than from price or chart inputs. Second, only the contaminated model (Pro 3.1*) beats no-change under any condition. Figure 6 shows the same comparison at the per-market level for D vs D_{UA}: D_{UA} wins the majority of markets for Flash (37/65) and Pro (32/65), confirming that the bias reduction reported in §4.4 is not driven by a few outlier markets.

J Contaminated Model Details

Pro 3.1* shows near-zero blind bias (+0.35 pp vs +2.0 pp clean average) and beats no-change by

	A	B	C	D	D_{UA}
Flash	−5.3	−4.9	−1.6	+1.0	−1.8
Pro	+3.5	+2.3	+5.9	+15.5	+12.3
GPT	+12.9	+5.6	+16.9	+19.4	+15.5
3.1*	−10.4	−8.2	−9.6	−9.4	–

Table 8: MAE vs no-change (%), territorial 7d. Chart-only (B) outperforms full context (D) for all clean models.

Figure 6: Per-market MAE under D (English) vs D_{UA} (supplemented with Ukrainian sources). Each dot is one territorial market. Points below the diagonal indicate D_{UA} outperforms D. D_{UA} wins the majority of markets for Flash (37/65) and Pro (32/65). Territorial markets, 7-day horizon.

10.4% blind, 9.4% with context – the only model to beat baseline under any condition. Despite this knowledge, its pro-capture push accuracy (27.9%) matches clean models, demonstrating that the directional signal in English military reporting overrides even direct outcome knowledge.

K Diplomatic Markets (Placebo)

D_{UA} hurts diplomatic predictions for Flash (+5.9% vs D) and Pro (+5.8%), while GPT is unchanged (−0.1%). Ukrainian military sources contain no diplomatic signal; English coverage of negotiations and sanctions is more informative. This confirms domain specificity (Figure 7).

L Linguistic Analysis of Offense-Dominant Framing

The quantitative results in §4 stand independently of the analysis below. We provide linguistic analysis of reasoning traces as qualitative illustration of the mechanism behind the statistical findings.

We analyze 444 reasoning traces (111 markets × 4 models, condition D, 7d horizon, first cutoff per market; ∼159,000 words) using regex-based proximity matching with manual validation. The analysis reveals eight dimensions of offense-dominant framing, all pointing in the same direction across all models.

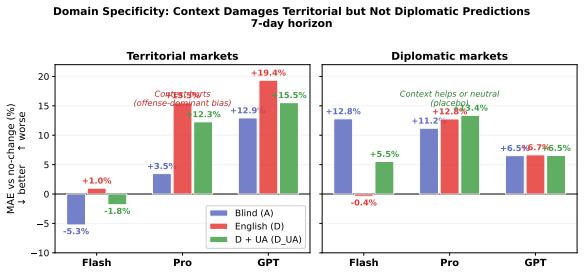

Figure 7: Domain specificity: English context damages territorial predictions (left) but not diplomatic ones (right). D_{UA} corrects territorial bias but hurts diplomatic predictions, confirming the intervention is domain-specific. 7-day horizon.

L.1 Agency: Who Acts

Russia is systematically framed as the agent. Across all models, Russia appears as grammatical subject 60–66% of the time (ratio 1.4–2.0× over Ukraine), is the first actor mentioned in 76–86% of texts, and receives 1.56–1.76× more total mentions. Ukraine's primary role is reactive.

L.2 Verb Semantics: What Each Side Does

Within 80 characters of each actor mention, 59–64% of verbs near Russia are offensive (advance, capture, assault, deploy) while Ukraine's verbs split between defensive (hold, resist, repel) and diplomatic. Russia receives 2.3–3.7× more offensive verbs than Ukraine across all four models.

L.3 Success/Failure Framing

The most extreme asymmetry. Russia's success-to-failure language ratio ranges from 6.7:1 to 14.1:1. Ukraine's ratio is 1.4–2.5:1. Russia is described in almost exclusively positive-outcome terms (advance, progress, gains, momentum) even when the model predicts those outcomes will not materialize. The contaminated model (Pro 3.1*) shows the highest ratio (14.1:1) despite knowing most territorial outcomes resolve against capture – framing and prediction are decoupled.

L.4 Territorial Lexicon

L.5 Conditional Erasure

Models reason about hypothetical Russian success but never about Ukrainian success:

Pro 3.1* produces 37 "if Russia succeeds" constructions – the most of any model – despite knowing most such outcomes do not occur.

Verb	Russia	Ukraine	Ratio
capture	143	16	8.9×
advance into	77	**0**	∞
seize	14	1	14×
encircle	15	2	7.5×
hold	2	3	0.7×

Table 9: Territorial verbs attributed to each actor (sum across 4 models, 444 traces). "Advance into" appears 77 times near Russia and **zero** times near Ukraine.

Pattern	GPT	Fl.	Pro	3.1*	Tot.
"If R succeeds…"	6	13	4	37	60
"If R fails…"	0	0	2	0	2
"If U succeeds…"	0	0	0	0	**0**
"If U fails…"	1	3	0	1	5

Table 10: Conditional framing in reasoning traces. Ukrainian success is never considered as a scenario.

L.6 Epistemic Framing

Russia "claims" (224 instances) while Ukraine "denies" (20 instances; Russia: 1). Ukraine's primary epistemic role is refuting Russian assertions rather than making its own. Russia is also "confirmed" 2× more than Ukraine, creating a paradox where Russian assertions are simultaneously more doubted and more validated.

L.7 Syntactic Subordination

"Despite Ukrainian resistance, Russia continues to advance" appears 30 times. The inverse –"Despite Russian challenges, Ukraine holds" – appears 6 times. Ukrainian action is systematically placed in concessive clauses; Russian action occupies the main clause. Ukrainian defense is framed as *overcome*; Russian offense as *persisting*.

L.8 Composite Scorecard

Every dimension – agency, verb semantics, success framing, conditional reasoning, epistemic credibility, syntactic structure – points in the same direction across all four models. The contaminated model, which *knows* most territorial outcomes resolve against offense, produces the most extreme framing on several dimensions. This decoupling of framing from knowledge confirms that the offense-dominant pattern is structural – embedded in how English-trained LLMs construct conflict narratives – rather than a reflection of model beliefs about outcomes.

Metric	GPT	Flash	Pro	3.1*
Mention ratio (R/U)	1.56	1.67	1.76	1.68
Offensive verb ratio	2.28	3.20	3.70	3.34
Russia S/F ratio	8.1	9.0	6.8	**14.1**
Ukraine S/F ratio	1.4	2.2	2.5	1.6
Subject ratio (R/U)	1.39	1.96	1.87	1.47
Russia first (%)	78.6	79.6	84.0	**85.9**
"If R succeeds"	6	13	4	**37**
"If U succeeds"	0	0	0	0

Table 11: Composite framing scorecard across all four models. Every metric shows the same direction. The contaminated model (3.1*) amplifies framing despite knowing outcomes.

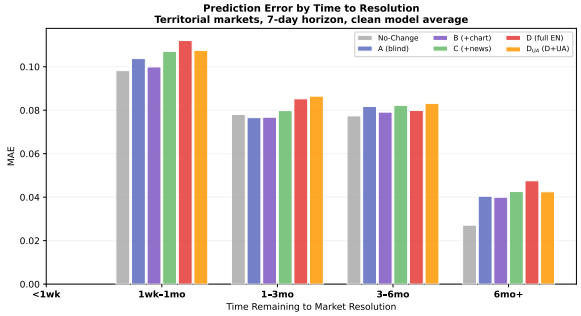

Figure 8: MAE by time remaining to market resolution. Context damage is strongest in the 1wk–1mo window where news flow is densest. Territorial markets, 7-day horizon, clean model average.

M Prediction Error by Time to Resolution

Figure 8 shows prediction error stratified by time remaining until market resolution. Context damage (D exceeding no-change) is concentrated in the 1-week-to-1-month window, where markets are most active and English news coverage is densest. At longer horizons (6+ months), all conditions converge as markets are less liquid and predictions are dominated by the prior.

N Case Study: The Zelenskyy Suit

Consider the Polymarket question "Will Zelenskyy wear a suit before June?" All three clean models predicted YES with high confidence (0.91–0.95), reasoning that a papal funeral demands formal attire – a logically sound inference from generic diplomatic norms. Any Ukrainian would have predicted differently. Zelenskyy has not worn a suit since February 24, 2022; the wartime military clothing is a deliberate political statement, not a wardrobe constraint. Returning to a suit would signal a fundamental shift in how Ukraine frames its wartime posture. The contaminated model, which

knows the outcome, predicted 0.33. The gap between 0.93 and 0.33 is not a reasoning failure – the logic is valid. It is an information ecosystem failure: the English-language corpus encodes "heads of state wear suits to funerals" but not "this particular head of state has made not wearing a suit a defining act of wartime leadership." The models reason fluently from the wrong world model.