COUNTERFACTUAL PLASTICITY: RISK-AWARE SELECTION OF MODEL UPDATES BEFORE EXECUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

An update can teach a language model new information while harming behavior that was previously reliable. When several update procedures are available, the question is not only whether to update, but which update to apply.

We introduce **Counterfactual Plasticity** (CP), a framework that predicts the likely benefit, collateral damage, and damage uncertainty of each candidate before it is run. CP calibrates the resulting selection rule with held-out split conformal risk control, then chooses the highest-benefit admissible candidate or abstains.

Controlled experiments show that benefit is easier to forecast than damage, and that simple forecast-based selection can fail when the candidate menu has real benefit and damage trade-offs. In frozen factual-editing evaluations, CP substantially reduces harmful selections in two higher-risk regimes. In an independent Qwen2.5-7B evaluation where the baseline is already benign, calibration selects $\lambda = 0$ and leaves the baseline decision unchanged. Together, these results show how a model can compare plausible updates before committing to one, while adding conservatism only when the evidence supports it.

1 INTRODUCTION

Modern language models are increasingly expected to change after pretraining. Factual correction, continual learning, personalization, domain adaptation, and model editing all modify a model that already possesses useful capabilities. Yet acquiring new information can also perturb unrelated knowledge or previously reliable behavior. This creates a basic plasticity problem: *how should a model be updated when several plausible updates differ in both benefit and collateral effects?*

Most update procedures are evaluated retrospectively: an intervention is applied, target behavior is measured, and collateral degradation is assessed afterward. But the same intended change can often be implemented through several candidate interventions, such as different learning rates, optimization durations, parameter scopes, or layer locations, with different benefit and damage trade-offs. We therefore ask whether these consequences can be estimated *before* execution and used to decide which candidate, if any, should become real.

We call this problem *counterfactual update selection*. Given a model state, a new experience, and a finite candidate set, the goal is to forecast each candidate's acquisition gain and collateral damage from pre-update information, then select one candidate or abstain. This treats updating as a decision over possible parameter-space futures rather than as a predetermined optimization step.

Our controlled experiments reveal a recurring asymmetry: acquisition is comparatively easy to forecast, whereas damage is more brittle across model families, distributional regimes, and candidate action spaces. Ensemble uncertainty is informative in several settings but can fail under structural action-space shift. In a heterogeneous controlled stress test, for example, the uncalibrated selector is harmful on 51.5% of episodes and frozen CP reduces this to 38.5% without eliminating the failure. This motivates calibrating the *decision rule* rather than trusting point forecasts alone, while also delimiting the scope of that calibration.

We introduce **Counterfactual Plasticity** (CP), a risk-aware decision layer over a finite candidate-update menu. CP constructs a pre-update fingerprint for each candidate, forecasts acquisition gain, collateral damage, and damage uncertainty, excludes candidates inconsistent with a calibrated damage criterion, and selects the highest-predicted-gain admissible candidate or abstains. In the final

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.

external evaluation, the conservatism of this rule is selected with split conformal risk control on held-out calibration edits, with the test set untouched.

We evaluate CP under frozen external factual-editing protocols with disjoint training, calibration, and one-shot test splits. In E1, on 90 held-out edits per model, CP reduces the harmful-decision rate of the uncalibrated ensemble selector (B1) from 24.4% to 7.8% on TinyLlama and from 23.3% to 8.9% on Gemma. The paired reductions are 16.7 and 14.4 percentage points, with both 95% confidence intervals entirely below zero. The utility cost differs by model: TinyLlama trades acquisition for lower risk and abstains on 17/90 edits, whereas Gemma retains essentially unchanged acquisition with no test abstentions. CP remains adaptive, selecting all six candidate procedures on TinyLlama and five of six on Gemma. In an independent E2 confirmation on Qwen2.5-7B, B1 is already benign (1/90 harms); split CRC selects $\lambda = 0$, so CP exactly matches B1 rather than claiming an unsupported incremental reduction.

The contribution is therefore not a new mechanism for constructing a particular edit, but a decision layer that forecasts and selects among *complete candidate parameter-update procedures before execution*. This distinguishes CP from methods that adapt how a single editing mechanism is internally constructed.

This distinction fixes both the action and the timing of the decision. An adaptive editor may use properties of an edit to choose an internal layer, parameter subset, or edit-specific construction within its own mechanism. CP is agnostic to how a candidate is constructed: adaptive editors can be members of $\mathcal{C}$, alongside ordinary optimization procedures. CP instead decides, from information available before execution, which complete procedure to instantiate on the current model state, or whether none is adequately supported. Likewise, CP is not a model-editing algorithm that claims a better post-edit representation; its object is the pre-execution policy over a fixed menu of possible updates. This separation permits the risk and utility decision to be studied without attributing improvements to a particular editing mechanism.

Our contributions are:

- We formulate **counterfactual update selection**: choosing among candidate parameter updates from pre-update information while balancing acquisition gain against collateral damage.
- We characterize an empirical forecasting asymmetry in which acquisition is comparatively predictable while damage is more brittle, including a structural action-space regime where uncertainty and frozen risk calibration fail to transfer reliably.
- We introduce **Counterfactual Plasticity**, combining pre-update forecasting, ensemble uncertainty, decision-level calibration, and abstention to select among complete candidate updates.
- We provide frozen external factual-editing evaluations in two risk-bearing E1 regimes and an independent Qwen2.5-7B E2 confirmation, showing both substantial reductions and appropriate non-intervention when the baseline policy is already benign.

These results are not a universal safety guarantee. The calibrated procedure is tied to its candidate-action distribution and stated evaluation assumptions, and arbitrary distribution shift can invalidate the conditions under which its risk calibration is meaningful.

2 Counterfactual Update Selection

For model parameters θ, a new experience x, and a finite menu of update procedures $\mathcal{C} = \{c_1, \ldots, c_K\}$, candidate c produces $\theta'_c = U(\theta, x; c)$. Candidates can differ in optimization strength, duration, parameter scope, or layer location. We must select one candidate before executing any of them.

Let ℓ_x be target loss and let $\mathcal{P} = \{p_1, \ldots, p_M\}$ be a protected set. The realized acquisition gain and collateral damage are

$$G(c) = \ell_x(\theta) - \ell_x(\theta'_c), \qquad D(c) = \frac{1}{M} \sum_{j=1}^{M} [\ell_{p_j}(\theta'_c) - \ell_{p_j}(\theta)].$$

Larger G is better and positive D is degradation. With prospectively fixed damage budget τ, a selected update is harmful when $H(c) = \mathbb{I}[D(c) > \tau]$. These outcomes are observable only after the candidate is run.

CP instead forms a pre-execution fingerprint $\phi(\theta, x, c)$ and predicts $(\widehat{G}_c, \widehat{D}_c, \widehat{U}_c)$ for every $c \in \mathcal{C}$. The resulting alternatives are *counterfactual updates*. ϕ contains only pre-update information; realized gains, damages, harm labels, and post-execution statistics are unavailable at decision time. Thus the task is not prediction error alone: a forecaster can err most on the high-gain candidates a selector prefers. We seek a policy

$$\pi : \{\phi(\theta, x, c)\}_{c \in \mathcal{C}} \to \mathcal{C} \cup \{\perp\}$$

that obtains acquisition while controlling $\mathbb{E}[H(\pi(x))]$; abstention $\perp$ executes no update and has zero harmful-update loss. Forecasting, risk estimation, and selecting the highest-utility admissible candidate are consequently distinct problems. The diagnostic oracle and full feature specification are given in Appendix A.

The protected set defines the operational meaning of collateral damage rather than an exhaustive notion of preservation. In particular, $D(c)$ averages loss changes over the prespecified protected items, and $H(c)$ records whether that average exceeds the fixed budget. The policy is evaluated once per requested experience, not once per candidate: candidates that are considered but not executed have no realized contribution to the decision loss. This is why candidate-level forecast quality alone is insufficient for the stated decision problem, and why a safe-looking candidate menu can still yield a risky selector when its ranking favors systematically underestimated updates.

3 Counterfactual Plasticity

CP uses pre-update fingerprints to forecast candidate gain and damage, then calibrates the *decision induced by these forecasts*, not a simultaneous guarantee for every candidate prediction. Let $\mu_G(c)$ rank candidates and let $\mu_D(c)$ and ensemble disagreement $\sigma_D(c)$ be the damage forecast and empirical uncertainty signal, respectively. (Architecture, features, and the interpretation limits of disagreement are in Appendix A.) For conservatism $\lambda \geq 0$, CP admits

$$\mathcal{A}_\lambda(x) = \{c \in \mathcal{C} : \mu_D(c) + \lambda \sigma_D(c) \leq \tau\}, \qquad \pi_\lambda(x) = \arg \max_{c \in \mathcal{A}_\lambda(x)} \mu_G(c),$$

and abstains when $\mathcal{A}_\lambda(x) = \varnothing$. This separates utility ranking from damage admissibility and makes uncertainty operational rather than a claimed posterior probability of harm.

Decision-level split CRC. For an edit i, the raw loss is

$$L_i(\lambda) = \mathbb{I}[D_i(\pi_\lambda(x_i)) > \tau]$$

when CP selects an update and 0 on abstention. Candidate switching means that raw policy loss need not be monotone along a scalar λ path. We therefore calibrate an all-supersets monotone closure over *forbidden sets*, not scalar-path shorthand. For $F \subseteq \mathcal{C}$, let $\pi(F)$ choose the highest-predicted-gain member of $\mathcal{C} \setminus F$ (or abstain if none remains), and define

$$\ell_0(y, F) = \mathbb{I}[D(y, \pi(F)) > \tau] \quad (= 0 \text{ on abstention}), \qquad \widetilde{\ell}(y, F) = \max_{F' \supseteq F} \ell_0(y, F').$$

Thus, for $F_1 \subseteq F_2$, $\widetilde{\ell}(y, F_1) \geq \widetilde{\ell}(y, F_2)$, while $\ell_0(y, F) \leq \widetilde{\ell}(y, F)$. For numerical λ, $F_\lambda(x) = \{c : \mu_D(x, c) + \lambda \sigma_D(x, c) > \tau\}$. This exact all-supersets closure is essential; it is not defined by $\sup_{\lambda' \geq \lambda} L_i(\lambda')$.

In E1 we enumerate all $2^6 = 64$ forbidden sets for every calibration edit and use the ordered family

$$\Lambda_{\mathrm{E1}} = \{0, 0.25, 0.50, \ldots, 15.00\}, \texttt{ FORCED_ABSTAIN},$$

with 61 numerical λ values. The terminal state forbids all candidates, executes no update, and has zero raw and closure loss. With 90 disjoint calibration edits and $\alpha = .10$, split CRC chooses the first eligible state using exactly

$$\frac{1 + \sum_{i=1}^{90} \widetilde{\ell}(y_i, F_\lambda(x_i))}{91} \leq .10.$$

It therefore accepts at most eight closure losses; `FORCED_ABSTAIN` ensures selection is defined if no numerical value is eligible. Forecast training, calibration, and final testing are disjoint. This E1 construction is distinct from the controlled v0.1 10-fold CV-CRC procedure.

Scope. Conditional on fixed E1 training data and a fixed pipeline (finite candidate set, features, forecasters, grid, deterministic policy, and damage/reset scoring rule), split CRC controls expected average harmful-update loss for a future edit exchangeable with the 90 calibration edits. The closure bound transfers this statement to the raw selected-update loss, with zero loss on abstention. It is not a per-edit, universal, realized-rate confidence, or online-deployment guarantee. It also depends on the finite candidate space and on offline calibration data containing every counterfactual candidate outcome. Appendix A gives the full construction, tie-breaking, proof details, and inference pseudocode.

Two details are worth making explicit. First, the guarantee concerns the random loss of the selected policy on a future exchangeable edit; it is not a claim that at most α of the 90 realized test decisions must be harmful. Finite test samples can fluctuate around their expectation. Second, abstention is part of the loss definition, not evidence that an unexecuted candidate was safe: it means that the policy commits no parameter update and therefore incurs zero harmful-update loss under this metric. Its practical cost is foregone acquisition, which is reported separately. Thus split CRC calibrates one precisely specified harm and abstention decision, rather than certifying all candidate predictions or all dimensions of model behavior.

4 When Does Risk-Aware Update Selection Matter?

We first ask when risk-aware selection is necessary rather than merely conservative. In controlled GPT-2 update episodes spanning facts, symbols, rules, and arithmetic, every candidate is executed from the same reset checkpoint and harm is defined as $D > \tau$ with $\tau = 0.01$. Forecasts use only pre-update fingerprints. These experiments are mechanistic development diagnostics; the controlled v0.1 policy uses a separate 10-fold CV-CRC procedure and does not inherit the E1 split-CRC guarantee.

Forecasting asymmetry. Acquisition is consistently easier to predict than collateral damage. In the original menu, ensemble damage uncertainty strongly tracks absolute damage error (Spearman $\rho \approx 0.87$ on calibration episodes and 0.85 on held-out episodes), showing that disagreement can be useful even when damage point estimates are imperfect. The decision problem is nevertheless harder than average prediction error: a selector can preferentially choose the high-gain candidates whose damage is most underestimated.

Benign menus. Risk calibration adds little when the action space is already safe. Under the original eight-candidate menu, the point-estimate selector achieves mean $G = 9.717$ with one harm in 30 episodes, while B1 achieves $G = 9.702$ with zero harms. Harder experiences with the same menu produce 0/400 B1 harms, and a densified update-strength menu produces only 1/400. Thus, neither harder experiences nor the mere presence of harmful candidates is sufficient; harmful updates must also be competitive under the forecasted utility landscape.

Structural ambiguity. We therefore introduce a heterogeneous menu varying learning rate, duration, parameter scope, layer location, and width. High-gain full-scope candidates now compete with safer localized updates, creating genuine risk pressure. Frozen B1 is harmful on $206/400 = 51.5\%$ of episodes, while frozen CP reduces this to $154/400 = 38.5\%$. All B1 harms are full-scope selections. Crucially, within that high-risk subspace the uncertainty and error association collapses to $\rho \approx 0.07$, and harmful full-scope outcomes are systematically underpredicted. Retraining on this regime sharply reduces harm, indicating a transport failure in damage forecasting rather than an inherent inability to separate every risky update.

The S2 result also sharpens what "structural" means here. The issue is not simply that a test episode is more difficult, nor that a menu contains more candidates. The heterogeneous menu introduces qualitatively different full-scope actions whose high predicted gain makes them competitive in the selector's ranking. Within precisely that region, damage is underpredicted and disagreement ceases to identify the largest errors. Frozen calibration can then exclude candidates according to a signal whose ordering no longer tracks the relevant failure mode. The partial reduction from 51.5% to 38.5% is therefore descriptive evidence of limited transfer, not evidence that CP solves the shifted regime; the improvement after retraining instead points to the need for calibration data representative of the available action space.

Figure 1 summarizes the key contrast. The negative cases are important: CP is not automatically beneficial when B1 is already benign, and frozen calibration need not transfer across a structurally changed action space. These lessons motivate the E0/E1 design below.

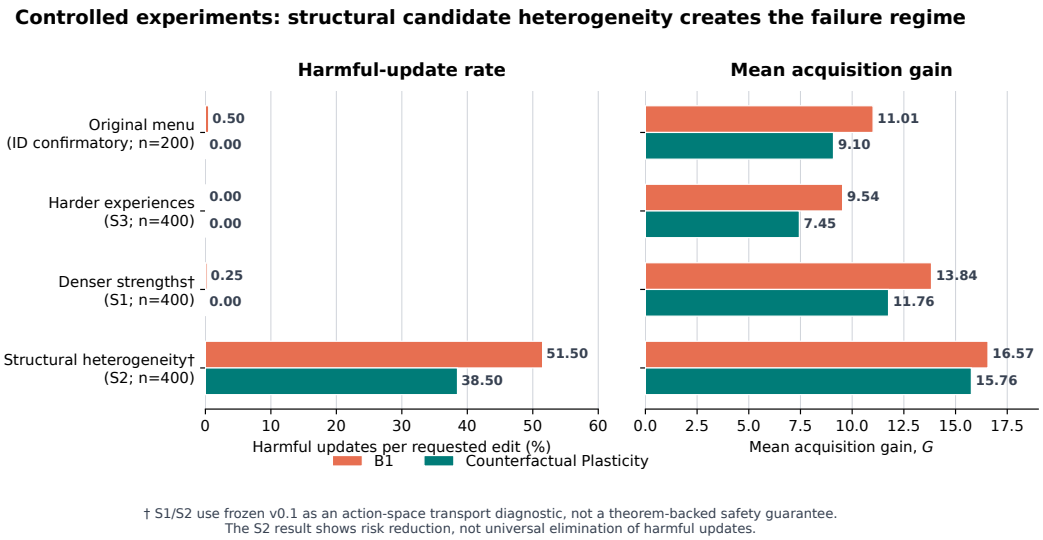


Figure 1: **Controlled regime summary.** Frozen B1 and CP are nearly harmless in the original menu, harder-experience S3, and denser-strength S1 settings. Structural candidate heterogeneity in S2 creates substantial harmful-update pressure (51.5% for B1); frozen CP reduces this to 38.5% but does not eliminate it. S1/S2 are transport diagnostics rather than theorem-backed E1 risk guarantees.

Additional B0/B1/B2 progression, S1/S3 diagnostics, and retraining analyses are reported in Appendix B.

5 Prospective External Factual Editing

We next evaluate whether such decision pressure appears in factual editing. The external study separates a descriptive development screen (E0) from a fresh frozen confirmatory evaluation (E1). E0 evaluates Qwen, TinyLlama, and Gemma on 60 edits under six fixed candidate procedures and 120 protected facts. Qwen is nearly benign, whereas TinyLlama and Gemma exhibit substantial benefit and damage trade-offs; those two regimes are therefore carried forward to E1. This is development-stage model/regime selection, not a prospective E0 claim, and no E0 example is reused in E1.

The E0-to-E1 separation is intentional. E0 asks whether the fixed menu creates enough decision pressure to make a risk-aware policy informative: in a nearly benign regime, reducing an already rare harm rate would be difficult to interpret and could be achieved by needless conservatism. TinyLlama and Gemma showed nontrivial trade-offs under the development screen, so they were chosen as the two frozen evaluation regimes. E1 then uses new edits, new protected items, disjoint forecast-training and calibration splits, and a locked test split. Consequently, E0 supports the choice of where to evaluate rather than the final risk claim, which rests on the separately frozen E1 protocol.

E2 is a separate, independent confirmation on Qwen2.5-7B rather than a continuation of E0. Its 180/90/90 training, calibration, and locked-test split is disjoint from E0, E1, and an excluded 12-edit E2 pilot. E2 uses the same six-procedure decision framing and primary harm definition, but runs Qwen in BF16; paraphrase and neighborhood measurements are secondary descriptive preservation metrics, not CRC-controlled losses.

5.1 FROZEN E1/E2 PROTOCOLS

For each model, E1 uses 360 fresh CounterFact edits and 180 fresh protected items, with the same identities across models but independent forecasting and calibration. Edits are grouped into 180 training, 90 calibration, and 90 locked one-shot test edits, with all six candidates for an edit kept in the same split. Before test outcomes are observed, we freeze the data manifests, model/tokenizer revisions, candidate mappings, feature schema, forecasters, seeds, baselines, $\tau = 0.01$, split-CRC target $\alpha = 0.10$, tie-breaking, abstention, and evaluation rules.

Split-CRC selects $\widehat{\lambda} = 2.25$ for TinyLlama and 3.00 for Gemma. The 90-edit calibration criterion permits at most eight closure losses: TinyLlama attains $9/91 = 0.0989$ regularized risk with eight closure losses, while Gemma attains $7/91 = 0.0769$ with six. The exact forbidden-set construction and its expected-risk scope are given in Section 3 and Appendix A.

E2 uses the same split sizes, protected-set size, candidate-menu cardinality, τ, and α, with independently frozen Qwen2.5-7B data, revisions, seeds, and BF16 execution. Its calibration selects $\widehat{\lambda} = 0$ with zero closure losses ($1/91 = 0.010989$ regularized risk).

Each edit and candidate pair is independently restored, fingerprinted, executed, and evaluated against all 180 protected items. The primary comparator is B1, which uses the same six-candidate action space and forecasting machinery as CP but without calibrated decision-level exclusion. Harm is measured per requested edit, with abstention assigned zero harmful-decision loss.

5.2 LOCKED EXTERNAL RESULTS

Table 1 gives the frozen test results.

Table 1: **Locked E1/E2 protocol, calibration, and primary external factual-editing results.** Each evaluation uses 180 training edits, 90 calibration edits, and 90 locked test edits; the protected set has 180 items and the action space has six frozen update procedures. Harm is $1[D > 0.01]$. Split CRC targets $\alpha = 0.10$. Harm rates are per requested edit (abstentions contribute zero harm); conditional CP harm is reported among executed updates.

Protocol and calibration

Model	Train / Cal. / Test	Protected	Candidates	λ	Closure losses; R_{CRC}
TinyLlama	180 / 90 / 90	180	6	2.25	$8/90$; $9/91 = 0.098901$
Gemma	180 / 90 / 90	180	6	3.00	$6/90$; $7/91 = 0.076923$
Qwen2.5-7B	180 / 90 / 90	180	6	0.00	$0/90$; $1/91 = 0.010989$

Locked test results

Model	B1		Counterfactual Plasticity			
	G	Harm	G	Harm	Harm—exec.	Abstain
TinyLlama	7.294038	24.44%	6.155598	7.78%	9.59%	18.89%
Gemma	21.311913	23.33%	21.307810	8.89%	8.89%	0%
Qwen2.5-7B	12.647885	1.11%	12.647885	1.11%	1.11%	0%

Denominator. The primary harm denominator is all 90 requested test edits. The conditional column uses only non-abstained edits and is supplementary.

In E1, CP substantially reduces harmful decisions relative to B1 on both models. On TinyLlama, harm falls from $22/90 = 24.4\%$ to $7/90 = 7.8\%$, a paired reduction of 16.67 percentage points (95% CI $[-24.44, -8.89]$). On Gemma, harm falls from $21/90 = 23.3\%$ to $8/90 = 8.9\%$, a reduction of 14.44 points (95% CI $[-22.22, -7.78]$).

The independent E2 Qwen2.5-7B result is deliberately different: B1 and CP both make one harmful update in 90 (1.1%), have identical mean acquisition ($G = 12.648$), and make no abstentions. Calibration selects $\widehat{\lambda} = 0$, so the paired CP versus B1 harm difference is 0.0 percentage points (95% CI $[0.0, 0.0]$). Thus E2 does not support an incremental H1 risk reduction; it shows that CP can leave an already benign selector unchanged.

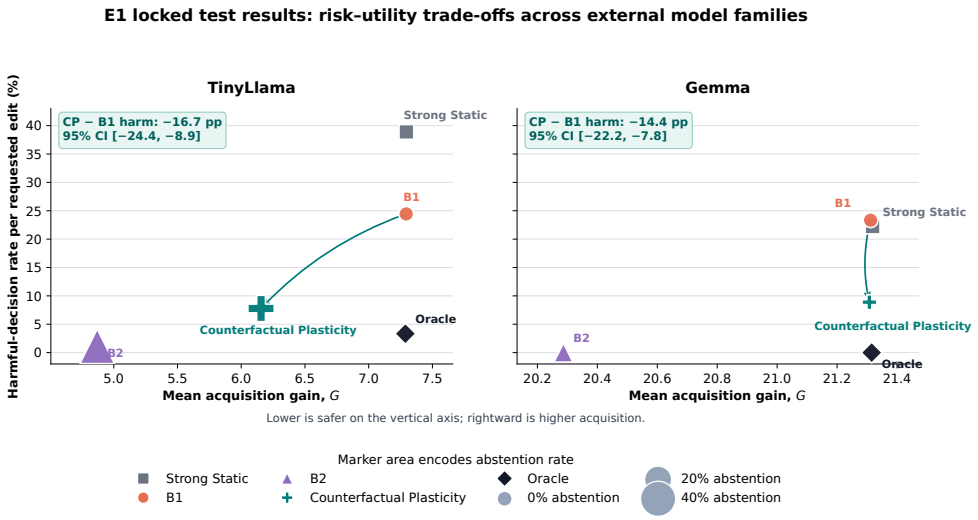


Figure 2: **Locked E1 risk and utility trade-offs.** Across 90 held-out edits per model, CP substantially reduces harmful decisions relative to B1. Marker area encodes abstention. The annotated CP versus B1 harm-rate differences use the preregistered paired bootstrap intervals. B2 is shown only as a descriptive prediction-level conformal baseline; no CP versus B2 inferential comparison was preregistered.

The utility cost is model dependent. TinyLlama loses 1.1384 mean acquisition relative to B1 (95% CI $[-1.7186, -0.6274]$) and abstains on 17/90 edits; among executed updates, harm is $7/73 = 9.6\%$. Gemma changes acquisition by only -0.00410 (95% CI $[-0.00621, -0.00238]$) and never abstains. CP does not collapse to a static rule: it uses all six candidates on TinyLlama and five of six on Gemma.

Why select per edit? The E1 policies use several procedures across the test edits, so one fixed update does not capture the choices made in these higher-risk regimes. E2 is the complementary case: its baseline is already benign and calibration leaves that policy unchanged. Per-edit selection is therefore most useful when the candidate menu creates meaningful instance-dependent trade-offs, rather than when additional exclusion is unsupported by the calibration regime.

The contrast between regimes is informative rather than a discrepancy to be averaged away. For TinyLlama, CP's lower harm rate is accompanied by fewer executed opportunities and lower mean acquisition, so its 17 abstentions make the risk reduction visibly costly under this utility measure. For Gemma, the same decision framework reduces observed harm while retaining nearly the same mean acquisition and executing every test edit. On Qwen, no additional exclusion is selected because B1 is already benign. These outcomes show that the value of calibrated exclusion depends on the model-specific alignment of candidate benefit, damage, and forecast uncertainty.

E2 also records paraphrase and neighborhood outcomes alongside the primary protected-loss measure. These broader preservation measurements are descriptive, not CRC-controlled, but they help check that the benign primary result is not being interpreted as a complete account of post-edit behavior.

Figure 2 summarizes the joint risk and utility trade-off.

Extended B2, oracle, selection, and forecasting diagnostics are in Appendix C.

6 Analysis and Ablations

The primary matched comparison is CP versus B1 because both share the same forecasters and six-candidate action space; the difference is decision-level calibration. Two diagnostics clarify the result.

Decision-level versus prediction-level protection. B2 conformalizes candidate damage predictions before selection and is more conservative than CP. On TinyLlama, B2 yields 1/90 harms, 35 abstentions, and mean $G = 4.870$, versus 7.8% harm, 17 abstentions, and $G = 6.156$ for CP. On Gemma, B2 yields 0/90 harms, five abstentions, and $G = 20.287$, versus 8.9% harm, no abstentions, and $G = 21.308$ for CP. These point estimates illustrate the utility/coverage cost of prediction-level protection, but no paired CP versus B2 inference was preregistered, so we make no superiority claim.

Forecasting and uncertainty. Across E1, gain R^2 is 0.994 on TinyLlama and 0.992 on Gemma, versus damage R^2 of 0.485 and 0.656; damage uncertainty is positively associated with absolute damage error ($\rho = 0.557$ and 0.459). E2 Qwen shows similarly strong forecast fit (gain $R^2 = 0.979$, damage $R^2 = 0.918$) and a positive uncertainty and error association ($\rho = 0.505$), yet selects $\lambda = 0$ because no additional exclusion is warranted. The S2 full-scope result ($\rho \approx 0.07$) shows that this relation can collapse exactly where structural risk concentrates. Ensemble disagreement is therefore a useful empirical signal, not a universal certificate.

High gain R^2 or a positive aggregate uncertainty and error association does not alone validate a selector: the S2 full-scope result shows that uncertainty can lose resolution exactly where risk concentrates. Forecast diagnostics are therefore evidence about a fixed regime, including its action space and selected subgroups, rather than a universal certificate.

These diagnostics reinforce the method's intended scope: calibration helps when naive selection faces real risk and the calibration regime remains relevant, but it cannot rescue arbitrary action-space shift. Extended tables, selection distributions, oracle statistics, and forecasting plots are in Appendix C.

7 Related Work

Model editing modifies pretrained models while trying to preserve unrelated behavior. Representative methods include MEND (Mitchell et al., 2022a), ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), and SERAC (Mitchell et al., 2022b). CP addresses a different decision: it treats a *complete candidate update procedure* as the action and forecasts which candidate should be executed before any candidate outcome is observed. It can therefore compare ordinary optimization procedures and editors that make their own internal choices, without claiming to replace either kind of method.

Continual-learning work studies the balance between learning and retaining previous behavior across sequences of tasks, while influence-style methods and update-effect predictors estimate how training examples or parameter changes affect model behavior (Koh & Liang, 2017). CP differs by coupling pre-update consequence forecasting to an explicit finite-action selection rule.

Ensemble methods provide useful epistemic diagnostics (Lakshminarayanan et al., 2017), but CP does not interpret ensemble disagreement as a calibrated probability of harm. Instead, conformal risk control (Angelopoulos et al., 2024) calibrates the bounded harmful-update loss induced by the final decision. Candidate switching makes raw loss non-monotone in scalar conservatism, motivating the all-supersets forbidden-set closure in Section 3. The resulting statement applies only under the fixed-pipeline and exchangeability assumptions already stated; it is not a universal safety guarantee.

8 Limitations and Discussion

CP reduces harmful selections in E1, but its validity is conditional on the specified candidate family and calibration regime. Split CRC controls expected average harmful-update loss for future edits exchangeable with calibration examples under a fixed pipeline; it is not a per-edit, realized-rate, online-deployment, or arbitrary-shift guarantee. The controlled S2 experiment shows why: structural action-space change can produce systematic damage underestimation and weaken the uncertainty signal exactly where risk concentrates.

Damage forecasting remains the central bottleneck. Acquisition is comparatively predictable, whereas collateral damage varies with model family, scope, and candidate structure. Conservative policies can always reduce harm by excluding powerful interventions, so low risk must be interpreted

jointly with utility and abstention. This trade-off is visible on TinyLlama (17/90 CP abstentions) but not on Gemma (0/90).

The method also inherits limitations from its finite action menu and protected set. If all candidates are harmful or ineffective, selection cannot recover a good intervention; and preserving 180 protected factual items does not imply preservation of all knowledge, reasoning, style, or safety-relevant behavior. Our external evidence remains limited to factual editing across two E1 risk-bearing regimes and one independent Qwen2.5-7B confirmation, so transfer to continual learning, preference optimization, personalization, or safety interventions remains open. E2's paraphrase and neighborhood outcomes broaden descriptive preservation measurement, but are not CRC-controlled. Finally, offline training/calibration requires evaluating every candidate outcome and protected loss vector; deployment executes only the selected candidate, but fingerprint extraction still incurs gradient cost.

These limitations also explain the controlled negative results. Calibration is most useful when multiple candidates genuinely compete and naive selection creates material risk; in benign menus it may only sacrifice utility, while under strong action-space shift it may fail to transfer. CP should therefore be viewed as a calibrated decision layer for a specified intervention regime, not as a universal model-update safety mechanism.

Extending CP to a new updater, scope, or protected-behavior definition requires re-establishing its pipeline: the menu determines the ranked counterfactuals, calibrated forbidden sets, and abstention options. Broader protected evaluation can also alter the calibration target and cost.

9 Conclusion

We introduced **Counterfactual Plasticity**, which frames post-training adaptation as a calibrated decision among candidate updates before any update is committed. Using pre-update fingerprints, it forecasts acquisition utility and collateral risk, accounts for uncertainty in the latter, and calibrates update selection against harmful-update risk.

Controlled experiments show why this distinction matters: acquisition utility is easier to forecast than collateral risk, and naive forecast-based selection can fail when high-gain, high-damage interventions compete. Prospective frozen-protocol factual editing shows that CP reduces harmful-update rates from 24.4% to 7.8% on TinyLlama and from 23.3% to 8.9% on Gemma; in an independent Qwen2.5-7B confirmation, it matches the already benign B1 policy (both $1/90$ harms) rather than imposing extra conservatism.

The method does not guarantee safety under distribution or action-space shift. Its broader contribution is a decision framing for model updating: when several plausible interventions are available, forecast their consequences and calibrate which update should become real.

The risk statement has a deliberately narrow operational meaning: under the frozen evaluation pipeline and exchangeability with its calibration edits, it concerns expected harmful-update loss of the selected policy, with abstention executing no update. It does not certify each individual edit, the correctness of every damage forecast, or preservation beyond the protected-loss definition. The controlled transport failure makes this boundary substantive rather than procedural. A calibrated threshold should travel with the candidate menu, feature/forecast pipeline, and protected evaluation on which it was calibrated.

The three-regime contrast shows why risk, utility, and abstention must be reported together. CP is not an argument for more conservative updating; it exposes and calibrates the trade-off among plausible parameter-space futures.

AI use statement

Generative AI tools were used for literature discovery, research ideation and discussion, code assistance, drafting and language editing, and manuscript formatting. All research questions, experimental design choices, implementation decisions, experimental runs, analyses, interpretations, and reported results were determined or verified by the authors. AI-generated suggestions, text, and code

were reviewed before inclusion or use, and reported experimental results were verified against the underlying experimental artifacts. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies model-update selection and the prediction of collateral effects from parameter updates. The experiments use existing language models and factual-editing benchmarks and do not involve human subjects or personally identifiable information. Although risk-aware update selection may reduce unintended model changes, it does not provide a guarantee of safe behavior under arbitrary deployment conditions. In particular, the proposed risk-control results apply only under the stated experimental assumptions and candidate-update distributions.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the candidate update procedures, forecasting features, training, calibration, and test splits, damage threshold, uncertainty estimation, split-CRC calibration procedure, decision rule, abstention mechanism, and evaluation metrics. Random seeds, model and tokenizer revisions, candidate mappings, and the frozen E1/E2 protocols are documented in the appendix and supplementary material. E2 provenance records the exact Qwen2.5-7B revision, BF16 execution, manifests, raw-data and analysis hashes, and its single locked-test evaluation after training and calibration were fixed. The code, experimental configurations, and supporting reproducibility artifacts are available at the following anonymous repository: `https://anonymous.4open.science/r/plasticity-forecasting-2C99`.

A METHOD DETAILS FOR COUNTERFACTUAL PLASTICITY

This appendix records the secondary mechanics underlying Sections 2 to 3. It is not required to apply the main-text decision rule, but specifies the frozen E1 construction; E2-specific provenance is provided in the supplementary material.

A.1 FINGERPRINTS, FORECASTS, AND DIAGNOSTIC ORACLE

For each candidate, the pre-update fingerprint combines target-state features (including target loss), global and layerwise gradient magnitude and distribution statistics, gradient-concentration measures, protected-gradient alignment, and candidate metadata such as learning rate, duration, parameter scope, and layer location. No realized post-update quantity, including gain, damage, harm, or a statistic calculated by executing the candidate, is part of $\phi(\theta, x, c)$.

Forecasting candidate consequences. Separate predictors are fit for $\widehat{G}_c = f_G(\phi(\theta, x, c))$ and $\widehat{D}_c = f_D(\phi(\theta, x, c))$. With B damage-forecast ensemble members $\widehat{D}_c^{(1)}, \ldots, \widehat{D}_c^{(B)}$, we use

$$\mu_D(c) = B^{-1} \sum_{b=1}^{B} \widehat{D}_c^{(b)}, \qquad \sigma_D(c) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\widehat{D}_c^{(b)} - \mu_D(c))^2},$$

and let $\mu_G(c)$ denote the acquisition forecast used for ranking. The forecasting architecture, preprocessing, ensemble seeds, and feature schema are frozen before E1 evaluation. Ensemble disagreement is an empirical candidate-dependent stability signal, not a calibrated posterior standard deviation; controlled experiments examine both its useful and failure regimes.

For diagnostic analysis only, an oracle observes every realized candidate outcome and selects $\arg\max_{c: D(c) \le \tau} G(c)$ when that set is nonempty; otherwise it selects $\arg\min_c D(c)$. It never abstains and can therefore be harmful when every candidate exceeds the budget. This oracle is not deployable.

A.2 FORBIDDEN-SET CALIBRATION DETAILS

Decision-level risk. For a forbidden set F, $\pi(F)$ selects the highest-μ_G member of $\mathcal{C} \setminus F$. Ties are resolved prospectively by lower predicted damage, then lower predicted damage standard deviation,

then candidate name. It abstains if $\mathcal{C} \setminus F = \varnothing$. The raw loss is

$$\ell_0(y, F) = \begin{cases} \mathbb{I}[D(y, \pi(F)) > \tau], & \pi(F) \neq \perp, \\ 0, & \pi(F) = \perp. \end{cases}$$

Monotone risk closure. The calibration loss is the all-supersets closure

$$\widetilde{\ell}(y, F) = \max_{F' \supseteq F} \ell_0(y, F').$$

It is set-monotone: if $F_1 \subseteq F_2$, then the supersets of F_2 are a subset of those of F_1, so $\widetilde{\ell}(y, F_1) \geq \widetilde{\ell}(y, F_2)$. Also F itself is a permitted superset, proving the pointwise raw-loss bound $\ell_0(y, F) \leq \widetilde{\ell}(y, F)$. This is why the closure accommodates non-monotonic candidate switching; a scalar-path maximum is neither its definition nor generally equivalent to it.

For each of the six-candidate E1 calibration edits, we exhaustively enumerate all $2^6 = 64$ forbidden sets to evaluate this closure. A numerical value λ maps to $F_\lambda(x) = \{c : \mu_D(x, c) + \lambda\sigma_D(x, c) > \tau\}$. The fixed E1 family has 61 numerical values from 0 to 15 in increments of 0.25, followed by `FORCED_ABSTAIN`; the latter maps to $F = \mathcal{C}$ and has zero raw and closure loss. The selected state is the first one satisfying the split-CRC criterion. This order is fixed before the evaluation and the terminal state is used only when no numerical state is eligible.

A.3 Split-CRC criterion and scope

With n_{cal} calibration edits, empirical closure risk is

$$\widehat{R}_{\text{cal}}(\lambda) = n_{\text{cal}}^{-1} \sum_{i=1}^{n_{\text{cal}}} \widetilde{\ell}(y_i, F_\lambda(x_i)).$$

For bounded loss, split CRC (Angelopoulos et al., 2024) selects the first state satisfying

$$\frac{n_{\text{cal}} \widehat{R}_{\text{cal}}(\lambda) + 1}{n_{\text{cal}} + 1} \leq \alpha.$$

In E1, $n_{\text{cal}} = 90$ and $\alpha = 0.10$, yielding $(1 + \sum_i \widetilde{\ell}_i)/91 \leq .10$, at most eight closure losses, and resolution $1/91 \approx 1.10\%$. Forecast models use the training split; this calibration split and the final test split are untouched by forecaster fitting, calibration, candidate-selection design, and hyperparameter revision.

Scope of the risk statement. Fixing training data and every pipeline element (candidate menu, features, forecasters, grid, tie-breaking, forced-abstain endpoint, reset protocol, and damage rule), and assuming the calibration and future edits are exchangeable conditional on that training data, bounded loss, no leakage, and no outcome-adaptive pipeline change, split CRC yields

$$\mathbb{E}[\widetilde{\ell}\{Y_{\text{new}}, F_{\widehat{\lambda}}(X_{\text{new}})\} \mid \text{fixed training data}] \leq \alpha.$$

The pointwise bound transfers this to expected average raw harmful-update loss. This E1 statement does not apply to the controlled v0.1 10-fold CV-CRC procedure, and it does not imply a conditional, per-edit, universal, realized-confidence, or arbitrary-shift guarantee. It relies on the specified finite candidate menu and requires offline access to every counterfactual candidate outcome for each calibration edit.

A.4 Inference pseudocode

Once forecasts and $\widehat{\lambda}$ are frozen, CP computes $\phi(\theta, x, c)$, $\mu_G(c)$, $\mu_D(c)$, and $\sigma_D(c)$ for every candidate; forms $\mathcal{A}_{\widehat{\lambda}}(x)$; selects its highest-$\mu_G$ member using the fixed tie-breaking rule; and otherwise returns $\perp$. Only that selected candidate is executed. All other counterfactual alternatives remain hypothetical. Algorithm 1 states the same procedure.

Algorithm 1 Counterfactual Plasticity

Require: Model θ, experience x, candidates $\mathcal{C}$, frozen forecast models, $\widehat{\lambda}$, and damage budget τ
1: **for** $c \in \mathcal{C}$ **do**
2: Compute $\phi(\theta, x, c)$ and obtain $\mu_G(c)$, $\mu_D(c)$, and $\sigma_D(c)$
3: **end for**
4: $\mathcal{A} \leftarrow \{c \in \mathcal{C} : \mu_D(c) + \widehat{\lambda}\sigma_D(c) \leq \tau\}$
5: **if** $\mathcal{A} = \varnothing$ **then**
6: **return** $\perp$
7: **else**
8: **return** the fixed-tie-break highest-μ_G candidate in $\mathcal{A}$
9: **end if**

B Extended Controlled Diagnostics

Controlled environment and v0.1 calibration. The controlled benchmark uses facts, symbols, rules, and arithmetic with independent checkpoint resets for every candidate. Fingerprints include target loss, gradient magnitude and layer profiles, gradient concentration, protected-gradient alignment, and candidate metadata. The original controlled v0.1 policy uses a separate 10-fold CV-CRC procedure over 90 development episodes; it is not the E1 split-CRC calibration statement.

Benign menus. On the original eight-candidate menu, the point-estimate selector reaches mean $G = 9.717$ with one harm in 30 held-out episodes, while B1 reaches $G = 9.702$ with zero harms. A purely acquisition-driven selector attains higher gain but harms 17/30 episodes. Increasing experience difficulty without changing the menu yields 0/400 B1 harms; a denser strength menu yields 1/400. These regimes show that available harmful candidates do not by themselves create policy risk when those candidates are not competitive under forecasted utility.

Structural transport failure. The S2 menu varies learning rate, duration, scope, layer location, and width. Three full-scope candidates combine high acquisition with harmful-update rates of roughly 17.0%, 37.0%, and 57.3%. Frozen B1 then produces 206/400 harms (51.5%), all from full-scope selections; frozen CP lowers this to 154/400 (38.5%). The within-full-scope uncertainty and error correlation is only $\rho \approx 0.07$. Retraining on S2 reduces harm substantially, indicating a transport failure in damage forecasting rather than a fundamental inability to separate every risky update. More aggressive scope-conditional bounds or direct harm classifiers can approach zero development harm largely by excluding the high-gain full-scope region, exposing the utility and risk trade-off.

C Extended External Protocol and Diagnostics

E0 development screen. E0 evaluates Qwen, TinyLlama, and Gemma on 60 shared factual edits with 120 protected items and six fixed candidate procedures. Qwen has 0/60 risk-pressure edits. TinyLlama has 16/60 risk-pressure edits and benefit and damage trade-offs on 46/60; Gemma has 48/60 risk-pressure edits and trade-offs on 59/60. All six oracle candidate identities occur for TinyLlama and Gemma, with no globally dominant candidate. E0 is descriptive development-stage screening only.

E1 outcome construction and audits. The 360 edits yield 2160 candidate outcomes per model and 4320 total. Every edit and candidate pair is independently restored, fingerprinted, executed, and scored using the full 180-item protected-loss vector. The rerun audits reproduce aggregate G and D exactly and contain no missing candidates, duplicate edit and candidate rows, or silent execution failures. The prospectively amended uniqueness rule uses the subject-rendered CounterFact rewrite prompt; after E0 exclusion and deduplication, 21,376 eligible examples remain before freezing the 360 edit and 180 protected manifests.

B2 and oracle reference. B2 protects candidate damage predictions before selection and is therefore more conservative than CP's decision-level calibration. On TinyLlama, B2 has 1/90 harms,

Table 2: Locked E1 descriptive comparison of B2 and CP. No paired CP versus B2 confidence interval was preregistered.

Model	B2 harm	B2 mean G	CP harm	CP mean G
TinyLlama	1/90 (1.1%)	4.870	7.8%	6.156
Gemma	0/90 (0.0%)	20.287	8.9%	21.308

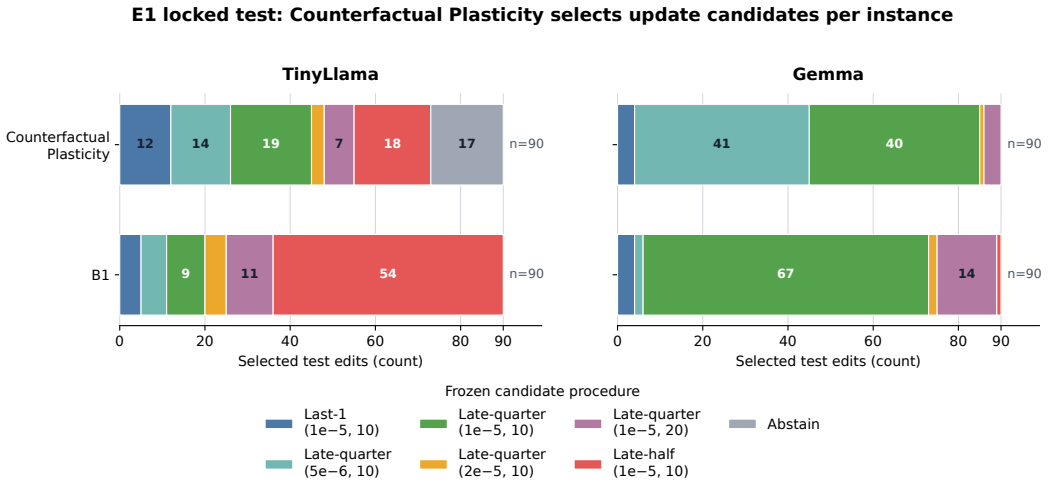


Figure 3: **Locked E1 candidate-selection distributions.** CP uses all six candidates on TinyLlama with 17 abstentions and five of six on Gemma with no abstention.

35 abstentions, and mean $G = 4.870$; on Gemma it has 0/90 harms, five abstentions, and mean $G = 20.287$. No paired CP versus B2 inference was preregistered. The diagnostic oracle selects the highest-gain candidate among those with realized $D \leq \tau$; if no safe candidate exists, it selects the minimum-damage candidate and never abstains. This fallback produces three TinyLlama oracle harms. Acquisition regret is $G(c^\star) - G(\pi(x))$ with $G(\perp) = 0$.

Adaptive selections. On TinyLlama, CP selects all six candidate procedures and abstains 17 times; on Gemma it selects five candidates and never abstains. Thus the lower harmful rate does not come from universal abstention or a single static update.

Forecasting diagnostics. On E1 test candidates, gain MAE is 0.134 for TinyLlama and 0.160 for Gemma, with $R^2 = 0.994$ and 0.992. Damage MAE is 0.094 and 1.882, with $R^2 = 0.485$ and 0.656. Candidate-level σ_D correlates with absolute damage error at $\rho = 0.557$ ($p = 2.25 \times 10^{-45}$) for TinyLlama and $\rho = 0.459$ ($p = 1.69 \times 10^{-29}$) for Gemma; mean σ_D is also larger on harmful than non-harmful candidate outcomes. The S2 full-scope result ($\rho = 0.072$) shows that this diagnostic relation need not transfer under structural shift.

Frozen hypotheses. H1 (lower harmful-decision rate than B1 in the same six-candidate space) is supported on both models. For H2, the observed behavior satisfies its intended qualitative content: positive useful acquisition without collapse to a single candidate or universal abstention. We avoid a binary claim absent an explicit frozen quantitative threshold. H3, the descriptive positive association between candidate-level damage uncertainty and absolute damage error, is supported on both models and is not itself a risk-control guarantee.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

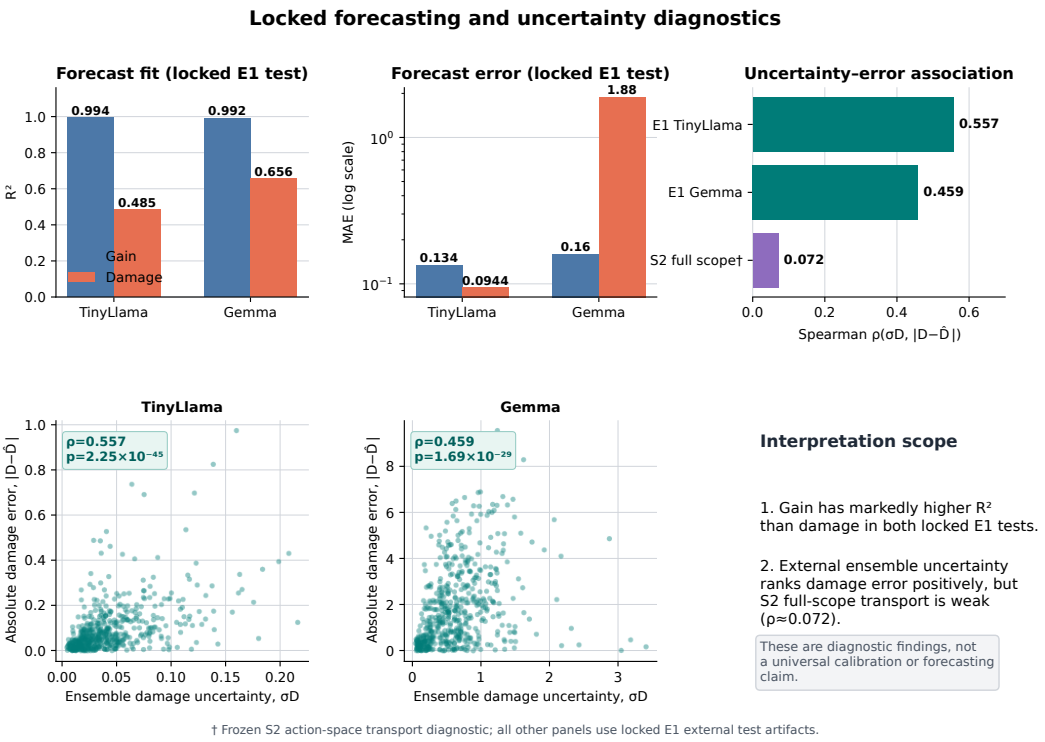


Figure 4: **Forecasting and uncertainty diagnostics.** E1 reproduces the benefit and damage predictability asymmetry and positive association between ensemble damage disagreement and absolute damage error, while the S2 full-scope transport regime shows a near-collapse of this association.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15817–15831, 2022b.