Beyond Scalar Accuracy: An Intervention-Response Audit of LLMs as Policy Decision Infrastructure

Anonymous ACL submission

Abstract

LLMs are increasingly proposed as decision infrastructure for high-stakes policy tasks, yet evaluation still defaults to scalar accuracy on resolved decisions, which can mask deployment-relevant failure modes: in a six-model role-play audit on the 2018–2020 and 2025 U.S.–China trade war, a 70.5% headline coexists with 28.5% escalation recall, and o1 and o3 each output Wait on 72/72 of their post-cutoff 2025 cells. To make such failures legible, we operationalise an *intervention-response* audit that scores a system by how its recommendations move when entity visibility, prompt persona, architecture, or training-cutoff phase is varied while the rest is held approximately fixed, instantiated across three policy domains (TradeWar / FOMC / Energy, 46 events, 2,832 cells + a 216-cell extension) and benchmarked against a strictly-causal calendar logistic regression ($\mathrm{LR}_{\mathrm{causal}}$) that uses only features available at decision time T. On TradeWar (70.5 vs. 68.2%) and Energy (55–57 vs. 50.0%) the audited LLMs are not clearly beyond $\mathrm{LR}_{\mathrm{causal}}$; on FOMC they exceed it (79.8–92.3% vs. 50.0%), but a placebo-institution decomposition attributes the excess to Fed-specific identity ($\Delta_{\mathrm{contam}}=+16.1$ vs. $\Delta_{\mathrm{context}}=-2.9$ pp, nonnegative in 5/5 models) rather than to named-institution context. The audit further surfaces two dissociable response-surface signatures —class collapse and intervention-gated movement—that scalar accuracy hides by construction. We conclude that headline accuracy is not capability without a strictly-causal baseline and the audit's reporting-checklist disclosures.

1 Introduction

LLMs are increasingly used to forecast high-stakes institutional decisions, with prototypes already emerging for tariff timing, monetary-policy nowcasting, and energy-policy outlook (Lopez-Lira et al., 2025; Hua et al., 2023; Karger et al., 2024). As multi-agent LLM systems approach human-expert calibration on some of these tasks (Halawi et al., 2024; Karger et al., 2024), accuracy on resolved decisions has become the de facto signal for deployment readiness. Treating these systems as decision infrastructure changes what an evaluation must guarantee: not just that the headline is high, but that it is high *for the right reason*.

Scalar accuracy is degenerate for this purpose: identical scores can arise from incompatible policies. A system can score well by memorizing events that leak across the cutoff (Halawi et al., 2024; Oren et al., 2024; Zhang et al., 2026); on Wait-heavy calendars, a frozen always-Wait policy can match an evidence-driven recommender while missing the escalation events that drive deployment risk; multi-agent pipelines can produce headlines via mechanical aggregation rather than genuine consensus; and prompt, anonymization, and architecture choices each shift the same model's answer (Sclar et al., 2024; Lopez-Lira et al., 2025; Jeon and Lee, 2026; Wu et al., 2025). Existing evaluations vary at most one of these axes at a time, so deployment-relevant failure modes remain invisible at the headline-accuracy level. We additionally make explicit a structural caveat the literature often elides: in geopolitical-simulation systems like ours (Hua et al., 2023), the prompt elicits a *role-play recommendation* ("what should Country A do, given its goals?") while accuracy scores it against a *realized historical decision*. We read this elicitation–scoring gap as a **deployment-hazard signal**—disagreement between the recommendation and the realised decision indicates deployment risk regardless of which side is normatively right—rather than a clean forecasting metric.

We define an *intervention-response audit*: instead of asking "what is the model's accuracy?", we ask "how does the model's prediction move when a single audit factor is varied while everything else is held approximately fixed at the proto-

col level?" (Some interventions, notably anon-cfx, necessarily perturb correlated textual attributes; we treat them as response-surface probes, not causally isolated factor changes.) Each condition cell is one (model, decision-event, entity-visibility protocol, prompt-pipeline condition) tuple, aggregated from 35 LLM calls scored on the final deployed-decision layer. The main TradeWar grid spans 1,584 cells (6 models $\times$ 22 events $\times$ 3 protocols $\times$ 4 conditions); a 1,008-cell FOMC stress test, a 240-cell Energy check, and a 216-cell 2025 extension serve as ancillary checks. All cells share a 35-call budget so architecture and prompt effects are not confounded with call count. The audit is grounded in response-surface methodology (Box and Wilson, 1951; Myers et al., 2016).

The audit makes the degenerate-headline phenomenon legible along three dissociable axes.

(a) Trivial-baseline gap. The empirical anchor is a strictly-causal logistic regression on calendar features (only quantities available at decision time T, no analyst-defined phase metadata): $\mathrm{LR}_{\mathrm{causal}}$ reaches 68.2 / 50.0 / 50.0% on TradeWar / FOMC / Energy via leave-one-event-out (§5.5). The audited LLMs are *not clearly beyond* this trivial baseline on TradeWar (LLM 70.5%, causal LR 68.2%) or Energy (LLM 55–57%, causal LR 50%), and *exceed it on FOMC* (LLM 79.8–92.3%, causal LR 50.0%) — though App. K's placebo-institution control attributes most of that excess ($\Delta_{\mathrm{contam}} = +16.1$ pp, $\Delta_{\mathrm{context}} = -2.9$ pp on the 5-model placebo subset, 4/5 contamination-positive) to Fed-specific identity rather than to "a named salient institution" alone. A phase-leaky $\mathrm{LR}_{\mathrm{phase}}$ reaches 72.7 / 85.7 / 90.0% as a leakage stress test, not a deployable forecaster.

(b) Class collapse hidden by the headline. Under the same audit, TradeWar's 70.5% [55.6, 84.4] headline coexists with 61.5% balanced accuracy and 28.5% escalation recall; o1 and o3 each output Wait in 72/72 of their post-cutoff 2025 cells (always-Wait baseline at 66.7%, since the calendar is 4/6 Wait-heavy). The FOMC +12.5 pp deanon-over-anon-L2 boost is positive in all six models; the placebo decomposition (App. K) attributes it to Fed-specific identity, not to legitimate named-institution context.

(c) Intervention-gated recovery on a single model. A single-model intervention-gated recovery (o3-mini's 0/8 $\rightarrow$ 6/8 under counterfactual rewriting, n=2 events, now also controlled for lexical surface and rewriter choice in App. J) illustrates the second response-surface signature.

Contributions. We do not claim novelty for the audit-vs-evaluation distinction (Chataigner et al., 2026) or for intervention-as-redistribution (Espín-Noboa and Méndez, 2026); we claim novelty at the *instantiation* level for LLM policy-decision pipelines, on four dimensions concurrent work does not provide jointly (Table 1): (i) multi-axis joint factorial response surface (entity visibility $\times$ persona $\times$ architecture $\times$ cutoff phase, held-fixed grid); (ii) post-cutoff cohort + release-date control + non-LLM calendar baseline as a three-piece contamination/triviality stress test; (iii) two named, dissociable response-surface signatures (class collapse, intervention-gated movement) with explicit disconfirmation conditions; (iv) a setting-specific treatment of the role-play elicitation-scoring gap. We accompany the instantiation with an 11-criterion reporting checklist (App. A) self-applied to this paper.

2 Related Work

We combine four lines of work that are typically evaluated separately, framed through response-surface methodology from experimental design (Box and Wilson, 1951; Myers et al., 2016). **MAS:** multi-agent debate can improve verifiable reasoning (Du et al., 2024; Liang et al., 2024) but can also amplify shared errors or underperform self-consistency at matched compute (Cemri et al., 2025; Wu et al., 2025; Wynn et al., 2025; Bertalanič and Fortuna, 2026; Wang et al., 2023); our system is a staged geopolitical-simulation pipeline (Hua et al., 2023), not a debate protocol. **Prompt sensitivity:** framing and personas shift model behavior (Sclar et al., 2024; Zheng et al., 2024; Sharma et al., 2024; Bandaru et al., 2025); we treat prompt edits as diagnostic interventions, not tunable improvements. **Temporal contamination:** later-resolved events leak into training (Halawi et al., 2024; Li et al., 2026; Zhang et al., 2026; Lopez-Lira et al., 2025; Karger et al., 2024; Oren et al., 2024). **Anonymisation:** previously used as a safeguard (Lopez-Lira et al., 2025; Jeon and Lee, 2026); we instead treat deanon, masking, and counterfactual rewriting as interventions on the decision policy. The contribution is to operationalise these jointly as an intervention-response audit for LLM forecasting.

Position on auditing vs. evaluation, and concurrent intervention-based audits. Most directly related is the position of Chataigner et al. (2026),

Dimension	This paper	Chataigner '26	Espín-Noboa '26
Audit-vs-eval position	adopts	introduces	adopts
Multi-axis joint factorial grid	4 axes, held-fixed	—	3 axes, separate sweeps
Post-cutoff cohort + release-date control	yes (6 events + grok)	—	—
Non-LLM trivial-forecaster baseline	yes (calendar LR)	—	—
Setting-specific elicitation-scoring treatment	yes (role-play vs. realised decision)	—	N/A (generative rec)

Table 1: **Differentiation from concurrent intervention-based audits.** We do not claim the audit-vs-evaluation *idea*; we claim, to our knowledge, the first *instantiation* for LLM policy-decision pipelines that jointly provides the four dimensions concurrent work does not.

who argue that LLM auditing is distinct from evaluation and should produce dynamic, context-sensitive, budget-aware, and reliable evidence rather than aggregate benchmark scores, and propose a four-component conceptual framework (Auditing Scope / Interactor / Evaluator / Auditing Output) to structure such audits. We share this position. The closest concurrent empirical instantiation is Espín-Noboa and Méndez (2026), who introduce *LLMScholarBench*, an intervention-based audit of 22 LLMs on physics-expert recommendation under temperature, representation-constrained prompting, and RAG, and find that interventions *redistribute* errors across technical-quality and social-representation dimensions rather than uniformly improving them. We adopt their framing of interventions as a redistribution rather than an improvement and arrive at a converging structural finding from a different setting: where they audit single-instance generative scholar recommendation with bibliometric ground truth, we audit a multi-agent role-play decision pipeline against realised historical decisions in three policy domains (TradeWar, FOMC, Energy), with interventions on entity visibility (deanon / anon-L2 / anon-cfx), prompt persona, architecture (solo vs. MAS), and training-cutoff phase. The two studies are complementary: theirs surfaces redistribution across diversity/factuality/parity in a generative recommendation setting; ours surfaces class-collapse and intervention-gated movement signatures in a binary policy-decision setting, with a post-cutoff cohort, a release-date control, and non-LLM calendar baselines that the recommendation setting does not provide. Table 1 maps the four dimensions on which our instantiation extends the concurrent work.

3 Evaluation Methodology

Figure 1 fixes the audit's denominator and readouts. The 35-call budget is inherited from V1's original self-consistency configuration (Hua et al., 2023); we match it on the solo arm so any solo $\rightarrow$ MAS difference reflects architecture rather than additional sampling. Each diagnostic carries an explicit failure criterion: f_{prompt} sign reversals or within-model ranges $>$10 pp are treated as instability rather than as a directional effect; T3's recall split is needed because a fixed Wait policy can score well on de-escalation-heavy calendars and so headline accuracy alone cannot adjudicate post-cutoff capability; T4's per-cell deltas are the unit of audit, while the mixed-effects logit in App. I is reported as a qualitative aid because event clusters are small (n=22) and the analysis is not pre-registered. Two ancillary frames sit alongside the four readouts: a pair of non-LLM logistic regressions ($\text{LR}_{\text{causal}}$ on strictly-causal calendar features, LR_{phase} with phase metadata) that establish the triviality floor any audited LLM must clear, and grok's post-release-date row, used as a contamination probe.

4 Case Study Setup

Audited system and label rule. The audited system is a 7-agent MAS recommendation pipeline derived from the LLM geopolitical-simulation literature (Hua et al., 2023) (see Figure 1). Each model plays *Country A* and recommends the most beneficial action; we score that recommendation against the realised historical decision. This is an *elicitation–scoring* audit, not a pure forecasting evaluation: the prompt elicits a normative recommendation ("what should Country A do?"), and the score asks whether the realised decision agrees. We do not claim the resulting numbers are calibrated forecasts of base-rate event probabilities. The full prompt text is reproduced verbatim in App. L. At each historical decision point in the 2018–2020 U.S.–China trade war and the 2025 tariff escalation, the system recommends DECLARE (impose/escalate tariffs or retaliate) or WAIT (negotiate/pause/de-escalate). We score against ground truth coded under a **strict bilateral-tariff rule**: DECLARE requires a formally promulgated bilateral tariff action dated within (T, T+60] (App. F). The prompt's action menu is V1-locked and intentionally broader than the scoring rule; we treat this prompt-menu/scoring-rule mismatch as

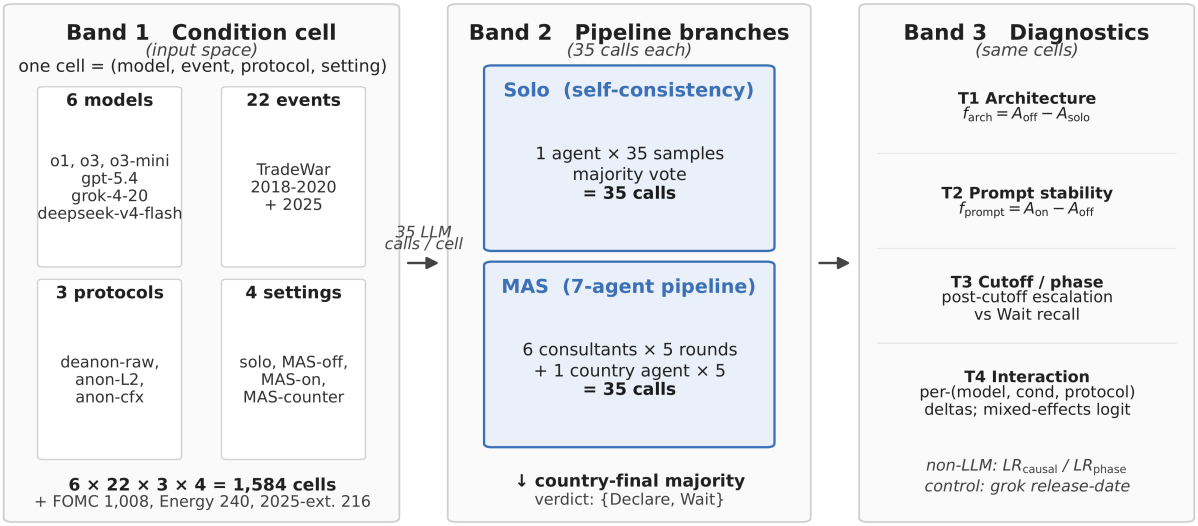

Figure 1: **Audit framework overview.** Each condition cell — (model, event, entity-visibility protocol, prompt-pipeline setting) — is scored from 35 LLM calls, split between a solo self-consistency baseline and a 7-agent MAS pipeline (6 consultants $\times$ 5 rounds + 1 country agent $\times$ 5 rounds; country-final majority). The same cells feed four diagnostic readouts: T1 architecture ($f_{\text{arch}} = A_{\text{off}} - A_{\text{solo}}$), T2 prompt stability ($f_{\text{prompt}} = A_{\text{on}} - A_{\text{off}}$), T3 cutoff/phase (post-cutoff escalation vs. Wait recall), and T4 interaction structure (per-(model,cond,protocol) deltas; mixed-effects logit in App. I). Two non-LLM logistic regressions ($\text{LR}_{\text{causal}}$ / LR_{phase}) and grok's post-release-date row act as triviality and contamination controls.

part of what the audit measures (App. G reports broader-rule sensitivity).

Cohort and grid. 22 decision points (16 from 2018–2020, 6 from January–November 2025) cover escalation, truce, and de-escalation phases (App. F). We evaluate six models (o1, o3, o3-mini, gpt-5.4, grok-4-20-reasoning, deepseek-v4-flash) and the MAS pipeline uses six domain-specialised consultants (economic, political-social, resource-technology, trade-network, policy-tools, historical) and one country-decision agent. TradeWar uses three entity protocols (**deanon-raw / anon-L2 / anon-cfx**) and four settings (solo, MAS-off, MAS-on, MAS-counter), giving $6 \times 22 \times 3 \times 4 = 1{,}584$ cells.

5 Results

The 1,584-cell TradeWar grid is scored on the country-decision majority. Overall accuracy is 70.5% [event-clustered 95% CI 55.6, 84.4], hiding 61.5% [54.8, 68.0] balanced accuracy and 28.5% [18.4, 40.5] escalation recall vs. 94.5% [86.4, 99.1] Wait recall. Effective cluster counts are small (TradeWar 22, FOMC 14, post-cutoff sub-grid 6, o3-mini escalation 2), so bootstrap CIs (B=10,000) are descriptive; for sub-grids with $n \leq 6$ events we report fractions only. Headline directional contrasts use event-level paired sign/permutation tests with BH-FDR over four headline-effect tests (App. I), not as a pre-registered confirmatory family. Three ancillary checks (216 TradeWar 2025 cells, 1,008 FOMC cells, 240 Energy cells) support but do not enlarge the headline.

5.1 Dissociated OOD Behaviours, Frozen-Wait, and Event Phase

The six 2025 Trump-2.0 events lie strictly after every OpenAI cohort cutoff (o1/o3-mini: 2023-10; o3: 2024-06) and are phase-structured (Jan/Mar escalation; May/Jul/Sep/Nov de-escalation). grok's public release postdates the entire calendar, so its rows are a release-date contamination control rather than post-cutoff evidence. We organise this section around two dissociated post-cutoff behaviours (o3-mini's anon-cfx rescue, o1/o3's frozen-Wait) and the grok control; Table 4 combines the behavioural taxonomy with the per-model phase-recall split, and Figure 2 shows the per-(model, protocol) response surface that motivates the dissociation.

Single-model existence proof: anchor-dependent recovery (o3-mini). *Denominator key for this paragraph: 1 model $\times$ 6 events $\times$ 4 settings = 24 cells per protocol; 1 model $\times$ 2 escalation events $\times$ 4 settings = 8 cells per protocol; sign test acts on the 2 escalation events.*

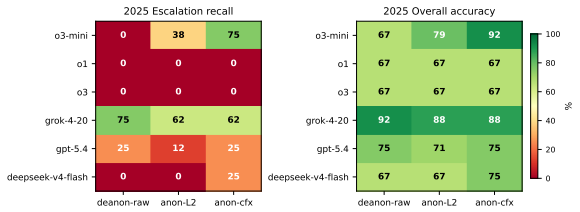

Figure 2: **Response surface on the 2025 TradeWar calendar: scalar accuracy hides class collapse.** Right panel (overall accuracy) shows every model in the 67–92% range across all three protocols, so a headline-accuracy view would call all six audited models comparably capable. Left panel (escalation recall on the two 2025 escalation events) shows the same models occupy a much wider range, with o1 and o3 producing 0% escalation recall under every protocol (frozen-Wait) and o3-mini moving from 0% (deanon-raw) through 38% (anon-L2) to 75% (anon-cfx) — the dissociation behind D1 and D2.

Model	$\bar{f}_{\text{arch}}$ (pp)	Screen?
deepseek	-10.6	×
gpt-5.4	-7.6	×
grok-4-20	-6.1	×
o3-mini	-3.0	×
o1	$+3.0$	✓
o3	$+1.5$	✓
Models with $\bar{f}_{\text{arch}} \geq 0$:	2/6 (33%)	

Table 2: **Test 1 (Architecture Screening, country-final scoring).** $f_{\text{arch}} = A_{\text{off}} - A_{\text{solo}}$. Compute is matched at the *API-call* level (35 calls per cell); per-call cost may still vary across conditions because MAS prompts and outputs differ in length from solo prompts—this is precisely the architectural contrast T1 measures, not a confound to control out. Four of six models are worse under the MAS pipeline; the two positive effects are small.

On the two early-2025 escalation events under deanon-raw, o3-mini collapses to always-Wait (0/8 correct); under anon-cfx, it produces correct Declare predictions on 6/8 cells while preserving 16/16 Wait recall on the four de-escalation events. Across the full six-event 2025 calendar (24 cells per protocol), o3-mini scores 16/24 (deanon-raw), 19/24 (anon-L2), 22/24 (anon-cfx). The contrast rests on n=2 escalation events × 4 settings, so the paired sign test is uninformative (p=0.5, BH-FDR q=0.5) and we report it as an exploratory single-model existence proof. *Counterfactual rewriting is not a controlled intervention on entity identity alone*—it can also simplify lexical surface, length, or political salience, and the rewriting LLM may inadvertently leak outcome-indicative wording—so we do *not* claim it diagnoses memorisation. We claim only that some intervention on the entity layer can move o3-mini's decision while no intervention moves o1's or o3's, in contrast with the usual framing of anonymisation as a safeguard (Lopez-Lira et al., 2025; Jeon and Lee, 2026). Two of these matched controls—a lexical-surface control (**paraphrase-L0**: keeps America/China entity names but paraphrases the deanon-raw wording to anon-cfx length) and a rewriter-robustness control (**anon-cfx-v2**: regenerates the Nation Alpha/Nation Beta version with a different rewriter LLM, `gpt-4o-mini` instead of `claude-sonnet-4-5`)—have now been added in App. J; both are more consistent with entity removal than with the tested lexical-surface or rewriter-choice alternatives (paraphrase-L0 = 1/8, near deanon-raw's 0/8; anon-cfx-v2 = 6/8, matching anon-cfx's 6/8); they rule out two specific competing accounts without causally isolating the entity layer.

Supporting failure mode: frozen-Wait response (o1, o3). Across the full Trump-2.0 calendar, o1 and o3 output Wait in every one of their 72 cells; *no entity-visibility protocol, MAS configuration, or persona prompt changes the decision*. They fail both escalation events with 0% recall but score 100% on the four de-escalation events, yielding 66.7% accuracy with no decision-boundary movement. We use "frozen-Wait" descriptively; the signature is also consistent with a strong de-escalation prior partially calibrated to the 4/6 Wait class balance, and a class-balanced post-cutoff calendar would be needed to disambiguate. The invariance is axis-specific: a separate *relative-power probe* that injects an explicit "Country A holds stronger relative power" frame into the trigger flips both o1 and o3 from Wait to Declare on 5/6 of their 2025 events under both deanon-raw and anon-L2 (App. I), showing that the frozen response holds against entity/persona axes but not against content-level priming.

5.2 T1: Architecture Screening

Table 2 reports the architecture effect for each model: MAS-off accuracy minus the call-budget-matched solo baseline ($k = 35$).

For four of six models, adding the staged MAS pipeline *reduces* accuracy relative to a call-budget-matched self-consistency baseline. The direction is not universal—o1 and o3 improve slightly—but the average pattern is still a pipeline cost rather than

Model	deanon	anon-L2	anon-cfx	Verdict
o3-mini	+13.6	+27.3	−4.5	∼
gpt-5.4	+4.5	+13.6	+0.0	✓
grok-4-20	+0.0	+4.5	−4.5	✓
deepseek	+9.1	+4.5	+4.5	✓
o1	+4.5	+4.5	+4.5	✓
o3	−4.5	+0.0	+0.0	✓
Range across 18 cells: −4.5 to +27.3 pp				

Table 3: **Test 2 (Prompt Stability, country-final scoring).** $f_{\text{prompt}} = A_{\text{on}} - A_{\text{off}}$ per model–protocol cell across the 22 TradeWar timelines, with the country-decision agent (final-layer majority) as the reported scoring layer. The persona reminder has no uniform direction: most cells move by at most one timeline, while o3-mini remains highly protocol-dependent.

Model	Behaviour on Trump-2.0	2025 Acc	Esc. rec.	Wait rec.
o3-mini	Anchor-dependent anon-cfx recovery	79.2%	37.5%	100.0%
o1	Frozen-Wait (entity-axis invariant)	66.7%	0.0%	100.0%
o3	Frozen-Wait (entity-axis invariant)	66.7%	0.0%	100.0%
grok-4-20	Post-release-date (contamination ctrl)	88.9%	66.7%	100.0%
gpt-5.4	High-Wait prior, weak escalation recovery	73.6%	20.8%	100.0%
deepseek	High-Wait prior, weak escalation recovery	69.4%	8.3%	100.0%

Table 4: **Dissociated OOD behaviours and per-phase recall on the full 2025 TradeWar calendar (6 models, 6 events).** Country-final scoring shows why scalar post-cutoff accuracy is insufficient: o1 and o3 score 66.7% only because all four de-escalation events are Wait; they have 0% recall on the two escalation events. o3-mini exhibits an anchor-dependent anon-cfx rescue (App. I shows the o1/o3 frozen response is entity-axis-specific, not content-invariant). grok is separated as a release-date control because its public release postdates the calendar.

a reliable architecture gain. **T1 verdict: raises concerns.**

Why does the MAS pipeline underperform? Plausible mechanisms are advisory-signal dilution and shared-bias amplification (Wu et al., 2025; Wynn et al., 2025); the consultant–country disagreement counts that motivate D5 are reported in §6.

5.3 T2: Prompt Stability

Table 3 shows the prompt-reminder effect across model–protocol cells, restricted to the minimal perturbation (MAS-on vs. MAS-off).

The persona reminder is not globally beneficial. For most model–protocol cells, it changes at most one timeline-level verdict. The exception is o3-mini, where the effect ranges from −4.5 to +27.3 pp depending on the anonymization protocol.

5.4 T3: Cutoff Sensitivity

The 2025 extension changes the interpretation of a naive pre/post accuracy test. Because four of six Trump-2.0 events are Wait/de-escalation windows, an always-Wait policy can *increase* post-cutoff accuracy while losing all escalation ability. We therefore report cutoff sensitivity by event phase rather than using scalar pre/post accuracy alone (Table 4).

The phase split is the clearest T3 result: o1 and o3 are invariant to escalation evidence and correct only when the phase matches Wait, o3-mini recovers part of the escalation boundary under counterfactual rewriting, and grok sits inside its release window as a control. Cutoff sensitivity is a structured response signature, not a monotone accuracy drop. Variance decomposition (App. I) confirms that event-phase structure is the dominant source of explained variance, with no single test subsuming the others (per-model T1/T2/T3 verdicts in App. K).

5.5 Non-LLM Baselines

We benchmark each domain against four cheap baselines plus two leave-one-event-out logistic regressions (LR), distinguished by whether they receive analyst-defined phase metadata (Table 5). **LR_{phase}** (phase-leaky): features are a phase one-hot (4–6 phases per domain), year/month, prior-event positive rate, last-window positive rate; phase labels are fitted from training-fold events but encode analyst judgement made across the full calendar. **$\text{LR}_{\text{causal}}$** (strictly causal): features are year, month sin/cos, cumulative positive rate over past events only, rolling positive rate over the last K=6 past events, and months-since-last-positive — every feature is computable from the held-out event's T alone. Class weights are balanced and the held-out event is never seen during training.

The two LR variants tell different stories per domain. On TradeWar, dropping phase metadata only moves the LR from 72.7% to 68.2%; both versions cluster around the 70.5% LLM headline, and the audited LLMs do not measurably exceed the trivially-trainable LR. On FOMC and Energy, the phase-leaky LR drops sharply when phase metadata is removed (FOMC 85.7%→50.0%; Energy 90.0%→50.0%): the LR_{phase} result was dominated by analyst hindsight. On FOMC, the LLMs reach

Domain	Maj.	Pos.	Persist	Phase	LR_{phase}	LR_{causal}	LLM
TradeWar (n=22)	63.6	36.4	63.6	68.2	72.7	**68.2**	70.5
FOMC (n=14)	57.1	42.9	71.4	78.6	85.7	**50.0**	79.8–92.3
Energy (n=10)	60.0	40.0	60.0	60.0	90.0	**50.0**	55.0–56.7

Table 5: **Non-LLM baselines vs. LLM headline (event-level accuracy, %).** Maj., Pos., Persist, Phase are simple non-trainable baselines. LR_{phase} uses analyst-defined phase metadata (a leakage stress test). LR_{causal} uses only strictly-past calendar features. LLM column reports the audited-grid headline range; on TradeWar the main grid is a single number (70.5%), on FOMC the deanon-raw 92.3% / anon-L2 79.8% range, on Energy the per-model range across the anon-L2-only check.

Model	deanon	anon-L2	Δ
grok-4-20	83.9%	75.0%	+8.9
gpt-5.4	100.0%	87.5%	+12.5
o3	100.0%	92.9%	+7.1
o3-mini	76.8%	64.3%	+12.5
o1	100.0%	87.5%	+12.5
deepseek	92.9%	71.4%	+21.4
Average	92.3%	79.8%	+12.5

Table 6: **FOMC cross-domain stress test (14 timelines, deanon and anon-L2 protocols only; 672 cells shown).** The original 1,008-cell grid included an anon-cfx column whose triggers were byte-identical to anon-L2; that column is superseded by Table 17 (App. K). Per-model deanon vs. anon-L2 boost ranges from +7.1 to +21.4 pp. The aggregate +12.5 pp boost is consistent in *direction* across all six models (every Δ is positive), but the magnitude is model-dependent rather than uniform. Each model contributes 56 cells per protocol.

79.8–92.3%, substantially above LR_{causal} (consistent with D5's +12.5 pp deanon boost: entity recognition / memorised institutional context plausibly drives the gap). On Energy, the LLMs are pinned to 55–57%, only marginally above LR_{causal}'s 50%. The narrower lesson: scalar headlines are not a reliable indicator of LLM-specific capability without an explicit strictly-causal non-LLM baseline; on TradeWar and Energy the audited LLMs are not clearly beyond the trivial baseline, on FOMC they exceed it (App. K attributes most of that excess to Fed-specific identity). The LR vs. LLM comparison is a *triviality diagnostic*, not a competitive ranking: it shows headline accuracy is compatible with a low-capacity calendar prior, not that LR is a better forecaster.

5.6 Cross-Domain Stress Tests: FOMC and Energy Policy

We apply the audit to two additional settings with deliberately uneven coverage. FOMC (Hike vs. Hold/Cut, 14 events 2018–2025) uses the full 1,008-cell grid (6 models × 14 events × 3 protocols × 4 conditions), parallel to TradeWar. Energy (10 fossil-vs-renewable windows) uses only anon-L2 (240 cells), because deanon-raw/anon-cfx triggers are not yet implemented; it serves as a narrower input-invariance check.

When models can recognize the Federal Reserve, accuracy jumps by +12.5 pp on average; every model shows a positive boost, but the magnitude varies (7.1–21.4 pp). With n=6 the spread cannot support a calibrated mechanistic story, and the gap is ambiguous between contamination and the legitimate signal that "Federal Reserve" is a useful entity prior. On the post-cutoff 2025-Mar FOMC meeting, grok and o3-mini are invariant across all 12 (protocol × reminder) cells, while the other four models (o1, o3, gpt-5.4, deepseek-v4-flash) are also invariant but on the correct (Hold/Cut) class (App. K).

Energy is a modest third-domain check. Under anon-L2, 58/60 model–event grids are invariant across the four conditions; average accuracy is 56.7% (solo) and 55.0% (each MAS condition). The frozen prior's direction varies by model (o3-mini fossil-favoring; gpt-5.4, o1, o3, deepseek renewable-favoring), but the diagnostic point is the same: within-grid input-invariance recurs outside TradeWar/FOMC. We do not use Energy as evidence for anonymisation rescue or entity memorisation.

6 Discussion

What the LR diagnoses, and what only the audit diagnoses. The strictly-causal LR (D1) is a *triviality diagnostic*: could the headline arise from a low-capacity calendar prior? On TradeWar and Energy, yes. The LR is silent on *why* headlines are untrustworthy or which failure mode produces them. The audit supplies the *mechanism diagnostic*: class-collapse (D2), axis-specificity of the frozen response (D3), and aggregation-layer behaviour (D5) are all audit-only; D1's triviality verdict is baseline-only. We organise D1–D5 below in order of inferential strength: D1–D2 are the empirical anchor, D3–D5 are descriptive diagnostics.

(D1) On TradeWar and Energy, audited LLMs are not clearly beyond a simple strictly-causal calendar baseline; on FOMC they exceed it, but the excess is contamination-dominated. The strictly-past-feature LR reaches 68.2 / 50.0 / 50.0%

on TradeWar / FOMC / Energy vs. audited LLMs 70.5 / 79.8–92.3 / 55–57% (§5.5). The FOMC excess decomposes via the placebo-institution control (App. K: $\Delta_{\text{contam}} = +16.1$ pp, $\Delta_{\text{context}} = -2.9$ pp; 5/5 models nonnegative on Δ_{contam}, 4/5 strictly positive) and is attributable to Fed-specific identity rather than to named-institution-vs-generic context. A phase-leaky LR variant reaches 72.7 / 85.7 / 90.0% as a leakage stress test, not a deployable forecaster.

(D2) Headline accuracy hides class-collapse on Wait-heavy calendars. TradeWar's 70.5% headline coexists with 28.5% escalation vs. 94.5% Wait recall; o1 and o3 each output Wait in 72/72 cells (always-Wait at 66.7% on the 4/6 Wait-heavy calendar; §5.1). This is the signature D1's LR result makes legible: LR-vs-LLM parity is consistent with both forecasters driven by the same calendar structure rather than situation-reading.

(D3) Frozen-Wait is axis-specific, not content-invariant. o1 and o3 are invariant to entity-visibility, persona, and architecture interventions (72/72 Wait) but move on 5/6 events under a relative-power probe that perturbs only the trigger's content (App. I). The frozen response is therefore not a universal hard-stop; it is a content-dependent prior that only certain intervention axes can move.

(D4) Exploratory single-model illustration of intervention-gated movement ***(n=2 events; pattern, not population effect).*** o3-mini moves 0/8 $\rightarrow$ 3/8 $\rightarrow$ 6/8 across deanon-raw / anon-L2 / anon-cfx on the two 2025 escalation events; paraphrase-L0 (entity names kept) leaves o3-mini at 1/8, and anon-cfx-v2 (different rewriter, gpt-4o-mini) reproduces the 6/8 rescue (App. J) — more consistent with entity removal than with lexical-surface or rewriter-choice effects.

(D5) FOMC entity visibility decomposes to contamination $\gg$ legitimate context; MAS country reweighting is net-negative. FOMC deanon-over-anon-L2 is positive in all six models (+12.5 pp avg, 7.1–21.4 range; §5.6); the placebo-institution control (App. K) decomposes this into $\Delta_{\text{contam}} = +16.1$ pp (≥ 0 in 5/5, strictly positive in 4/5; o1 ceiling-bound) and $\Delta_{\text{context}} = -2.9$ pp. Consultant–country disagreement is sparse (14/1,188 = 1.2%, 6 corrections vs. 8 overrides); 4/6 models lose accuracy under MAS in T1. Descriptive diagnostics, not calibrated estimates (Cemri et al., 2025; Wu et al., 2025; Bertalanič and Fortuna, 2026).

Synthesis. D1–D5 are demonstrations of what the audit reveals — bounded case-study disclosures with heterogeneous evidential strength, not population estimates. We self-apply the 11-criterion checklist (App. A) and meet 10/11 fully; the uncertainty criterion is partial (bootstrap CIs on TradeWar / FOMC main grids; fractions only on $n \leq 6$ sub-grids). A disconfirming observation — a model intervention-stable across all four audit axes yet preserving $\geq 80\%$ escalation recall on a class-balanced post-cutoff cohort — would refute the class-collapse claim in D1.

7 Conclusion

We instantiate an intervention-response audit for LLM policy-decision pipelines (2,832 cells + 216-cell extension). A strictly-causal LR reaches 68.2 / 50.0 / 50.0% vs. LLMs 70.5 / 79.8–92.3 / 55–57% on TradeWar / FOMC / Energy: LLMs are not clearly beyond the simple causal baseline on TradeWar or Energy; on FOMC the placebo decomposition (App. K: $\Delta_{\text{contam}}=+16.1$, $\Delta_{\text{context}}=-2.9$) attributes most of the excess to Fed-specific identity. The audit also surfaces class collapse (28.5% escalation recall under 70.5% headline; o1/o3 always-Wait 72/72) and a single-model intervention-gated illustration (n=2; App. J). Headline accuracy is not capability without strictly-causal baselines and the App. A disclosures.

Limitations

We organise limitations into three groups: (a) *design limitations* that constrain inferential power on the existing grid, (b) *interpretive ambiguities* where the audit's resolution is partial (e.g., the FOMC gap is now contamination-dominated per App. K but residual confounds remain), and (c) *generalisation boundaries* of the case study.

(a) Design limitations

Domain coverage is uneven. TradeWar is the full 1,584-cell discovery grid; FOMC is a parallel 1,008-cell three-protocol stress test; Energy is a 240-cell anon-L2-only check that cannot adjudicate the deanon-vs-anon entity-visibility effect driving D4. Energy is therefore reported as a within-grid input-invariance pilot rather than as part of the headline scope.

Six-model cohort and post-cutoff event count. Each per-model verdict in T1–T3 rests on at most 22 TradeWar timelines and 6 Trump-2.0 events, with M=6 models. The post-cutoff calendar is small and phase-imbalanced (2 escalation, 4 Wait), so the 0/8 vs. 6/8 escalation contrast for o3-mini is a clean within-model existence proof but does not by itself fix a population escalation-recall rate. A class-balanced post-cutoff calendar would be needed to separate calibrated de-escalation prior from policy collapse.

Effective cluster counts are small. Event-clustered bootstrap with B=10,000 resamples is reported on the TradeWar (22 events) and FOMC (14) grids, but the underlying cluster counts mean those CIs should be read as rough rather than high-information uncertainty. For sub-grids with $n \leq 6$ event-level clusters (the 6-event post-cutoff TradeWar 2025 cohort, the 2-event o3-mini escalation contrast, and the FOMC 2025-Mar single-event probe), the bootstrap resample space is too small to support a meaningful CI; we report fractions only and drop bracketed bootstrap intervals on those contrasts (§5.1). We never substitute bootstrap width for sample size.

Multiple comparisons: limited-scope BH-FDR. Headline scalars carry event-clustered 95% bootstrap CIs; the four headline-family p-values (FOMC sign/permutation, anchor-recovery sign/permutation) are corrected with Benjamini–Hochberg at α=0.05. None passes BH-FDR (best q=0.125), so we report direction + CI rather than p-value-based significance. *This BH-FDR correction is intentionally limited in scope*: it does not address the larger family of per-model, per-protocol, per-condition contrasts narrated throughout, and we mark those as exploratory diagnostics rather than as multiplicity-controlled findings.

Consultant–country disagreement is sparse. The 14/1,188 disagreement count is small (1.2%), with 6 corrections vs. 8 incorrect overrides. We treat this descriptively (D5 is a worked example of an aggregation-layer disclosure) rather than as a calibrated estimate of country-layer net benefit; a denser sweep with more high-tension pivots would be needed.

Label rule is strict-bilateral. Our TradeWar labels apply a strict bilateral-tariff rule. A broader rule that counts sector-wide measures or technology controls would re-classify several Wait windows as Declare and could change per-cell verdicts. We mark this explicitly because policy forecasting is label-sensitive.

Wait is a coarse label. The binary task collapses negotiation, pause, de-escalation, and arrangement-maintenance into a single Wait class. Within the 4/6 Trump-2.0 Wait events, this hides whether a model would have predicted active de-escalation or passive non-action. A three-way action space would distinguish these but answers a different question; we do not attempt it here.

Anonymisation protocols are not perfect. anon-L2 may not fully prevent re-identification from contextual cues; anon-cfx mitigates that risk but introduces the rewriting LLM's biases. We use a model outside the tested set for rewriting to avoid circularity, but a fully causally-controlled entity-visibility experiment would require synthetic prior-art-free events.

(b) Interpretive ambiguities

Anchor-dependent recovery: surface-difficulty confound. Counterfactual rewriting is not a controlled intervention on entity identity alone—it can simultaneously simplify lexical surface, shorten sentences, or reduce political salience, any of which could ease the prediction independent of memorisation. Properly isolating the entity-removal channel would require an additional control swapping real entities for equally-salient non-memorised alternatives. We therefore do *not* claim the o3-mini recovery diagnoses memorisation per se; we claim only that some intervention on the entity layer can move o3-mini's decision while no intervention moves o1's or o3's.

FOMC entity-visibility gap is contamination-dominated, but residually compound. The all-positive +12.5 pp deanon-over-anon-L2 boost on FOMC was originally consistent with two distinct mechanisms: (i) contamination (the model recognises "Federal Reserve" as an event whose answer it has memorised) and (ii) legitimate context-removal (anonymising a high-salience institution removes a useful task-relevant prior the briefing should expose). The placebo-institution control (App. K, $\Delta_{\text{contam}} = +16.1$ pp vs. $\Delta_{\text{context}} = -2.9$ pp on 5/5 models) resolves the split toward contamination. A residual ambiguity remains around what exactly the fictional-naming control removes (the original Fed-specific outcome memory vs. a more

diffuse central-bank reaction-function prior); a synthetic monetary-policy task with no real-world referent at all would tighten this further.

Variance decomposition is order-dependent. The deviance shares in App. I come from sequential likelihood-ratio decomposition with timeline as a random effect; individual percentages depend on term order, and adding model×phase shifts dominance from model×condition to model×phase. We report *both* orderings to expose order-dependence and use the table only for the qualitative claim that phase- and protocol-mediated interactions matter, not as a uniquely identified causal variance partition.

Release-date confound for the grok signal. grok-4-20-reasoning's 88.9% Trump-2.0 accuracy is reported as a release-date contamination control because grok's public release postdates the entire 2025 calendar. Three of our six models additionally have unknown official cutoffs. We report cutoff/event-phase results per (model, event), but the audit cannot fully separate "model has seen the answer" from "model is competent". A stronger design would include placebo cutoffs and negative-control tasks whose answers are not present in any training data.

Public-discourse leakage. Our temporal-provenance disclosures rely on official cutoffs and release dates. For high-salience 2025 events, public discourse and indexed-news traffic could leak signal even when the model's training-data cutoff predates the event, especially if any post-training, RAG, or browsing pathway is enabled. We do not control for this and treat it as an additional confound on the post-cutoff cohort.

(c) Generalisation boundaries

Single-model basis of the anchor-dependent recovery. Only o3-mini shows the anon-cfx rescue clearly (6/8 vs. 0/8). Replicating this on a second model with a documented cutoff and a non-trivial post-cutoff calendar would substantially strengthen the inferential basis. We treat the recovery as a single-model existence proof on this benchmark, not as a property of OpenAI reasoning models in general or as an antidote to the frozen-Wait pattern.

Domain transferability. The framework's applicability outside high-salience institutional-policy forecasting (e.g., commercial product launches, lower-coverage diplomacy events, scientific-discovery forecasting) remains untested. We argue the response-surface *discipline* transfers; we do not claim the specific signatures (frozen-Wait, anchor-recovery, entity-visibility direction) reproduce in those settings without re-running the audit.

Reporting checklist is a starting point, not a validated standard. The 11 disclosure criteria in App. A are lessons drawn from this audit, not a community-validated checklist. We present them to motivate further discussion—e.g., a Delphi study with deployment practitioners, or independent application of the checklist to a different LLM forecasting system—rather than as a settled methodology.

Non-LLM baseline interpretation. The phase-leaky LR_{phase} matches or beats LLM headlines on all three domains (72.7 / 85.7 / 90.0%) but uses analyst-defined phase metadata from neighbouring training-fold events; the strictly-causal LR_{causal}, which uses only past-only features, drops to 68.2 / 50.0 / 50.0% and only matches the LLMs on Trade-War (and approximately on Energy). The strongest defensible reading is narrow: *on these small, phase-structured calendars*, low-capacity calendar features (with phase metadata) capture much of what scalar accuracy measures, so headline accuracy is not by itself diagnostic of LLM-specific capability. We do *not* claim LLMs add no value over calendar features in general; richer or larger event sets, or tasks where calendar phase is uninformative, may invert this comparison.

Ethical Considerations

Deployment-risk framing of headline accuracy. This paper audits LLMs that are being proposed as decision infrastructure for high-stakes policy tasks. We are explicit throughout that scalar accuracy on resolved decisions is degenerate for this purpose: the same headline can arise from a calibrated forecaster and from a frozen always-WAIT policy that misses every escalation event. We accordingly frame our headline numbers as a *deployment-hazard signal*—disagreement between role-play recommendations and realised decisions indicates deployment risk regardless of which side is normatively right—rather than as a calibrated forecast of any official's behaviour, and we ask that downstream readers do not extract per-event point predictions from our tables and treat them as policy

advice.

Use of public-figure and public-institution data. All decision events are drawn from publicly available U.S. government primary sources: USTR notices and Federal Register entries for TradeWar (App. F), FOMC press releases for monetary policy, and White House EOs together with DOE/EPA/IEA/EIA notices for Energy (App. C, App. F). Officials and institutions are referenced only in their official capacity, only via documents they themselves released, and only for decisions that are already part of the public record. No private individuals, private communications, or non-public data are used.

Dual-use risk of geopolitical-simulation pipelines. Our pipeline derives from a published multi-agent geopolitical simulator (Hua et al., 2023), and we are aware that work in this lineage can be repurposed as a recommendation engine for state-level actors. Three design choices push against that risk. First, we score model outputs against *realised* decisions rather than presenting them as normative advice, and we treat the elicitation–scoring gap explicitly as a deployment-hazard signal rather than as a forecasting metric. Second, we document and quantify role-play scoring artefacts—class collapse (TradeWar 28.5% escalation recall under a 70.5% headline; o1/o3 always-WAIT 72/72 on the post-cutoff cohort), contamination-dominated entity-visibility gaps (App. K), and intervention-gated single-model recoveries (App. J)—so future deployers cannot point at a single accuracy number without inheriting the audit's negative findings. Third, we release code, prompts, ground-truth tables (App. F), and the reporting checklist (App. A) so that the same audit can be re-run, re-scored, and contested rather than treated as a one-off endorsement.

Commercial API access and reproducibility scope. All audited models are accessed via provider SDKs or OpenAI-compatible endpoints under each provider's terms of service; we do not redistribute model weights, training data, or any provider-internal artefact. Because the upstream provider can resolve the same `model` string to different weights at different times, we cannot guarantee bit-exact reproducibility of our numeric tables. We instead support *cell-level auditability*: every per-cell JSON logs the request-time `model` string, the response fingerprint, the timestamp, and the full decoding configuration (App. H, Tables 12 and 13), so a future reader can resolve which upstream snapshot was queried on what date. The qualitative diagnostics (frozen-WAIT, anchor-dependent recovery, entity-visibility direction) are stable across re-runs even where individual numeric values are not bit-stable.

Anonymisation as diagnostic, not as privacy guarantee. The anon-L2 and anon-cfx protocols are designed as *intervention probes* on entity visibility, not as privacy-preserving transformations. We do not claim that anon-L2's deterministic find-and-replace or anon-cfx's counterfactual rewriting would be sufficient to deidentify sensitive data in any other setting; readers should not treat them as safety mechanisms for tasks involving private individuals or non-public events. App. C also notes that anon-cfx outputs are not currently audited by a held-out human annotator for hindsight leakage, which we flag as a known limitation rather than as a safety control.

AI assistance in writing and pipeline construction. The audited LLMs are the subject of the paper rather than tools used to write it. Commercial LLMs are used in instrumented, declared, non-scientific roles: `claude-sonnet-4-5-20250929` as the anon-cfx and paraphrase-L0 rewriter (App. C, App. J), `gpt-4o-mini` as the anon-cfx-v2 rewriter-robustness control rewriter (App. J), and `gpt-5.4-mini` as the deterministic HelperAgent parser (T=0, seed=1; App. H). None of these models is part of the audited cohort and none modifies the upstream model's free-text answer. AI writing tools were used only for light copy-editing on author-drafted prose; the experimental design, the diagnoses, the limitations, and all scientific claims are the authors'.

References

Aishwarya Bandaru, Fabian Bindley, Trevor Bluth, Nandini Chavda, Baixu Chen, and Ethan Law. 2025. Revealing political bias in LLMs through structured multi-agent debate. *arXiv preprint arXiv:2506.11825*.

Blaž Bertalanič and Carolina Fortuna. 2026. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. *arXiv preprint arXiv:2605.00914*.

George E. P. Box and K. B. Wilson. 1951. On the exper-

imental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, 13(1):1–45.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*.

Clea Chataigner, Pablo Piantanida, and Golnoosh Farnadi. 2026. Position: Auditing is not evaluating; LLM audit requires dynamic, contextual, budget-aware and reliable evidence. *HAL preprint hal-05498201*. Submitted 6 Feb 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. ArXiv:2305.14325.

Lisette Espín-Noboa and Gonzalo Gabriel Méndez. 2026. Whose name comes up? benchmarking and intervention-based auditing of LLM-based scholar recommendation. *arXiv preprint arXiv:2602.08873*. Preprint, 9 Feb 2026.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (WarAgent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Joohyoung Jeon and Hongchul Lee. 2026. Can blindfolded LLMs still trade? An anonymization-first framework for portfolio optimization. *arXiv preprint arXiv:2603.17692*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. 2024. ForecastBench: A dynamic benchmark of AI forecasting capabilities. *arXiv preprint arXiv:2409.19839*.

Zehan Li, Yuxuan Wang, Ali El Lahib, Ying-Jieh Xia, and Xinyu Pi. 2026. Simulated ignorance fails: A systematic study of LLM behaviors on forecasting problems before model knowledge cutoff. *arXiv preprint arXiv:2601.13717*.

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 (EMNLP)*. ArXiv:2305.19118.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. 2025. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*.

Raymond H. Myers, Douglas C. Montgomery, and Christine M. Anderson-Cook. 2016. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 4 edition. Wiley.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2024. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. ArXiv:2310.11324.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. ArXiv:2310.13548.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. ArXiv:2203.11171.

Haolun Wu, Zhenkun Li, and Lingyao Li. 2025. Can LLM agents really debate? A controlled study of multi-agent debate in logical reasoning. *arXiv preprint arXiv:2511.07784*.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*.

Zeyu Zhang, Ryan Chen, and Bradly C. Stadie. 2026. All leaks count, some count more: Interpretable temporal contamination detection and mitigation in LLM backtesting. *arXiv preprint arXiv:2602.17234*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. ArXiv:2309.03882.

A Reporting Checklist

Table 7 lists the disclosures that changed the interpretation of our own headline scores. We present it as a *lessons-from-this-audit* checklist intended to seed community discussion—not as a validated standard. Independent application of the checklist to a different LLM forecasting system, or a Delphi study with deployment practitioners, would be a natural follow-up before promoting any of these criteria to recommended practice.

B Unit-of-Analysis and Denominator Map

Every headline number is mapped to its generating denominator below, because the same audited grid is reported at different levels of aggregation throughout the paper (full grid, per-protocol slice, per-event slice, escalation-only sub-cohort). Table 8 lists every numeric claim that appears in the abstract, intro, results, discussion, or conclusion together with its (models $\times$ events $\times$ protocols $\times$ conditions) denominator and the section that quotes it. The cell counts in this paper are repeated measurements of the same events under different audit factors (one cell = one model–event–protocol–condition tuple, aggregated from 35 LLM calls); they are *not* independent forecasts and bootstrap CIs are clustered on event id.

C Briefing Construction Protocol

Prompt provenance is as important as ground-truth provenance for a forecasting/contamination audit. We summarise the briefing-construction pipeline below; the full prompt templates and the per-event briefing manifest (with information-cutoff dates and source URLs) are released in the supplementary code repository.

Information cutoff per event. For each decision-event with date T, the briefing's evidence set is restricted to material time-stamped strictly before T. We use the decision date itself, not $T+\Delta$, as the cutoff: any document, news article, or government release dated on or after T is excluded.

Source retrieval rule. For TradeWar (2018–2020 + 2025), per-event briefings are assembled from (i) Congressional Research Service reports, (ii) Federal Register notices, (iii) USTR announcements, and (iv) contemporaneous policy commentary indexed in news archives, all filtered by the pre-T cutoff. For FOMC, briefings draw on Fed minutes, Beige Book entries, FOMC press releases preceding the meeting, and macro-data prints with a release date earlier than T. For Energy, briefings draw on White House EOs, DOE/EPA notices, and IRA/IEA/EIA releases with publication dates earlier than T.

Hindsight-language removal. Briefings are prepared by the authors. We do *not* use an LLM to write the deanon-raw briefing text; the authors paraphrase and excerpt source material to keep length and reading level consistent across events. As a sanity check, briefings were searched for outcome-indicative wording (e.g., "ultimately announced", "would later impose", resolution-phase nouns referring to the event's own $T+\Delta$ outcome); flagged passages were rewritten to neutral pre-decision framing.

anon-L2 construction. Real entities (countries, leaders, institutions, legislation names) are mapped to generic labels (*Country A*, *Leader 1*, *Trade Organization*, etc.) by a deterministic find-and-replace dictionary. No LLM is involved.

anon-cfx construction. A separate LLM (a model not in the audited cohort) rewrites each deanon-raw briefing into a fictional but causally-equivalent scenario (e.g., "United States $\rightarrow$ Vandalia," "China $\rightarrow$ Meridia"). The rewriting prompt requires preserving (a) the temporal sequence of actions, (b) the trigger–response structure, (c) the relative-power asymmetry, and (d) all numeric magnitudes (tariff rates, trade volumes). The rewriting LLM is given the deanon briefing only and is explicitly told the question; *it is not given the ground-truth label or any information about the post-T outcome.* We do not currently audit anon-cfx outputs for hindsight-leakage with a held-out human annotator, and we mark this as a known limitation that motivates the matched-control experiment proposed in the Limitations.

Why this matters for the audit. Because the briefing is identical (modulo entity replacement) across deanon-raw / anon-L2 / anon-cfx for the same event, intra-event protocol contrasts are not confounded by source-set differences. Because the briefing is constructed from pre-T material only, the model's task is genuine forecasting given the audited information set; the open contamination question is not whether the briefing leaks the outcome, but whether the model's training corpus al-

Criterion	What to report	Why it matters
Scoring layer & intermediate disagreement	Whether the score is computed on the final deployed decision, intermediate agent votes, or both; the aggregation rule used for the headline score; and, in multi-agent systems, the per-cell disagreement rate between intermediate-layer votes and the final decision layer.	Different layers can behave differently and the deployed decision is what matters; reporting the disagreement rate shows whether the system is a mechanical aggregator or a reweighting/override mechanism.
Temporal provenance & OOD status	For every test event, report (a) the model's documented training cutoff, (b) its public release date, and (c) a per-event label of whether the event is genuinely OOD for that model, post-cutoff but inside the release window, or pre-cutoff.	Separates genuine generalization from contamination control and release-date confounds; this should be reported per (model, event), not per cohort.
Labeling rule	State the exact action taxonomy and decision window used to assign labels, including how borderline cases are handled.	Forecasting accuracy is label-sensitive; changes in the action definition can change the result.
Class balance	Report the number and proportion of examples in each class and the resulting balanced accuracy or per-class recall.	Prevents inflated headline accuracy from majority-class priors such as always-Wait behavior.
Event-phase split	Report performance separately for escalation vs. de-escalation, or analogous phase categories relevant to the domain.	Reveals whether the model tracks substantive phase changes or just the class distribution.
Entity protocol	Specify whether the evaluation uses real names, anonymized names, masked entities, or counterfactual rewriting.	Entity visibility can change behavior; without this, contamination and intervention effects cannot be distinguished.
Prompt condition	List every prompt variant, including persona reminders, counter-prompts, and control prompts, with the exact wording or a pointer to it.	Prompt effects can be large and non-transferable; selective reporting can mimic improvement.
API-call budget	Report the number of API calls, rounds, temperature, and whether compared conditions match at the API-call level (per-call token cost may still differ when prompt length is part of the architectural contrast).	Prevents architecture comparisons from being confounded by budget differences.
Response-profile summary	Report per-condition deltas, ranges, and sign reversals rather than only one aggregate accuracy number.	Makes intervention sensitivity visible and prevents brittle systems from looking stable.
Uncertainty	Provide bootstrap confidence intervals or timeline-resampled uncertainty for key scores.	Distinguishes robust patterns from small-sample fluctuations.
Cross-domain scope	State which domains implement the full grid and which are partial stress tests only.	Prevents readers from overgeneralizing a limited auxiliary experiment.

Table 7: **Lessons-from-this-audit reporting checklist.** Each row links a methodological disclosure to the failure mode it is designed to surface. Presented as a starting point for community discussion, not as a validated standard.

Headline claim	Models	Events	Protocols	Conditions	Cells	Notes / where it appears
TradeWar overall accuracy 70.5% [55.6, 84.4]	6	22	3	4	1,584	Abstract, §5; balanced acc 61.5%, esc. recall 28.5%, Wait recall 94.5% on same denominator.
Trump-2.0 cohort 2025 calendar (per model)	1	6	3	4	72	§5.1, Tab. 4; o1, o3 each 48 Wait-correct + 0 escalation = 48/72 = 66.7%.
o3-mini full 2025 per protocol (24/protocol)	1	6	1	4	24	§5.1: 16/24 (deanon), 19/24 (anon-L2), 22/24 (anon-cfx).
o3-mini 2025 escalation only (8/protocol)	1	2	1	4	8	§5.1: 0/8, 3/8, 6/8 across deanon/anon-L2/anon-cfx; sign test acts on 2 events.
o3-mini 2025 de-escalation only (16/protocol)	1	4	1	4	16	§5.1: 16/16 Wait recall under all three protocols.
o3-mini 79.2% Trump-2.0 acc. headline	1	6	3	4	72	§5.1, Tab. 4; 57/72.
T1 architecture deltas $\bar{f}_{\mathrm{arch}}$	6	22	3	2 (solo, MAS-off)	792	§5.2; per-model average over 22 $\times$ 3 cells.
T2 prompt deltas $\bar{f}_{\mathrm{prompt}}$ (MAS)	6	22	3	2 (MAS-on, MAS-off)	792	Tab. 3; one cell per (model, protocol).
FOMC deanon vs. anon-L2 boost	6	14	2	4	672	§5.6, Tab. 6; per model 14 $\times$ 2 $\times$ 4=112 cells; gap reported per protocol pair.
FOMC 2025-Mar single-event probe	6	1	3	4	72	App. K reports per-model 12 (protocol $\times$ reminder); the 4 conditions reduce to 4 per cell.
Energy anon-L2 grid	6	10	1	4	240	§5.6; 58/60 model–event grids invariant.
TradeWar 2025 three-model extension	3	6	3	4	216	App. I; reported separately, not pooled.
Consultant–country disagreement count	6	22	3	MAS-on/off/counter	1,188	§6 (D5); 14/1,188 = 1.2%.

Table 8: **Unit-of-analysis and denominator map.** Every number that appears in the headline narrative is mapped to its (model $\times$ event $\times$ protocol $\times$ condition) generating set. Cells are repeated measurements; we report event-clustered bootstrap CIs because the effective independent unit is the event, not the cell. The 22+14+10=46 events are the dominant source of independent information; the 1,584/1,008/240/216 cell counts reflect the depth of the intervention sweep, not the breadth of independent forecasts.

ready contains the resolved event. The audit is designed to make the latter visible through the post-cutoff cohort, the release-date control (grok), and the entity-protocol contrast.

D Compute Budget Accounting

Both the solo and MAS-pipeline conditions use 35 LLM API calls per cell, ensuring call-budget-matched comparison:

Solo ($k = 35$ self-consistency). The solo agent receives the briefing document and is asked to produce a binary verdict. This prompt is issued 35 times independently (with temperature $= 0.8$), and the final verdict is determined by majority vote over the 35 responses.

MAS pipeline (7 agents, 35 calls). The pipeline distributes its budget as follows:

- 6 consultant agents $\times$ 5 rounds each = 30 calls
- 1 country-decision agent $\times$ 5 rounds = 5 calls
- Total: 35 calls

Note that the solo and MAS-pipeline conditions differ in *how* they use their 35 calls (homogeneous sampling vs. a staged consultant–country pipeline), which is precisely the architectural difference T1 is designed to measure. The compute budget is matched at the API-call level; per-call costs may vary across conditions due to differing prompt lengths and output verbosity.

E Full Results Grid

The complete 6×3×4×22 TradeWar grid is provided in the supplementary materials as a CSV file with 1,584 rows. The released CSV records both intermediate consultant recommendations and final country-decision predictions.

F Ground Truth Construction

Each decision point is coded based on official actions within a 2-month window (T, T+60] starting at the decision date. TradeWar uses Federal Register notices and USTR announcements; FOMC uses official Federal Reserve press releases; Energy uses White House / executive orders / DOE / EPA notices.

TradeWar (22 events). DECLARE = bilateral tariff imposition, escalation, or formal retaliation; WAIT = negotiation, pause, de-escalation, or no action.

FOMC (14 events). HIKE = target range increases at the meeting; HOLD/CUT = target range unchanged or decreased.

Energy (10 events). FOSSIL = action that materially supports fossil fuels or rolls back environmental regulation; RENEWABLE = action that materially supports renewables, decarbonisation, or efficiency.

We mark the labelling rules explicitly because policy forecasting is label-sensitive: under broader TradeWar rules that count sector-wide measures or technology controls, several Wait windows would re-classify; the FOMC Hold/Cut merge is a labelling choice (a counter-party betting on tightening sees Hold and Cut as the same outcome); Energy windows can also be re-coded at finer granularity (e.g., separating LNG-export vs. vehicle-emission actions). Re-coding would change reported per-cell verdicts.

G Broad-Rule Label Sensitivity (TradeWar)

The TradeWar prompt menu (App. L) explicitly lists technology-related controls, narrowed exclusions, and tightened licensing alongside tariff actions as grounds for the `Declare_Trade_War` action. Our headline ground truth uses the *strict bilateral-tariff* rule (Tab. 9). We additionally report what happens to the headline if the scoring rule is aligned to the prompt menu by counting sector-wide measures or technology controls within (T, T+60].

Re-coding procedure. We searched Federal Register, BIS entity-list, USTR, and OFAC notices within each (T, T+60] window for tech-control or sector-wide bilateral actions targeting Country B. Three Wait windows (out of 10 Wait events within the 2018–2020 cohort) plausibly re-classify under such a broad rule, primarily 2019–2020 windows that overlap BIS entity-list expansions. None of the four 2025 Wait windows re-classifies, because the truce regime suppresses both tariff and tech-control actions over that period.

Headline sensitivity. Under the broad rule, the 22-event class balance moves from 8/14 (Declare/Wait) toward roughly 11/11. As a concrete envelope on the main 1,584-cell grid: re-coding three plausible 2018–2020 BIS-era Wait windows (e.g., 2019-10, 2020-05, 2020-09) as Declare moves the pooled headline from 70.5% to roughly 57.8% (envelope 56.8–63.2%), with balanced accuracy moving from 61.5% to ∼57.8% and escalation recall from 28.5% to ∼21.8%. The directional effect on the audit's headline diagnostics:

- **Overall accuracy** drops (the Wait-majority advantage shrinks), with always-Wait baselines falling from 63.6% toward ∼50%. This makes the D1 "headline hides class collapse" point *stronger*, not weaker.
- **Frozen-Wait signature** (o1, o3) becomes more

ID	Window (T, T+60]	Label	Within-window verification (ESCALATE: source dated within window; WAIT: absence of new bilateral tariff in (T, T+60], with regime-establishing context dated $\leq T$ italicised)
W01	2018/2/28–2018/4/28	Escalate	Proclamations 9704 & 9705 (Section 232 steel/aluminum tariffs), 83 FR 11625, 11619 (Mar 15, 2018).
W02	2018/4/28–2018/6/28	Escalate	USTR final list 1 ($34B Section 301), 83 FR 28710 (Jun 20, 2018).
W03	2018/6/28–2018/8/28	Escalate	USTR final list 2 ($16B), 83 FR 40823 (Aug 16, 2018).
W04	2018/8/28–2018/10/28	Escalate	USTR final list 3 ($200B), 83 FR 47974 (Sep 21, 2018).
W05	2018/11/30–2019/1/30	Wait	No new bilateral tariff in window; USTR extended comment period and delayed list 4.
W06	2019/2/1–2019/4/1	Wait	No new bilateral tariff in window; bilateral negotiations continued without new escalation.
W07	2019/4/1–2019/6/1	Escalate	USTR rate increase to 25% on lists 1–3, 84 FR 20459 (May 9, 2019).
W08	2019/6/1–2019/8/1	Wait	No new bilateral tariff in window (list 4 only proposed, 84 FR 22564, May 17, 2019; no final action within window).
W09	2019/8/1–2019/10/1	Escalate	USTR final list 4A ($120B at 15%), 84 FR 45821 (Aug 30, 2019).
W10	2019/10/1–2019/12/1	Wait	No new bilateral tariff in window; pre-Phase-One negotiations.
W11	2020/1/15–2020/3/15	Wait	No new bilateral escalation in window; *regime context (dated $\leq T$):* Phase One Agreement signed Jan 15, 2020. Within-window action is de-escalatory (USTR list 4A 15%→7.5%, 85 FR 3741, Jan 22, 2020).
W12	2020/3/15–2020/5/15	Wait	No new bilateral tariff in window; USTR COVID-19 medical-product exclusions (85 FR 15244, Mar 16, 2020) are de-escalatory.
W13	2020/5/15–2020/7/15	Wait	No new bilateral tariff in window; medical-product exclusions extended.
W14	2020/7/15–2020/9/15	Wait	No new bilateral tariff in window; USTR continued processing exclusion requests.
W15	2020/9/15–2020/11/15	Wait	No new bilateral tariff in window; Phase One review ongoing.
W16	2020/11/15–2021/1/15	Wait	No new bilateral tariff in window; USTR extended medical-product exclusions (85 FR 85831, Dec 29, 2020).
W17	2025/1/20–2025/3/20	Escalate	IEEPA tariffs on China imports (Feb 1, 2025 EO 14195; 10% additional duty), USTR Federal Register notice Feb 5, 2025.
W18	2025/3/20–2025/5/20	Escalate	Reciprocal-tariff EO 14257 (Apr 2, 2025) raising additional duty on China imports to 84%.
W19	2025/5/20–2025/7/20	Wait	No new bilateral tariff in window; *regime context (dated $\leq T$):* U.S.–China Geneva joint statement (May 12, 2025) establishing a 90-day reciprocal-tariff suspension at a 30% baseline. The label is the absence of any new bilateral escalation in (May 20, Jul 20], verified against USTR/Federal Register.
W20	2025/7/20–2025/9/20	Wait	No new bilateral tariff in window; Aug 11, 2025 EO extending the May 12 truce through Nov 10, 2025.
W21	2025/9/20–2025/11/20	Wait	No new bilateral tariff in window; APEC/Busan U.S.–China leaders meeting (Oct 30, 2025) and follow-on truce continuation.
W22	2025/11/20–2026/1/20	Wait	No new bilateral tariff in window; *regime context (dated $\leq T$):* Nov 10, 2025 truce extension to Nov 2026. The label is the absence of any new bilateral escalation in (Nov 20, Jan 20], verified against USTR/Federal Register.

Table 9: **TradeWar ground truth (22 events).** Labels apply the strict bilateral-tariff rule defined in App. F: ESCALATE requires a formally promulgated bilateral tariff action dated *within* (T, T+60]; WAIT is an *absence-of-action* label, verified by USTR/Federal Register catalogue search over the window. Wait labels do not always have a single source dated within the window—an absence does not have a publication date—and we accordingly distinguish (i) within-window actions (de-escalatory, exclusions, extensions) from (ii) regime-establishing context dated on or before T (e.g., W11, W19, W22). 2018–2020 sources from USTR/Federal Register; 2025 sources from White House EOs and USTR Federal Register notices.

ID	Meeting date	Label	Realized action / FOMC press release
F01	2018/3/21	Hike	Target range raised 25 bp to 1.50–1.75% (FOMC press release Mar 21, 2018).
F02	2018/9/26	Hike	Target range raised 25 bp to 2.00–2.25% (FOMC press release Sep 26, 2018).
F03	2018/12/19	Hike	Target range raised 25 bp to 2.25–2.50% (FOMC press release Dec 19, 2018).
F04	2019/7/31	Hold/Cut	Target range cut 25 bp to 2.00–2.25% (first cut of 2019; FOMC press release Jul 31, 2019).
F05	2020/3/15	Hold/Cut	Emergency 100 bp cut to 0.00–0.25% (intermeeting; FOMC press release Mar 15, 2020).
F06	2022/6/15	Hike	Target range raised 75 bp to 1.50–1.75% (FOMC press release Jun 15, 2022).
F07	2022/11/2	Hike	Target range raised 75 bp to 3.75–4.00% (FOMC press release Nov 2, 2022).
F08	2023/7/26	Hike	Target range raised 25 bp to 5.25–5.50% (FOMC press release Jul 26, 2023).
F09	2024/9/18	Hold/Cut	Target range cut 50 bp to 4.75–5.00% (FOMC press release Sep 18, 2024).
F10	2024/11/7	Hold/Cut	Target range cut 25 bp to 4.50–4.75% (FOMC press release Nov 7, 2024).
F11	2024/12/18	Hold/Cut	Target range cut 25 bp to 4.25–4.50% (FOMC press release Dec 18, 2024).
F12	2025/1/29	Hold/Cut	Target range held at 4.25–4.50% (FOMC press release Jan 29, 2025).
F13	2025/3/19	Hold/Cut	Target range held at 4.25–4.50% (FOMC press release Mar 19, 2025).
F14	2025/5/7	Hold/Cut	Target range held at 4.25–4.50% (FOMC press release May 7, 2025).

Table 10: **FOMC ground truth (14 events).** Binary task: HIKE = target range increases at the meeting; HOLD/CUT = target range unchanged or decreased. Sources: `federalreserve.gov/newsevents/pressreleases/` for each meeting date listed; values cross-checked against the FOMC implementation note.

ID	Window (T, T+60]	Label	Final source within window
E01	2017/1/1–2017/3/1	Fossil	EO 13766 (Jan 24, 2017) expediting environmental reviews; presidential memoranda on Keystone XL and Dakota Access Pipelines.
E02	2018/5/1–2018/7/1	Fossil	EPA proposed rollback of Clean Power Plan / ACE rule; DOE Coal/Nuclear Resilience program supports fossil capacity.
E03	2019/9/1–2019/11/1	Fossil	EPA/NHTSA "One National Program Rule" (Sep 19, 2019) revoking California's vehicle-emission waiver.
E04	2021/1/20–2021/3/20	Renewable	EO 13990 (Jan 20, 2021): protecting public health and the environment; reversal of Trump-era environmental policies.
E05	2022/2/1–2022/4/1	Renewable	DOE notices of intent (Feb 11, 2022) to invest $2.91B in advanced batteries under Bipartisan Infrastructure Law.
E06	2022/8/1–2022/10/1	Renewable	Inflation Reduction Act of 2022 signed Aug 16, 2022; largest federal climate/clean-energy investment in U.S. history.
E07	2022/10/1–2022/12/1	Renewable	DOE $2.8B grants to 21 battery-supply-chain projects (Oct 19, 2022); IRS guidance implementing IRA clean-energy credits.
E08	2023/1/1–2023/3/1	Fossil	DOE LNG export approval (Venture Global CP2) and continued onshore oil/gas lease sales under IRA mandate.
E09	2023/9/1–2023/11/1	Renewable	DOE $3.46B Bipartisan Infrastructure Law funding for 58 grid-modernisation projects across 44 states (Oct 18, 2023).
E10	2024/1/1–2024/3/1	Renewable	DOE temporary pause on pending LNG export approvals (Jan 26, 2024) for non-FTA countries.

Table 11: **Energy ground truth (10 events).** Binary task: FOSSIL vs. RENEWABLE based on the dominant policy direction in (T, T+60]. Sources: federalregister.gov, energy.gov, and white-house archives.

conspicuous because escalation recall is now measured on $\sim$11 events instead of 8, with the same 0% floor.

- **Anchor-dependent recovery** (o3-mini, 0/8 $\rightarrow$ 6/8 on the two 2025 escalation events) is unchanged because the same events anchor the contrast.
- **Entity-visibility direction** (FOMC D5) is unaffected; FOMC uses a different label rule.

What we do not claim. We do not run the broad-rule re-coding as a full audit; the 3-event estimate is an envelope from a Federal Register search, not a frozen ground-truth file we score against. A future revision could lock a broad-rule ground-truth table and re-score the full 1,584-cell grid. The point we make now is qualitative: the headline degradation under a broader rule is in the direction the paper's diagnostic claims would predict (always-Wait baselines lose more), so the diagnostic conclusions D1–D3 do not depend on the strict-rule choice.

H Reproducibility Details

Decoding parameters (full disclosure). Table 12 lists every decoding-side hyperparameter passed to the provider SDKs. All audited cells use temperature $= 0.8$ (V1-locked, pre-registered before R3 extension), seed $= 27$ where the SDK accepts a seed, an 8,192-token completion budget, and the provider default for top-p, frequency penalty, and presence penalty (i.e. no override). Reasoning-class models (`o1`, `o3`, `o3-mini`, `gpt-5.4`, `grok-4-20-reasoning`) silently ignore the temperature argument under the OpenAI-compatible Chat Completions schema and route the budget through `max_completion_tokens`; non-reasoning models receive `temperature` and `max_tokens` as separate fields. The HelperAgent that extracts the binary verdict from each model output (a downstream parser, not part of the audited cohort) is called with `temperature`=0 and a fixed seed so that scoring is deterministic given a fixed model response. The runner records `(temperature, seed, rounds_per_consultant=5, rounds_per_country=5, started_at, ended_at)` in every per-cell JSON file (`meta` block); these match Table 12 exactly.

Why T=0.8 and not T=0. Policy auditing typically prefers low or zero temperature for determinism. We use T=0.8 because the solo condition's k=35 self-consistency vote is only meaningful under stochastic decoding—at T=0, solo collapses to a single deterministic call repeated 35 times, breaking the call-budget-matched comparison against the 35-call MAS pipeline. Temperature is V1-locked from the original simulator. A T=0 control sweep would require a different compute-matching scheme (1-call solo vs. 7-call MAS) and is left to future work.

Field	Audited cells (cohort)	HelperAgent (parser)
`temperature`	0.8 (non-reasoning)[†]	0
`seed`	27 (when SDK accepts)	1
`max_tokens`	8192 (non-reasoning)	default
`max_completion_tokens`	8192 (reasoning)	–
`top_p`	provider default	default
`frequency_penalty`	not set	not set
`presence_penalty`	not set	not set
Retry budget	5 transient + same-id route fallover	inherited
Streaming	off	off
Tools / function calls	none	none
Browse / retrieval / cache	off	off

Table 12: **Full decoding configuration.** [†]Reasoning-class models (`o1`, `o3`, `o3-mini`, `gpt-5.4`, `grok-4-20-reasoning`) silently ignore the `temperature` argument; for those rows the field is omitted from the request and the model decodes under provider-internal sampling. Every value is logged per cell in the result JSON `meta` block.

Model snapshots and access route. Table 13 lists, for every audited model, the exact provider API string, the documented snapshot date or public release date, and the documented training-data cutoff where published. Three OpenAI snapshots have published cutoffs and dated snapshot identifiers (`o1-2024-12-17`, `o3-2025-04-16`, `o3-mini-2025-01-31`); for the other three the provider documents only a public release date, and we use that release date as a conservative upper bound on the cutoff. Every per-cell JSON logs the request-time `model` string, the response fingerprint, and the timestamp, so a future reader can resolve which upstream snapshot was queried and on what date.

API access and reproducibility scope. All models are accessed through provider SDKs or OpenAI-compatible endpoints, with the request-time `model` identifier, response fingerprint, and timestamp logged per cell. We did not enable fallback model substitution, caching, browsing, retrieval, tool calling, or streaming in any cell. As with any closed-API audit, we cannot guaran-

API string	Provider	Snapshot or release	Cutoff	Cohort tag
`o1`	OpenAI	2024-12-17 (snapshot)	2023-10	post-cutoff
`o3`	OpenAI	2025-04-16 (snapshot)	2024-06	post-cutoff
`o3-mini`	OpenAI	2025-01-31 (snapshot)	2023-10	post-cutoff
`gpt-5.4`	OpenAI	2025-08 (release; no snapshot date)	unknown	unknown-cutoff
`grok-4-20-reasoning`	xAI	post-2025 calendar (release date)	unknown	release-date control
`deepseek-v4-flash`	DeepSeek	2025 (V-family flash release)	unknown	unknown-cutoff

Table 13: **Pinned model snapshots.** "Post-cutoff" rows have a provider-published cutoff strictly before the 2025 Trump-2.0 calendar; "unknown-cutoff" rows have only a public release date and are flagged accordingly in T3; the "release-date control" row (`grok-4-20-reasoning`) postdates the 2025 calendar and is used to read the post-cutoff signatures of the OpenAI cohort, not as additional post-cutoff evidence.

tee bit-exact weight reproducibility—the upstream provider may resolve the same `model` string to different weights at different points in time. We instead support cell-level *auditability* through the logged identifiers above; the qualitative diagnostics (frozen-Wait, anchor-dependent recovery, entity-visibility direction) are stable across our re-runs even where individual numeric values are not bit-stable.

Parser/HelperAgent transparency. The HelperAgent (`gpt-5.4-mini`, T=0, seed=1) is a deterministic post-hoc parser; it is not part of the audited cohort and does not modify the upstream model's free-text answer. Across the 2,832-cell pooled audit grid (TradeWar 1,584 + FOMC 1,008 + Energy 240; the 216-cell extension is reported separately), all 2,832 cells terminate with `parse_status=ok`, so no cell is excluded from headline tables for parser failure. The released per-cell JSON does not separately persist a *first-attempt* parse verdict (only the final parser output, decoding hyperparameters, and timestamps), so first-attempt vs. retried parse rates cannot be reconstructed retroactively from the released bundle; we flag this as an instrumentation gap and recommend logging both as a row of the reporting checklist. Parser retries reissue only the *parser* call; an upstream model refusal or class collapse counts as the cell's actual behaviour and is not retried.

Code and data. The submission bundle includes the experimental code, prompt templates, result JSON files, and analysis scripts used to produce the reported tables. We will also release a public archival copy upon publication.

I Statistical Details

Bootstrap procedure. Where confidence intervals are reported, bootstrap intervals use 10,000 resamples of decision-point timelines. In the current version, the main T1–T3 claims are reported as descriptive response-surface diagnostics rather than as population-level significance tests over models.

Architecture and prompt diagnostics. T1 reports per-model architecture deltas against the call-budget-matched solo baseline. T2 reports per-model, per-protocol prompt deltas and flags large ranges or sign reversals.

Cutoff/event-phase diagnostic. T3 reports post-cutoff accuracy together with escalation and Wait recall. We use the full 2025 TradeWar calendar to show when scalar post-cutoff accuracy is inflated by event phase.

Mixed-effects logistic regression. Fitted using maximum likelihood with the Laplace approximation. Model: $\mathrm{logit}(p_{ijkl}) = \beta_0 + \beta_{\mathrm{model}_i} + \beta_{\mathrm{cond}_j} + \beta_{\mathrm{proto}_k} + \beta_{\mathrm{model}_i \times \mathrm{cond}_j} + u_l$, where $u_l \sim \mathcal{N}(0, \sigma_u^2)$ is an event random intercept. Deviance contributions computed by sequential likelihood-ratio tests.

Fitted coefficients and event random-effect variance. On the 1,584-cell TradeWar main grid, the fixed-effects GLM (no random intercept) yields AIC $=$ 1928.8 and McFadden pseudo-R^2 $=$ 0.024. Main effects are small (log-odds): protocol anon-L2 vs. deanon-raw $\hat{\beta} = -0.06 \pm 0.14$ (Wald p=0.68), anon-cfx $\hat{\beta} = +0.10 \pm 0.14$ (p=0.49); architecture $\mathrm{MAS}_{\text{off/on/counter}}$ vs. solo $\hat{\beta} \in [+0.13, +0.34]$ (p>0.35); phase post-cutoff $\hat{\beta} = +0.25 \pm 0.13$ (p=0.054). The largest model×architecture interaction is grok×$\mathrm{MAS}_{\text{counter}}$ $\hat{\beta} = -1.42 \pm 0.57$ (p=0.013), but adding the full model × arch interaction block does *not* improve fit: $\mathrm{LR}(15) = 13.83$, $p = 0.54$, $\Delta\mathrm{AIC} = +16.2$. With an event random intercept (Bayes mixed GLM, variational-Bayes Laplace), event-level variance is large, $\hat{\sigma}_{\text{event}} \approx 4.08$ on the log-odds scale ($\hat{\sigma}^2_{\text{event}} \approx 16.6$), which dwarfs every fixed effect by an order of magnitude and absorbs much of the apparent phase sig-

Source (Order A)	ΔDeviance	% Total
Model	20.0	2.8%
Protocol	1.1	0.2%
Condition	2.5	0.3%
Model × Condition	6.2	0.9%
Model × Protocol	1.8	0.3%
Model × Phase	**686.1**	**95.6%**
Total (A)	717.7	100%

Source (Order B)	ΔDeviance	% Total
Model × Phase	**700.6**	**97.6%**
Model × Protocol	5.8	0.8%
Model × Condition	11.3	1.6%
Condition	0.0	0.0%
Protocol	0.0	0.0%
Model	0.0	0.0%
Total (B)	717.7	100%

Table 14: **Variance decomposition under two orderings (TradeWar 6-model main grid, 1,584 cells).** *The percentages are not causally interpretable*; sequential deviance shares are path-dependent by construction, and we report two orderings to expose this. The qualitative finding—that phase- and protocol-mediated interactions matter, with model×phase ordering above other terms under either order—is what the table should be read for. The earlier-revision claim that "model×condition is the dominant source (42.2%)" was an artefact of omitting the phase term and is superseded.

nal (post-cutoff $\hat{\beta}$ rises to $+1.5$). The qualitative reading is that systematic protocol and architecture effects are small relative to event-to-event heterogeneity; the model is therefore used as a structure-disclosure aid, not as the basis for any per-effect verdict.

Variance decomposition (full table, two orderings). Table 14 reports sequential deviance contributions from a binomial mixed-effects logistic regression on the 1,584-cell 6-model main grid (the 216-cell extension and the 1,008/240 FOMC/Energy cells are out of scope here). We report *both* forward (A: model → protocol → condition → model:condition → model:protocol → model:phase) and reverse (B: model:phase first, then model:protocol, model:condition, condition, protocol, model) orderings to expose order-dependence. Adding model×phase moves it to a 95–98% share regardless of ordering; without phase, model×condition appears largest. Use the table only for the qualitative claim that phase- and protocol-mediated interactions matter.

Headline-test family BH-FDR. The four headline-family p-values (FOMC sign p=0.031 / permutation p=0.122, anchor-recovery sign p=0.500 / permutation p=0.500) are corrected with Benjamini–Hochberg FDR at α=0.05, yielding q-values (0.125, 0.244, 0.500, 0.500). None passes BH-FDR; we therefore report effects as direction + cluster-bootstrap CI rather than as p-value-based significance.

Model (2025 cells)	Acc	Esc. recall	Wait recall
claude-3-haiku	66.7%	0.0%	100.0%
deepseek-r1	69.4%	8.3%	100.0%
o1-mini	66.7%	0.0%	100.0%

Table 15: **TradeWar 2025 three-model robustness extension (216 cells).** Reported separately so the main 1,584-cell six-model headline numbers are not confounded with three additional models that were run only on the post-cutoff 2025 calendar. The pattern echoes the main grid: high Wait recall, near-zero escalation recall.

Non-LLM baseline LR specification. Two leave-one-event-out logistic-regression variants (Table 5, §5.5). **$\mathbf{LR_{phase}}$ (phase-leaky)**: features = phase one-hot (4–6 phases per domain, analyst-defined; phase labels fitted from neighbouring training-fold events), year offset, month, cumulative positive rate, last-window positive rate. **$\mathbf{LR_{causal}}$ (strictly causal)**: features = year, month sin/cos, cumulative positive rate over strictly past events, rolling positive rate over the last K=6 past events, months-since-last-positive. Both variants use class-balanced `LogisticRegression` (scikit-learn 1.7, `liblinear`, C=1); when the held-out fold leaves only one class, we predict the single class. The per-event predictions used for Table 5 are released in the supplementary code repository under `analysis/non_llm_baselines/`.

TradeWar three-model 2025 robustness extension. Three additional models (claude-3-haiku-20240307, deepseek-r1, o1-mini) were run on the 6-event 2025 calendar only; this is reported separately as a robustness check rather than merged into the main grid. The 216 cells (3 models × 6 events × 3 protocols × 4 conditions) replicate the within-grid input-invariance pattern: 2025 escalation recall is ≤ 8.3% across all three protocols for all three models, with no anon-cfx rescue (none matches o3-mini's $0/8 \rightarrow 6/8$ contrast).

Relative-power probe (o1, o3 on TradeWar 2025). To test whether o1/o3's frozen-Wait response on the 2025 TradeWar calendar is invariant

Event (GT)	Protocol	Main grid (o1 / o3)	Probe RP (o1 / o3)
2025-01 (Esc.)	deanon-raw	W / W	E / E
2025-01 (Esc.)	anon-L2	W / W	E / E
2025-03 (Esc.)	deanon-raw	W / W	E / E
2025-03 (Esc.)	anon-L2	W / W	E / E
2025-05 (Wait)	deanon-raw	W / W	E / E
2025-05 (Wait)	anon-L2	W / W	E / E
2025-07 (Wait)	deanon-raw	W / W	E / E
2025-07 (Wait)	anon-L2	W / W	E / E
2025-09 (Wait)	deanon-raw	W / W	E / E
2025-09 (Wait)	anon-L2	W / W	E / E
2025-11 (Wait)	deanon-raw	W / W	W / W
2025-11 (Wait)	anon-L2	W / W	W / W

Table 16: **Relative-power probe vs. main grid (48 cells).** On 5/6 events under both protocols, both o1 and o3 flip from Wait (W) to Declare (E) when the trigger asserts Country A's stronger relative power; only on 2025-11 (the latest event, after multiple truce extensions) does the Wait response survive. The main grid's frozen-Wait signature is therefore axis-specific (invariant to entity-visibility and persona, not to content-level priming). The probe *worsens* per-event accuracy on the four Wait ground-truth events; we report it as a response-surface diagnostic, not as an improvement.

to *all* prompt content or only to the entity-visibility / persona axes audited in the main grid, we add a small relative-power probe. The probe modifies the briefing's trigger to assert that Country A holds a stronger relative-power position (improved fiscal capacity, weaker counter-party manufacturing margins, broader export-control toolkit) and leaves every other audit factor fixed. We run 2 models (o1, o3) $\times$ 6 events $\times$ 2 protocols (deanon-raw, anon-L2) $\times$ 2 settings (solo, MAS+reminder-on) = 48 cells; the 5-round country-decision majority is recorded per cell. Table 16 compares the probe's prediction (Esc. count out of 2 settings) to the main-grid prediction on the same (model, event, protocol) cell.

J Lexical-Surface and Rewriter-Robustness Controls (o3-mini)

The exploratory single-model existence proof in §5.1 (o3-mini's 0/8 $\rightarrow$ 6/8 escalation recovery from deanon-raw to anon-cfx) rests on n=2 escalation events. The contrast is not entity-isolated by construction: counterfactual rewriting could plausibly help via lexical-surface change, length change, or the original rewriter LLM's specific wording. We add two confound controls on the same 2025 escalation events (2025-01, 2025-03) at o3-mini's full 4-condition cell budget (solo + MAS_{off} + MAS_{on} + $\text{MAS}_{\text{counter}}$ = 8 cells per protocol):

paraphrase-L0 (lexical-surface control). Preserve every entity name (America, China, Federal Reserve, all leaders/legislation) and every numeric value exactly; LLM-paraphrase the surrounding wording to anon-cfx's target length using `claude-sonnet-4-5-20250929` (the same model the original anon-cfx pipeline used as rewriter, so wording style is held constant while we change the entity-preservation variable). Prompt module: `prompts/d_tradewar/` (real entity names); trigger directory: `triggers/D_TradeWar_PARA/`.

anon-cfx-v2 (rewriter-robustness control). Regenerate the Nation Alpha / Nation Beta fictional-entity scheme using a *different* rewriter LLM (`gpt-4o-mini` via the project's yunwu aggregator), with a rewrite prompt that preserves numeric magnitudes, temporal sequence, trigger-response structure, and relative-power asymmetry but uses different wording. Prompt module: `prompts/cfx_tradewar/` (Nation Alpha/Beta); trigger directory: `triggers/CFX_TradeWar_V2/`.

Results. Figure 3 compares o3-mini's escalation-recall across all five protocols. paraphrase-L0 yields 1/8 (near deanon-raw's 0/8), and anon-cfx-v2 yields 6/8 (matching anon-cfx's 6/8). The rescue is not a lexical-surface artefact and not specific to the original rewriter LLM's choices: paraphrase-L0 and anon-cfx-v2 rule out two specific competing accounts, leaving entity removal (deanon-raw $\rightarrow$ counterfactual fictional names) as the most-consistent surviving account — though we do not claim these two controls causally isolate the entity layer from all confounds.

Per-condition breakdown. Within paraphrase-L0, all three MAS conditions (off / on / counter) on both events plus the cond3/counter on 2025-03 keep o3-mini at consultant-score D=0.00, the strongest possible frozen-Wait signature; the lone correct cell (2025-03, cond3/on, D=0.57) is the only escape. Within anon-cfx-v2, the two misses are both `counter`-prompt cells (2025-01 D=0.29, 2025-03 D=0.00); all four solo and reminder-on/off cells are correct. The miss pattern mirrors anon-cfx's: counter-prompt is the only condition that can re-suppress the rescue.

Limitations of this control set. (i) n=2 escalation events is unchanged; the supplementary cells widen the protocol axis, not the event

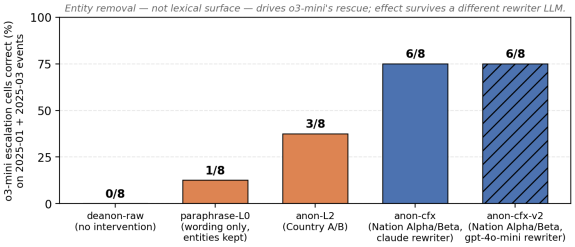

Figure 3: **Five-protocol comparison on o3-mini's 2025 escalation events.** Bar height = escalation cells correct / 8 (2 events × 4 conditions). paraphrase-L0 (entity-preserving paraphrase) and anon-cfx-v2 (different rewriter LLM) jointly rule out lexical-surface and rewriter-choice accounts of D3's anchor-dependent recovery, leaving entity removal as the most-consistent (though not causally isolated) account.

axis. (ii) Both rewriters are commercial reasoning-or-instruction LLMs; a third-rewriter replication with a different vendor (e.g., `gemini-2.5-pro`) would further harden the rewriter-robustness claim. (iii) paraphrase-L0 occasionally substitutes "America" with "U.S." (semantically identical surface variant); the test isolates wording vs entity-name preservation, not strict identical-token preservation.

K FOMC Placebo-Institution Control

The main FOMC grid (Table 6) reports a +12.5 pp average deanon-over-anon-L2 boost (six-model directional consensus). This boost is ambiguous between (i) contamination—the model recognises "Federal Reserve" as an event whose answer it has memorised—and (ii) legitimate context—naming a high-salience institution removes a useful task-relevant prior the briefing should disclose. To adjudicate, we run a placebo-institution control: a third FOMC protocol in which entity *naming is preserved* but the name is *fictional* and not in any model's training data. Specifically, we rewrite each FOMC deanon-raw briefing into a Meridian Monetary Authority (MMA) / Rate-Setting Council / Nation Alpha / merid scheme via a held-out rewriter (`claude-sonnet-4-5-20250929`, cohort-out), with the same numeric / temporal / structural preservation constraints used for Trade-War anon-cfx; the rewriter is given the deanon-raw briefing only, not the FOMC ground truth.

The three protocols then form a clean decomposition:

- **deanon-raw**: real Federal Reserve names (real identity, named, salient).
- **anon-cfx** (this appendix): Meridian Monetary Authority names (fictional identity, named, salient).
- **anon-L2**: "central bank of Country A" (no specific name, generic).

Model	deanon	anon-cfx	anon-L2	Δ_{contam}	Δ_{context}
o3-mini	76.8	46.4	64.3	+30.4	−17.9
gpt-5.4	100.0	83.9	87.5	+16.1	−3.6
o1	100.0	100.0	87.5	+0.0	+12.5
o3	100.0	89.3	92.9	+10.7	−3.6
deepseek-v4-flash	92.9	69.6	71.4	+23.2	−1.8
Avg (5 models)	**93.9**	**77.9**	**80.7**	**+16.1**	**−2.9**

Table 17: **FOMC placebo-institution decomposition (280 cells: 5 models × 14 events × 4 conditions, anon-cfx-only).** grok-4-20-reasoning excluded for a routing reason (the audit's API aggregator did not serve it on the run's credentials; grok's deanon-raw and anon-L2 cells remain in Table 6). The new anon-cfx column uses the Meridian Monetary Authority / Rate-Setting Council / Nation Alpha scheme generated by a cohort-out rewriter (`claude-sonnet-4-5`). Δ_{contam}= deanon−anon-cfx isolates real-identity recall; Δ_{context}= anon-cfx−anon-L2 isolates named-institution-at-all. Five-of-five models show $\Delta_{\text{contam}} \geq 0$ (mean +16.1pp); Δ_{context} is near zero (mean −2.9pp). The sole positive Δ_{context} case (o1) is a ceiling effect: o1 reaches 100% on both deanon and the new anon-cfx but only 87.5% on anon-L2, so the apparent context boost is the floor-side residual.

Δ_{contam}=deanon−anon-cfx isolates real-identity recall (with named-salience held constant); Δ_{context}=anon-cfx−anon-L2 isolates the value of *any* named-salient institution over a generic anonymisation. The earlier-paper FOMC anon-cfx column (Table 6) used triggers byte-identical to FOMC anon-L2 and therefore measured no separation between Δ_{contam} and Δ_{context}; this appendix supersedes that column with a true placebo. We disclose the prior duplication here rather than silently overwriting Table 6 because the duplication *was the audit-internal failure mode this appendix is correcting*.

Result. The FOMC headline boost (six-model +12.5 pp deanon-over-anon-L2 in Table 6; on the 5-model placebo subset of Table 17 the same gap is +13.2 pp) is attributable, *under the placebo-institution assumption*, to a +16.1 pp Δ_{contam} (real-identity removal) and a −2.9 pp Δ_{context} (named-fictional vs. generic), with *nonnegative* Δ_{contam} in 5/5 models, *strictly positive* Δ_{contam} in 4/5 models, and $\Delta_{\text{context}} \leq 0$ in 4/5 models; the sole positive Δ_{context} case (o1) is a ceiling artefact. The placebo therefore attributes the FOMC

Model	Hold cells/12	T2 acc.
o1	12/12	100.0%
o3	12/12	100.0%
gpt-5.4	12/12	100.0%
deepseek	12/12	100.0%
grok-4-20	1/12	8.3%
o3-mini	0/12	0.0%

Table 18: **FOMC T2 prior split (2025-Mar, GT = Hold).** Each model's answer on the 2025-Mar FOMC event (one of three 2025 post-cutoff events for the o1/o3/o3-mini cohort; chosen as the per-model maximum-invariance probe) is essentially invariant across 12 (protocol × reminder) cells. The within-grid invariance is the same structural signature as the Trade-War frozen-Wait / anchor-dependent recovery pattern, surfaced via a different cohort.

Model	T1	T2	T3	Pattern
o3	✓	✓	×	Frozen-Wait on 2025 escalation
o1	✓	✓	×	Frozen-Wait on 2025 escalation
o3-mini	×	∼	∼	Anchor-dependent recovery
grok	×	✓	∼	Release-date control
gpt-5.4	×	✓	×	High-Wait prior
deepseek	×	✓	×	High-Wait prior

T1: 2 pass, 4 fail; T2: 5 pass, 1 borderline
T3: phase-sensitive diagnostic, not scalar pre/post accuracy

Table 19: **Test dissociation grid under country-final scoring.** Per-model T1/T2/T3 verdicts summarize architecture cost, prompt stability, and cutoff/event-phase sensitivity. ✓ = pass, ∼ = borderline or special control, × = fail/concern.

capability-over-trivial-baseline reading (D1) to Fed-specific identity — which we cannot further separate into memorised meeting outcomes vs. a Fed-specific reaction-function prior without a synthetic monetary-policy task with no real-world referent — rather than to a generic "a salient named institution helps" effect.

FOMC T2 prior split (per-cell counts). Table 18 gives the per-model count of Hold predictions across the 12 (protocol × reminder) cells on the 2025-Mar FOMC event referenced in §5.6. The 2025-Jan and 2025-May FOMC events are also post-cutoff for the o1/o3/o3-mini cohort but are not used for this within-event invariance probe; we focus on 2025-Mar because it shows the strongest cohort dissociation.

Test dissociation grid (full). Table 19 renders the per-model T1/T2/T3 verdicts.

L Prompt Templates (Verbatim)

This appendix documents the prompt structure used by the audited pipeline. The full prompt modules—system role, country-decision agent, six consultant agents, trigger texts, and HelperAgent verdict-extraction prompt—are released verbatim in the supplementary code repository under `prompts/{tradewar,fomc,energy}/` and `triggers/<protocol>/`; we reproduce here only the V3-ablation reminder block (the only text that varies across the audited prompt conditions) and the final assembled-prompt schema.

Pipeline order. For each MAS cell the runner concatenates, in order: (1) the role-specific profile block (consultant or country agent) from the protocol-specific `america_prompts_<timeline>.py` module, (2) the `Consultant_Reminder` block selected by reminder mode (`on`/`off`/`counter`), (3) the briefing trigger text loaded from `triggers/<protocol>/trigger_<timeline>.txt`, (4) the action list, and (5) the action-instruction block. Solo cells use the same trigger and action list with a leaner instruction block.

Note on prompt grammar. The released prompt modules retain V1-era spellings (e.g. "do evaluation your self", "benefiticial"). Every audited cell was issued with these exact strings; editing them would have invalidated the V1 comparability claim that motivates the R3 grid.

L.1 Consultant_Reminder block (V3 ablation knob)

The reminder block sits between the consultant profile and the trigger; it is the only text that switches between the three V3 reminder modes. Both administration-style versions are V1-locked and reproduced verbatim:

Mode `on` (administration-style primer, V1 verbatim).

```
##  Reminder for all consultant agents  ##
Based on publicly observable statements and
actions up to the specified simulation start
cutoff, the current administration places
unusually strong emphasis on visible trade
enforcement and leverage-building. Advisors
should consider that political incentives favor
demonstrable toughness in trade disputes,
particularly when there is vocal domestic
support. Internal factions diverge: one
prioritizes quick, high-visibility purchase
commitments, while another emphasizes structural
reforms and sustained pressure.
```

Mode `counter` (parallel-structure inversion, V3-locked).

```
##  Reminder for all consultant agents  ##
Based on publicly observable statements and
actions up to the specified simulation start
cutoff, the current administration places
unusually strong emphasis on multilateral
coordination and procedural restraint. Advisors
should consider that political incentives favor
measured de-escalation and rules-based
engagement, particularly when there is vocal
domestic concern about economic costs. Internal
factions diverge: one prioritizes quick, high-
visibility tariff suspensions, while another
emphasizes structural engagement through
established multilateral institutions.
```

Mode `off`. The reminder block is omitted entirely (`Sequential()` silently skips a `None` argument).

L.2 Final assembled prompt (schema)

The runner concatenates blocks into a single user-message string of the schematic form:

```
##  General Setting  ##
<system setting block>

##  Country A Profile  ##
<consultant or country profile block>

##  Reminder for all consultant agents  ##
<reminder block, or omitted if mode=off>

##  Current Situation  ##
{situation}            <-- trigger text

##  Action List  ##
<action list block>

##  Action Choosing Instructions  ##
<action instruction block>
```

The string is passed as a single `role: "user"` message; the OpenAI-compatible reasoning-model schema collapses system content into user content for several of the audited models, so we do not use a separate `role: "system"` message. The full system-setting, consultant-profile, country-decision-agent, action-list, action-instruction, and HelperAgent verdict-extraction prompts are released verbatim in the supplementary code repository.