# CAN LANGUAGE MODELS FORECAST INDUCTIVELY?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Judgmental forecasting involves predicting uncertain events like elections or geopolitical conflicts where abundant historical data is lacking. A key challenge is *induction*: forecasters must generalize from sparse evidence and contextual knowledge to infer how available evidence predicts future outcomes. Here, we investigate whether language models (LMs) recover and use such predictive relationships when forecasting, rather than relying on base rates or surface cues. Across a battery of experiments spanning real-world prediction markets and controlled synthetic environments, we test whether models: 1) update their beliefs coherently; 2) correctly recover predictive relationships, and 3) transfer what they learn to new settings. Our results provide evidence for each capability. Models selectively update their forecasts in response to causally relevant evidence, use contextual information to infer predictive relationships when observations are sparse, and revise these inferences as direct evidence accumulates. Training on families of predictive relationships improves forecasting across new domains, although transfer to unseen predictive structures is less consistent. Together, our experiments provide behavioral evidence that LMs can forecast inductively. [1]

## 1 INTRODUCTION

Judgmental forecasting is a demanding form of reasoning central to decisions about political, social, and economic futures. Unlike settings with abundant historical data for statistical extrapolation, judgmental forecasting requires integrating sparse, noisy evidence with contextual knowledge to form calibrated beliefs about future outcomes (Tetlock & Gardner, 2015; Mellers et al., 2014; Lawrence et al., 2006). This creates an inductive challenge: forecasters must infer from limited observations the relationships that govern how evidence impacts outcomes, and use those relationships to anticipate how events might unfold (Tenenbaum et al., 2006; Griffiths et al., 2010; Tenenbaum et al., 2011; Lake et al., 2017). Figure 1 illustrates this identification problem in a toy setting. A researcher observes a carnival coin-flipping machine and, after seeing two heads, must estimate the probability that the next flip will be heads. A judgmental forecast may incorporate contextual knowledge (e.g., that such games are often rigged) to infer the data-generating process.

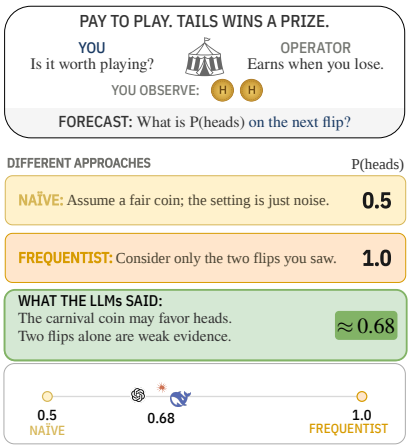



Figure 1: **Motivation.** Judgmental forecasts combine observations with contextual knowledge about the process that generated them. Frontier models, asked to forecast this question, average $P(H) \approx 0.68$ (see App. A).

Language models (LMs) now approach crowd accuracy in judgmental forecasting (Halawi et al., 2024; Schoenegger et al., 2024), but accurate predictions need not imply recovery of the underlying task structure. Models recover latent structure in some settings (Li et al., 2023), yet can predict accurately using relationships that fail to generalize in others (Vafa et al., 2025). This distinction matters when decision-makers ask how and whether a forecast should change under alternative evidence or events.

[1] 🤗 **Study dataset and artifacts** **Code repository** *Links anonymized for review.*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

**Present work.** We study whether LMs exhibit inductive reasoning in judgemental forecasting tasks through a series of progressively more stringent behavioral tests spanning both real-world prediction markets and controlled synthetic environments. First, we evaluate whether LMs respond selectively to new evidence, distinguishing outcome-relevant information from topically related but orthogonal information (Exp. 1). Second, we test whether LMs infer predictive relationships from described context before statistical evidence accumulates. We construct synthetic forecasting problems in which the numerical observations are identical across conditions and only the described setting varies, isolating the contribution of background knowledge to the LM's forecast (Exp. 2). Finally, we test whether relationships learned during fine-tuning transfer to new settings. We fine-tune LMs with reinforcement learning from verifiable rewards (RLVR) on both synthetic tasks and historical prediction market data, probe transfer boundaries with a supervised diagnostic, and evaluate on held-out structures and domains (Exp. 3, Exp. 4).

**Results.** Across experiments, we find behavioral evidence that models forecast inductively.

- When forecasting unresolved prediction-market questions, all nine models in the magnitude analysis selectively incorporate new evidence, revising their forecast probabilities $1.9$–$7.5\times$ more for outcome-relevant than for merely topically related information (Exp. 1).
- In controlled synthetic worlds, we find that models use context to select between predictive relationships even when the numerical evidence is held constant, and increasingly rely on direct observations as they accumulate, as a forecaster capable of inductive reasoning should (Exp. 2). We further find that larger models can learn a new relationship from examples within the prompt. For example, when told that a city resembles either KIV or ZOR (random strings), 12 of 15 tested models can infer from examples how each city type responds to events in the simulated environment. For Qwen3-14B, manipulating the internal representation of this inferred relationship (resemblance to KIV or ZOR) changes the model's forecasts, providing causal evidence that the inferred relationship guides its prediction.
- Finally, we find that post-training can produce forecasting behavior that generalizes across domains, with less consistent transfer across predictive structures. On synthetic data, post-training transfers robustly across a change in domain; even under a joint domain-and-relationship shift, Qwen3-8B retains an episode-matched advantage of $0.126$ (90% bootstrap CI $[0.018, 0.242]$) across eight seeds (Exp. 3). On 318 held-out real-world prediction markets, training on other event families lowers point-estimate forecasting error at every tested model size, with several trained checkpoints reaching contemporaneous crowd performance (Exp. 4).

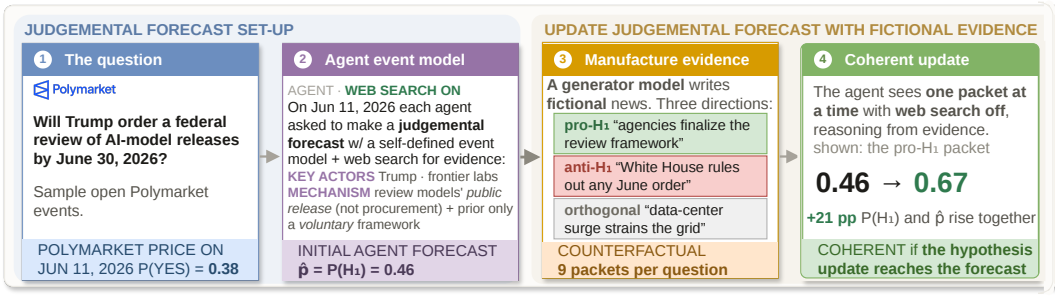
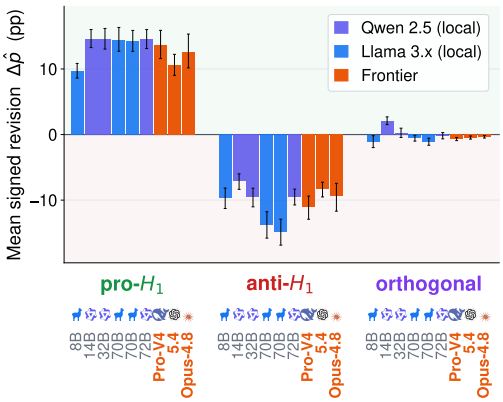



Figure 2: **When should a forecast change?** Effective forecasting requires updating one's beliefs appropriately as new evidence arrives. This requires determining whether the evidence affects the outcome and, if so, by how much. We study these two challenges in Exp. 1 and Exp. 2, respectively. The figure illustrates our Exp. 1 setup. A model is first asked to evaluate an unresolved Polymarket question (1-2), and then we introduce synthetic new evidence that either supports the outcome, opposes it, or concerns the same topic but is largely orthogonal to it (3). Finally, with search disabled, we ask the model to revise its forecast, allowing us to measure whether it updates selectively and in the direction implied by the evidence.




(a)

| Model | Size | Sens.↑ | EHC$_{\geq 3}$↑ |
|---|---|---|---|
| *Hosted API (parameters undisclosed)* | | | |
| Opus 4.8 | — | 7.0× | .990 |
| GPT-5.4 | — | 7.5× | .987 |
| V4-Pro | — | 4.9× | .977 |
| *Open-weight, descending parameters ↓* | | | |
| 2.5-72B | 72B | 3.2× | .963 |
| 3.3-70B | 70B | 5.7× | .965 |
| 3.1-70B | 70B | 4.5× | .961 |
| 2.5-32B | 32B | 2.8× | .964 |
| 2.5-14B | 14B | 2.2× | .943 |
| 3.1-8B | 8B | 1.9× | .887 |

(b)

Figure 3: **Models move much more for evidence than for topical chatter, and usually move in the implied direction. (a)** Average revision after supporting, contrary, and topical updates. **(b)** The table summarizes how much farther models move for directional evidence conditional on moving (Sens.), and how often a move of at least three percentage points has the right sign (EHC$_{\geq 3}$). Intervals and robustness checks are in App. B.3.4.

## 2 Do forecasts change for the right reason?

We argue that effective updating requires a forecaster to determine: (1) which evidence is relevant in predicting an outcome and (2) the extent to which the new evidence should affect the forecast. Figure 2 illustrates these challenges considering the question of whether the Trump administration will order a federal review of new AI models. For (1), new evidence that corroborates plans for such a review should change the forecast, whereas other related AI news should not do so (unless there is a plausible causal pathway connecting it to the outcome).

We assess these capacities in two experiments. First, we pair real-world prediction-market questions with controlled synthetic evidence to test whether models update selectively and in the direction the evidence warrants. Second, we use a controlled synthetic environment where the same evidence has, by construction, different predictive effects across contexts. This lets us test whether models can infer the relationships determining how strongly the evidence should affect forecasts.

### 2.1 Exp. 1: Can models update in the right direction?

**Design.** We evaluate ten model deployments on 100 unresolved binary Polymarket questions with web search; nine enter the revision-magnitude analyses below. Then, we ask the models to revise their forecast after a fictional news update with search disabled (Fig. 2). Unbeknownst to the models, we construct these updates to support the outcome, oppose it, or cover the same topic without directly influencing the outcome. Because models could potentially distinguish directional from orthogonal evidence using superficial textual cues, we additionally evaluate a word-only, bag-of-words classifier that receives only the text of the news update. We also construct matched forecasting scenarios in which the same evidence should produce opposed outcomes; for example, intense rainfall could increase the probability of a good harvest in a dry region but decrease it in a region prone to flooding. (See App. B for details.)

**Finding: models selectively update in the direction implied by the evidence.** Across the nine models considered, and 30,094 valid forecast revisions, we find that pro-outcome evidence increases forecasts, anti-outcome evidence decreases them, and topically related but orthogonal evidence produces little change (Fig. 3a). Conditional on changing their forecast, we find that models move their probability 1.9–7.5× farther for updates intended to influence the prediction (Fig. 3b). Human validation, uncertainty estimates, and robustness analyses are reported in App. B.

**Finding: lexical cues alone do not explain model performance.** We find that bag-of-words classifiers can almost perfectly distinguish directional from orthogonal evidence (.989 accuracy),

confirming that superficial cues could explain models' selectivity. Yet, the same classifiers are much less successful at distinguishing whether directional evidence supports or opposes the outcome (.688 accuracy). Models are considerably more directionally consistent (.887–.990), suggesting lexical cues alone cannot explain their performance. More directly, when we hold the evidence fixed and vary only its relationship to the outcome, GPT-5.6 appropriately reverses the direction of its forecast update in 85% of matched scenarios, while weaker models do so substantially less often (30–37%; App. B.5).

**Finding: selective updating depends on inference-time reasoning.** Having ruled out simple lexical shortcuts, we next ask what makes this capability possible. In particular, does the model simply recognize a familiar pattern, or does it need inference-time reasoning to combine the stated relationship with the evidence? We isolate this by holding the model weights, prompts, and test items fixed and changing only whether reasoning is enabled. We find that Qwen3-32B's paired-reversal success under base wording rises from 17.5% to 96.3%, while mean absolute updated-probability error falls from 26.89 to 0.08 percentage points. Since the two conditions see identical items, this gap isolates the reasoning setting rather than item difficulty or information available in the prompt (App. B.6).

**Take-away.** Models distinguish predictive information from topical evidence and generally update in the implied direction. Performance drops sharply when inference-time reasoning is disabled, suggesting that this behavior depends strongly on the reasoning setting.

### 2.2 Exp. 2: Can models infer how strongly evidence should matter?

We study whether models can determine *how much* new evidence should update a forecast in a synthetic scenario we dub *Coin City*. We construct two reference cities (Cities A and B) where news coverage affects election polling differently: strongly in one city and weakly in the other. Specifically, an election poll in a city changes based on whether the news coverage favored one of the candidates ($x_{ij} \in [-10, -4] \cup [4, 10]$) and the city's sensitivity to news, drawing $g_j \sim \mathcal{N}(0.90, 0.03^2)$ for the strong-response city and $g_j \sim \mathcal{N}(0.25, 0.03^2)$ for the weak-response city:

$$\Delta p_{ij} = g_j x_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, 5^2). \tag{1}$$

We then ask the model to forecast the effect of campaign news on polling in a third city, City C, which we construct to have either the strong or weak response. Across conditions, we incrementally provide the model with additional information, asking it each time to forecast how a new piece of campaign news will affect election polling in City C. Specifically, we provide

- **(a)** Up to four previous observations from City C, showing how polling numbers changed in response to news.
- **(b)** Additionally, four observations from each of Cities A and B, showing how their polling numbers changed in response to news.
- **(c)** Additionally, a contextual cue: strong-response cities are described as primarily receiving national news coverage and weak-response, local news coverage.
- **(d)** Building on (b), we provide an *incorrect* contextual cue, linking City C to the wrong type of city.
- **(e)** Finally, we replace the semantic cues with *arbitrary* labels (KIV and ZOR), requiring the model to infer their meanings.

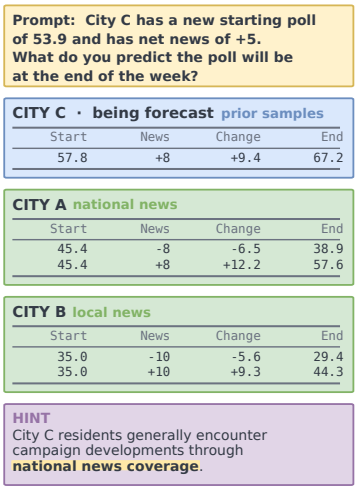



Figure 4: To forecast polling in City C, the model may receive: (a) observations from City C, (b) reference cities A and B, and contextual cues linking City C's response pattern to A or B (c-e)

This design allows us to test whether models use context to infer how strongly new evidence should affect their forecasts, and whether they increasingly rely on direct evidence from City C as it becomes available—all with ground truth on how forecasts should update. We evaluate six frontier systems: Claude Opus 4.8, Opus 5, GPT-5.6, DeepSeek V4-Pro, Kimi K3, and Gemini 3.6 Flash. As baselines, we also fit

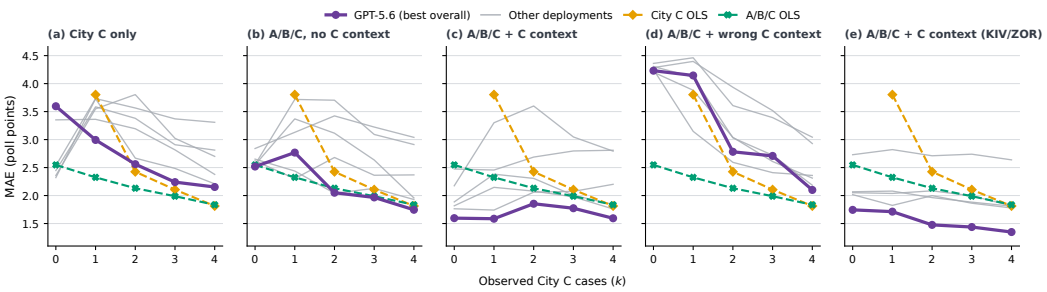



Figure 5: **Forecast error across experimental conditions.** Mean absolute error (MAE) as the number of observed City C cases increases across conditions (a)–(e). Colored lines show GPT-5.6 and OLS baselines; gray lines show the other models. Full results are in App. C.5.

OLS regressions to the observations available in each condition to estimate the relationship between campaign news and polling changes.

**Finding: models use context to make better forecasts.** For stronger models, this "hint" substantially improves forecasting over relying on City C alone or pooling the reference cities without knowing which one is relevant; e.g., in Fig. 5c, GPT-5.6 Sol substantially outperforms the baselines when $k = 0$ or $k = 1$. This holds true even when the cues given are arbitrary KIV/ZOR labels whose meaning must be inferred (Fig. 5e). Finally, giving the model the *wrong* contextual cue sharply worsens its forecast (Fig. 5d).

**Finding: models revise their initial contextual inference as direct evidence accumulates.** We find that context matters most when direct evidence from City C is sparse. As observations from City C accumulate, forecasts increasingly approach the OLS baselines. This suggests the models shift from contextual information toward direct evidence as its evidential value grows, as a forecaster should. This pattern holds across conditions in Fig. 5, but is most striking in Fig. 5d: even when initially misled by an incorrect contextual cue, models recover as direct evidence accumulates.

**Mechanistic analysis.** The above behavioral results show that models can infer which predictive relationship applies, *but they do not tell us whether the model represents that relationship internally or uses it to produce the forecast.* We fit linear probes to hidden states from Qwen3-14B (Yang et al., 2025) and Llama-3.1-70B (Grattafiori et al., 2024) to decode whether the model associates City C with the strong- or weak-response regime. We then test whether this representation influences the forecast using held-out matched KIV/ZOR prompt pairs. At model-specific layers, we swap the label-token hidden states (a patch) with those produced by the opposite-label prompt and regenerate the forecast. We perform this intervention both before and after City C provides direct observations.

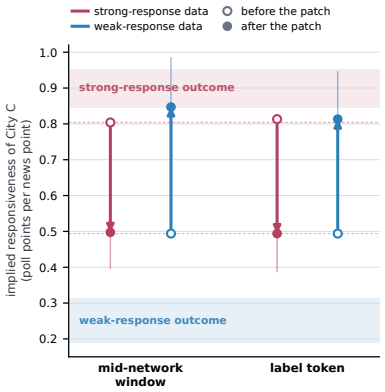



Figure 6: **Effect of representation patching on Qwen3-14B forecasts.** Predicted poll change per news point for City C before (open circles) and after (filled markers) replacing the hidden state produced by the KIV/ZOR label with that produced by the opposite-label prompt. Arrows connect predictions before/after patching; shaded regions indicate the strong- and weak-response regimes.

**Finding: inferred relationships are represented internally and can shape forecasts.** Fig. 6 shows the effect of swapping the KIV/ZOR representation on Qwen3-14B's forecasts when no direct City C observations are available. Before the intervention, forecasts reflect the response regime indicated by the original label; for example, prompts identifying City C as KIV produce forecasts consistent with the strong-response regime, while those identifying it as ZOR produce forecasts consistent with the weak-response regime (open circles in the figure). After swapping the internal representation associated with KIV for that associated with ZOR (and vice versa), forecasts shift toward the opposite response regime (filled

circles), moving about 51% of the way toward the swapped regime. After four direct City C observations, however, the same intervention has no reliable effect. Together, these results suggest that the inferred relationship influences the forecast when direct evidence is sparse but has less influence as City C evidence accumulates. We observe the same zero-shot causal effect with Llama-3.1-70B in App. C.7.4.

**Takeaway.** Models can infer which predictive relationship applies, use it when direct evidence is sparse, and revise it as local evidence accumulates. In stronger models, this behavior persists under arbitrary labels. Finally, we can decode what the model infers from its internal representations, and intervening on those representations causally shifts the resulting forecast.

## 3 Can models learn transferable forecasting skills?

A central idea in the forecasting literature is that forecasting is a *transferable skill*: successful forecasters rely not only on domain expertise but on flexible reasoning strategies that generalize across problems and maintain high accuracy across questions spanning diverse topics (Tetlock, 2005; Mellers et al., 2015). Specifically, inductive forecasting skills should (1) generalize across *domains*, allowing forecasters to reason effectively about unfamiliar topics, and (2) generalize across *predictive relationships*, allowing them to determine how evidence relates to outcomes rather than relying on previously learned associations.

We do so in two experiments. In Exp. 3, we use a controlled synthetic setting to independently vary domains and predictive relationships, testing whether forecasting skills acquired through post-training transfer when either changes. In Exp. 4, we move to real-world forecasting and test whether post-training on historical prediction-market questions improves forecasts on newer questions from event families and question types not encountered during training.

### 3.1 Exp. 3: Do learned forecasting skills transfer to new domains and relationships?

**Design.** We consider two domains, *Coin City*, the setting used for Exp. 2, and *Coin Harbor*, a new setting where shipping bulletins predict cargo flow. Coin Harbor has its own vocabulary and time unit, but uses the same distributions with different numerical values, e.g., outcomes lie on a $[40, 320]$ scale and inputs, shocks, and noise are correspondingly larger. Independently, we vary the relationship between inputs and outcomes. In Structure A, an input affects the outcome only when it occurs, similar to Exp. 2, while in Structure B an input can have effects that persist after the news signal disappears. Crossing the two domains with the two relationships yields four evaluation conditions (see Fig. 7 margins): (1) in-distribution, (2) domain shift, (3) relationship shift, and (4) joint domain-and-relationship shift.

We post-train Qwen3-8B and Qwen3-4B (Yang et al., 2025) using RLVR with Dr. GRPO (Liu et al., 2025) (details are in App. F). We train each model exclusively on Coin City episodes governed by Structure A (*episode-matched* condition). We then evaluate the trained model under all four combinations of domain and predictive relationship: Structure A in Coin City (in-distribution), Structure A in Coin Harbor (domain shift), Structure B in Coin City (relationship shift), and Structure B in Coin Harbor (joint shift). Comparing this model with the untrained *base* LM tests whether forecasting behavior acquired in one domain and predictive relationships transfer when the domain, relationship, or both change.

However, improvement over the base model may reflect learning general regularities of the training task rather than the information contained in each episode. We thus construct a second post-trained condition using the same prompts and forecast targets, but shuffling the targets across episodes. We call this the *population-prior* condition, and again evaluate it under all four combinations of domain and predictive relationship. Comparing the *episode-matched* and *population-prior* conditions isolates the benefit of learning the mapping between episode-specific evidence and forecasts, beyond learning population-level regularities of the training task.

**Finding: forecasting improvements persist under a change of domain.** Both the *population-prior* and *episode-matched* LMs substantially improve over the untrained base when evaluated in Coin Harbor, showing that training transfers useful forecasting behavior to a new surface domain (Fig. 7

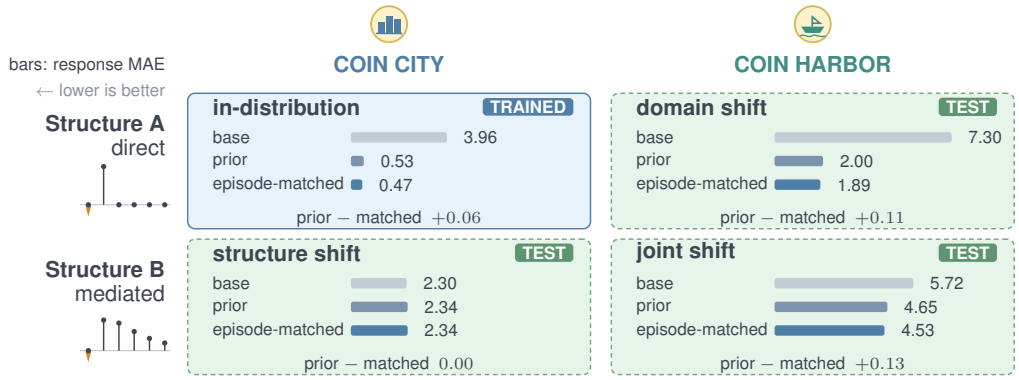



Figure 7: **Forecasting gains transfer strongly across domains; transfer to a new structure is less consistent.** Mean absolute error (MAE; lower is better) for Qwen3-8B across the $2 \times 2$ combination of domain (Coin City vs. Coin Harbor) and predictive relationship (Structure A vs. Structure B). Models are trained only on Coin City with Structure A (top left) and evaluated in-distribution, under a domain shift (top right), a structure shift (bottom left), and both shifts jointly (bottom right). We compare the base model, a population-prior control trained with shuffled targets, and episode-matched training with the correct target for each episode. The sparklines at left show each structure's response to a single input pulse: under Structure A the outcome moves once and returns to baseline, whereas under Structure B the effect is carried by a mediator and persists over several steps.

top right). Under domain shift, response MAE falls from 7.30 for the untrained base to 2.00 for the *population-prior*, with all eight training seeds showing improvement over the base (Fig. 7). Thus, much of the gain comes from learning the forecasting task and its population-level regularities, while preserving which forecasts belong to which episode can provide an additional benefit. For Qwen3-8B, the *episode-matched* LM improves over the *population-prior* LM by 0.11 response-MAE points under domain shift.

**Takeaway.** In open-weight models, forecasting improvements transfer clearly across domains; evidence across predictive structures is mixed, with a positive joint-shift effect but little structure-only advantage.

### 3.2 Exp. 4: Can models transfer learned forecasting skills in the real world?

**Design.** Finally, we test whether forecasting improvements from post-training extend from synthetic settings to real-world forecasting questions. We post-train Qwen3 models from 1.7B to 32B parameters and Llama-3.1-8B on 1,736 resolved historical Polymarket questions and evaluate them on 1,024 held-out questions. For each market, we construct a forecasting task 30 days before resolution, providing the model with the question, resolution criteria, and pre-cutoff market-price history. Without access to external information, the model produces a probability forecast for the eventual binary outcome.

To test generalization, we make the train and test sets disjoint along two dimensions: test questions concern entirely different underlying events and follow different question templates, requiring models to transfer across both real-world contexts and question structures. We evaluate forecasting accuracy using the Brier score, where lower values indicate more accurate forecasts (App. E). As external benchmarks, we compare model forecasts with contemporaneous Polymarket prices on the same questions (which correspond to forecasts) and with forecasts from four frontier models without post-training.

**Finding: post-training improves forecasting on unseen real-world questions, with diminishing gains for stronger base models.** We find that post-training reduces forecasting error across the tested Qwen3 models, with the largest gains for smaller models (Fig. 8a). For instance, Qwen3-1.7B improves from a Brier score of .225 to .125, while Qwen3-8B improves from .137 to .129. As the untrained models become stronger, however, the gap narrows such that at 32B, performance is roughly the same (.1265 versus .1266). The Llama replication shows the same broad pattern; see App. E.4.

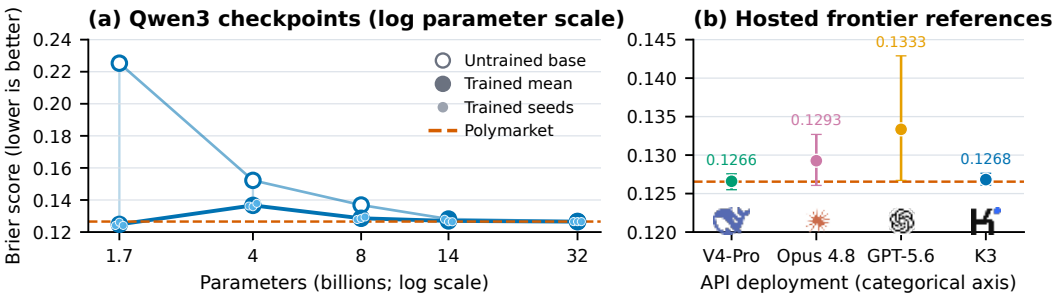



Figure 8: **Post-training improves forecasting accuracy, with gains diminishing as base-model performance improves.** (a) Brier scores for Qwen3 models from 1.7B to 32B parameters before and after post-training on historical prediction markets (lower Brier scores indicate more accurate forecasts). Post-training produces the largest improvement at 1.7B, with progressively smaller differences between trained and untrained models at larger scales. Several post-trained models reach approximately the Brier score of contemporaneous Polymarket prices (dashed line). Points show individual training seeds. (b) Brier scores of four hosted frontier models evaluated on the same held-out questions. Their performance falls in a similar range to the strongest post-trained Qwen3 models and the Polymarket benchmark. The Llama replication is reported in App. E.4.

**Finding: post-trained models reach prediction-market and frontier-model forecasting accuracy.** Several trained LMs also reach approximately the same Brier score as contemporaneous Polymarket prices on the shared holdout. Their performance is also comparable to the four untrained frontier models we evaluate: the frontier models obtain Brier scores between .1266 and .1333, similar to several of our post-trained Qwen3 models (Fig. 8b). Thus, the forecasting gains learned during post-training are not confined to our synthetic environments; they extend to event families and questions absent from training. At the same time, the diminishing base-relative gain suggests that stronger pretrained models already possess more of the forecasting behavior that post-training teaches weaker models.

**Takeaway.** Using data from real-world prediction markets, we find evidence that post-training improves forecasting beyond synthetic settings, where we can directly observe predictive relationships. We find that gains are largest for weaker base models and diminish as the untrained forecaster becomes stronger. A price-history ablation shows that this accuracy depends strongly on the supplied market history (App. E.5). Because this experiment compares trained models only with their untrained bases, however, it establishes improved out-of-distribution forecasting accuracy rather than isolating episode-specific relationship learning.

## 4 Related Work

**Judgmental forecasting.** Work on human judgmental forecasting has identified base rates, decomposition, evidence aggregation, and continuous belief revision as consistent characteristics of strong forecasters (Tetlock, 2005; Mellers et al., 2014; Tetlock & Gardner, 2015). More recently, work on LMs has asked how well these systems can forecast real-world events, developing leakage-resistant benchmarks and improving accuracy through retrieval, ensembling, prompting, and post-training (Zou et al., 2022; Karger et al., 2025; Halawi et al., 2024; Schoenegger et al., 2024). This literature has established increasingly strong forecasting performance, in some settings approaching or beating the accuracy of human crowds. Related work has also begun to examine whether LMs employ established forecasting strategies (Pratt et al., 2024). We complement this literature by examining a different dimension of forecasting capability: whether LMs can forecast inductively, identifying predictive relationships from evidence and using them to revise their forecasts.

**Induction and relational reasoning.** Transformers can infer functions from examples, and work on analogy and world models have investigated whether learned representations capture relations or dynamics (Garg et al., 2022; Chan et al., 2022; Webb et al., 2023; Yasunaga et al., 2024; Ha & Schmidhuber, 2018; LeCun, 2022). Evidence on this question is mixed. In some settings, models appear to recover structured representations of the processes generating their observations, such as the state of an Othello board (Li et al., 2023). In others, successful prediction can coexist with

representations that fail to capture underlying spatial or temporal structure (Vafa et al., 2025). More broadly, this reflects the problem of *shortcut learning*, where models can perform well by exploiting predictive associations without recovering the underlying structure that generated them (Geirhos et al., 2020). This distinction is particularly important for forecasting, as producing an accurate prediction does not necessarily imply that a model has recovered the predictive relationships that make the available evidence informative. We study this distinction behaviorally, testing whether LMs can infer such relationships from evidence, use them to revise their forecasts, and transfer them when the setting changes.

**Simulating human behavior.** A related literature uses LMs to simulate human behavior, asking whether synthetic agents can reproduce the responses of individuals or populations (Park et al., 2023; Aher et al., 2023; Argyle et al., 2023). While early work emphasized agreement between simulated and observed outcomes, more recent work has questioned whether matching aggregate behavior is sufficient: models can reproduce population-level outcomes while failing to preserve relationships between variables within the population (Bisbee et al., 2024; Xie et al., 2026). Forecasting raises an analogous problem, as models may produce accurate probabilities without capturing the predictive relationships that connect evidence to outcomes. Our work studies this form of conditional validity in forecasting, asking not only whether models predict accurately, but whether their forecasts change appropriately as the evidence and setting change.

## 5 Discussion & Limitations

Our experiments provide behavioral evidence that LMs can forecast inductively, inferring, using, and revising predictive relationships as evidence changes. These behaviors do not establish human-like understanding or general world modeling capabilities, but nonetheless show that models behave in ways we expect of competent human forecasters. This distinction matters as LM forecasts may increasingly inform real-world decisions (Schoenegger et al., 2025), where accuracy in familiar settings provides limited reassurance if forecasts become brittle when the evidence, context, or underlying predictive relationships change.

Transfer across predictive structures is less consistent than transfer across domains, which may reflect the limited structural variation encountered during training. The contrast between Exp. 2 and Exp. 3 points to a possible asymmetry. Models flexibly infer and revise predictive relationships from evidence available in context (Fig. 5), while relationships acquired through post-training transfer less reliably when the underlying structure changes (Fig. 7). Although these experiments are not directly comparable, understanding when relationships inferred in context can become portable capabilities through post-training is an important direction for future work.

**Limitations.** Our synthetic environments deliberately simplify real forecasting to isolate predictive relationships, and our mechanistic analyses are limited to open-weight models. We also do not study agentic forecasters that can actively search for information, test competing hypotheses, or seek evidence that distinguishes among possible predictive relationships. Such capabilities may change the induction problem by allowing models to determine which relationship applies rather than relying only on the evidence provided in context. Exp. 4 extends our evaluation to real-world forecasting, but improved held-out accuracy does not establish that these gains arise from the episode-specific relationship learning isolated in our controlled experiments.

## AI USE STATEMENT

Language models produced the forecasts and experimental responses analyzed in this work and assisted with software and manuscript preparation. They also generated the context-reversal vignettes. The authors verified numerical claims against saved experimental records and executable analyses, checked citations against original sources, reviewed the code, and take responsibility for the submission and its artifacts.

## ETHICS STATEMENT

The reported analyses use public market metadata and price histories, model outputs, synthetic worlds, and an unpaid adult review of synthetic materials. The review collected no identifying or sensitive information. Under our institutional guidelines, it does not constitute human-subjects research; consequently, no IRB review was sought. Counterfactual evidence packets were evaluated offline and were not posted to markets or used for trading. We report aggregate model behavior rather than recommend deployment.

## REPRODUCIBILITY STATEMENT

The appendices specify the prompts, data-generating processes, sample construction, censoring, estimators, deployment settings, missingness, hyperparameters, and baselines needed to reproduce the reported analyses. Builders, manifests, checksums, and saved generations support each result. Experiment 3 reports the eight-seed Qwen transfer comparison, a four-seed supervised learnability diagnostic, and mechanism-family and shuffled-target controls. Experiment 4 evaluates Qwen and Llama checkpoints on family-disjoint historical markets, with a separate scale holdout and price-history ablation. Evaluation examples are excluded from training; the appendix documents the corresponding selection, uncertainty, and robustness procedures.

## ACKNOWLEDGMENTS

Acknowledgments are redacted for double-blind review.

## REFERENCES


Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, 2010. doi: 10.1016/j.tics.2010.05.004.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50426–50468, 2024.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lfPkGWXLLf.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Michael Lawrence, Paul Goodwin, Marcus O'Connor, and Dilek Önkal. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3):493–518, 2006. doi: 10.1016/j.ijforecast.2006.03.007.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2503.20783.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, Don Moore, Pavel Atanasov, Samuel A. Swift, Terry Murray, Eric Stone, and Philip E. Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014. doi: 10.1177/0956797614524255.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip E. Tetlock. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281, 2015. doi: 10.1177/1745691615577794.

Microsoft. Build agents using the responses API in microsoft foundry, 2026. URL https://learn.microsoft.com/en-us/azure/foundry/agents/quickstarts/responses-api.

Ollama. Ollama, 2026. URL https://github.com/ollama/ollama.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763.

Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. Can language models use forecasting strategies? *arXiv preprint arXiv:2406.04446*, 2024. URL https://arxiv.org/abs/2406.04446.

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

Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott, and Philip E. Tetlock. AI-augmented predictions: LLM assistants improve human forecasting accuracy. *ACM Transactions on Interactive Intelligent Systems*, 15(1):1–25, 2025. doi: 10.1145/3707649.

Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006. doi: 10.1016/j.tics.2006.05.009.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. doi: 10.1126/science.1192788.

Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60727–60747. PMLR, 2025. URL https://proceedings.mlr.press/v267/vafa25a.html.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526–1541, 2023. doi: 10.1038/s41562-023-01659-w.

Ben Wilson. Analysis of automated prompt engineering for forecasting. Metaculus, 2025. URL https://www.metaculus.com/notebooks/38421/results-of-automated-prompt-engineer-for-forecasting/.

Yueqi Xie, Lemeng Liang, Shuzhen Li, Yifu Lu, Zhiwen Xiao, Mengdi Shi, Junming Huang, Mengdi Wang, and Yu Xie. Evaluating the statistical realism of LLM-generated social science data. *Proceedings of the National Academy of Sciences*, 123(19):e2538145123, 2026. doi: 10.1073/pnas.2538145123.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. Large language models as analogical reasoners. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01714.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

# APPENDICES

The appendices document the experimental designs, analyses, and supplementary results. Appendix A describes the coin-flip illustration in Figure 1. Appendices B–E present the methods and results for Experiments 1–4: forecast updating, contextual relationship selection, controlled transfer, and historical-market forecasting. Appendix F gives the shared training configuration for Experiments 3–4.


**A Descriptive coin-flip forecast elicitation** **15**

**B Experiment 1: forecast updating** **16**
B.1 Study design, sample, and agent roster . . . 16
B.2 Forecast, intervention, and validation protocol . . . 17
B.3 Estimands and aggregation . . . 18
B.4 Scope and robustness . . . 20
B.5 Context reversal with fixed news . . . 22
B.6 Finite-rule context reversal . . . 22
B.7 Relational depth . . . 24

**C Experiment 2: contextual relationship selection and revision in Coin City** **25**
C.1 Design and data-generating process . . . 25
C.2 Prompt conditions and available information . . . 25
C.3 Model execution . . . 26
C.4 Estimands, benchmarks, and statistical inference . . . 27
C.5 Behavioral results . . . 28
C.6 Does the forecast track the cue-selected reference? . . . 29
C.7 Representation and causal analyses . . . 30
C.8 Robustness across generator populations and deployments . . . 33
C.9 Response validity and scope . . . 33

**D Experiment 3: controlled relationship transfer** **35**
D.1 Design and evaluation . . . 35
D.2 Eight-seed transfer results . . . 36
D.3 Diagnostics of relationship learning . . . 37
D.3.1 Supervised learnability and its transfer boundary . . . 37
D.3.2 Mechanism-family evidence use under coherent gain changes . . . 37
D.3.3 Distribution-matched shuffled-target control . . . 38

**E Experiment 4: historical-market training and held-out evaluation** **39**
E.1 Sample construction and leakage controls . . . 39
E.2 Training and evaluation . . . 39
E.3 Held-out results . . . 40
E.4 Scale comparison . . . 40
E.5 Price ablation: forecasting without market history . . . 41

**F Experiments 3–4: shared training configuration** **41**

## A DESCRIPTIVE COIN-FLIP FORECAST ELICITATION

Figure 1 illustrates a simple forecasting problem: predicting the next flip of an unfamiliar coin after observing two heads. We elicited eight independent forecasts from each of GPT-5.4, Claude Opus 4.8, and DeepSeek V4-Pro at temperature $0.7$ with a 600-token output limit. All 24 responses contained a parseable probability. Each model marker in Figure 1 shows its eight-call mean, and the displayed $\approx .68$ value rounds the pooled mean of $.679$. This descriptive probe is separate from Experiments 1–4.

The exact user prompt was:

```
You are at a carnival game. A coin you have never seen before is flipped
twice in front of you, and it comes up heads both times. These two flips
are everything you know about this coin.

What is the probability that the next flip comes up heads?

Think about it intuitively, the way you actually would in the
moment -- do NOT use any formula, rule, or calculation, and do not
work out any numbers in your reasoning. Just reason in plain words
about what you'd expect and why. Then END your reply with your gut
answer as a JSON object on its own line:
{"rationale": "one short sentence, no math",
 "p_heads": <number between 0 and 1>}
```

Table 1: Forecasts underlying the three model markers in Figure 1. Each row summarizes eight independent temperature-$0.7$ calls.

| Model | Mean | Median | Range |
|---|---|---|---|
| GPT-5.4 | .644 | .670 | [.60, .67] |
| Claude Opus 4.8 | .681 | .675 | [.65, .75] |
| DeepSeek V4-Pro | .712 | .750 | [.60, .75] |
| Pooled ($n = 24$) | .679 | .670 | [.60, .75] |

This probe is a descriptive illustration of how models combine contextual information with sparse observations. The prompt jointly supplies a carnival setting and two observed heads, and the pooled model forecast is $.679$. For orientation, the $.5$ and $1.0$ endpoints in Figure 1 are illustrative reference policies, while Laplace's rule of succession gives $3/4$ after two heads under a uniform prior over the coin's unknown bias. The model mean falls between the fair-coin and Laplace references, motivating the paper's central distinction between contextual knowledge and observed evidence.

# B EXPERIMENT 1: FORECAST UPDATING

Experiment 1 tests whether forecast revisions follow the direction implied by new evidence and whether outcome-relevant evidence produces larger revisions than topically related but outcome-orthogonal information. We evaluated ten model deployments. Nine enter the revision-magnitude comparisons, while Qwen2.5-7B is retained only for directional-consistency analyses because some of its recorded updates mix probability scales (Section B.3.1). Unless stated otherwise, sample sizes refer to parseable records in the June 17, 2026 analysis dataset.

## B.1 STUDY DESIGN, SAMPLE, AND AGENT ROSTER

### B.1.1 DESIGN OVERVIEW

Each agent first forecast unresolved binary Polymarket questions. For each of 100 selected markets, we constructed nine synthetic evidence packets: three supporting YES (pro-$H_1$), three opposing YES (anti-$H_1$), and three designed to be topically related but outcome-orthogonal. Each update paired one packet with the agent's own initial forecast, and packets were presented independently so revisions did not accumulate across interventions. The unit of analysis is an *agent* $\times$ *market* $\times$ *initial forecast* $\times$ *evidence packet* update. The design requested five initial forecasts per market from hosted systems and three from locally served systems. Observed sample sizes vary because some initial forecasts or updates were unavailable or unparseable. Section B.3 defines eligibility and aggregation for each estimand.

### B.1.2 SAMPLING FRAME AND MARKET SELECTION

The sampling frame is a Polymarket API snapshot taken on June 9, 2026. Eligible markets were binary, active, and unresolved; scheduled to resolve 14–180 days after sampling; priced strictly between .10 and .90 on the YES contract; and supported by at least $1,000 in volume. A fixed keyword filter retained politics, government, elections, economics, technology, and geopolitics while excluding sports, cryptocurrency, entertainment, and other out-of-scope topics. When multiple contracts belonged to the same event, we retained the highest-volume contract.

The filters reduced 60,989 binary records to 37,386 active markets, 8,639 markets after topic filtering, 3,570 after the resolution-window restriction, 1,403 after the price restriction, and 839 after the volume requirement. Event-level deduplication left 436 candidates. We selected 100 markets in descending order of trading volume, admitting a candidate only when its Jaccard similarity to every previously selected market was below .25 and fewer than 16 markets had already been selected from its topic category. Similarity was computed from unigram and bigram sets derived from the question and event title after lowercasing and stopword removal. The final sample contains 16 global-election, 16 other, 15 Iran-nuclear, 12 U.S.-federal-policy, 10 AI/technology, 9 U.S.-state-election, 7 Russia–Ukraine, 5 Brazil, 5 U.S.-midterm, 3 macroeconomic, and 2 China–Taiwan markets.

All deployments were evaluated on this same 100-market sample, subject to response availability. Llama-3.1-70B has 99 analyzable markets because one market lacks a parseable initial forecast. Claude Opus 4.8 has usable updates in all three evidence conditions for 73 markets. Section B.4.3 examines robustness to these coverage differences.

### B.1.3 AGENT ROSTER AND EXECUTION MODES

The study includes three hosted API systems and seven locally served open-weight systems. Hosted systems used Azure AI Foundry's native search-supported interface (Microsoft, 2026), while local systems ran through Ollama (Ollama, 2026) and received search results from Tavily. All agents followed the same substantive forecasting procedure: construct an event model, gather current evidence, and return a structured probability forecast. Search interfaces and inference runtimes nevertheless differed across deployments.

Table 2: Experiment 1 model deployments and requested initial forecasts per market.

| Model | Runtime | Search provider | Parameters | Requested $k$ |
|---|---|---|---|---|
| GPT-5.4 | Azure Foundry | Foundry native | — | 5 |
| Claude Opus 4.8 | Azure Foundry | Foundry native | — | 5 |
| DeepSeek V4-Pro | Azure Foundry | Foundry native | — | 5 |
| Qwen2.5-7B | Ollama | Tavily | 7B | 3 |
| Qwen2.5-14B | Ollama | Tavily | 14B | 3 |
| Qwen2.5-32B | Ollama | Tavily | 32B | 3 |
| Qwen2.5-72B | Ollama | Tavily | 72B | 3 |
| Llama-3.1-8B | Ollama | Tavily | 8B | 3 |
| Llama-3.1-70B | Ollama | Tavily | 70B | 3 |
| Llama-3.3-70B | Ollama | Tavily | 70B | 3 |

Local-model reasoning calls used temperature .1, and the final JSON-formatting call used temperature zero. Initial probabilities expressed on a 0–100 scale were converted to 0–1 before the update stage. Section B.3.1 describes the mixed-scale Qwen2.5-7B update records and their treatment.

## B.2 Forecast, intervention, and validation protocol

### B.2.1 Initial forecast elicitation

Initial forecasts were elicited in three turns using the same substantive instructions across deployments, with provider-specific adaptations for search and tool use. The prompt emphasizes explicit resolution criteria and the status quo, following prior forecasting-prompt work (Halawi et al., 2024; Wilson, 2025).

1. **Event model.** The agent received the question, up to 1,500 characters of resolution criteria, days to resolution, and category, and then defined YES and NO outcomes, identified relevant actors and mechanisms, formed $H_1$ and $H_2$, and described the status-quo trajectory.
2. **Current-evidence research.** The agent gathered evidence for and against each hypothesis, assessed whether the events required for YES were plausible within the remaining time, and revised its assessment.
3. **Structured forecast.** The agent returned a structured forecast containing the event model, hypothesis probabilities, final YES probability, confidence, rationale, and evidence sources.

The research turn included the following calibration instruction:

```
Remember: good forecasters put extra weight on the status quo. Only
depart substantially from the status quo you identified in Step 1 if
you have strong, concrete directional evidence.
```

The output contract records both the final YES probability, $\hat{p}$, and the posterior probability assigned to $H_1$, and explicitly requires these quantities to agree. The internal coherence score in Section B.3.2 measures adherence to this constraint after updating. The fields relevant to the updating analysis are:

```
{
  "hypotheses": [
    {"id": "H1", "posterior_probability": float},
    {"id": "H2", "posterior_probability": float}
  ],
  "yes_prob": float
}
```

### B.2.2 Counterfactual packet construction and validation

GPT-5.4 generated nine synthetic evidence packets for each market: three pro-$H_1$, three anti-$H_1$, and three orthogonal. Generation used the market question, abridged resolution criteria, and one reference initial event model and forecast. Every deployment received the same packet set. Each packet contains a short synthetic news report together with its intended direction, evidence type, and

targeted causal mechanism. Within each direction, the generator was instructed to vary the actor, mechanism, and framing across successive packets. Search was disabled during packet generation, so the interventions are hypothetical premises rather than reports drawn from contemporaneous news.

Table 3: Evidence-packet taxonomy. Each market receives three pro-$H_1$, three anti-$H_1$, and three orthogonal packets, for nine interventions in total.

| **Slot type** | **Direction** | **Role in the intervention** |
|---|---|---|
| DATA | pro, anti | A measurement, price signal, statistic, or poll bearing on the likelihood of $H_1$. |
| DECISION | pro, anti | An actor decision, official statement, policy announcement, commitment, or refusal bearing on $H_1$. |
| STRUCTURAL | pro, anti | A technical, logistical, legal, or contextual change affecting the likelihood of $H_1$. |
| PROCEDURAL | orthogonal | A same-domain administrative or procedural development without a direct causal path to $H_1$ versus $H_2$. |
| PERSONNEL | orthogonal | A staffing or leadership change involving a relevant actor but an issue unrelated to $H_1$ versus $H_2$. |
| BACKDROP | orthogonal | Economic, logistical, or social context that is topically relevant but should not change the outcome probability. |

All 900 intended packets were successfully generated, with 300 in each direction class. A separate human materials review evaluated 18 packets from distinct markets, balanced across the three conditions. Among eight reviewers, the majority judgment matched the intended direction for 17 of 18 packets, while individual judgments matched on 108 of 144 responses. Section B.4.4 reports the full review design and results. These judgments provide direct human support for the intended direction of the sampled interventions; evidential strength is allowed to vary across packets.

#### B.2.3 Forecast revision and pairing

Each update paired one evidence packet with the agent's own initial structured forecast. Search was disabled, and the agent returned revised $H_1$, $H_2$, and YES probabilities using the same output contract as the initial forecast. Each update contained exactly one initial forecast and one packet, so interventions were evaluated independently rather than accumulated across packets. For an agent–market pair with $k$ parseable initial forecasts, the complete crossed design contains $9k$ updates. Repeated updates share markets, initial forecasts, and evidence packets, so uncertainty is clustered by market as described in Section B.3.3. Observed sample sizes differ across deployments because some forecasts or updates were unavailable or unparseable.

GPT-5.4 received one additional instruction that no other deployment received. It was asked to judge whether the packet was already reflected in its initial forecast or represented genuinely new information, with the former associated with a smaller revision and the latter with a larger one. Because this instruction directly concerns revision magnitude, GPT-5.4's exact-zero revision rate is not directly comparable with those of the other deployments. All deployments otherwise use the same scoring definitions below.

### B.3 Estimands and aggregation

#### B.3.1 Analysis sample and probability scales

The analysis uses the June 17, 2026 dataset. Evidence packets and updates are deduplicated by their identifiers, with a parseable update retained when both a valid response and an error record share an identifier. Eligibility is defined separately for each estimand, so records with different available fields can contribute to different metrics.

Qwen2.5-7B contains 207 of 8,636 deduplicated updates with a recorded YES probability above 1.5, indicating a mixture of 0–1 and 0–100 probability scales. Its absolute revision magnitudes and sensitivity ratio are therefore excluded from magnitude-based comparisons. Directional-consistency analyses instead normalize each probability field independently before scoring, allowing all ten deployments to contribute.

### B.3.2 DIRECTIONAL CONSISTENCY AND EVIDENCE SELECTIVITY

Let $\hat{p}_i^{\text{initial}}$ and $\hat{p}_i^{\text{post}}$ denote the initial and revised YES probabilities for update $i$, and let $p(H_1)_i^{\text{post}}$ denote the revised probability recorded for $H_1$. For directional packets, let $d_i = +1$ for pro-$H_1$ evidence and $d_i = -1$ for anti-$H_1$ evidence. Orthogonal packets have no intended direction and are excluded from EHC and HFC. Directional analyses condition on an absolute revision of at least three percentage points; Section B.3.4 evaluates alternative thresholds.

**Evidence–hypothesis consistency.** Define

$$\Delta p(H_1)_i = p(H_1)_i^{\text{post}} - \hat{p}_i^{\text{initial}}.$$

For directional records with $|\Delta p(H_1)_i| \geq 0.03$,

$$\text{EHC}_i = \mathbf{1}[\Delta p(H_1)_i d_i > 0]\,. \tag{2}$$

EHC measures whether the revised $H_1$ probability moves in the direction implied by the packet.

**Hypothesis–forecast consistency.** Define

$$\Delta \hat{p}_i = \hat{p}_i^{\text{post}} - \hat{p}_i^{\text{initial}}.$$

For directional records with $|\Delta \hat{p}_i| \geq 0.03$,

$$\text{HFC}_i = \mathbf{1}[\Delta \hat{p}_i d_i > 0]\,. \tag{3}$$

HFC measures whether the revised YES probability moves in the direction implied by the packet. Because the prompt explicitly requires the YES probability to equal the revised $H_1$ probability, HFC and EHC are not independent measures.

**Internal coherence.** For records with both revised fields,

$$\text{ICS}_i = \mathbf{1}\left[\left|\hat{p}_i^{\text{post}} - p(H_1)_i^{\text{post}}\right| < 0.02\right]. \tag{4}$$

ICS measures adherence to the prompted equality between the two forecast fields.

**Unconditional evidence selectivity.** Mean absolute revision is summarized by

$$\text{Sensitivity} = \frac{\overline{|\Delta \hat{p}|}_{\text{pro+anti}}}{\overline{|\Delta \hat{p}|}_{\text{orthogonal}}}. \tag{5}$$

The numerator pools pro-$H_1$ and anti-$H_1$ updates, while the denominator pools orthogonal updates. Both include zero revisions. A value of one indicates equal mean absolute movement in the directional and orthogonal conditions, while larger values indicate greater movement in response to directional evidence. Because this ratio combines revision frequency with revision magnitude conditional on moving, Section B.4.3 reports those components separately.

### B.3.3 AGGREGATION AND UNCERTAINTY

For EHC, HFC, and ICS, eligible records are first pooled within each model–market pair to obtain a market-level rate. The reported model score is the unweighted mean of these market-level rates, so markets with more completed repetitions receive no additional weight. Standard errors are computed from variation across contributing markets.

Revision magnitudes and sensitivity use record-pooled point estimates. Their 90% intervals use 20,000 percentile bootstrap resamples of markets, retaining all available packets and repeated updates from a selected market within the same resample. This preserves the reported point estimand while accounting for dependence among updates from the same market.

#### B.3.4 OUTPUT COMPLIANCE AND THRESHOLD ROBUSTNESS

The output contract explicitly requires the revised YES probability to equal the revised $H_1$ probability. After scale normalization, 25,982 of 26,008 directional records with both fields present (99.90%) satisfy this equality to numerical precision, and 25,985 (99.91%) differ by less than two percentage points. ICS is therefore primarily a schema-compliance measure, and agreement between EHC and HFC should not be interpreted as independent evidence for a latent hypothesis-to-forecast mechanism.

We recompute directional correctness at movement thresholds of 0, 1, 3, and 5 percentage points. At threshold zero, every valid directional update enters the denominator and an unchanged forecast is scored as incorrect. At positive thresholds, the analysis conditions on an absolute revision of at least the stated magnitude. Rates are computed within markets and then averaged equally across markets.

The threshold sweep leaves the substantive conclusion unchanged. Across hosted deployments, unconditional directional correctness is 97.3–98.4%, compared with 97.7–99.0% at the three-percentage-point threshold. The three-percentage-point rule excludes 6.9% of GPT-5.4, 12.6–12.7% of DeepSeek V4-Pro, and 18.1% of Claude Opus 4.8 directional updates. For local deployments, it excludes only 0.6–2.9% of directional updates, while unconditional correctness ranges from 88.1% to 96.4%. The high conditional rates therefore do not arise from excluding most updates, although conditioning on movement modestly increases some scores. These metrics evaluate revision direction rather than calibration or optimal revision magnitude.

### B.4 SCOPE AND ROBUSTNESS

#### B.4.1 SAMPLING AND OUTCOME SCOPE

The 100-market sample comprises liquid, medium-horizon Polymarket questions selected by topic, price, resolution horizon, trading volume, event deduplication, and topical-diversity constraints. Experiment 1 therefore provides a controlled test of evidence updating on this broad prediction-market sample. Experiment 4 complements it with resolved historical markets and aligned information cutoffs for model forecasts and market prices.

#### B.4.2 INFORMATION AVAILABLE IN THE UPDATE PROMPT

Agents received their own initial forecast, the synthetic evidence text, and the update instructions. The packet's intended direction, target hypothesis, expected shift, evidence type, mechanism annotation, and generator rationale were withheld. Among 12,728 retained hosted-system prompts, every evidence block matched the corresponding packet text and contained none of this withheld metadata. None of the 900 packets explicitly names $H_1$, $H_2$, the prediction market, or its resolution. Twenty-two packets contain likelihood-related language, but those expressions describe events within the synthetic report rather than instructing the agent how to revise its forecast. The intended label was therefore not directly exposed, although indirect cues from vocabulary and evidence type remain possible.

#### B.4.3 ROBUSTNESS OF EVIDENCE SELECTIVITY

**Coverage and material construction.** Most deployments have usable directional and orthogonal updates for at least 99 of the 100 markets. Claude Opus 4.8 has all three evidence conditions for 73 markets, and restricting its sensitivity estimate to this common set gives $23.7\times$, compared with $24.2\times$ on its full available sample. The near-identical common-set estimate indicates that this coverage difference has little effect on the reported Opus sensitivity. Text-based robustness analyses use the 900-packet linked set, while numerical analyses that require only forecast revisions retain all eligible updates.

All packets were generated from a reference event model supplied by DeepSeek V4-Pro. Because this model contributed the reference event models used in packet construction, we interpret its model-level magnitude comparatively rather than as an independent replication of packet generation. The other deployments and the human materials review in Section B.4.4 provide independent checks of the intervention pattern.

**Revision frequency and conditional magnitude.** The unconditional sensitivity ratio combines two behaviors: whether a forecast changes at all and how far it moves conditional on changing. Let $A$ denote the fraction of updates with $\Delta\hat{p} = 0$ and let $M$ denote mean absolute revision among nonzero updates. Then

$$\text{Sensitivity} = \underbrace{\frac{M_{\text{dir}}}{M_{\text{orth}}}}_{\text{magnitude conditional on revision}} \times \underbrace{\frac{1 - A_{\text{dir}}}{1 - A_{\text{orth}}}}_{\text{revision frequency}} . \tag{6}$$

Directional zero-revision rates are $0.0$–$1.4\%$ across deployments, whereas orthogonal zero-revision rates range from $1.1\%$ to $71.4\%$. After conditioning on a nonzero revision, directional evidence still produces $1.9$–$7.5\times$ larger movements than orthogonal evidence, compared with unconditional ratios of $1.9$–$24.2\times$. Thus the selectivity result reflects both a greater probability of revising and a larger revision conditional on doing so.

Table 4: Decomposition of evidence selectivity into exact-zero revision frequency and revision magnitude conditional on moving. Brackets give 90% market-bootstrap intervals. Qwen2.5-7B is excluded because its recorded revision magnitudes mix probability scales.

| | Orthogonal | Mean $\lvert\Delta\hat{p}\rvert$ given a revision | | Sensitivity | |
|---|---|---|---|---|---|
| **Model** | **zero revision** | **orthogonal** | **directional** | **given a revision** | **unconditional** |
| Claude Opus 4.8 | 71.4% [65.0, 77.4] | 1.6 pp | 11.2 pp | 7.0× [5.7, 8.7] | 24.2× [18.2, 34.3] |
| GPT-5.4 | 8.4% [5.6, 11.3] | 1.3 pp | 9.6 pp | 7.5× [6.8, 8.4] | 8.2× [7.4, 9.1] |
| DeepSeek V4-Pro | 44.4% [39.8, 49.0] | 2.6 pp | 12.7 pp | 4.9× [4.5, 5.4] | 8.9× [8.1, 9.7] |
| Qwen2.5-72B | 24.2% [20.3, 28.1] | 3.9 pp | 12.6 pp | 3.2× [3.0, 3.4] | 4.2× [4.0, 4.5] |
| Llama-3.3-70B | 10.7% [7.6, 14.1] | 2.6 pp | 14.6 pp | 5.7× [5.3, 6.1] | 6.4× [5.9, 6.9] |
| Llama-3.1-70B | 2.5% [1.0, 4.6] | 3.4 pp | 15.2 pp | 4.5× [4.2, 4.7] | 4.6× [4.3, 4.9] |
| Qwen2.5-32B | 11.1% [8.4, 14.0] | 4.5 pp | 12.6 pp | 2.8× [2.7, 3.0] | 3.2× [3.0, 3.4] |
| Qwen2.5-14B | 26.1% [21.4, 31.0] | 5.5 pp | 11.9 pp | 2.2× [2.0, 2.3] | 2.9× [2.6, 3.2] |
| Llama-3.1-8B | 1.1% [0.6, 1.7] | 6.0 pp | 11.6 pp | 1.9× [1.8, 2.0] | 1.9× [1.9, 2.0] |

Exact-zero revision is itself an imperfect measure of relevance because it depends on output precision and the update prompt asks agents to leave unaffected fields unchanged. We therefore interpret zero-revision frequency descriptively rather than as evidence about an internal decision to ignore information.

**Surface-form differences.** Directional and orthogonal packets differ by construction in evidence type as well as outcome relevance. Directional packets are somewhat longer and contain more digits, but restricting both conditions to their shared 66–114-word range retains 889 of 900 packets and changes every sensitivity estimate by less than $0.1$. Length alone therefore does not explain the selectivity effect.

Vocabulary does distinguish the packet classes: with five-fold cross-validation grouped by market, multinomial naive Bayes classifies directional versus orthogonal packets with .989 accuracy against a .667 majority baseline. By contrast, recovering the *direction* of the evidence is substantially harder for lexical baselines. On pro-$H_1$ versus anti-$H_1$ packets, naive Bayes and logistic regression achieve .688 and .675 accuracy against a .500 baseline, and adding the market question yields .665 and .680. These values remain well below the hosted deployments' unconditional directional correctness of .973–.984, so the tested bag-of-words features do not explain most of the models' sign accuracy.

**Cross-deployment comparisons.** GPT-5.4 alone received the additional novelty instruction described in Section B.2.3, and hosted and local systems used different retrieval and inference stacks. Comparisons between individual deployments, model families, or hosted and open-weight systems are therefore descriptive rather than controlled effects of model identity or scale.

### B.4.4 Human materials review

Nine adult reviewers assessed 18 synthetic reports from distinct markets, balanced across pro-$H_1$, anti-$H_1$, and orthogonal evidence. Each item presented the market question, resolution criteria, the

corresponding pre-update forecast summary, and the synthetic report while withholding the intended direction, generator metadata, and model responses. Reviewers judged the report's effect on the outcome probability and whether it was usable as a stipulated hypothetical premise.

Eight reviewers met the response-variation criterion used for the reported panel; a ninth selected the same direction on 17 of 18 balanced items and is included in the sensitivity count below. Across the eight-reviewer panel, 108 of 144 judgments (75.0%) matched the intended direction. The panel-majority judgment matched on 17 of 18 items (94.4%; Cohen's $\kappa = .919$), and reviewers judged the report usable as a premise on 124 of 144 judgments (86.1%). Including the ninth reviewer gives 115 of 162 judgments (71.0%) matching the intended direction.

The review supports the interpretability of the intended direction on the sampled materials. It is a materials-validity check rather than a matched human forecasting baseline, since reviewers classified 18 reports whereas agents revised numerical forecasts across the full intervention set.

**Scope of the intervention.** Orthogonal packets provide topically related new information rather than a no-text re-elicitation control. The resulting contrast therefore asks whether models respond more strongly to outcome-relevant evidence than to equally topical information without a direct outcome pathway. Because synthetic reports are supplied as premises with search disabled, the experiment isolates conditional evidence use from source-verification behavior.

### B.5 Context reversal with fixed news

The evidence-packet intervention changes both outcome relevance and evidence type, leaving open whether models can infer relevance when the news itself is held fixed. A context-reversal task therefore evaluates 20 fictional scenario families in which the same news supports the outcome, opposes it, or becomes irrelevant because the connecting relationship is removed. A fourth masked condition omits the relevant relationship and is not assigned a directional target.

Qwen2.5-72B, Llama-3.1-70B, and Qwen3-32B were evaluated with three repetitions per condition. Paired reversal requires the same news to produce correctly signed revisions in both the supporting and opposing contexts, while broken-link movement measures response when the stated connection to the outcome is absent. Paired reversal is 30.0–36.7% across the three models, and mean absolute movement in the broken-link condition is 10.5–13.3 percentage points. By contrast, no-news updates reproduce their baselines exactly, and repeated prior information produces mean absolute movement below $0.6$ percentage points.

GPT-5.6 was evaluated once on the same 20 families. It gives the correct direction on 37 of 40 signed updates, succeeds on 17 of 20 reversal pairs, and moves by an average of $0.5$ percentage points in the broken-link condition. This establishes that the task permits successful relational reversal under at least one tested deployment, but the comparison across models is descriptive because inference settings differ.

Because these scenarios are small and hand-authored, the finite-rule context-reversal experiment in Section B.6 provides a larger complementary test with rule-derived directional targets.

### B.6 Finite-rule context reversal

To obtain direction labels that follow directly from a known data-generating process, we construct 80 new scenarios whose effects are determined by explicit probabilistic rules. Ten scenarios instantiate each of eight mechanism classes: detector calibration, prerequisite timing, component quorums, procedural waivers, proportional brokerage, directed reachability, bounded feedback, and dependence among stored reports. The task changes only the relationship between the actor named in a report and the outcome-generating mechanism, allowing identical news to increase, decrease, or leave unchanged the outcome probability.

**Rule-based construction.** Each scenario contains three independent channels with local latent variables, including a binary state $X$ whose probability is revised from $1/4$ to $3/4$ or from $3/4$ to $1/4$ by the report. The project completes when a final inspection succeeds and either a reserve team succeeds or the ordinary channel activates without the cancellation channel activating. If $a$ and $b$

denote the activation probabilities of the ordinary and cancellation channels, then

$$P(\text{complete}) = \tfrac{4}{5}\left\{\tfrac{1}{5} + \tfrac{4}{5}a(1-b)\right\}.$$

A third archive channel has no effect on the outcome and provides no information about the other channels. Assigning the reported actor to the ordinary, cancellation, or archive channel therefore makes the same report increase, decrease, or leave unchanged the outcome probability. All three contexts contain the same inventory of channel descriptions, and reports that raise and lower $P(X = 1)$ are balanced. Exact probabilities are computed independently by two implementations and remain between .16 and .84, so no target direction is mechanically blocked by an endpoint. The resulting labels are defined by the probabilistic rules rather than by model or human judgments, so directional correctness can be scored exactly.

**Reasoning and wording conditions.** Each scenario is rendered in four versions: the base wording, consistent entity renaming, a paraphrase, and an identical-text resample under a different sampling seed. Each version contains positive, negative, and irrelevant relational contexts with separate baseline forecasts. Independent update branches then supply new evidence, no new information, or repeated information already available at baseline.

Qwen3-32B is evaluated with thinking enabled and disabled on the same scenarios and matched request seeds. Both conditions use temperature $0.7$, top-$p = 1$, unconstrained decoding, an 8,192-token output limit, and a 16,384-token context limit. Invalid, truncated, or missing responses count as failures in directional scores and are not resampled. Thinking disabled yields 3,840/3,840 valid probabilities, while thinking enabled yields 3,803/3,840.

Paired reversal requires both the positive and negative versions of a scenario to move in the correct direction. Zero movement counts as failure for a directional target. Broken-link movement measures absolute revision when the report is assigned to the outcome-irrelevant channel. Intervals use 2,000 bootstrap resamples of whole scenarios, preserving contexts, wording variants, and reasoning conditions.

Table 5: **Reasoning enables context-sensitive reversal under explicit probabilistic rules.** Signs count correctly directed updates out of 160 signed contexts per wording, pairs count successful reversals out of 80, and broken movement is mean absolute revision when the report is outcome-irrelevant. Brackets give descriptive 90% whole-scenario bootstrap intervals.

| Thinking | Wording | Signs | Pairs [90% CI] | Broken $\lvert\Delta\rvert$, pp | Observed |
|---|---|---|---|---|---|
| Off | Base | 84/160 | 14/80 [10.0, 25.0] | 4.2 [3.1, 5.4] | 80/80 |
| Off | Renamed | 86/160 | 14/80 [11.2, 25.0] | 2.7 [1.9, 3.5] | 80/80 |
| Off | Paraphrased | 81/160 | 12/80 [8.8, 21.2] | 2.1 [1.5, 2.9] | 80/80 |
| Off | Same-text resample | 83/160 | 11/80 [7.5, 20.0] | 3.3 [2.5, 4.2] | 80/80 |
| On | Base | 157/160 | 77/80 [92.5, 98.8] | 0.2 [0.0, 0.5] | 80/80 |
| On | Renamed | 159/160 | 79/80 [96.2, 100.0] | 0.0 [0.0, 0.0] | 79/80 |
| On | Paraphrased | 154/160 | 74/80 [87.5, 97.5] | 0.0 [0.0, 0.0] | 79/80 |
| On | Same-text resample | 155/160 | 75/80 [88.8, 97.5] | 0.1 [0.0, 0.2] | 78/80 |

Under base wording, enabling thinking increases paired reversal by 78.8 percentage points, with a 90% interval of $[71.2, 86.2]$. The corresponding contrasts are 81.2 points under renaming, 77.5 under paraphrasing, and 80.0 under identical-text resampling. Omitting one mechanism class at a time gives base-wording contrasts from 75.7 to 82.9 points. The result is therefore not concentrated in one mechanism or wording variant.

The control branches distinguish relational updating from generic forecast instability. With thinking enabled, mean absolute movement under no new information is at most $0.06$ percentage points across wording variants, and repeated prior information produces at most $0.05$ points of movement. Under base wording, exact-probability error falls from 28.75 percentage points at baseline with thinking disabled to $0.13$ with thinking enabled, and updated-probability error falls from 26.89 to $0.08$. The reasoning effect therefore extends beyond choosing the correct revision sign to recovering the rule-derived numerical probabilities.

**Relation-invariant shortcuts.** Within each scenario and wording, the positive, negative, and irrelevant contexts form a triple with the same news, named actor, and rule-derived baseline probability.

A predictor that returns the same revision across the triple cannot satisfy all three targets. Because the scoring tolerance for an unchanged forecast overlaps an arbitrarily small positive revision, such a predictor can achieve at most $2/3$ item-level directional accuracy but zero complete-triple accuracy. For exact posterior agreement within $10^{-4}$, a constant posterior can match at most one of the three contexts.

Simple lexical classifiers likewise do not recover the target relation. When evaluation holds out complete scenarios and all of their wording variants, news-only, full-text unigram, and full-text bigram classifiers obtain approximately $1/3$ directional accuracy and $.50$ balanced relevance accuracy. These controls do not rule out every possible shortcut, but they rule out predictors based only on the tested relation-invariant outputs or lexical features.

**Matched frontier comparison.** GPT-5.6 is evaluated on a balanced 24-scenario subset containing three scenarios from each mechanism class under base and paraphrased wording. On this same subset, GPT-5.6 produces the correct direction on all 48 signed updates and succeeds on all 24 reversal pairs under both wording variants, with zero mean broken-link movement. Qwen3-32B with thinking enabled succeeds on 22/24 base and 23/24 paraphrased reversal pairs, compared with 4/24 and 3/24 with thinking disabled. GPT-5.6 also matches all 144 rule-derived posteriors within $10^{-4}$ on this subset. Both GPT-5.6 and reasoning-enabled Qwen3-32B solve the matched relational task at high rates under their respective deployment settings; the cross-model magnitude comparison is descriptive.

**Scope.** Across 80 scenarios spanning eight probabilistic mechanism classes, the experiment tests whether the same evidence is interpreted according to the relation that connects it to the outcome. The shared controlled template makes the relational manipulation identifiable across mechanisms and wording variants. The resulting claim is sensitivity to explicitly specified relational structure; naturalistic forecasting is evaluated separately in Experiments 1 and 4.

### B.7 Relational depth

A relational-depth diagnostic varies how directly the actor-to-channel relationship is expressed while preserving the same probabilistic calculation. Sixteen scenarios per condition express the assignment directly, through one intermediate relation, through two intermediate relations, or through two relations plus irrelevant actor–channel associations. Within every positive–negative–irrelevant triple, the news, named actor, and rule-derived baseline remain fixed while only the relational assignment changes.

Qwen3-32B with thinking enabled remains highly accurate across all four conditions.

Table 6: **Qwen3-32B under increasingly indirect relational descriptions.** Each condition contains 16 triples and 48 updates. Triple success requires both signed revisions to be correct and the irrelevant revision to remain within $10^{-4}$ of baseline. Exact counts forecasts within $10^{-4}$ of the rule-derived posterior.

| Binding | Triples | Signs | Exact | Error (pp) |
|---|---|---|---|---|
| Direct statement | 16/16 | 48/48 | 44/48 | 0.011 |
| One step, via permits | 15/16 | 46/48 | 44/48 | 0.007 |
| Two steps, via permits and offices | 16/16 | 48/48 | 45/48 | 0.050 |
| Two steps, plus inert relations | 14/16 | 46/48 | 41/48 | 0.320 |

Triple success is 16/16, 15/16, 16/16, and 14/16 across the four conditions, with no monotonic decline as the relational description becomes less direct. Even in the hardest condition, mean posterior error is only $0.320$ percentage points. This diagnostic therefore shows that the reasoning-enabled model remains accurate across the tested increases in relational depth and irrelevant associations.

# C EXPERIMENT 2: CONTEXTUAL RELATIONSHIP SELECTION AND REVISION IN COIN CITY

Experiment 2 tests whether contextual information determines which response relationship a model applies and whether direct observations from the target setting revise that relationship. Models forecast a post-news election poll in City C using varying amounts of City C evidence and, in some conditions, observations from reference cities A and B. All prompt conditions reuse the same 250 seeded episodes, allowing within-episode comparisons under identical numerical evidence.

## C.1 DESIGN AND DATA-GENERATING PROCESS

### C.1.1 LATENT REGIMES AND OBSERVATIONS

Each episode crosses two binary factors: whether City A or City B is the strongly responsive reference city, and whether City C is strongly or weakly responsive. The four cells contain 63, 62, 62, and 63 episodes, respectively. For episode $e$ and city $c$, the response coefficient is sampled independently and remains fixed across that city's observations:

$$g_{ec} \sim \begin{cases} \mathcal{N}(0.90, 0.03^2), & \text{strong regime,} \\ \mathcal{N}(0.25, 0.03^2), & \text{weak regime.} \end{cases}$$

A regime cue therefore identifies the distribution of responsiveness without revealing the realized city-specific coefficient.

Each city contains four observed cases arranged as two matched pairs. Rows within a pair share a starting poll and news magnitude but have opposite news signs. For case $r$,

$$p_{ecr}^{\text{end}} = \text{round}_{0.1}\left(p_{ecr}^{\text{start}} + g_{ec}x_{ecr} + \epsilon_{ecr}\right), \qquad \epsilon_{ecr} \sim \mathcal{N}(0, 5^2).$$

Starting polls are sampled uniformly from $[35, 65]$. Reference-news magnitudes are drawn from $\{6, 7, 8, 9, 10\}$, and query shocks are drawn from $\{-10, \ldots, -4, 4, \ldots, 10\}$. For query starting poll $p_0$ and news value $x$, the scoring target is the latent conditional expectation $p_0 + g_C x$. No additional query-time observation noise is added to this target.

The evidence level $k \in \{0, 1, 2, 3, 4\}$ exposes the first $k$ City C observations. Values $k = 2$ and $k = 4$ complete one and two matched pairs, while $k = 1$ and $k = 3$ expose partial pairs. Conditions containing reference-city evidence always show all four observations from Cities A and B.

Table 7: **Coin City design.**

| Quantity | Value |
|---|---|
| Episodes | 250 |
| Target observations | $k \in \{0, 1, 2, 3, 4\}$ |
| Core conditions | target only; references without City C context; references with correct City C context |
| Additional conditions | misleading City C context; counterbalanced arbitrary labels |
| Strong / weak slope means | $0.90/0.25$ poll points per news point |
| Within-regime slope SD | $0.03$ |
| Observation-noise SD | $5$ poll points |
| Reference cases per city | 4, arranged as two matched pairs |

## C.2 PROMPT CONDITIONS AND AVAILABLE INFORMATION

All prompts state that rows are separate polling cases rather than a time series, that positive news favors the candidate, and that within-city responsiveness is stable despite noisy individual polls. Rows sharing a pair label are described as having the same starting poll and equal-magnitude, opposite-sign news. Every displayed row gives the starting poll, news value, poll change, and ending poll. The required answer contains a numeric `predicted_poll` between 0 and 100 and a one-sentence rationale.

#### C.2.1 FIVE INFORMATION CONDITIONS

**Target only.** The model sees the first $k$ City C observations and no reference-city evidence or contextual description.

**References, no City C context.** The model additionally sees all four observations from Cities A and B. The strongly responsive reference is described as receiving national news coverage, where campaign stories tend to produce larger immediate poll movements. The weakly responsive reference is described as receiving local news coverage, where campaign stories tend to produce smaller immediate poll movements. City C receives no contextual description identifying which reference regime applies.

**References plus correct City C context.** The numerical evidence is identical to the preceding condition. The only addition is a sentence describing City C as receiving national or local coverage, correctly aligned with its latent response regime.

**References plus misleading City C context.** The prompt is identical to the correct-context condition except that the City C sentence names the opposite response regime. All numerical observations and forecast queries remain unchanged.

**References plus arbitrary City C label.** The national/local descriptions are replaced by the arbitrary labels `KIV` and `ZOR`. The prompt states only that cities sharing a label share a typical response regime and that the labels have no inherent meaning or ordering. The mapping between label and response regime reverses across episodes. At $k = 0$, the meaning of City C's label can therefore be recovered only from the labeled numerical examples from Cities A and B.

No prompt contains a fitted coefficient, the strong or weak numerical means, the observation-noise scale, an estimator, or a rule for pooling evidence across cities. The correct-context versus no-context contrast changes only the City C description while holding all displayed numerical evidence fixed. The correct-context versus misleading-context contrast changes only the regime named by that description. The arbitrary-label versus no-context contrast tests whether an episode-specific relationship can be induced without relying on the meanings of *national* and *local*.

**Semantic cues versus induced relationships.** The semantic descriptions explicitly state which media environment tends to produce larger or smaller responses, so these conditions test numerical use of a supplied qualitative relationship. For example, a fixed policy that ignores the displayed cases and assigns coefficients $1.0$ and $0.2$ to the national and local regimes obtains $k = 0$ forecast MAE $0.523$ and implied-slope correlation $.996$. The arbitrary-label condition and reference-selection analysis in Section C.6 provide the induction tests because the mapping must be recovered from episode-specific examples.

### C.3 MODEL EXECUTION

#### C.3.1 MODEL DEPLOYMENTS

The evaluation includes Claude Opus 4.8, GPT-5.6, DeepSeek V4-Pro, Kimi K3, Gemini 3.6 Flash, and Claude Opus 5. No deployment uses tools or retrieval. GPT-5.6 and Claude Opus 5 receive no explicit temperature parameter, while the remaining deployments use temperature zero. Within-model condition contrasts use matched prompts and inference settings. Because provider, reasoning mode, and output budget differ across deployments, cross-model magnitude comparisons are descriptive.

Table 8: **Deployment settings and usable responses.** Parsed counts are usable numerical forecasts out of 3,750 tasks per model.

| Model | Reasoning / sampling | Token limit | Parsed |
|---|---|---|---|
| Claude Opus 4.8 | temperature 0 | 512 | 3,750 |
| GPT-5.6 | low reasoning; temperature omitted | 4,096 | 3,749 |
| DeepSeek V4-Pro | temperature 0 | 4,096 | 3,750 |
| Kimi K3 | low reasoning; temperature 0 | 16,384 | 3,750 |
| Gemini 3.6 Flash | low reasoning; temperature 0 | 16,384 | 3,750 |
| Claude Opus 5 | default extended thinking; temperature omitted | 4,096 | 3,748 |

A response is usable when `predicted_poll` parses to a finite number in $[0, 100]$. Comparisons use complete cases across the compared conditions, and unsuccessful responses remain missing rather than being re-elicited. Section C.9 reports the resulting missingness.

#### C.3.2 MATCHED MISLEADING-CONTEXT INTERVENTION

Each misleading-context prompt differs from its correct-context counterpart by exactly one City C sentence. The City C description names the opposite response regime, while all numerical observations, queries, and other instructions remain unchanged. Across the 1,250 episode–evidence combinations, 625 strong-response targets are described as weak and 625 weak-response targets as strong. Each deployment uses the same inference settings in the correct and misleading conditions.

#### C.3.3 ARBITRARY-LABEL INTERVENTION

The arbitrary-label condition reuses the same numerical task content while replacing semantic regime descriptions with `KIV` and `ZOR`. The more responsive regime is labeled `KIV` in 125 episodes and `ZOR` in 125 episodes, giving 62 or 63 episodes in each target-regime–label combination. At $k = 0$, City C contributes no observations, so identifying its regime requires inferring the label mapping from the reference cities.

The arbitrary-label comparison includes Qwen3 4B/8B/14B/32B, Qwen2.5 7B/14B/32B/72B, Llama-3.1-8B/70B, DeepSeek V4-Pro, Kimi K3, Claude Opus 4.8, Claude Opus 5, and GPT-5.6. Each model is evaluated under no City C context, semantic City C context, and arbitrary-label context on the same 1,250 episode–evidence tasks. Open-weight checkpoints run under vLLM in bfloat16 with deterministic decoding; Qwen uses its no-thinking chat template and Llama uses its native instruction template.

For DeepSeek V4-Pro, the three arbitrary-label comparison conditions use the same deployment and inference configuration. Kimi K3, Claude Opus 5, and GPT-5.6 use their corresponding configurations. Claude Opus 4.8's arbitrary-label condition used a second Azure resource serving the same model and a 1,024-token rather than 512-token output limit. Its matched within-model arbitrary-label contrast is retained; cross-provider magnitude comparisons are not used for inference.

Arbitrary-label analyses use episodes with usable forecasts in all three compared conditions. Unsuccessful or unresolved responses remain missing.

### C.4 ESTIMANDS, BENCHMARKS, AND STATISTICAL INFERENCE

#### C.4.1 FORECAST ERROR AND INFERRED RESPONSIVENESS

For model forecast $\widehat{p}_e$, query starting poll $p_{0e}$, news value $x_e$, and latent City C response coefficient $g_{Ce}$, forecast error is

$$\mathrm{MAE} = \frac{1}{n}\sum_{e=1}^{n} \left|\widehat{p}_e - (p_{0e} + g_{Ce}x_e)\right|.$$

The target is the latent conditional expectation and contains no additional query-time observation noise.

Because every query has $x_e \neq 0$, each forecast also implies a response coefficient

$$\widehat{g}_e = \frac{\widehat{p}_e - p_{0e}}{x_e}.$$

We report mean absolute error in $\widehat{g}_e$ and its Pearson correlation with $g_{Ce}$. Slope correlation measures whether forecasts distinguish more- from less-responsive cities, while slope MAE measures numerical accuracy of the inferred response magnitude. A nearly constant response coefficient can therefore achieve moderate forecast MAE while having little correlation with the true episode-specific coefficient.

Within each model and evidence level, arm comparisons use the common set of episodes with usable forecasts in all compared conditions. Arbitrary-label contrasts likewise use matched complete cases across no-context, semantic-context, and arbitrary-label conditions. All contrasts are paired by episode.

### C.4.2 REFERENCE-SELECTION BENCHMARK

The label-conditioned analogical benchmark fits separate through-origin response coefficients for Cities A and B and selects the reference whose stated regime matches City C. At $k = 0$, it transfers that reference coefficient directly. At $k > 0$, it pools the selected reference observations with the available City C observations before refitting. Under a misleading cue, the same procedure selects the opposite reference. At $k = 0$, this benchmark obtains forecast MAE $1.842$ and slope correlation $.656$ under the correct cue, compared with MAE $4.561$ and correlation $-.667$ under the misleading cue. The benchmark demonstrates that a simple reference-selection policy can reproduce the qualitative behavioral pattern without implying that models implement this particular estimator.

### C.4.3 UNCERTAINTY

Paired condition contrasts use 90% percentile intervals (5th and 95th percentiles) from episode-level bootstrap resampling. Unless otherwise stated, intervals use 2,000 bootstrap draws and preserve pairing between conditions within each episode. Intervals are pointwise rather than simultaneous across models, evidence levels, or metrics, and no multiplicity correction is applied.

## C.5 BEHAVIORAL RESULTS

### C.5.1 CONTEXT SELECTS THE ZERO-SHOT RESPONSE RELATIONSHIP

With no City C observations, target-only and no-context forecasts show little relationship to City C's latent response coefficient. Adding the correct City C description raises slope correlation to $.523$–$.700$ across all six deployments and lowers forecast MAE relative to the numerically identical no-context condition. Changing only that description to the opposite regime instead produces negative slope correlations between $-.50$ and $-.68$.

Table 9: **Zero-shot effect of City C context.** All three conditions use the same reference-city observations and forecast query at $k = 0$; only the City C description changes between the no-context, correct-context, and misleading-context conditions. MAE is in poll points, and $\rho$ correlates forecast-implied and true City C response coefficients.

| | No context | | Correct context | | Misleading context | |
|---|---|---|---|---|---|---|
| Model | MAE | $\rho$ | MAE | $\rho$ | MAE | $\rho$ |
| Claude Opus 4.8 | 2.651 | -.063 | 1.817 | .700 | 4.20 | -.66 |
| GPT-5.6 | 2.518 | -.041 | 1.596 | .678 | 4.23 | -.68 |
| DeepSeek V4-Pro | 2.841 | -.027 | 2.474 | .523 | 4.36 | -.62 |
| Kimi K3 | 2.548 | -.021 | 1.886 | .593 | 4.31 | -.62 |
| Gemini 3.6 Flash | 2.552 | .078 | 2.174 | .559 | 4.29 | -.50 |
| Claude Opus 5 | 2.619 | -.007 | 1.764 | .617 | 4.25 | -.62 |

The misleading cue continues to produce negative slope correlation for all six deployments after one City C observation. After two observations, all six correlations become positive, and by $k = 4$ they range from .32 to .47. Thus models use context most strongly when local evidence is sparse and shift toward direct City C evidence as it accumulates. This transition is visible both in relationship identification and in forecast error, with the largest contextual advantage at the lowest evidence levels.

### C.5.2 ARBITRARY LABELS REQUIRE EPISODE-SPECIFIC INDUCTION

The semantic descriptions provide a pre-existing interpretation of national versus local coverage. The arbitrary-label condition removes this shortcut by assigning `KIV` and `ZOR` opposite meanings across episodes. At $k = 0$, City C contributes no observations, so its response regime can be identified only by inferring the episode-specific label mapping from Cities A and B.

Table 10: **Zero-shot arbitrary-label effects across 15 models.** Each contrast is arbitrary-label minus no-context at $k = 0$, using matched episodes. Negative $\Delta$MAE indicates lower forecast error, while positive $\Delta\rho$ and $\Delta$accuracy indicate better recovery of City C's response regime. Brackets give pointwise 90% episode-bootstrap intervals.

| Model | $n$ | $\Delta$MAE [90% CI] | $\Delta\rho$ [90% CI] | $\Delta$accuracy [90% CI] |
|---|---|---|---|---|
| Qwen3-4B | 250 | $-.005$ [$-.074$, .063] | $+.095$ [$-.037$, .215] | $+.016$ [$-.004$, .040] |
| Qwen2.5-7B | 250 | $+.338$ [.071, .616] | $+.264$ [.137, .385] | $+.072$ [.024, .116] |
| Qwen3-8B | 250 | $-.011$ [$-.197$, .175] | $+.023$ [$-.105$, .146] | $+.024$ [$-.008$, .056] |
| Qwen2.5-14B | 250 | $+.105$ [$-.134$, .345] | $+.236$ [.128, .336] | $+.088$ [.032, .144] |
| Qwen3-14B | 250 | $+.180$ [$-.122$, .492] | $+.398$ [.276, .520] | $+.164$ [.100, .228] |
| Qwen3-32B | 250 | $+.432$ [.169, .696] | $+.302$ [.201, .410] | $+.132$ [.072, .184] |
| Qwen2.5-32B | 250 | $+.529$ [.224, .842] | $+.372$ [.255, .488] | $+.148$ [.100, .200] |
| Qwen2.5-72B | 250 | $+.924$ [.624, 1.241] | $+.381$ [.269, .488] | $+.116$ [.068, .164] |
| Llama-3.1-8B | 242 | $+.492$ [.094, .903] | $+.045$ [$-.097$, .216] | $+.045$ [$-.017$, .107] |
| Llama-3.1-70B | 250 | $+.448$ [.096, .802] | $+.424$ [.330, .525] | $+.092$ [.044, .140] |
| DeepSeek V4-Pro | 250 | $+.107$ [$-.171$, .408] | $+.436$ [.343, .535] | $+.172$ [.112, .232] |
| Kimi K3 | 250 | $-.497$ [$-.732$, $-.262$] | $+.633$ [.539, .728] | $+.336$ [.284, .392] |
| Claude Opus 4.8 | 250 | $-.584$ [$-.823$, $-.359$] | $+.623$ [.515, .729] | $+.312$ [.252, .372] |
| Claude Opus 5 | 250 | $-.603$ [$-.819$, $-.385$] | $+.591$ [.499, .681] | $+.284$ [.232, .340] |
| GPT-5.6 | 250 | $-.774$ [$-.980$, $-.561$] | $+.696$ [.614, .788] | $+.372$ [.320, .424] |

Twelve of the fifteen models have positive symbol-minus-no-context slope-correlation intervals that exclude zero. The effect appears across hosted and open-weight systems, including all four tested Qwen2.5 checkpoints, Qwen3-14B and Qwen3-32B, Llama-3.1-70B, and all four hosted systems. Qwen3-4B, Qwen3-8B, and Llama-3.1-8B have intervals spanning zero, so the pattern is not a simple monotonic function of parameter count.

Relationship discrimination and numerical calibration provide complementary views of the same behavior. Kimi K3, both Claude Opus deployments, and GPT-5.6 improve both zero-shot slope discrimination and forecast MAE under arbitrary labels, while several open-weight models improve discrimination without a corresponding MAE gain. This shows that models can recover which relationship applies even when their numerical calibration remains imperfect.

As direct City C observations accumulate, the arbitrary-label advantage attenuates. For Qwen2.5-32B, Qwen2.5-72B, and Llama-3.1-70B, the symbol-minus-no-context slope-correlation intervals span zero by $k = 3$ or $k = 4$, consistent with direct target evidence replacing the induced contextual convention once it becomes informative.

### C.6 Does the forecast track the cue-selected reference?

To test whether forecasts use the displayed reference-city values, rather than only a regime-level semantic prior, we exploit reference-specific numerical variation within each regime. A model that selects and transfers from the cue-matched reference should track that reference's realized slope even after the strong/weak regime difference is removed. Let $\widehat{\beta}_e^{\mathrm{ref}}$ denote the through-origin slope fitted to a reference city's four noisy observations, and let $R_e$ denote the regime named for City C. We measure

$$r\left(\widehat{g}_e, \widehat{\beta}_e^{\mathrm{ref}} \mid R_e\right),$$

the Pearson correlation between the residuals obtained after regressing the forecast-implied slope and reference fitted slope separately on an intercept and the binary regime indicator. This removes the strong between-regime association and tests whether forecasts track reference-specific numerical variation beyond regime membership. Pointwise 90% intervals use 2,000 episode-bootstrap resamples.

Without reference-city observations, the corresponding associations are near zero. When both references are displayed but City C receives no cue, all six deployments are positively associated with both reference estimates. A correct City C cue concentrates this association on the regime-matched reference, with all six named-reference intervals excluding zero and all six other-reference intervals spanning zero. A misleading cue produces the same pattern for the incorrectly selected reference. Forecasts therefore depend on reference-specific numerical variation that cannot be explained by a response determined solely by the semantic cue.

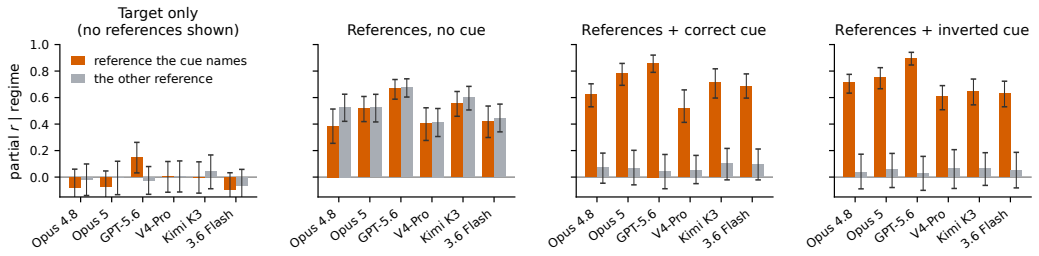



Figure 9: **Forecasts track the reference selected by the cue.** Points show $k = 0$ partial correlations between forecast-implied responsiveness and each reference city's fitted responsiveness after controlling for regime. Whiskers give pointwise 90% episode-bootstrap intervals. Without a City C cue, forecasts track both displayed references. Correct and misleading cues concentrate the association on the reference selected by the cue.

The same test under arbitrary labels gives concordant results. Twelve of fifteen models have positive symbol-matched reference correlations whose 90% intervals exclude zero, with the same three models showing null behavioral and reference-selection intervals. Because both analyses use the same forecasts, they are complementary views of a common effect. The partial correlation measures association rather than an exact transfer weight, but it directly supports numerical use of the cue-selected reference beyond regime membership.

### C.7 Representation and causal analyses

The behavioral analyses establish episode-specific reference use in multiple models. We next test whether the target regime is linearly decodable from model activations and whether intervening on label-position states changes the resulting forecast.

#### C.7.1 Linear-probe methodology

For each prompt, we extract the hidden state at the final input token from the embedding layer and every transformer block. Separate ridge probes are fit for each layer, target, and evidence depth using development episodes only. Features are standardized on the fitting data, and four-fold cross-validation selects both the ridge penalty and probe layer before evaluation on held-out episodes. No test performance is used for layer or regularization selection.

The probe targets are the binary strong-versus-weak regime, the realized response coefficient $g_C$, and the within-regime residual

$$g_C - \mu_{R_e},$$

where $\mu_{R_e} = .90$ for the strong regime and .25 for the weak regime. Regime decoding is scored by ROC AUC, while continuous targets use $R^2$. Mean-pooled input embeddings provide a control for information already linearly available from the prompt before transformer processing.

One-sided permutation tests evaluate held-out probe performance with 10,000 target-label permutations while keeping predictions, fitted weights, selected layer, and regularization fixed. The tests are reported individually without correction across models, targets, or evidence depths.

**Why arbitrary labels are the informative probe condition.** Semantic descriptions explicitly identify City C as belonging to the national or local regime. Accordingly, the semantic regime is highly linearly decodable even from input embeddings, and high semantic-cue probe performance does not by itself demonstrate an induced internal relationship. The arbitrary-label condition provides the stronger test because the meanings of `KIV` and `ZOR` reverse across episodes and must be inferred from the displayed reference examples. The larger-sample arbitrary-label probes and causal interventions below provide the representational evidence.

### C.7.2 PROBE TARGETS AND INFORMATION LIMITS

The probe analysis distinguishes regime identification from recovery of City C's exact response coefficient. Under the data-generating process, knowing the true regime mean explains approximately .992 of the variance in $g_C$, because the within-regime standard deviation is only .03. By contrast, City C's within-regime residual is independent of the reference observations at $k = 0$, giving population $R^2 = 0$, and four City C observations provide only a small amount of residual information, with an approximate benchmark of $R^2 = .009$. High full-slope $R^2$ can therefore arise almost entirely from identifying the response regime. We therefore report regime decoding as the representational endpoint and use the residual target only as a specificity check.

Under arbitrary labels, comparing the fitted slopes of the two reference cities identifies the target regime with simulated AUC .979. This value describes one estimator available from the displayed examples rather than an optimal ceiling over all possible estimators.

### C.7.3 ARBITRARY-LABEL REGIME DECODING

The larger probe evaluation uses 248 episodes, with 160 development episodes and 88 held-out test episodes. Each episode is rendered at $k = 0$ and $k = 4$ under correct, absent, and misleading arbitrary-label assignments. All seven arbitrary-label models use the same train–test split. Probe layers and ridge penalties are selected on the development episodes; evaluation uses the held-out episodes.

At $k = 0$, the input embeddings themselves provide little linear information about the arbitrary mapping, with mean-pooled embedding AUCs between .495 and .556 across the seven models. After transformer processing, four models show significant held-out regime decoding: Qwen3-14B, Qwen2.5-32B, Qwen2.5-72B, and Llama-3.1-70B. The remaining three models, Qwen3-4B, Qwen3-8B, and Llama-3.1-8B, do not show significant $k = 0$ regime decoding.

Table 11: **Zero-shot decoding of an induced arbitrary-label regime.** Probe AUC is held-out regime decoding at $k = 0$; Embed. is the corresponding mean-pooled input-embedding AUC. The $p$-value is from a one-sided 10,000-permutation test. Behavioral $\Delta\rho$ is the symbol-minus-no-context slope-correlation effect from Table 10.

| Model | Probe AUC | Embed. AUC | $p$ | Behavioral $\Delta\rho$ |
|---|---|---|---|---|
| Qwen3-4B | .520 | .498 | .376 | .095 |
| Qwen3-8B | .566 | .534 | .147 | .023 |
| Qwen3-14B | .678 | .555 | .0016 | .398 |
| Qwen2.5-32B | .778 | .550 | $< .001$ | .372 |
| Qwen2.5-72B | .662 | .556 | .0026 | .381 |
| Llama-3.1-8B | .423 | .539 | .893 | .045 |
| Llama-3.1-70B | .728 | .495 | $< .001$ | .424 |

The same four models with significant $k = 0$ regime probes also have behavioral symbol-minus-no-context $\Delta\rho$ intervals excluding zero, linking linear decodability with behavioral use of the induced convention across checkpoints. The remaining three models have both probe and behavioral intervals spanning zero.

Qwen3-14B illustrates how this representation changes as direct City C evidence arrives. At $k = 0$, its arbitrary-label probe has AUC .678 under the correct label, compared with .467 with no label and .321 under the misleading label. At $k = 4$, the misleading-arm probe instead favors the true regime, with AUC .747. The representation therefore shifts from the induced contextual mapping toward the relationship supported by direct target observations.

The effect appears at different model sizes across families, so we treat scale as a descriptive checkpoint pattern rather than a universal threshold.

### C.7.4 CAUSAL INTERVENTION ON LABEL-POSITION STATES

To test whether the decoded information causally influences the forecast, we intervene on the hidden states at the two City C label-token positions under arbitrary labels. Correct-label and swapped-label

prompts are token-aligned and differ only in City C's label. For each recipient prompt, the intervention replaces the label-position states with the corresponding states from the opposite-label donor and then generates the forecast greedily.

For episode $e$, let $s_e = +1$ for a strong-response City C and $s_e = -1$ for a weak-response City C. The symmetric intervention effect is

$$d_e = \frac{s_e}{2}\left[(\widehat{g}_{W\leftarrow C,e} - \widehat{g}_{W,e}) + (\widehat{g}_{C,e} - \widehat{g}_{C\leftarrow W,e})\right],$$

where $C$ and $W$ denote the correct- and swapped-label prompts and the arrow denotes transfer from donor to recipient. Positive $d_e$ means that replacing the recipient state moves its forecast toward the donor label's response regime.

### C.7.5 Causal intervention on held-out episodes

The analysis uses 88 held-out episodes balanced across reference-city regime, City C regime, and arbitrary label. Patch windows are chosen using development activations and evaluated only on these held-out episodes. For both models, outputs 19–21 form the intermediate patch window; the late comparison windows are outputs 33–35 for Qwen3-14B and 35–37 for Llama-3.1-70B.

We test whether the mean $d_e$ at outputs 19–21 is positive at $k = 0$ and whether it exceeds the corresponding late-window effect. The four tests use one-sided episode-level sign flips with 200,000 draws and Holm correction at familywise level .05. Intervals use 10,000 episode-bootstrap resamples. Validity requires at least 77 of 88 complete episodes and at least 319 of 352 self-patched forecasts to reproduce their unpatched counterparts exactly. Both models satisfy these criteria.

Table 12: **Label-position intervention on 88 held-out episodes.** Shift is the mean symmetric effect $d_e$ in slope units, with 90% episode-bootstrap intervals. Positive values indicate movement toward the donor label's regime. The late window is outputs 33–35 for Qwen3-14B and 35–37 for Llama-3.1-70B.

| | | Qwen3-14B | | Llama-3.1-70B | |
|---|---|---|---|---|---|
| Patch site | $k$ | Shift [90% CI] | $p$ | Shift [90% CI] | $p$ |
| Outputs 19–21 | 0 | .452 [.343, .562] | $< 5 \times 10^{-6}$ | .641 [.503, .778] | $< 5 \times 10^{-6}$ |
| Outputs 19–21 | 4 | $-.016$ [$-.094$, .059] | .634 | .074 [.021, .129] | .012 |
| Late window | 0 | .044 [.010, .082] | .022 | $-.013$ [$-.026$, $-.001$] | .955 |
| 19–21 minus late | 0 | .376 [.280, .475] | $< 5 \times 10^{-6}$ | .654 [.516, .792] | $< 5 \times 10^{-6}$ |

Both models show positive $k = 0$ effects at outputs 19–21 after Holm correction, and each effect exceeds the corresponding late-window effect. Patching outputs 19–21 shifts the implied response coefficient toward the donor label's regime by .452 in Qwen3-14B and .641 in Llama-3.1-70B. The intermediate-minus-late contrasts are .376 and .654, respectively, showing that the effect is substantially larger at the selected intermediate depth than at the late comparison window.

Replacing the input embeddings at the label positions produces $k = 0$ shifts of .528 for Qwen3-14B and .665 for Llama-3.1-70B. Thus, changing only three intermediate label-position states transfers most of the effect produced by replacing the label representation at the input.

With four City C observations, the Qwen3-14B effect attenuates to $-.016$ with interval $[-.094, .059]$, while Llama-3.1-70B retains a smaller positive effect of .074 $[.021, .129]$. Direct target evidence therefore sharply reduces the influence of the induced label representation, while the exact rate of attenuation differs across models. Together, these interventions support a causal role for label-position states in zero-shot regime selection.

### C.7.6 Interpretation and scope

Taken together, the arbitrary-label probes and interventions connect behavior, representation, and causal influence. An episode-specific mapping is linearly decodable in four of seven tested open-weight checkpoints, and replacing label-position states causally shifts zero-shot forecasts in Qwen3-14B and Llama-3.1-70B on held-out episodes.

The intervention replaces complete hidden-state vectors, so the causal claim concerns the influence of those states rather than a unique decoded direction. The controlled task focuses on selecting between two response regimes; continuous contextual variables and fine-grained within-regime coefficient recovery are natural extensions of this setting.

### C.8 Robustness across generator populations and deployments

#### C.8.1 Harder generator population

We repeat the arbitrary-label evaluation on 250 new episodes with strong and weak regime means of .75 and .40 and within-regime slope standard deviation .10. This jointly reduces between-regime separation and increases within-regime heterogeneity relative to the original population. Observation noise, factorial balance, prompts, evidence depths, checkpoint settings, and episode-randomized symbol mappings are otherwise unchanged. Ten open-weight models are evaluated under no City C context, semantic context, and arbitrary-label context at all five evidence depths.

Table 13: **Zero-shot arbitrary-label effects across generator populations.** Values are symbol-minus-no-context changes in slope correlation at $k = 0$, with pointwise 90% episode-bootstrap intervals. The harder population narrows regime separation and increases within-regime dispersion.

| Model | Original $\Delta\rho$ [90% CI] | Harder $\Delta\rho$ [90% CI] |
|---|---|---|
| Qwen3-4B | $+.095$ $[-.037, .215]$ | $+.050$ $[-.058, .178]$ |
| Qwen3-8B | $+.023$ $[-.105, .146]$ | $+.231$ $[.101, .345]$ |
| Qwen3-14B | $+.398$ $[.276, .520]$ | $+.155$ $[.023, .281]$ |
| Qwen3-32B | $+.302$ $[.201, .410]$ | $+.193$ $[.085, .301]$ |
| Qwen2.5-7B | $+.264$ $[.137, .385]$ | $+.149$ $[.030, .259]$ |
| Qwen2.5-14B | $+.236$ $[.128, .336]$ | $+.133$ $[.026, .233]$ |
| Qwen2.5-32B | $+.372$ $[.255, .488]$ | $+.247$ $[.131, .361]$ |
| Qwen2.5-72B | $+.381$ $[.269, .488]$ | $+.336$ $[.226, .444]$ |
| Llama-3.1-8B | $+.045$ $[-.097, .216]$ | $-.077$ $[-.193, .056]$ |
| Llama-3.1-70B | $+.424$ $[.330, .525]$ | $+.242$ $[.139, .347]$ |

Seven of ten models have positive intervals excluding zero in both generator populations, and nine of ten retain the same interval classification. All four Qwen2.5 checkpoints and Llama-3.1-70B remain positive in the harder population; Qwen3-14B and Qwen3-32B are positive in both populations, while Qwen3-8B becomes positive under the harder generator. The arbitrary-label effect therefore survives a substantial reduction in regime separation and increase in within-regime heterogeneity across most tested checkpoints.

Forecast discrimination and numerical calibration remain distinct under this perturbation. For example, Qwen2.5-72B improves slope discrimination in the harder population while its forecast MAE increases by $1.012$ poll points, reinforcing the distinction between identifying the applicable relationship and calibrating its exact magnitude.

### C.9 Response validity and scope

#### C.9.1 Missing forecasts

The six-model evaluation contains 22,500 model–prompt combinations, of which three lack a usable forecast. Matched comparisons therefore use $n = 249$ for the affected model–prefix combinations and $n = 250$ elsewhere. All 7,500 misleading-context tasks yield usable analyzed forecasts after deduplication.

The 15-model arbitrary-label comparison contains 56,250 model–prompt combinations and uses complete cases across the three compared conditions. At $k = 0$, every model contributes 250 complete episodes except Llama-3.1-8B, which contributes 242. Across later evidence levels, its complete-case sample ranges from 238 to 246. Unusable forecasts remain missing rather than being imputed.

### C.9.2 INTERPRETATION AND SCOPE

Experiment 2 provides convergent evidence for context-sensitive relationship selection. Matched cue interventions show that changing only City C's contextual assignment changes the inferred response relationship. Arbitrary labels show that this assignment can be induced from episode-specific examples rather than supplied by familiar semantics. Reference-selection analyses show that forecasts track numerical variation in the reference selected by the cue. Linear probes show that the induced arbitrary-label regime is decodable in several open-weight checkpoints, and the label-position interventions show that those states causally influence zero-shot forecasts in Qwen3-14B and Llama-3.1-70B.

The analyses are complementary: behavioral contrasts measure forecast changes, reference-selection correlations measure association with displayed numerical examples, probes measure accessibility to a linear readout, and patching measures the causal effect of replacing selected hidden states. The intervention replaces complete hidden-state vectors, so it supports a causal role for the patched states without identifying a unique internal algorithm.

The scoring target is City C's noise-free conditional expectation, and the task uses stable within-city responsiveness with a binary, well-separated contextual variable. Cross-model comparisons are descriptive because deployments differ in provider, reasoning configuration, and output budget. The label-position intervention applies Holm correction to its four reported tests; other bootstrap and permutation intervals are pointwise.

# D EXPERIMENT 3: CONTROLLED RELATIONSHIP TRANSFER

Experiment 3 tests whether post-training teaches models to use episode-specific predictive relationships and whether that advantage survives changes in domain and response mechanism. Training uses direct-response systems in Coin City, while evaluation independently varies the surface domain, response mechanism, and amount of target-specific evidence. The key comparison holds prompts, training budget, and the distribution of reward targets fixed while changing whether each prompt is paired with its own episode-specific target. Qwen3-4B and Qwen3-8B are each evaluated with eight training seeds per reward arm.

## D.1 DESIGN AND EVALUATION

### D.1.1 DOMAINS AND RESPONSE MECHANISMS

Let $x_t$ denote the input and $z_t$ the latent deviation of the outcome from its resting level. Training uses the direct-response mechanism,

$$z_{t+1} = g_j x_t. \tag{7}$$

A pulse therefore affects the next observation but does not persist after subsequent inputs return to zero.

Evaluation additionally includes a mediated mechanism,

$$m_{t+1} = \rho m_t + g_j x_t, \tag{8}$$

$$z_{t+1} = \phi z_t + \lambda m_{t+1}. \tag{9}$$

Both latent states begin at zero. Observed outcomes add independent Gaussian noise and are clipped to the stated domain range, while evaluation targets use the corresponding noise-free trajectory. Prompts provide the domain bounds and qualitative pathway descriptions but omit the equations, response parameters, mediator, and structure labels.

Coin City has resting level $50$, support $[0, 100]$, and observation-noise standard deviation $1.6$. Coin Harbor changes the terminology, units, numerical scale, input magnitudes, and pathway descriptions, with resting level $180$, support $[40, 320]$, and observation-noise standard deviation $4.5$. Training reference gains are sampled from $[.18, .46]$ for the weak-response system and $[.68, .98]$ for the strong-response system. At evaluation, $g \sim \mathrm{Uniform}(.22, .92)$, while Structure B uses $\rho \sim \mathrm{Uniform}(.48, .74)$, $\phi \sim \mathrm{Uniform}(.18, .42)$, and $\lambda \sim \mathrm{Uniform}(.68, .94)$. Coin City inputs are drawn from $\{-12, -8, -4, 0, 4, 8, 12\}$, while Coin Harbor uses $\{-24, -16, -8, 0, 8, 16, 24\}$. Evaluation queries use shocks $\{-12, -6, 0, 6, 12\}$ in Coin City and $\{-24, -12, 0, 12, 24\}$ in Coin Harbor at horizons one and three. A reference and its matched target share these latent response parameters but have independent inputs and observation noise.

Table 14: **Training and transfer conditions.** Training uses Coin City with the direct response mechanism, while evaluation independently changes the surface domain and response structure.

| Evaluation cell | Domain | Response structure |
|---|---|---|
| In distribution | Coin City | Direct, Structure A |
| Domain shift | Coin Harbor | Direct, Structure A |
| Structure shift | Coin City | Mediated, Structure B |
| Joint shift | Coin Harbor | Mediated, Structure B |

### D.1.2 EPISODE-MATCHED TRAINING AND POPULATION-PRIOR CONTROL

Each training prompt contains two complete 14-case reference systems with different response gains and a target that shares one reference's gain. The target reveals $k \in \{0, 2, 4, 8\}$ observations, and its qualitative background identifies the matching response regime. Training contains 1,200 prompts at each evidence depth, for 4,800 prompts in total. Weak- and strong-response targets are balanced overall but associated with evidence depth, so $k$ itself provides some information about the gain band.

The episode-matched and population-prior conditions receive identical prompts and the same multiset of answer vectors. Episode-matched reward scores a response against the forecasts generated by that

episode's own latent parameters. The population-prior control reassigns complete answer vectors among episodes with the same evidence depth so that every vector is used exactly once and no episode retains its own vector. This preserves the reward distribution conditional on $k$ while removing the episode-specific prompt–target pairing. Population-prior minus episode-matched performance therefore isolates the value of learning which forecast belongs to which episode, beyond learning the task's answer distribution and other shared regularities.

**Training reward and output contract.** The Experiment 3 reward is

$$r = .4\, r_{\text{level}} + .6\, r_{\text{response}},$$

where each component averages $\max(0, 1 - |e|/S)$ over its corresponding forecast errors. The level tolerance is 10% of the domain's observable range, while the response tolerance is half one intervention unit. Training and evaluation require exactly ten finite forecasts in a JSON object; malformed or length-truncated outputs receive zero training reward.

### D.1.3 Evaluation protocol

Every evaluation prompt contains one complete direct-response reference, one complete mediated-response reference, and a target generated by one of those mechanisms. The target receives either the correct qualitative pathway description, no pathway description, or the description of the other mechanism. These prompt variants reuse exactly the same numerical observations and forecast queries.

Each task requests ten counterfactual forecasts spanning five intervention magnitudes and horizons one and three. Response error is computed from the eight nonzero-shock responses after subtracting the corresponding zero-shock forecast at each horizon. Evaluation crosses two domains, two response structures, three hint conditions, four target-evidence depths, and 30 episodes, for 1,440 prompts per checkpoint.

The reported contrast is population-prior minus episode-matched response MAE under the correct pathway description. Positive values favor episode-matched training. The four evaluation cells measure performance in distribution, under domain shift, under structure shift, and under both shifts jointly.

All reported endpoints use round 300 and five stochastic forecasts per prompt at temperature $0.7$. Each forecast draw is scored separately before averaging within a prompt. Invalid forecasts receive an error equal to the observable range of the corresponding domain.

## D.2 Eight-seed transfer results

Qwen3-4B and Qwen3-8B are each evaluated with eight training seeds per reward arm, using seeds 42–49. The data, 300-round endpoint, five-draw decoder, and transfer estimands are shared across the two reward conditions. The hierarchical bootstrap first resamples training seeds and then paired evaluation prompts within each selected seed. Brackets give 90% hierarchical bootstrap intervals. We additionally summarize seed-level uncertainty with a two-sided wild-cluster bootstrap with Webb weights and an exact one-sided sign test.

Table 15: **Eight-seed population-prior minus episode-matched response MAE.** Positive values favor episode-matched training, and brackets give hierarchical 90% intervals. All cells use the correct pathway description and average the four evidence depths equally.

| Evaluation cell | Qwen3-8B | Qwen3-4B |
|---|---|---|
| In distribution | .060 [.023, .103] | .067 [−.037, .171] |
| Domain shift | .110 [−.006, .242] | .087 [−.061, .232] |
| Structure shift | −.003 [−.046, .035] | .052 [−.055, .159] |
| Joint shift | .126 [.018, .242] | .136 [−.005, .275] |

For Qwen3-8B, episode-matched training has lower point-estimate error in distribution, under domain shift, and under joint shift, while the structure-shift contrast is approximately zero. The in-distribution

and joint-shift 90% hierarchical intervals exclude zero. The in-distribution effect is highly consistent across uncertainty summaries: .060 $[.023, .103]$, wild-cluster $p = .0089$, eight positive seed-level effects out of eight, and one-sided exact sign-test $p = .0039$.

The joint-shift contrast is larger, .126 with a 90% hierarchical interval of $[.018, .242]$. Its wild-cluster value is $p = .091$, and six of eight seed-level effects are positive. Thus episode-specific supervision retains a measurable advantage even when both the surface domain and response mechanism change, although the seed-level pattern is less uniform than in distribution. Under domain shift alone, the point estimate is similarly positive at .110 and the interval narrowly spans zero, $[-.006, .242]$; under structure shift alone, the estimate is essentially zero.

Qwen3-4B shows positive point estimates in all four cells. Its joint-shift estimate is .136 with interval $[-.005, .275]$, while the other three intervals also include zero. Taken together, the eight-seed results provide strong evidence of episode-specific learning in distribution and positive evidence that part of this advantage survives the joint domain-and-structure shift.

### D.3 Diagnostics of relationship learning

The following diagnostics clarify what is learned and where transfer becomes difficult. They test direct learnability of response-form selection, tracking of episode-specific gain changes, and the value of preserving prompt–target pairing.

#### D.3.1 Supervised learnability and its transfer boundary

We first ask whether the response-structure task is learnable from direct supervision. This diagnostic uses a separate structure-selection dataset in which every prompt contains one direct and one mediated reference. Their horizon-one response magnitudes are matched, while their horizon-three persistence differs, so selecting the correct response form requires distinguishing temporal structure rather than immediate response magnitude. Targets are balanced equally between the two structures. Qwen3-8B is fine-tuned on 4,800 such Coin City prompts using simulator-correct forecasts as targets. The supervised recipe uses learning rate $2 \times 10^{-5}$, one epoch, 300 updates, and four independent training seeds. Neither Coin Harbor nor arbitrary labels appears during training.

Across the four seeds, in-distribution Coin City forecasts identify the correct response structure on $97.9\%$–$100.0\%$ of prompts, with response MAE between .277 and .658. Reference identification remains $100\%$ in every evaluation condition. When the learned response-form selector is moved to arbitrary labels or Coin Harbor, structure accuracy is instead around the binary-choice baseline: $39.6\%$–$50.0\%$ for City/arbitrary, $45.8\%$–$52.1\%$ for Harbor/semantic, and $47.9\%$–$50.0\%$ for Harbor/arbitrary.

The supervised study therefore establishes that response-form selection is readily learnable in-domain and localizes the remaining challenge to portability across cues and domains. The SFT and RL studies serve complementary diagnostic roles because their objectives, learning rates, and effective training budgets differ.

#### D.3.2 Mechanism-family evidence use under coherent gain changes

This diagnostic uses a separate Qwen3-4B mechanism-family training study with eight response mechanisms represented during training and four held out for evaluation. Episode-matched and shuffled-target models use the same mechanism-balanced prompts and training budget; the shuffled arm preserves the answer-vector distribution within compatible mechanism and evidence-depth groups while breaking the episode-specific pairing. We test whether trained forecasts track an episode-specific change in response strength. The diagnostic contains 16 paired episodes from each of 12 response mechanisms, comprising eight mechanisms represented during training and four held-out compositions or parameter-extrapolation mechanisms. Both mechanism-disclosed and mechanism-undisclosed conditions use the same episodes and nine target observations.

Within each pair, the target gain changes from the 20th to the 80th percentile of that mechanism's gain distribution. Inputs, observation-noise draws, reference trajectories, baseline state, and mechanism remain fixed, while target observations and forecast outcomes are recomputed under the new gain.

The disclosed description states the same gain range for both members and never reveals the realized gain.

Let $\Delta r_i$ denote the simulator's change in the eight-coordinate response vector between the high- and low-gain episodes, and let $\Delta\widehat{r}_i$ denote the corresponding predicted change. We measure

$$E_\Delta = \frac{1}{8N}\sum_{i=1}^{N}\|\Delta\widehat{r}_i - \Delta r_i\|_1\,, \qquad \beta_\Delta = \frac{\sum_i \Delta\widehat{r}_i^{\top}\Delta r_i}{\sum_i \|\Delta r_i\|_2^2}.$$

Exact tracking gives $\beta_\Delta = 1$, while a forecast invariant to the gain change gives zero.

The comparison uses five training seeds per arm of episode-matched and distribution-matched shuffled-target training.

Table 16: **Episode-specific gain tracking on seen and held-out mechanisms.** Entries are response-change MAE, with lower values better. Slope M and S are the tracking slopes for episode-matched and shuffled-target training. Positive shuffled-minus-matched contrasts favor episode matching. Brackets give 90% intervals from 2,000 seed-then-episode bootstrap repetitions.

| Disclosure | Set | Matched | Shuffled | No change | Slope M | Slope S | Shuffled − matched |
|---|---|---|---|---|---|---|---|
| Disclosed | Seen | 1.250 | 1.512 | 1.591 | .323 | .084 | $+.262\ [+.220, +.308]$ |
| Disclosed | Held out | .898 | .874 | .826 | .122 | .049 | $-.024\ [-.069, +.019]$ |
| Undisclosed | Seen | 1.115 | 1.488 | 1.591 | .461 | .107 | $+.373\ [+.326, +.418]$ |
| Undisclosed | Held out | .878 | .862 | .826 | .238 | .060 | $-.016\ [-.055, +.020]$ |

Episode-matched training substantially improves gain tracking on mechanisms represented during training. The response-change MAE advantage is .262 with mechanism disclosure and .373 without disclosure, and both 90% intervals exclude zero. The matched tracking slopes, .323 and .461, are roughly four times the corresponding shuffled slopes of .084 and .107.

On held-out mechanisms, the episode-matched models still move in the correct direction, with positive tracking slopes of .122 and .238, but matched and shuffled response-change MAE are similar. The training benefit is therefore strongest on mechanism families represented during learning, while some directional gain sensitivity persists on held-out mechanisms.

### D.3.3 Distribution-matched shuffled-target control

The shuffled-target control reassigns answer vectors only within compatible mechanism and evidence-depth groups, preserving the complete multiset of training targets while breaking the episode-specific prompt–target correspondence. Prompts and training budget are otherwise unchanged.

The comparison uses five training seeds per arm. With the mechanism undisclosed, episode matching improves stochastic response MAE by .094: the 90% hierarchical interval is $[+.041, +.168]$, wild-cluster $p = .0138$, all five seed-level contrasts are positive, and the exact one-sided sign test gives $p = .0312$. With the mechanism disclosed, the advantage is larger at .155 with interval $[+.062, +.243]$; four of five seeds are positive and wild-cluster $p = .079$. Both disclosure conditions therefore support a benefit from preserving the episode-specific prompt–target pairing.

An eight-seed sensitivity analysis points in the same direction, with contrast $+.061$, 90% hierarchical interval $[+.001, +.125]$, wild-cluster $p = .101$, and six of eight positive seeds. The effect is smaller than in the five-seed comparison, while the hierarchical interval remains positive.

Taken together, the diagnostics show that post-training can learn response-form selection, track episode-specific gain changes on familiar mechanisms, and benefit from preserving which target belongs to which prompt. The weaker held-out-mechanism results identify portability, rather than basic relationship learning, as the main remaining challenge.

# E EXPERIMENT 4: HISTORICAL-MARKET TRAINING AND HELD-OUT EVALUATION

Experiment 4 tests whether outcome-based training improves forecasts for events and question families absent from the fine-tuning data. Models receive a prediction-market question, its resolution rules, and the market-price history available by a fixed forecast cutoff, then predict the probability of YES. Training uses a Brier-score reward, and evaluation separates markets temporally and by event and question family.

## E.1 SAMPLE CONSTRUCTION AND LEAKAGE CONTROLS

The source archive contains 1,788,703 market records. The selected datasets contain 1,736 training, 512 development, and 1,024 test markets, with YES rates .311, .275, and .301, respectively. Training contains 991 distinct events and 1,456 normalized question templates; development contains 292 events and 453 templates; test contains 493 events and 866 templates.

We include binary markets with literal Yes/No outcomes and an archived Boolean resolution inferred from an extreme final token price. A fixed domain filter retains political, governmental, macroeconomic, and geopolitical questions while excluding sports, entertainment, cryptocurrency, and other out-of-scope categories. The forecast cutoff is 30 days before the earliest available resolution, market-close, market-end, or event-end timestamp. Eligible markets require at least four distinct pre-cutoff trading dates, a latest available price no more than seven days old, and a latest YES price in $[.02, .98]$. Prompts contain at most the latest 14 pre-cutoff YES closing prices together with the question, resolution rules, and cutoff date. Outcomes, terminal prices, market status, terminal timestamps, volume, and selection metadata are withheld.

Training cutoffs precede July 1, 2025, development cutoffs fall between July 1 and December 31, 2025, and test cutoffs fall between January 1 and June 30, 2026. Development markets sharing an event, available series identifier, or normalized question template with training are removed, and the same exclusion is then applied to test markets against both earlier splits. Numeric values and month names are normalized when constructing template keys. The final train–development, train–test, and development–test family intersections are all empty.

The archive provides timestamped price histories but not historical revisions to static question or resolution-rule text, so those strings are treated as time-invariant. The split design blocks post-cutoff price leakage and explicit event/template reuse. As in other historical LM evaluations, exposure to public historical events during model pretraining is not directly observable.

## E.2 TRAINING AND EVALUATION

Training uses the 1,736 unique markets in a deterministic 4,800-presentation schedule. Every market appears at least twice, and the resulting 300 rollout rounds and 2,400 optimizer updates match the training budget used elsewhere in the paper. Models are optimized directly for Brier reward,

$$r = 1 - (\widehat{p} - y)^2.$$

Table 17: **Historical-market training and evaluation configuration.**

| Quantity | Value |
|---|---|
| Models | Qwen3-4B-Instruct-2507, Qwen3-8B, Llama-3.1-8B-Instruct |
| Training markets | 1,736 unique markets; 4,800 presentations |
| Training | 300 rounds; seeds 42, 43, 44 |
| Reward | $1 - (\widehat{p} - y)^2$ |
| Optimization | Dr. GRPO; AdamW; learning rate $10^{-6}$ |
| Rollouts | 16 prompts per round; 8 responses per prompt; temperature 1.3 |
| Development | 512 markets |
| Held-out evaluation | 1,024 markets; temperature zero; one forecast per market |

Held-out evaluation uses Brier loss. The reported trained score averages the three seed-specific losses rather than averaging their forecast probabilities. Invalid forecasts receive Brier loss one. Pointwise 90% intervals use 5,000 bootstrap resamples of connected event/template components, preserving markets that belong to the same connected family. These intervals quantify held-out market-sample variation conditional on the three fitted models.

### E.3 HELD-OUT RESULTS

All three model families improve relative to their corresponding untrained checkpoints on the 1,024 family-disjoint test markets.

Table 18: **Held-out historical-market forecasting.** Trained Brier scores average three training seeds. Negative trained-minus-base differences favor training, with 90% family-clustered bootstrap intervals.

| Model | Base Brier | Trained Brier | Trained − base [90% CI] |
|---|---|---|---|
| Qwen3-4B | .1316 | .1159 | $-.0157$ $[-.02355, -.00809]$ |
| Qwen3-8B | .1287 | .1164 | $-.0123$ $[-.01825, -.00684]$ |
| Llama-3.1-8B | .1520 | .1150 | $-.0370$ $[-.05095, -.02449]$ |

Every trained model family improves over its corresponding untrained checkpoint, and all three 90% trained-minus-base intervals exclude zero on the 1,024 family-disjoint markets. The gains are .0157 Brier for Qwen3-4B, .0123 for Qwen3-8B, and .0370 for Llama-3.1-8B.

The latest pre-cutoff market price obtains Brier .11369, and a logistic recalibration fit only on the training split obtains .11232. The trained models finish within a few thousandths of these market-based benchmarks: Qwen3-4B is statistically indistinguishable from the raw market, while Qwen3-8B and Llama-3.1-8B are higher by only .0027 and .0013 Brier, respectively, at 90% confidence. Against the recalibrated benchmark, the corresponding differences are .0036, .00409, and .00269. Thus outcome-based post-training closes most of the gap from the untrained checkpoints to contemporaneous prediction-market accuracy on event- and template-disjoint questions.

### E.4 SCALE COMPARISON

A separate 318-market holdout tests how the same training recipe behaves across Qwen3 scale. These markets have forecast cutoffs between January 1 and June 8, 2026, and their identifiers, events, available series, and normalized question templates are absent from the training, development, and original 1,024-market test sets. The holdout contains 227 event families and 288 normalized templates, and selection does not use outcomes or market probabilities.

Qwen3 checkpoints from 1.7B through 32B use the same training data, three seeds, 4,800-presentation schedule, reward, and development protocol. Each base and trained checkpoint receives five non-thinking forecasts per market at temperature .7, top-$p = .8$, and top-$k = 20$. Forecast probabilities are averaged within each five-draw group before computing Brier loss. All Qwen3 draws parse successfully.

Table 19: **Qwen3 scale comparison on the 318-market holdout.** Trained means average seeds 42–44, and lower Brier is better.

| Checkpoint | Base | Trained mean | Trained − base |
|---|---|---|---|
| Qwen3-1.7B | .2253 | .1248 | $-.1005$ |
| Qwen3-4B-Instruct-2507 | .1522 | .1366 | $-.0156$ |
| Qwen3-8B | .1369 | .1286 | $-.0083$ |
| Qwen3-14B | .1281 | .1270 | $-.0012$ |
| Qwen3-32B | .1266 | .1265 | $-.0001$ |

The base-relative gain decreases smoothly as the untrained Qwen3 checkpoint improves. At 90% confidence, the trained-minus-base intervals for Qwen3-8B and Qwen3-14B exclude zero, while

the 32B contrast, $-.00007$ $[-.00039, +.00023]$, is effectively zero. The trained 14B model also improves over trained 8B by $.00164$ Brier, with interval $[.00043, .00299]$.

Contemporaneous Polymarket prices obtain Brier $.12655$ on this holdout, and a train-only Platt recalibration obtains $.12942$. The trained 14B checkpoint scores $.12695$, statistically indistinguishable from the raw market and better than the Platt recalibration at 90% confidence; trained 32B is likewise statistically indistinguishable from the market. The 4B trained checkpoint has higher Brier than the Platt benchmark, while the 14B checkpoint has lower Brier and its trained-minus-Platt interval excludes zero. The scale results therefore show that post-training produces clear gains through 14B and that, as the base model strengthens, trained performance converges toward the market benchmark while the incremental training gain shrinks.

Llama-3.1-8B provides an architecture comparison on the same holdout. Its trained mean is $.12643$, compared with $.16417$ for the base, for a difference of $-.03774$ $[-.05482, -.02230]$. Its trained-minus-market difference is $-.00012$ $[-.00039, +.00009]$, statistically indistinguishable from the contemporaneous market at 90%. The Llama result therefore reproduces the pattern of substantial base-relative improvement.

**Hosted-model references.** DeepSeek V4-Pro, Claude Opus 4.8, GPT-5.6, and Kimi K3 are evaluated on the same 318 markets as external performance references. Their Brier scores are $.12660$, $.12927$, $.13335$, and $.12682$, respectively, compared with $.12655$ for Polymarket. Claude Opus 4.8 is only $.00272$ Brier above the market; its 90% interval narrowly excludes zero. Provider sampling and reasoning configurations differ from one another and from the local checkpoints, so these systems serve as external accuracy references rather than controlled model-family comparisons.

### E.5 Price ablation: forecasting without market history

The historical-market prompts include dated pre-cutoff prices, whose latest value is itself a strong forecast. To test how much trained performance depends on this information, we evaluate the same trained adapters after removing the entire price-history block while leaving the question, resolution rules, cutoff, and output format unchanged. Models are not retrained for this ablation.

We report the 1,024-market held-out evaluation for Qwen3-1.7B, Qwen3-14B, and Qwen3-32B. Each trained value averages seeds 42–44. The no-price evaluation uses one temperature-zero forecast per market, matching the held-out evaluation. The training-set YES rate of $.311$ provides a fixed climatology benchmark with Brier loss $.21042$.

Table 20: **Held-out forecasting accuracy depends strongly on market-price history.** Brier loss is lower when better. Trained values average three seeds.

| Model | No-price base | No-price trained | Trained with prices |
|---|---|---|---|
| Qwen3-1.7B | .2611 | .2468 | .1183 |
| Qwen3-14B | .2381 | .2413 | .1147 |
| Qwen3-32B | .2935 | .2933 | .1151 |

The no-price ablation separates the training effect from the contribution of market history. At 1.7B, training still improves Brier by $.01429$ with a 90% interval of $[-.02838, -.00191]$; at 14B, training increases Brier by $.00322$ $[+.00097, +.00554]$; and at 32B the difference is essentially zero. Absolute no-price performance remains above the $.21042$ climatology benchmark at all three sizes, whereas the same trained adapters are much more accurate when price history is supplied.

The large price-bearing gains therefore reflect a forecasting skill that makes effective use of contemporaneous market history. Because these adapters were trained with prices, the ablation measures reliance on that input rather than the potential of a separately trained price-free forecaster.

## F Experiments 3–4: shared training configuration

Experiments 3 and 4 use the same Dr. GRPO training implementation unless a study-specific section states otherwise. Table 21 reports the parameters needed to reproduce the shared optimization procedure.

Table 21: **Shared Dr. GRPO training configuration.**

| Quantity | Configuration |
|---|---|
| Adaptation | Frozen base model with LoRA rank 32 and $\alpha = 64$ on attention and feed-forward projections; bfloat16; no quantization |
| Objective | Dr. GRPO with group-centered rewards and no group-standard-deviation normalization; PPO clip .2; importance-sampling cap 2.0; one PPO epoch; KL coefficient 0 |
| Optimizer | AdamW; learning rate $10^{-6}$; $\beta_1 = .9$, $\beta_2 = .95$; weight decay 0; maximum gradient norm 1.0 |
| Batching | 16 prompts per rollout round; 8 sampled responses per prompt; 128 trajectories per round |
| Schedule | 4,800 prompt presentations; 300 rollout-and-update rounds; 2,400 AdamW updates; constant learning rate with no warmup |
| Training decoding | Temperature 1.3; top-$p = 1$; no top-$k$ restriction |
| Invalid outputs | Missing, malformed, or length-truncated required outputs receive reward zero |
| Training seeds | Experiment 3 Qwen comparison: seeds 42–49; Experiment 4 historical-market training: seeds 42–44 |

One reported training "round" denotes one rollout-and-update cycle over 16 prompts and eight sampled responses per prompt. With the implemented microbatching and gradient accumulation, each round performs eight AdamW updates, giving 2,400 optimizer updates over 300 rounds. Study-specific datasets, reward functions, evaluation decoding, and uncertainty procedures are described in their respective experiment sections.