

Decomposing Self-Reflection in LLM Agents: Diagnostic Scaffolds, Uncertainty Labels, and Action Diversity

Poli Nemkova
Columbia University
poli.nemkova@columbia.edu

Haeshitha Indukuri
University of North Texas
HaeshithaIndukuri@my.unt.edu

Abstract

Self-reflective agents typically bundle monitoring language with a downstream control policy, making it difficult to identify why reflection changes behavior. We present a six-condition ablation that separates evidence re-exposure, diagnostic questions, uncertainty labels, differentiated follow-up actions, and diagnosis–action alignment. Our testbed is an LLM forecasting agent evaluated with Llama-3.3-70B on 310 armed-conflict escalation cases, plus supplementary GPT-4o and 12-country transfer probes. Diagnostic questions produce comparable F1 to free reflection (0.296 vs. 0.297), and adding a seven-label taxonomy while holding the action fixed changes F1 by only +0.008. The closest matched comparison for action differentiation is taxonomy-only versus random action assignment (F vs. D), which holds the typed monitor and taxonomy fixed while replacing the shared-action policy with randomly varying substantive actions: F1 rises from 0.304 to 0.338. Type-matched routing reaches 0.379, but we detect no significant advantage over random assignment (C vs. D, $p = 0.714$). This underpowered contrast does not establish equivalence; it prevents assigning the gain specifically to matching. Brier scores remain flat and transfer is unreliable. Thus, richer reflective traces do not establish better agent control. Our results motivate controls matched for information, vocabulary, and actions when attributing gains in reflection pipelines and identify simple action perturbations as a necessary baseline for type-specific policies.

1 Introduction

Self-reflection is now a common control pattern in LLM agents: a system inspects an intermediate result, describes what may be wrong, selects a follow-up operation, and tries again (Yao et al., 2023; Shinn et al., 2023; Renze and Guven, 2024). Yet evaluations often compare this entire bundle with no reflection. Such comparisons cannot tell

whether a gain comes from seeing the input again, producing a longer trace, naming a failure type, or executing a different action. This is a credit-assignment problem for agent architecture, not merely a prompting detail.

We study that problem in an LLM forecasting agent that first predicts from structured evidence, then monitors the forecast, selects a reasoning action, and revises its decision. A six-condition ladder separates five commonly bundled components: evidence re-exposure, diagnostic questions, uncertainty labels, differentiated actions, and diagnosis–action alignment. The design has direct analogues in tool-using agents: a failure label such as *missing evidence* has not demonstrated value unless it is separated from the retrieval or replanning action that follows it.

The matched controls change the apparent conclusion. The two prompt variants produce nearly identical predictions, although the contrast is not perfectly factor-matched (F1 = 0.296 vs. 0.297, $p = 1.000$). Adding a seven-label taxonomy while collapsing all non-confident states to one shared action adds only 0.008 F1. We detect no advantage of type-matched over random action assignment ($p = 0.714$), and GPT-4o reverses their ordering in a supplementary probe. With 65 positives this is not evidence of equivalence; the control instead prevents attributing the gain specifically to diagnosis–action alignment. Flat Brier scores and a transfer reversal further rule out a broad claim of improved forecasting skill.

We contribute (1) a controlled ablation template with information-, vocabulary-, and action-matched conditions for reflective agents; (2) evidence that diagnostic structure and uncertainty terminology are insufficient without a distinct behavioral consequence; and (3) evidence that action variation can alter decisions even when correct diagnosis–action alignment is unestablished. Fluent self-diagnosis should not receive credit for an

agent-level gain until simpler action perturbations have been excluded.

2 Related Work

Agent architectures interleave reasoning with actions, observations, and feedback. ReAct couples reasoning traces to external actions (Yao et al., 2023), while Reflexion adds verbal feedback across attempts (Shinn et al., 2023). Interactive evaluations such as AgentDojo test agents that execute tools over changing, potentially adversarial observations (Debenedetti et al., 2024); tool-agent trace studies further identify distinct failure types at the action interface (Xiong et al., 2025). Together, these works locate reliability in the reasoning–action loop, but end-to-end success does not identify which internal component caused a change.

Metacognitive prompting asks models to inspect their reasoning (Wang and Zhao, 2024), and unconstrained reflection can reinforce an initial commitment (Liang et al., 2024). We therefore treat reflection as a monitor–control pipeline and perform component-level credit assignment. Unlike uncertainty estimation based on confidence or output distributions (Hüllermeier and Waegeman, 2021), we manipulate the connection between a verbal diagnosis and follow-up behavior.

3 Reflection Agent and Controlled Ablation

Agent architecture. The agent has three stages: a forecaster produces a prediction and confidence; a monitor re-examines that state and the evidence; and a controller selects a follow-up reasoning action before one final revision. Conditions intervene only in the monitor and controller, leaving the forecaster, evidence, and revision opportunity fixed. Thus, the experiment tests which internal control signal changes the agent’s terminal decision rather than comparing unrelated prompting pipelines.

Task and data. Given a 60-day observation window, the model predicts whether ACLED-recorded fatalities (Raleigh et al., 2010) will exceed 130% of the preceding two-week baseline during the next 14 days. The primary test set contains 310 cases (62 each from Ethiopia, Myanmar, Somalia, Sudan, and Ukraine), ordered temporally within each country; the earliest 60% form the training period and the latest 40% the test period. The test set spans late 2023–early 2026 and contains 65

Condition	Questions	Taxonomy	Action
Single-shot (A)	–	–	none
Free reflection (E)	no	partial [†]	model-selected
Diagnostic refl. (B)	yes	no	shared
Taxonomy only (F)	yes	yes	shared
Random action (D)	yes	yes	random
Type-matched (C)	yes	yes	matched

Table 1: Six conditions separate diagnostic questions, uncertainty labels, action differentiation, and diagnosis–action alignment. Letters preserve correspondence with the original experiment logs. [†]Condition E contains residual uncertainty-related terminology in auxiliary evidence blocks but is not presented with the explicit seven-type taxonomy (Appendix C).

positive cases (21.0%). The temporal split prevents future cases from informing earlier forecasts. Each condition receives the same evidence bundle: lagged ACLED statistics, socioeconomic indicators, XGBoost and random-forest scores and disagreement, source quality assessments, and LDA topic signals. Appendix A gives further construction details.

Conditions. Table 1 shows the intervention chain. All reflection conditions receive the same current forecast and evidence and permit one revision. The five diagnostic questions cover evidence volume, recency, model agreement, counterevidence, and forecast weaknesses. The taxonomy contains six substantive uncertainty types plus CONFIDENT; each maps to a fixed reasoning action. *Taxonomy only* keeps the labels but maps every non-confident diagnosis to comparative reasoning. *Random action* draws a substantive type after each monitor call (the case–iteration level) using a fixed seed of 42, then executes its associated action; it does not sample one global random mapping. This preserves the action machinery while breaking diagnosis–action alignment. Full definitions and mappings appear in Appendix B.

Models and evaluation. The primary experiments use Llama-3.3-70B (Grattafiori et al., 2024) at temperature 0.2. Each condition is run three times; binary decisions are aggregated by majority vote and confidence by the mean. We report positive-class F1 and recall because positives are rare, but also calibrated Brier score to distinguish thresholded behavior from probabilistic quality. Brier scores use five-fold cross-validated Platt scaling (Platt, 1999). We use paired McNemar tests (McNemar, 1947) for binary predictions and 2,000

stratified bootstrap resamples for descriptive F1 intervals. GPT-4o (OpenAI, 2024) and a 300-case, 12-country transfer set are supplementary, single-run probes and are not treated as confirmatory replications.

4 Results

Diagnostic scaffolding changes no decisions.

Free reflection and diagnostic reflection produce comparable results ($F1 = 0.297$ vs. 0.296 ; $\text{recall} = 0.508$ for both), with indistinguishable case-level decisions (McNemar $p = 1.000$). We treat this contrast as suggestive rather than a clean single-factor test: the free- and diagnostic-reflection prompts differ not only in the presence of the five diagnostic questions but also in action-set breadth and residual uncertainty-related terminology (Appendix C). Accordingly, the near-identical outcomes show that the full prompt variants behave similarly, but do not isolate the causal contribution of the five questions alone. Our cleanest test of whether action differentiation itself changes decisions is instead the taxonomy-only-versus-random contrast below (F vs. D), which holds the typed monitor and taxonomy fixed while replacing the shared-action policy with randomly varying substantive actions; the type-matched-versus-random contrast (C vs. D) instead holds the action space itself fixed and isolates diagnosis–action alignment specifically.

Taxonomy vocabulary is also insufficient.

Adding seven uncertainty labels while holding the substantive action fixed changes F1 from 0.296 to 0.304 ($\Delta = +0.008$), with strongly overlapping bootstrap intervals and unchanged recall. The model’s use of more specific epistemic language does not by itself improve discrimination.

Action variation is associated with changed decisions, but type-specificity is unestablished.

Holding the typed monitor and taxonomy fixed, moving from one shared action to randomly varying actions raises F1 from 0.304 to 0.338 (taxonomy-only vs. random action, F vs. D)—our cleanest descriptive evidence that action differentiation is associated with changed behavior, independent of diagnosis–action matching. Random action assignment reaches $F1 = 0.338$, a directional increase of 0.042 over diagnostic reflection (McNemar $p = 0.497$). Type-matched assignment reaches 0.379 , but its 0.041 advantage over random assignment is also nonsignificant ($p = 0.714$). With 65

positive cases, this cannot establish equivalence or exclude a meaningful matching benefit. Even type-matched versus single-shot is nonsignificant ($p = 0.202$). Thus varying the action can alter the operating point—recall rises from 0.385 to 0.677 —but this study cannot isolate correct matching as the cause. Positive-prediction rate (Table 2) moves in step with recall across all six conditions (0.371 – 0.540), the more direct behavioral signature of a shifted operating point than accuracy-based tests can provide on their own.

Calibration and transfer bound the claim.

All six conditions have post-calibration Brier scores between 0.165 and 0.166 . Thus, the observed F1 ordering is compatible with changed classification behavior rather than better probability estimates. On 12 unseen countries, single-shot, diagnostic reflection, and type-matched action achieve $F1 = 0.260$, 0.290 , and 0.280 , respectively: generic reflection slightly exceeds type-matched action and neither significantly improves on the baseline. Country-level gains in the primary set are concentrated in Myanmar and Ukraine (Section 4.1), further limiting generalization.

4.1 Where Do the Behavioral Changes Occur?

Country concentration.

Table 3 shows that the aggregate ordering is not geographically uniform. In Myanmar, free and diagnostic reflection recover modestly from a single-shot F1 of zero, but taxonomy vocabulary changes almost nothing ($0.154 \rightarrow 0.162$). Random action assignment then reaches 0.316 , and type-matched assignment 0.353 . Ukraine shows a sharper version of the same pattern: the three shared-action conditions remain near 0.10 , while random and matched actions reach 0.432 and 0.500 . These two countries supply most of the aggregate increase. By contrast, all six conditions are nearly indistinguishable in Ethiopia, and Somalia has no monotonic ablation ordering. Sudan is the only country where taxonomy-only reflection (0.370) already approaches type-matched action (0.373). Both Myanmar and Ukraine also show the weakest single-shot performance in the primary set ($F1 = 0.000$ and 0.167 ; Table 3), so any subsequent behavioral change has more room to register as a large relative gain there, independent of whether the underlying mechanism is meaningful. With only 8–18 positive cases per country, these patterns are descriptive rather than evidence of country-specific mechanisms.

Condition	F1	Recall	Precision	Pos.-pred. rate	Brier
Single-shot (A)	0.278	0.385	0.218	0.371	0.166
Free reflection (E)	0.297	0.508	0.210	0.508	0.165
Diagnostic refl. (B)	0.296	0.508	0.209	0.510	0.166
Taxonomy only (F)	0.304	0.508	0.217	0.491	0.166
Random action (D)	0.338	0.585	0.238	0.516	0.166
Type-matched (C)	0.379	0.677	0.263	0.540	0.166
<i>Always escalation (hard)</i>	<i>0.347</i>	<i>1.000</i>	<i>0.210</i>	<i>1.000</i>	<i>0.790</i>
<i>Constant $p = 0.21$ (soft)</i>	—	—	—	—	<i>0.166</i>

Table 2: Primary results ($n = 310$, Llama-3.3-70B, three-run majority vote). Recall, precision, and positive-prediction rate are majority-vote quantities; precision and positive-prediction rate for the six reflection conditions are computed from the reported F1 and recall. Bootstrap 95% CIs on F1 (omitted here for space) are [0.200,0.353], [0.233,0.359], [0.232,0.359], [0.221,0.383], [0.272,0.400], and [0.322,0.436] for A, E, B, F, D, and C respectively; paired significance claims use McNemar tests. Two trivial baselines bound the six conditions: a hard *always-escalation* rule matches or exceeds five of six on F1 with a poor Brier score (0.790), while a *soft* rule that outputs the constant base rate ($p = 0.21$) for every case—never crossing a decision threshold, hence no F1/recall entry—achieves a Brier score matching the six real conditions almost exactly. The six conditions’ near-constant precision (0.209–0.263, against a 21.0% base rate) and Brier scores (0.165–0.166) together indicate that F1 gains here largely reflect a shift in operating point (positive-prediction rate 0.371–0.540) rather than calibrated discrimination above the base rate.

	A	E	B	F	D	C
Ethiopia	.148	.148	.148	.148	.167	.167
Myanmar	.000	.150	.154	.162	.316	.353
Somalia	.427	.430	.467	.430	.394	.475
Sudan	.233	.357	.347	.370	.333	.373
Ukraine	.167	.100	.091	.100	.432	.500

Table 3: F1 by country. The largest changes under differentiated actions occur in Myanmar and Ukraine; Ethiopia and Somalia are comparatively flat.

Tier	n	A	E	B	D	C
Catchable	29	.682	.792	.816	.792	.926
Borderline	165	.125	.128	.088	.222	.195
Deceptive	116	.179	.200	.217	.203	.255

Table 4: Exploratory F1 by pre-defined difficulty tier. Taxonomy-only (F) is omitted because tier-level estimates with few positives are unstable.

Difficulty does not yield a simple routing advantage. Cases were grouped before evaluation using observable properties: *catchable* cases have strong momentum, *borderline* cases lie near the escalation threshold, and *deceptive* cases contain a misleading recent momentum dip. Type-matched action is strongest on catchable cases, but random action exceeds it on the much larger borderline tier (0.222 vs. 0.195). The deceptive tier shows only a modest matched-action advantage. Although exploratory, Table 4 again argues against treating correct typed assignment as a uniformly beneficial mechanism.

The GPT-4o probe supports the action-variation interpretation. In a single run on the same 310 cases, diagnostic reflection reaches $F1 = 0.272$ and taxonomy-only reflection 0.304 ($p = 0.773$), while random action reaches 0.345 ($p = 0.025$ vs. diagnostic reflection). Crucially, type-matched action scores 0.328, below random assignment. This probe suggests that varying actions can matter on a second backbone, but it does not replicate an advantage for correct type–action alignment. Because the probe lacks repeated runs, omits free reflection, and has a missing recall value for the matched condition, we treat it as supplementary rather than confirmatory.

5 Discussion

These results separate claims often collapsed under *self-reflection*. More analysis is not sufficient: five diagnostic questions do not change predictions relative to free reflection. More specific uncertainty language is not sufficient either: the taxonomy-only condition behaves like its taxonomy-free control. A fluent diagnostic trace should therefore not be treated as evidence of improved agent control.

Matched controls change the interpretation. Comparing single-shot prediction with the full type-matched agent suggests a 0.101 F1 gain, but that contrast changes evidence exposure, diagnostic structure, terminology, and follow-up action simultaneously. Intermediate controls attribute only +0.019 to free reflection, essentially zero to struc-

tured questions, and +0.008 to taxonomy vocabulary with a shared action. Random and matched assignment both move performance, without a detected difference. This is not equivalence—smaller matching benefits remain possible—but it means the full-agent contrast cannot identify matching as the mechanism. Reflection components therefore require information-, vocabulary-, and action-matched controls before a gain is credited to the agent’s diagnosis.

Changed behavior is not yet improved skill.

Action variation may break a default no-escalation commitment, add prompt diversity, or merely increase positive predictions. Rising recall with flat calibrated Brier scores supports an operating-point shift, not a general improvement in probability estimation. Future work should compare precision–recall curves and raw probabilities at matched recall or positive-prediction rates. Cost also matters: the type-matched agent uses six model calls per case versus two for the single-shot pipeline. Performance-per-cost and type-specific control remain unestablished.

Implications beyond forecasting. The same confound arises whenever an agent labels a failure and changes what it does: *missing evidence* triggers retrieval, *navigation error* triggers backtracking, or *incorrect assumption* triggers a new test. Evaluations should hold observations fixed, collapse labels onto a shared recovery, and randomize or rotate recoveries before crediting a matched policy. They should also distinguish task success from generic increases in action rate, search depth, or token use. Here actions are prompt instructions and cannot alter state or acquire evidence. Tool calls could amplify both correct and incorrect matching, so matched and randomized controls remain necessary. The transfer reversal likewise shows that diagnoses alone do not validate a recovery policy.

6 Conclusion

Self-reflection is an agent pipeline, not one intervention. Questions and uncertainty labels show little measurable benefit alone; differentiated actions are associated with changes in decisions, but we detect no matching advantage over random assignment. This is not equivalence: it means the full-agent gain cannot yet be attributed to matching, calibrated Brier score, or reliable transfer.

Reflection studies should control information,

language, action diversity, and matching before crediting self-diagnosis. In forecasting and interactive agents alike, a failure label is useful only if its recovery beats simpler action perturbations.

Limitations

The primary test set contains only 65 positive cases, limiting power for pairwise and country-level comparisons. In particular, the nonsignificant matched-versus-random test cannot establish equivalence or rule out a practically meaningful matching benefit. All primary experiments use one domain and one principal backbone; GPT-4o and transfer results are single-run probes with incomplete condition coverage. The hand-built taxonomy is a design artifact, and the available actions are prompted reasoning strategies rather than genuine retrieval or source adjudication. One mapped action in particular, `INSUFFICIENT_EVIDENCE` $\rightarrow$ *request additional sources* (Table 5), cannot retrieve anything the agent does not already have; because this diagnosis is followed deterministically under type-matching but only by chance under random assignment, it is a plausible confound for the matched-versus-random contrast in either direction, independent of whether correct diagnosis–action alignment has genuine value. Consequently, conclusions about action-matched controls are so far demonstrated in one non-interactive agent domain. Condition D also uses one seeded randomization schedule rather than multiple randomization seeds, limiting inference about variability across random assignments. Post-hoc Brier scores cannot replace threshold sweeps or precision–recall analysis. Finally, temporally adjacent observation windows may share historical evidence even though labels and forecasts respect temporal order; this reflects repeated real-world forecasting but reduces case independence.

Ethical Considerations

Conflict forecasts can produce harmful false alarms or missed escalations and may be misused to justify political or operational decisions. The systems studied here are research prototypes, not decision-support tools. Their weak transfer performance is direct evidence against deployment without local validation, human review, and clear accountability for downstream decisions.

References

Edoardo DeBenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for

llm agents. *Advances in neural information processing systems*, 37:82895–82920.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

OpenAI. 2024. *GPT-4o system card*. Technical report, OpenAI.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Matthew Renze and Erhan Guven. 2024. Self-reflection in large language model agents: Effects on problem-solving performance. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

Qian Xiong, Yuekai Huang, Ziyu Jiang, Zhiyuan Chang, Yujia Zheng, Tianhao Li, and Mingyang Li. 2025. *Butterfly effects in toolchains: A comprehensive analysis of failed parameter filling in LLM tool-agent systems*. In *Findings of the Association*

for *Computational Linguistics: EMNLP 2025*, pages 16712–16729. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

A Additional Experimental Detail

Forecast construction. Each case uses information available through the end of its 60-day observation window. The forecast target compares fatalities in the subsequent 14 days with 130% of the fatalities in the final 14 days of the observation window. Cases are ordered before splitting within country; no future target value is included in an evidence bundle. All six conditions operate on identical serialized bundles and differ only in their reflective instructions and action selection.

Transfer set. The 300-case transfer probe contains 25 cases each from the Democratic Republic of the Congo, Belarus, Poland, Israel, Palestine, Lebanon, Iran, Yemen, Syria, Colombia, Haiti, and Mexico, using the same label construction. No system choice was tuned on these outcomes.

B Condition and Action Definitions

The diagnostic protocol asks the model to assess evidence volume and recency, agreement between statistical models, plausible counterevidence, and the principal weakness of the current forecast. The conditions differ as follows:

Single-shot (A)

One forecast, without reflection.

Free reflection (E)

Reconsiders the evidence without the five questions.

Diagnostic reflection (B)

Answers the questions, then either exits or uses comparative reasoning; no uncertainty labels are shown.

Taxonomy only (F)

Assigns a named type, but every non-confident type receives the same comparative-reasoning instruction as B.

Random action (D)

Draws uniformly among six substantive

types and executes the corresponding action after each monitor call. The draw therefore occurs at the case-iteration level, not as a single random mapping shared by all cases. Python’s random generator is seeded once at 42 when the script starts.

Type-matched action (C)

Diagnoses a type from the evidence and executes its fixed mapped action.

Uncertainty type	Mapped action
Insufficient evidence	Request sources
Conflicting sources	Comparative reasoning
Distribution shift	Context recalibration
Low-quality data	Reweight sources
Temporal inconsistency	Prioritize recency
Model ambiguity	Sample alternatives
Confident	Exit

Table 5: The fixed type–action map. Actions are reasoning instructions; the agent cannot retrieve new sources in this experiment.

B.1 Example reflection trace

The following shortened, verbatim excerpt is from the supplied Groq run log for case_MM_094_20231008. It illustrates why fluent reflection is not itself evidence of correct control: the monitor and action misinterpret the direction of the same model probability while producing a superficially coherent revision.

Initial forecast: no_escalation, confidence 0.67. “The recent 14-day fatality count (548) is lower than the prior 14 days (701) ... The ML model scores also suggest a low probability of escalation, with a consensus probability of 0.076.”

Monitor: CONFLICTING_SOURCES, confidence 0.62. “The forecast prediction of no escalation is based on a decreasing momentum (0.85) ... but this is contradicted by the ML consensus probability (0.076).”

Action output: comparative reasoning. “I trust the ML consensus probability and the high average daily fatalities over the qualitative prediction and the decreasing momentum.” Revised prediction: escalation, confidence 0.72.

Synthesis: escalation, confidence 0.70. “The high average daily fatalities (51.9) and the recent increase in fatalities in Week 9 (143) drove the change.”

The probability 0.076 supports *no escalation*, rather than contradicting it. The trace therefore makes behavioral change visible while showing

that neither a named diagnosis nor a confident revision guarantees a coherent recovery. This distinction is central to the matched and randomized-action comparison.

C Executable Prompt Templates

The following prompt builders are reproduced verbatim from the evaluation scripts. Dynamic evidence, model scores, country names, and prior outputs are interpolated at runtime. Conditions C and D use the same typed monitor; D overrides the parsed type with a seeded random draw before control. Condition F also uses the typed monitor but collapses every non-CONFIDENT diagnosis to comparative reasoning. The system message for every call is: *You are a careful conflict analyst. Follow the output format exactly.* Some source-code comments retain legacy internal condition labels; throughout the paper, the canonical condition labels are those in Table 1.

C.1 Common forecast prompt

```
284 FEW_SHOT_EXAMPLES = ""
285 EXAMPLES OF REASONING (historical cases — no future data shown):
286 ---
287 ---
288 EXAMPLE 1 — Sudan, stable baseline, sudden spike incoming
289 Signals: baseline_2w=32, trend=stable, momentum=0.95, high socioeconomic stress,
290       new armed group movements reported in recent weeks
291 PREDICTION: escalation | CONFIDENCE: 0.72
292 REASONING: Despite stable fatality trend, new armed actor activity and high
293 socioeconomic stress suggest latent tension. Momentum near 1.0 gives no clear
294 directional signal, but qualitative indicators warrant elevated risk.
295 ---
296 ---
297 EXAMPLE 2 — Myanmar, sharply increasing trend
298 Signals: baseline_2w=41, trend=increasing, momentum=2.1,
299       battles expanding to new districts, civilian targeting increasing
300 PREDICTION: escalation | CONFIDENCE: 0.88
301 REASONING: Strong momentum (2.1x) combined with geographic spread and new
302 event types (civilian targeting) strongly indicate escalation. Multiple
303 converging signals justify high confidence.
304 ---
305 ---
306 EXAMPLE 3 — Ethiopia, decreasing trend
307 Signals: baseline_2w=671, trend=decreasing, momentum=0.62,
308       recent ceasefire negotiations, reduced battle events
309 PREDICTION: no_escalation | CONFIDENCE: 0.85
310 REASONING: Sharply falling momentum (0.62x) and active peace negotiations
311 suggest de-escalation. High absolute fatality count is the regional baseline,
312 not a current spike. No new actor or geographic expansion signals.
313 ---
314 ---
315 EXAMPLE 4 — Ukraine, borderline case
316 Signals: baseline_2w=84, trend=stable, momentum=1.15,
317       minor geographic shifts, no new actors
318 PREDICTION: no_escalation | CONFIDENCE: 0.58
319 REASONING: Modest momentum increase (1.15x) without qualitative shifts in
320 actor dynamics or geography is insufficient for confident escalation prediction.
321 Slight upward trend warrants caution but not a positive prediction.
322 ---
323 ""
324
325
326 def _build_forecast_prompt(region: str, target: str, evidence_text: str, summary: str,
327                           ml_evidence: Optional[Any] = None,
328                           topic_evidence: Optional[Any] = None) -> str:
329     if target == "eruption":
330         label_desc = "violent eruption"
331         definition = (
332             "Eruption: a sudden, sharp spike in violence — significantly more fatalities, "
333             "new armed actors, or major geographic expansion compared to the recent baseline."
334         )
335         eruption_prior = ""
336     BASE RATE WARNING — READ BEFORE REASONING:
337     Violent eruptions are RARE events. In historical data across conflict zones, eruptions
338     occur in roughly 5–10% of 14-day windows. This means the DEFAULT prediction is
339     no_eruption unless you have strong, specific evidence to the contrary.
340
341     GATE: You may only predict eruption if you can satisfy BOTH of the following:
342     1. QUANTITATIVE SIGNAL: fatalities in the most recent 14 days are at least 2× the
343       60-day weekly average, OR ML consensus probability > 0.40 for eruption.
344     2. QUALITATIVE SIGNAL: at least one of — new armed actor entering the conflict,
345       major geographic expansion to a new region, or a documented trigger event
346       (offensive launched, ceasefire collapse, mass displacement spike).
347
348     If you cannot satisfy both gates with specific numbers/names from the evidence,
349     your prediction MUST be no_eruption. Set confidence accordingly (0.65–0.80 for
350     no_eruption when evidence is stable; only exceed 0.70 for eruption if both gates clear).
351     ""
352     else:
353         label_desc = "conflict escalation"
354         definition = (
355             "Escalation: a meaningful worsening of conflict — could be rising fatalities, "
```

```

356     "new frontlines opening, new armed groups engaging, increased civilian targeting, "
357     "or a qualitative shift in conflict intensity. NOT just a small fatality fluctuation."
358 )
359     eruption_prior = ""
360
361     # Build ML evidence block — injected as an additional signal stream
362     ml_block = ""
363     if ml_evidence is not None:
364         label_ev = getattr(ml_evidence, target, None)
365         if label_ev is not None:
366             ml_block = f'""
367 ML MODEL EVIDENCE (GDELT-based statistical models — treat as one additional signal):
368 {label_ev.to_prompt_str()}
369
370 NOTE: ML model scores reflect patterns in news media volume, tone, and conflict event types
371 in the 7–60 days before this prediction window. They are NOT ground truth.
372 High consensus probability supports escalation; low consensus supports stability.
373 Model disagreement (high std) indicates ambiguous or conflicting signals.
374 ""
375
376     # Build topic evidence block
377     topic_block = ""
378     if topic_evidence is not None:
379         topic_block = f'""
380 NEWS THEME EVIDENCE (LDA topic model on GDELT articles):
381 {topic_evidence.to_prompt_str()}
382
383 NOTE: Dominant topics reflect what news media was covering in the 30 days before
384 this prediction window. A DISTRIBUTION_SHIFT flag means the news themes are
385 unusual compared to baseline patterns — this may indicate a novel situation
386 where historical patterns are less reliable.
387 ""
388
389     return f'""You are an expert conflict forecasting analyst specializing in {region}.
390
391 TASK: Predict {label_desc} in {region} in the NEXT 14 DAYS.
392
393 DEFINITION: {definition}
394 {eruption_prior}
395 IMPORTANT: Escalation is NOT defined by a simple fatality ratio threshold.
396 You must consider ALL of the following signals:
397 - Trend direction: is violence increasing, stable, or decreasing?
398 - Momentum: is the recent 14-day period higher or lower than prior 14 days?
399 - Event type shifts: are there new types of violence (battles, civilian targeting)?
400 - Geographic spread: are new areas being affected?
401 - Actor dynamics: are new armed groups entering or existing ones escalating?
402 - Socioeconomic stress: do indicators suggest underlying instability?
403
404 {FEW_SHOT_EXAMPLES}
405
406 EVIDENCE (60-day observation window)
407 {evidence_text}
408
409 SYSTEM SUMMARY:
410 {summary}
411 {ml_block}{topic_block}
412 STEP-BY-STEP REASONING:
413 1. What is the recent 14-day fatality count vs prior 14 days and vs 60-day weekly average?
414    State exact numbers.
415 2. Are there qualitative shifts in event types, actors, or geography?
416 3. What do socioeconomic indicators suggest about underlying stability?
417 4. What do the ML model scores suggest, and do they agree or disagree with the qualitative signals?
418 5. Do the news themes (topics) suggest anything unusual about this period?
419 6. What is the strongest case FOR {label_desc} in the next 14 days?
420 7. What is the strongest case AGAINST {label_desc}?
421 {"8. ERUPTION GATE CHECK: Do you satisfy BOTH the quantitative AND qualitative gate above?" if target == "eruption" else "8. Weigh all signals and
422    ↪ make a final judgment."}
423
424 OUTPUT FORMAT (follow exactly):
425 PREDICTION: escalation OR no_escalation
426 CONFIDENCE: 0.0 to 1.0
427 KEY_EVIDENCE: comma-separated list of the 3-5 most important signals
428 CONTRADICTING_EVIDENCE: signals pointing the other way, or "none"
429 REASONING: 2-4 sentences explaining your judgment across multiple signal types.
430 ""

```

C.2 Typed monitor prompt (C, D, and F)

```

432 def _build_monitor_prompt(
433     region: str,
434     target: str,
435     forecast: ForecastState,
436     evidence_text: str,
437     ml_evidence: Optional[Any] = None,
438     topic_evidence: Optional[Any] = None,
439 ) -> str:
440     opposite = "no_escalation" if forecast.prediction == "escalation" else "escalation"
441
442     # Build ML signal block for monitor

```

```

443 ml_monitor_block = ""
444 if ml_evidence is not None:
445     label_ev = getattr(ml_evidence, target, None)
446     if label_ev is not None:
447         direction = "supports" if label_ev.prob_mean > 0.35 else "contradicts" if label_ev.prob_mean < 0.15 else "is ambiguous about"
448         ml_monitor_block = f""
449 ML MODEL SIGNAL:
450 Consensus probability: {label_ev.prob_mean:.3f} (std={label_ev.prob_std:.3f}) — this {direction} the forecast prediction
451 Model votes: {label_ev.vote_positive}/{label_ev.vote_total} predict positive
452 ML uncertainty flag: {label_ev.uncertainty_flag} — {label_ev.uncertainty_reason}
453 Per-model: {' '.join(f"{s.model_name}=p{s.probability:.2f}" for s in label_ev.scores)}
454
455 Consider: Does the ML signal agree with the qualitative forecast? If not, does this disagreement
456 represent a genuine CONFLICTING_SOURCES signal worth flagging?
457 ""
458
459 # Build topic shift signal block for monitor
460 topic_monitor_block = ""
461 if topic_evidence is not None:
462     topic_monitor_block = f""
463 NEWS THEME SIGNAL:
464 Dominant topics: {' '.join(t.to_prompt_str() for t in topic_evidence.top3_topics)}
465 Distribution shift: {topic_evidence.shift_flag} (cosine_dist={topic_evidence.cosine_dist:.3f})
466 Shift reason: {topic_evidence.shift_reason}
467
468 Note: cosine_dist > 0.35 is the threshold for a meaningful topic distribution shift.
469 A shift alone does NOT automatically mean DISTRIBUTION_SHIFT uncertainty — it is only
470 one of four required triggers. Check whether it materially changes the forecast direction.
471 ""
472
473 # Eruption-specific gate reminder for monitor
474 eruption_gate_block = ""
475 if target == "eruption" and forecast.prediction == "escalation":
476     eruption_gate_block = f""
477 ERUPTION GATE AUDIT (prediction=eruption was made — verify both gates were satisfied):
478 Gate 1 (quantitative): Are recent 14d fatalities  $\geq 2 \times$  the 60-day weekly average?
479 OR is ML eruption prob_mean > 0.40?
480 Gate 2 (qualitative): Is there a named new actor, geographic expansion, or
481 documented trigger event in the evidence?
482 If EITHER gate is not clearly satisfied with specific numbers/names, this forecast
483 is likely a FALSE POSITIVE and the monitor should flag CONFLICTING_SOURCES or
484 recommend downgrading confidence significantly.
485 ""
486
487 return f""You are a critical evaluator stress-testing a conflict forecast for {region}.
488 Your job is to find weaknesses — not to validate the forecast.
489
490 FORECAST TO EVALUATE
491 Prediction: {forecast.prediction}
492 Stated confidence: {forecast.confidence}
493 Reasoning: {forecast.reasoning}
494 Key evidence used: {' '.join(forecast.key_evidence_used) or 'not specified'}
495 Contradicting evidence noted: {' '.join(forecast.contradicting_evidence) or 'none'}
496
497 STRUCTURED EVIDENCE (use this for all quantitative counting in Q1–Q5)
498 {evidence_text}
499 {ml_monitor_block}{topic_monitor_block}{eruption_gate_block}
500 IMPORTANT: When counting [acled_weekly] rows in Q1, count ONLY lines tagged
501 [acled_weekly] in the STRUCTURED EVIDENCE section above. Do NOT count lines
502 from any NEWS NARRATIVE SUMMARY section — that section is supplementary context
503 only and should not affect evidence volume assessment.
504 If a NEWS NARRATIVE SUMMARY is present, use it to inform Q4 (counter-argument)
505 and Q5 (weakness check) only — it may name specific actors, events, or anomalies
506 that are useful for those questions.
507 ANSWER THESE DIAGNOSTIC QUESTIONS FIRST (be specific, use numbers from the evidence):
508 Q1. EVIDENCE VOLUME: How many weeks of ACLED data are present? How many total events?
509 Is there enough data to make a reliable forecast at all?
510 Q2. RECENCY DELTA: What are the exact fatality counts for the most recent 14 days vs
511 the prior 14 days? Compute the % difference. Is it >40% AND does it contradict
512 the 60-day baseline direction? (Small differences <40% are NORMAL noise.)
513 Q3. ML AGREEMENT: Does the ML consensus probability agree or disagree with the
514 qualitative prediction? Specifically: if prediction=escalation, is prob_mean > 0.35?
515 If prediction=no_escalation, is prob_mean < 0.35? A strong contradiction (e.g.
516 prob_mean=0.6 but prediction=no_escalation) is meaningful; a weak one is not.
517 Q4. COUNTER-ARGUMENT: State the single strongest specific argument for {opposite}.
518 Can you make it using actual numbers? If not, the forecast is probably well-supported.
519 Q5. WEAKNESS CHECK: Is there one specific, concrete weakness — a data gap, a
520 contradicting data point, an unusual pattern — that would materially change
521 the forecast if resolved? Name it precisely or say "no material weakness found."
522
523 THEN choose the PRIMARY uncertainty type using this decision tree. Work through it
524 IN ORDER and stop at the FIRST condition that is clearly and specifically true:
525
526 1. Is the evidence window <30 days OR total event count <5?
527 → INSUFFICIENT_EVIDENCE
528
529 2. Do specific named evidence items directly contradict each other with actual numbers?
530 Examples: "recent 14d fatalities +80% BUT ML prob_mean=0.08" OR "ceasefire reported
531 BUT fatalities rising week-over-week". Vague contradiction does NOT qualify.
532 → CONFLICTING_SOURCES

```

533
534 3. From Q2: is the recent 14d vs prior 14d difference STRICTLY >40% AND does it point
535 in the OPPOSITE direction to the 60-day baseline trend? Both conditions required.
536 → TEMPORAL_INCONSISTENCY
537
538 4. Did Q5 identify data that is patchy, missing key periods, or from unreliable sources
539 (low source quality score, high tier4_pct)?
540 → LOW_QUALITY_DATA
541
542 5. From Q4: did you FAIL to construct a convincing counter-argument using actual numbers?
543 AND from Q3: does the ML signal agree with the qualitative prediction (or is neutral)?
544 AND from Q5: no material weakness was found?
545 → CONFIDENT ← use this whenever the evidence coherently supports the forecast
546
547 6. Is this case structurally unusual in a way that specifically undermines the forecast?
548 This requires ONE of the following measurable triggers — you must name which one:
549 a) Topic cosine_dist_from_train_mean > 0.35 (news themes are genuinely anomalous)
550 b) A single week has fatalities > 3 × the 60-day weekly average AND it is the most
551 recent week (not a historical outlier)
552 c) A named armed actor type appears in the evidence that has NOT appeared in the
553 prior 60 days for this region (genuinely novel actor dynamics)
554 d) The observation window overlaps with a documented structural break: ceasefire
555 signed, major offensive launched, government collapse — name it specifically
556 If you cannot cite one of (a)–(d) with a specific value or name, do NOT use this type.
557 → DISTRIBUTION_SHIFT
558
559 7. ONLY if you can write two specific sentences — one strong FOR, one strong AGAINST —
560 both using actual numbers, and neither is clearly stronger than the other:
561 → MODEL_AMBIGUITY (CONFLICTING_SIGNALS must contain specific numbers)
562
563 8. A material weakness exists but genuinely doesn't fit any category above:
564 → UNKNOWN_UNCERTAINTY (UNKNOWN_DESCRIPTION required — be specific)
565
566 CALIBRATION RULES — read carefully:
567 - CONFIDENT does NOT mean perfect certainty. Use it when evidence coherently points
568 one direction, even at moderate confidence (0.55–0.72). Most cases should exit here.
569 - TEMPORAL_INCONSISTENCY requires BOTH: >40% delta AND direction contradicts baseline.
570 Do NOT use it for normal week-to-week fluctuation. This is the most over-used type.
571 - DISTRIBUTION_SHIFT requires ONE of four specific measurable triggers (cosine_dist>0.35,
572 recent week >3 × average, novel named actor, documented structural break). A region
573 being "volatile" or "complex" does NOT qualify. If you can't name the specific trigger
574 with a value, use CONFIDENT with slightly lower confidence instead.
575 - CONFLICTING_SOURCES requires named, specific contradictions with numbers. "The
576 situation is complex" does not qualify.
577 - MODEL_AMBIGUITY is a last resort. If you are using it because things are "uncertain",
578 choose INSUFFICIENT_EVIDENCE or CONFIDENT with lower confidence instead.
579 - Do NOT anchor to the forecast's stated confidence. Assess independently.
580
581 OUTPUT FORMAT (follow exactly, no extra text):
582 UNCERTAINTY_TYPE: <one type from the list above>
583 UNKNOWN_DESCRIPTION: <required if UNKNOWN_UNCERTAINTY, else "n/a">
584 CONFIDENCE: <0.0 to 1.0 — your independent assessment>
585 EVIDENCE_QUALITY: <0.0 to 1.0>
586 EVIDENCE_GAPS: <specific gaps found, or "none">
587 CONFLICTING_SIGNALS: <specific conflicts found, or "none">
588 RECOMMENDED_ACTION: <one of: request_additional_sources / comparative_reasoning / context_recalibration / reweight_sources / recency_prioritization
589 ↔ / sample_alternative_forecasts / exit>
589 REASONING: 2-4 sentences referencing specific numbers from the evidence.
590 ACTION_REASONING: 1-2 sentences explaining why this action addresses the weakness.
591 ""
592
593

C.3 Diagnostic reflection prompt (B)

```

430 def _build_monitor_prompt(
431     region: str,
432     target: str,
433     forecast: ForecastState,
434     evidence_text: str,
435     ml_evidence: Optional[Any] = None,
436     topic_evidence: Optional[Any] = None,
437 ) -> str:
438     """
439     CONDITION B — Generic Reflection Monitor (no uncertainty taxonomy).
440
441     This is the ablation variant for the typed-vs-generic experiment.
442     The monitor prompts the LLM to reflect on forecast weaknesses and revise
443     confidence, but WITHOUT the structured uncertainty taxonomy or decision tree.
444     Everything else (evidence blocks, parser-compatible output format) is unchanged.
445
446     Condition mapping:
447     A = baseline (single-shot, no reflection) — baseline_results CSV
448     B = THIS FILE (generic reflection, no typing) — run with --run-id 201+
449     C = typed monitor-only — monitor_only_ablation CSV
450     D = full agent (typed monitor + loop) — metacognitive_results CSV
451     """
452     opposite = "no_escalation" if forecast.prediction == "escalation" else "escalation"
453
454     # Build ML signal block — kept identical to original for fair comparison

```

```

455 ml_monitor_block = ""
456 if ml_evidence is not None:
457     label_ev = getattr(ml_evidence, target, None)
458     if label_ev is not None:
459         direction = "supports" if label_ev.prob_mean > 0.35 else "contradicts" if label_ev.prob_mean < 0.15 else "is ambiguous about"
460         ml_monitor_block = f""
461 ML MODEL SIGNAL:
462 Consensus probability: {label_ev.prob_mean:.3f} (std={label_ev.prob_std:.3f}) — this {direction} the forecast prediction
463 Model votes: {label_ev.vote_positive}/{label_ev.vote_total} predict positive
464 ML uncertainty flag: {label_ev.uncertainty_flag} — {label_ev.uncertainty_reason}
465 Per-model: {' '.join(f"{s.model_name}=p{s.probability:.2f}" for s in label_ev.scores)}
466 ""
467
468 # Build topic shift signal block — kept identical to original
469 topic_monitor_block = ""
470 if topic_evidence is not None:
471     topic_monitor_block = f""
472 NEWS THEME SIGNAL:
473 Dominant topics: {' '.join(t.to_prompt_str() for t in topic_evidence.top3_topics)}
474 Distribution shift: {topic_evidence.shift_flag} (cosine_dist={topic_evidence.cosine_dist:.3f})
475 Shift reason: {topic_evidence.shift_reason}
476 ""
477
478 # Eruption gate block — kept identical to original
479 eruption_gate_block = ""
480 if target == "eruption" and forecast.prediction == "escalation":
481     eruption_gate_block = f""
482 ERUPTION GATE AUDIT (prediction=eruption was made — verify both gates were satisfied):
483 Gate 1 (quantitative): Are recent 14d fatalities  $\geq 2 \times$  the 60-day weekly average?
484     OR is ML eruption prob_mean > 0.40?
485 Gate 2 (qualitative): Is there a named new actor, geographic expansion, or
486     documented trigger event in the evidence?
487 If EITHER gate is not clearly satisfied, this forecast may be a FALSE POSITIVE.
488 ""
489
490 # ----- #
491 # GENERIC REFLECTION PROMPT — no taxonomy, no decision tree #
492 # The LLM reflects freely on forecast quality and revises confidence. #
493 # ----- #
494 return f""You are a critical evaluator stress-testing a conflict forecast for {region}.
495 Your job is to find weaknesses — not to validate the forecast.
496
497 FORECAST TO EVALUATE
498 Prediction: {forecast.prediction}
499 Stated confidence: {forecast.confidence}
500 Reasoning: {forecast.reasoning}
501 Key evidence used: {' '.join(forecast.key_evidence_used) or 'not specified'}
502 Contradicting evidence noted: {' '.join(forecast.contradicting_evidence) or 'none'}
503
504 STRUCTURED EVIDENCE
505 {evidence_text}
506 {ml_monitor_block}{topic_monitor_block}{eruption_gate_block}
507 ANSWER THESE DIAGNOSTIC QUESTIONS (be specific, use numbers from the evidence):
508 Q1. EVIDENCE VOLUME: How many weeks of ACLED data are present? How many total events?
509     Is there enough data to make a reliable forecast at all?
510 Q2. RECENCY DELTA: What are the exact fatality counts for the most recent 14 days vs
511     the prior 14 days? Compute the % difference. Does it agree with the forecast direction?
512 Q3. ML AGREEMENT: Does the ML consensus probability agree or disagree with the prediction?
513 Q4. COUNTER-ARGUMENT: State the single strongest specific argument for {opposite}.
514     Can you make it using actual numbers?
515 Q5. WEAKNESS CHECK: Is there one specific, concrete weakness that would materially change
516     the forecast if resolved? Name it precisely or say "no material weakness found."
517
518 Based on your answers above, assess the overall reliability of this forecast.
519 Consider: Is the evidence sufficient? Are there contradictions? Does the forecast
520 hold up under scrutiny? Revise the confidence up or down as appropriate.
521
522 If the forecast appears well-supported with no material weaknesses, set
523 RECOMMENDED_ACTION to "exit". Otherwise set it to "comparative_reasoning".
524
525 OUTPUT FORMAT (follow exactly, no extra text):
526 UNCERTAINTY_TYPE: MODEL_AMBIGUITY
527 UNKNOWN_DESCRIPTION: n/a
528 CONFIDENCE: <0.0 to 1.0 — your independent assessment after reflection>
529 EVIDENCE_QUALITY: <0.0 to 1.0>
530 EVIDENCE_GAPS: <specific gaps found, or "none">
531 CONFLICTING_SIGNALS: <specific conflicts found, or "none">
532 RECOMMENDED_ACTION: <exit OR comparative_reasoning>
533 REASONING: 2-4 sentences referencing specific numbers from the evidence.
534 ACTION_REASONING: 1-2 sentences explaining your confidence adjustment.
535 ""
536
537

```

C.4 Free-reflection prompt (E)

```

430 def _build_monitor_prompt(
431     region: str,
432     target: str,
433     forecast: ForecastState,

```

```

434 evidence_text: str,
435 ml_evidence: Optional[Any] = None,
436 topic_evidence: Optional[Any] = None,
437 ) -> str:
438     opposite = "no_escalation" if forecast.prediction == "escalation" else "escalation"
439
440     # Build ML signal block for monitor
441     ml_monitor_block = ""
442     if ml_evidence is not None:
443         label_ev = getattr(ml_evidence, target, None)
444         if label_ev is not None:
445             direction = "supports" if label_ev.prob_mean > 0.35 else "contradicts" if label_ev.prob_mean < 0.15 else "is ambiguous about"
446             ml_monitor_block = f""
447
448     ML MODEL SIGNAL:
449     Consensus probability: {label_ev.prob_mean:.3f} (std={label_ev.prob_std:.3f}) — this {direction} the forecast prediction
450     Model votes: {label_ev.vote_positive}/{label_ev.vote_total} predict positive
451     ML uncertainty flag: {label_ev.uncertainty_flag} — {label_ev.uncertainty_reason}
452     Per-model: {' '.join(f"{s.model_name}=p{s.probability:.2f}" for s in label_ev.scores)}
453
454     Consider: Does the ML signal agree with the qualitative forecast? If not, does this disagreement
455     represent a genuine CONFLICTING_SOURCES signal worth flagging?
456     ""
457
458     # Build topic shift signal block for monitor
459     topic_monitor_block = ""
460     if topic_evidence is not None:
461         topic_monitor_block = f""
462
463     NEWS THEME SIGNAL:
464     Dominant topics: {' '.join(t.to_prompt_str() for t in topic_evidence.top3_topics)}
465     Distribution shift: {topic_evidence.shift_flag} (cosine_dist={topic_evidence.cosine_dist:.3f})
466     Shift reason: {topic_evidence.shift_reason}
467
468     Note: cosine_dist > 0.35 is the threshold for a meaningful topic distribution shift.
469     A shift alone does NOT automatically mean DISTRIBUTION_SHIFT uncertainty — it is only
470     one of four required triggers. Check whether it materially changes the forecast direction.
471     ""
472
473     # Eruption-specific gate reminder for monitor
474     eruption_gate_block = ""
475     if target == "eruption" and forecast.prediction == "escalation":
476         eruption_gate_block = f""
477
478     ERUPTION GATE AUDIT (prediction=eruption was made — verify both gates were satisfied):
479     Gate 1 (quantitative): Are recent 14d fatalities  $\geq 2 \times$  the 60-day weekly average?
480     OR is ML eruption prob_mean > 0.40?
481     Gate 2 (qualitative): Is there a named new actor, geographic expansion, or
482     documented trigger event in the evidence?
483     If EITHER gate is not clearly satisfied with specific numbers/names, this forecast
484     is likely a FALSE POSITIVE and the monitor should flag CONFLICTING_SOURCES or
485     recommend downgrading confidence significantly.
486     ""
487
488     return f""You are a critical evaluator stress-testing a conflict forecast for {region}.
489     Your job is to find weaknesses — not to validate the forecast.
490
491     FORECAST TO EVALUATE
492     Prediction: {forecast.prediction}
493     Stated confidence: {forecast.confidence}
494     Reasoning: {forecast.reasoning}
495     Key evidence used: {' '.join(forecast.key_evidence_used) or 'not specified'}
496     Contradicting evidence noted: {' '.join(forecast.contradicting_evidence) or 'none'}
497
498     STRUCTURED EVIDENCE (use this for all quantitative counting in Q1–Q5)
499     {evidence_text}
500     {ml_monitor_block}{topic_monitor_block}{eruption_gate_block}
501     IMPORTANT: When counting [acled_weekly] rows in Q1, count ONLY lines tagged
502     [acled_weekly] in the STRUCTURED EVIDENCE section above. Do NOT count lines
503     from any NEWS NARRATIVE SUMMARY section — that section is supplementary context
504     only and should not affect evidence volume assessment.
505     If a NEWS NARRATIVE SUMMARY is present, use it to inform Q4 (counter-argument)
506     and Q5 (weakness check) only — it may name specific actors, events, or anomalies
507     that are useful for those questions.
508     # ---- NO-QUESTIONS BASELINE (Condition E) ----
509     # This variant strips all diagnostic scaffolding from the monitor.
510     # No Q1-Q5 questions, no decision tree, no calibration rules.
511     # The LLM receives only the evidence and must assess its epistemic
512     # state entirely from scratch, with no structured guidance.
513     # This isolates the contribution of the diagnostic questions themselves
514     # from the contribution of the uncertainty type taxonomy.
515
516     Review the evidence and the forecast above.
517     Based solely on your assessment of the evidence quality and your confidence
518     in the forecast, classify the epistemic state of this forecast.
519
520     Consider: Is the evidence sufficient? Are there contradictions? Is this
521     case structurally typical or unusual? How confident are you in the prediction?
522
523     OUTPUT FORMAT (follow exactly, no extra text):
524     UNCERTAINTY_TYPE: <one type from the list above>
525     UNKNOWN_DESCRIPTION: <required if UNKNOWN_UNCERTAINTY, else "n/a">
526     CONFIDENCE: <0.0 to 1.0 — your independent assessment>

```

```

524 EVIDENCE_QUALITY: <0.0 to 1.0>
525 EVIDENCE_GAPS: <specific gaps found, or "none">
526 CONFLICTING_SIGNALS: <specific conflicts found, or "none">
527 RECOMMENDED_ACTION: <one of: request_additional_sources / comparative_reasoning / context_recalibration / reweight_sources / recency_prioritization
    ↳ / sample_alternative_forecasts / exit>
528 REASONING: 2-4 sentences referencing specific numbers from the evidence.
529 ACTION_REASONING: 1-2 sentences explaining why this action addresses the weakness.
530 """
531
532

```

C.5 Action and synthesis prompts

```

594 def _build_action_prompt(
595     region: str,
596     target: str,
597     action: str,
598     uncertainty_type: str,
599     forecast: ForecastState,
600     epistemic_state: EpistemicState,
601     evidence_text: str,
602     iteration: int = 0,
603 ) -> str:
604     """Build a targeted re-reasoning prompt based on the diagnosed action."""
605
606     action_instructions = {
607         "request_additional_sources": f"""
608 The forecast was flagged for INSUFFICIENT EVIDENCE.
609 Focus on extracting maximum signal from the available evidence — do not ask for more data.
610 - Go through each weekly ACLED period individually: list the week, event count, fatalities
611 - Identify the directional trend: is violence increasing, stable, or decreasing week-over-week?
612 - Find the most recent 7-day period and treat it as the primary signal
613 - Flag any extreme outlier weeks that distort the average
614 - Make a committed prediction based ONLY on the pattern you find in available data.
615 You must produce a definitive direction. "Insufficient evidence" is not an excuse to hedge.
616 """,
617         "comparative_reasoning": f"""
618 The forecast was flagged for CONFLICTING SOURCES.
619 Conflicting signals identified: {' '.join(epistemic_state.conflicting_signals) or 'see evidence'}
620 - List each conflict as: "[Signal A] says X vs [Signal B] says Y"
621 - For each conflict, state which signal is MORE RECENT and which comes from a MORE RELIABLE source
622 - Make an EXPLICIT judgment: "I trust [A] over [B] because [specific reason]"
623 - After resolving each conflict, make a single committed prediction
624 Do NOT end with "the picture is mixed". You must choose a direction and commit to it.
625 """,
626         "context_recalibration": f"""
627 The forecast was flagged for DISTRIBUTION SHIFT — this case looks atypical for {region}.
628 - State specifically WHAT is atypical: scale, actor mix, geographic pattern, or topic shift
629 - Determine whether the anomaly makes escalation MORE or LESS likely than baseline patterns suggest
630 - Ask: does the anomaly represent a genuine structural break (new phase of conflict) or a
631 temporary spike that will revert? Evidence for each?
632 - Produce a recalibrated prediction that accounts for the structural change
633 You must reach a committed direction. "This is unusual" is not a conclusion.
634 """,
635         "reweight_sources": f"""
636 The forecast was flagged for LOW QUALITY DATA.
637 - Identify which specific evidence items are low-reliability (low source quality score, tier4 sources)
638 - List what remains after removing or heavily discounting those items
639 - Reason ONLY from the high-reliability evidence items
640 - State explicitly: "After downweighting [X], the remaining signal suggests [direction]"
641 - Acknowledge the increased uncertainty in your confidence score (expect 0.50–0.65 range)
642 Make a committed direction based on reliable evidence only, even if confidence is modest.
643 """,
644         "recency_prioritization": f"""
645 The forecast was flagged for TEMPORAL INCONSISTENCY — recent trend contradicts older baseline.
646 Gaps identified: {' '.join(epistemic_state.evidence_gaps) or 'see evidence'}
647 - State the exact recent 14-day fatality count and the exact prior 14-day count
648 - State the 60-day baseline average weekly fatalities
649 - Make an EXPLICIT judgment: "The recent 14-day trend OVERRIDES / DOES NOT OVERRIDE the baseline because..."
650 - If recent trend overrides: state the new prediction direction clearly
651 - If baseline holds: explain what in the recent data is noise vs signal
652 Produce a final direction. Recency weighting is a tool to resolve ambiguity, not add more.
653 """,
654         "sample_alternative_forecasts": f"""
655 The forecast was flagged for MODEL AMBIGUITY — evidence could support either prediction.
656 - CASE FOR escalation: state 2-3 specific numbers from the evidence that support it
657 - CASE FOR no_escalation: state 2-3 specific numbers from the evidence that support it
658 - Now weigh them: which case uses MORE RECENT evidence? Which uses HIGHER RELIABILITY sources?
659 - Make a final judgment: "On balance, [direction] is more supported because [specific reason]"
660 - Report confidence in the 0.52–0.62 range to reflect genuine ambiguity
661 You must commit to a direction. Refusing to decide is not acceptable output.
662 """,
663     }
664
665     instruction = action_instructions.get(action, "Re-examine the evidence carefully and produce a revised forecast.")

```

```

671
672     return f""""You are a conflict analyst for {region} performing targeted re-analysis (iteration {iteration + 1}).
673
674 ORIGINAL FORECAST
675 Prediction: {forecast.prediction} | Confidence: {forecast.confidence}
676 Reasoning: {forecast.reasoning}
677
678 METACOGNITIVE DIAGNOSIS
679 Uncertainty type: {uncertainty_type}
680 Evidence gaps: {' '.join(epistemic_state.evidence_gaps) or 'none'}
681 Conflicting signals: {' '.join(epistemic_state.conflicting_signals) or 'none'}
682
683 EVIDENCE
684 {evidence_text}
685
686 TARGETED RE-ANALYSIS INSTRUCTION
687 {instruction}
688
689 IMPORTANT: Do not simply restate the original reasoning.
690 You must produce NEW insight grounded in specific numbers from the evidence.
691 If after re-analysis you agree with the original, say so explicitly AND state what
692 evidence would need to change for you to flip — this is still useful new information.
693
694 OUTPUT FORMAT (follow exactly):
695 REVISED_PREDICTION: escalation OR no_escalation
696 REVISED_CONFIDENCE: 0.0 to 1.0
697 CHANGED: yes OR no
698 CHANGE_REASONING: <if changed: the specific number/signal that drove the change;
699                 if not changed: the specific number/signal that confirmed the original>
700 KEY_INSIGHT: <one sentence — the single most important NEW finding. Must contain a number
701              or a named signal. "The situation remains uncertain" is NOT acceptable.>
702 REASONING: 2-4 sentences with specific numbers from the evidence.
703 """"
704
705
706 def _build_synthesis_prompt(
707     region: str,
708     target: str,
709     original_forecast: ForecastState,
710     epistemic_state: EpistemicState,
711     action_taken: str,
712     action_output: str,
713     evidence_text: str,
714 ) -> str:
715     # Extract KEY_INSIGHT and REVISED_CONFIDENCE from action output to anchor synthesis
716     _action_insight = ""
717     _action_revised_conf = ""
718     if action_output:
719         for _line in action_output.splitlines():
720             if _line.strip().upper().startswith("KEY_INSIGHT:"):
721                 _action_insight = _line.split(":", 1)[1].strip()
722             if _line.strip().upper().startswith("REVISED_CONFIDENCE:"):
723                 _action_revised_conf = _line.split(":", 1)[1].strip()
724
725     return f""""You are synthesizing an updated conflict forecast for {region} after metacognitive correction.
726
727 ORIGINAL FORECAST
728 Prediction: {original_forecast.prediction} | Confidence: {original_forecast.confidence}
729 Reasoning: {original_forecast.reasoning}
730
731 UNCERTAINTY DIAGNOSED
732 Type: {epistemic_state.uncertainty_type}
733 Gaps: {' '.join(epistemic_state.evidence_gaps) or 'none'}
734 Conflicts: {' '.join(epistemic_state.conflicting_signals) or 'none'}
735
736 CORRECTIVE ACTION TAKEN: {action_taken}
737 KEY INSIGHT FROM RE-ANALYSIS: {_action_insight or '(see full action output below)'}
738 REVISED CONFIDENCE FROM RE-ANALYSIS: {_action_revised_conf or '(see full action output below)'}
739
740 FULL ACTION OUTPUT:
741 {action_output}
742
743 FULL EVIDENCE
744 {evidence_text}
745
746 TASK: Produce a single synthesized forecast that integrates the corrective re-analysis above.
747
748 SYNTHESIS RULES — follow strictly:
749 1. Start from the KEY INSIGHT and REVISED CONFIDENCE above, not the original forecast.
750    If the re-analysis produced a revised confidence, your output confidence should be
751    close to it (within ±0.08) unless the full evidence contradicts it.
752 2. If the re-analysis changed the prediction direction, adopt the new direction UNLESS
753    you can cite a specific number in the evidence that overrules it — state that number.
754 3. If the re-analysis CHANGED=no but found a specific weakness, reduce confidence by
755    0.05–0.10 from the original to reflect that weakness — do not keep confidence flat.
756 4. If the re-analysis CHANGED=no and found NO weakness (evidence was solid), you may
757    raise confidence by up to 0.05, and the loop should naturally exit next iteration.
758 5. Do NOT average the original and revised confidences. Make a judgment call.
759 6. Do NOT restate the original reasoning verbatim. The synthesis must add new clarity.
760

```

```

761 CRITICAL FLIP RULE: If changing prediction (no_escalation ↔ escalation), you must
762 name one specific piece of evidence with a number that justifies the flip.
763
764 OUTPUT FORMAT (follow exactly):
765 PREDICTION: escalation OR no_escalation
766 CONFIDENCE: 0.0 to 1.0
767 CHANGED_FROM_ORIGINAL: yes OR no
768 CHANGE_REASONING: <specific evidence cited, or "no change: [reason]">
769 KEY_EVIDENCE: <the 2-3 signals that most drove this forecast>
770 CONTRADICTING_EVIDENCE: <what you discounted and why, or "none">
771 REASONING: 2-4 sentences that would not appear in the original forecast.
772 ""
773
774
775 # -----
776 # Parser helpers
777 # -----
778

```

D Code and Data Availability

The three condition scripts, executable prompt builders, and an example trace accompany this submission and are also available at <https://github.com/PolNemkova/agent-self-reflection>. ACLED records are not redistributed; the release includes preprocessing code and case indices for reconstructing the evaluation from data obtained under the provider's terms.