WHEN CAN FIRST-ORDER MODELS OF FINE-TUNING BOUND FORGETTING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a language model on new data can make it forget facts that it should keep. We ask whether measurements taken at the start of a fine-tuning run can bound, for each protected fact, the probability that the run makes the model forget it. In LoRA fine-tuning with stochastic gradient descent on models from 0.6B to 14B parameters, a first-order response model estimated by finite-difference probes predicts changes of per-fact margins with correlation 0.974–0.998. Predictions of forgetting built on this model nevertheless failed; in Qwen3-0.6B, forgetting requires parameter changes far outside the region in which the model was validated. The probes can, however, bound the probability that a margin first falls below a boundary near zero: we derive Freedman and Azuma first-passage bounds for a linear surrogate of the margin and test on new runs whether they hold for the model. The bounds contain a term R that measures how much the response coefficients change during the run. The simplified Freedman bound, which sets $R = 0$, certified most facts but was violated in 14 of 112 conditions, and in these conditions every fact on which it was violated had $R \geq a$, where a is the distance of the fact's margin to the boundary. The complete Freedman bound certifies only facts with $R < a$, and it held in every condition. On the violated facts, the spread of the margin across test runs was a median of 14.6 times the prediction of the response model, so the failures are breakdowns of the model, and in our data they occurred only where $R \geq a$. We found this pattern post hoc and tested it in two preregistered confirmatory studies with 43 new conditions: the complete bound held in all of them, and the simplified bound failed there on only 3 facts, each with $R \geq a$. A screen that decided for a whole condition whether to use the simplified bound failed on new starting states. First-order models of fine-tuning can thus bound forgetting on the facts whose response coefficients change by less than their distance to the boundary.

1 INTRODUCTION

Fine-tuning a language model on new data can make it forget facts that it answered correctly before (Kirkpatrick et al., 2016; Luo et al., 2023), and sequential knowledge editing has the same effect (Gupta et al., 2024). Existing methods predict which examples an update will forget and rank them by risk (Jin & Ren, 2024; 2025). A ranking, however, leaves open how likely a given fact is to survive. Before an update, one would like a guarantee for each protected fact: an upper bound on the probability that the run makes the model forget it.

This paper asks when such a bound can be computed from measurements taken at the start of the run. We study LoRA fine-tuning (Hu et al., 2022) with stochastic gradient descent (SGD) and gradient clipping, and for each protected fact we track its margin, the log-probability gap between the correct answer and the strongest alternative. A fact is forgotten if its margin falls below zero within the horizon. Local models of training, such as influence functions and linearizations (Koh & Liang, 2017; Pruthi et al., 2020; Malladi et al., 2023), are accurate for small parameter changes, and our first-order response model is accurate in the same sense: it predicts margin changes under held-out perturbations with correlation 0.974–0.998 across 6 checkpoints from 0.6B to 14B parameters. Yet five predictions of forgetting built on this model failed on new data. Predicting what happens over a run therefore requires more than local accuracy.

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.

A coupling argument separates the two sources of error and shows what to bound instead (Section 4). A local model that is run forward errs by at most its accumulated local error plus the probability that the process leaves the region in which the model was checked. In Qwen3-0.6B, forgetting a fact requires a parameter change about 18 times larger than the largest change at which we validated the model, so the course of forgetting lies outside what the model describes. The model can, however, bound the probability that a margin first reaches a boundary near zero. We derive such first-passage bounds for a linear surrogate of the margin from Freedman's inequality (Freedman, 1975) and the Azuma inequality, and we test on new runs whether they hold for the model itself. Both contain a term R that measures how much the response coefficients change during the run, and all their inputs are computed before the run (Figure 1). Setting $R = 0$ gives a simplified Freedman bound that certifies most facts.

Our main finding concerns this simplification. Across 112 conditions, the simplified Freedman bound was violated in 14, and every one of the 34 facts on which it was violated had $R \geq a$, where a is the distance of the fact's margin to the boundary (Figure 2a). R comes from the probes and a from a reference run, so both are known before the run. The complete Freedman bound, which keeps R, certifies only facts with $R < a$, and both it and the Azuma bound held in every condition. The failures go beyond the simplification: a bound with the correct coefficients at each time also failed on facts with $R \geq a$, and on the violated facts the spread of the margin across test runs was a median of 14.6 times the prediction. The failures are therefore breakdowns of the first-order model, and they occurred only on facts with $R \geq a$, although most such facts survived (Section 6). Because we found this pattern after an interim look at the data, we preregistered two confirmatory studies on new conditions (Section 7). In their 43 conditions, the complete bound held everywhere, and the simplified bound failed on only 3 facts, each with $R \geq a$, so the location of the failures rests mainly on the post hoc data.

Before this finding, we had tested a different repair: a screen that uses the simplified bound only in conditions whose certified facts have a large median standardized margin. It removed every failure on earlier conditions but failed on new starting states, where the complete bound, the Azuma bound and the per-fact rule held. This contrast is consistent with failures that are confined to individual facts.

Contributions.

1. Evidence that an accurate first-order response model fails to predict forgetting over a fine-tuning run, and a coupling argument for what local probes can bound instead (Sections 3 and 4).
2. Freedman and Azuma first-passage bounds on per-fact forgetting with a coefficient-change term R, and the finding that every observed failure of the first-order model occurred on a fact with $R \geq a$, a condition known before the run (Sections 4 and 6).
3. Prospective studies on five models from 0.6B to 14B parameters, CounterFact facts and a doubled horizon, with two confirmatory tests, a comparison with fitted thresholds and simulation, and a screen that failed (Sections 7 and 8).

2 Related work

Forgetting during fine-tuning. Prior work measures, forecasts, explains or reduces forgetting; we bound its probability for each fact before the run. Example forgetting was first measured during standard training (Toneva et al., 2019). In language models, forgetting has been measured empirically (Luo et al., 2023), related to fine-tuning loss by scaling laws (Kalajdzievski, 2024), and found smaller under LoRA than under full fine-tuning (Biderman et al., 2024). New knowledge also affects what the model already knows (Gekhman et al., 2024), and sequential editing leads to gradual and then abrupt forgetting (Gupta et al., 2024). Forecasting methods rank the upstream examples that an update will forget, from logit changes (Jin & Ren, 2024) or low-rank task-example associations (Jin & Ren, 2025), rather than bound a probability. Theory explains forgetting through the overlap of neural tangent kernels (Doan et al., 2021) and in linear regression (Evron et al., 2022). Mitigation changes the update: elastic weight consolidation (Kirkpatrick et al., 2016), replay of maximally interfered examples (Aljundi et al., 2019), and null-space (Fang et al., 2025) or low-curvature (Ikram et al., 2026) projection. Our bounds leave the update unchanged.

Local models of training. Influence functions (Koh & Liang, 2017), gradient tracing (Pruthi et al., 2020) and kernel linearizations of fine-tuning (Malladi et al., 2023) describe training locally, and interpretability methods built on them are usually validated on short horizons (Bereska & Gavves, 2024; Mueller et al., 2025). We ask what such a local model supports over a full fine-tuning run.

Bounds and risk control. Our bounds apply maximal martingale inequalities (Freedman, 1975) at a fixed horizon; time-uniform versions (Howard et al., 2020) would allow monitoring during the run. Generalization bounds for language models bound a population risk after training (Lotfi et al., 2024), whereas we bound a per-fact event during a run. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023), risk-controlling prediction sets (Bates et al., 2021), conformal risk control (Angelopoulos et al., 2024), Learn-then-Test (Angelopoulos et al., 2021) and selective prediction (Geifman & El-Yaniv, 2017) calibrate thresholds on exchangeable data. Our conditions were chosen by study design, which rules out exchangeability, so we fixed the thresholds of our rules on earlier conditions and tested them on new ones.

3 Setting and the first-order response model

Two-task fine-tuning. A model is first fine-tuned on task A, which contains the protected facts, and then on task B. We use two data sources. With *synthetic facts*, task A teaches 320 associations between invented entities and one of four answer labels, each in two paraphrase templates. We protect 64 of them and query each with a third template reserved for evaluation. Task B teaches 128 new associations of the same form. With *CounterFact* (Meng et al., 2022), the 64 protected facts are true facts that the pretrained model already answers correctly, so task A is skipped, and task B trains on 256 counterfactual rewrites of other subjects, which is knowledge editing by fine-tuning. Llama-3.2-1B and Gemma-3-12B denote the instruction-tuned checkpoints. All models are adapted with LoRA of rank 8 on all linear layers (Hu et al., 2022) over 4-bit weights (Dettmers et al., 2023). Task A uses AdamW. A run proceeds only if the accuracy on the protected facts at the start of task B reaches a fixed threshold (0.75 for synthetic facts, 0.95 for CounterFact). Task B uses SGD with batch size 8 (4 for CounterFact), learning rate η, and gradient clipping at norm $G_0 = 1$. We call the parameters at the start of task B the *starting state*. A *condition* is a model, a data source, a task-B learning rate and a seed; a *configuration* is the model, data source and learning rate of a condition. The seed determines task-A training and the minibatch orders of the test runs defined below.

Margin and forgetting. For protected fact i with correct answer y_i among four candidates, the margin after t task-B steps is $Y_i(t) = \log p_{\theta_t}(y_i \mid x_i) - \max_{y \neq y_i} \log p_{\theta_t}(y \mid x_i)$. It is positive exactly when the model answers correctly. We evaluate margins on the grid $G = \{0, 10, \dots, T\}$ with horizon $T = 40$ steps ($T = 80$ in one study), and a *forgetting event* occurs if $Y_i(t) < 0$ for some $t \in G$. The *reference run* is one task-B run with a fixed minibatch order, and $m_i(t)$ is the margin along it. A fact is *eligible* if $\min_{t \in G} m_i(t) > 0$. Other facts are already forgotten on the reference run, and our claims concern eligible facts only. A *test run* is a further task-B run from the same starting state with its own minibatch order. We use 32 test runs per condition (24 in two earlier conditions) to measure how often each fact is forgotten.

First-order response model. Let $b_1, \dots, b_k$ ($k = 8$) be an orthonormal basis of the leading directions of the margin gradients of 16 protected facts at the starting state; the basis is computed from the starting state alone. For a perturbation time $s \in G$ and a later time t, the response coefficients $q_{i,t,s} \in \mathbb{R}^k$ are central finite differences, $q_{i,t,s,j} = \big(Y_i^{+j}(t) - Y_i^{-j}(t)\big)/2\epsilon$, where $Y_i^{\pm j}$ is the margin when $\pm\epsilon b_j$ ($\epsilon = 0.02$) is added to the parameters at step s of the reference run and the same minibatch order is replayed up to step t. The first-order response model predicts that a perturbation $\sum_j u_j b_j$ added at step s changes the margin at time t by $q_{i,t,s}^\top u$. We call the finite-difference measurements the *probes*. The probes also estimate, at each perturbation time s, the covariance Σ_s of the clipped task-B gradients projected on the basis. Each probe replays the run from a perturbation time to T, so the probes cost $O(k\,|G|\,T)$ SGD steps, which grows with the square of the horizon. All quantities used by the bounds are computed before the first test run (Figure 1).

Accuracy of the response model. We test the model on held-out perturbations, random combinations of the basis directions whose effect is measured separately and compared with the prediction

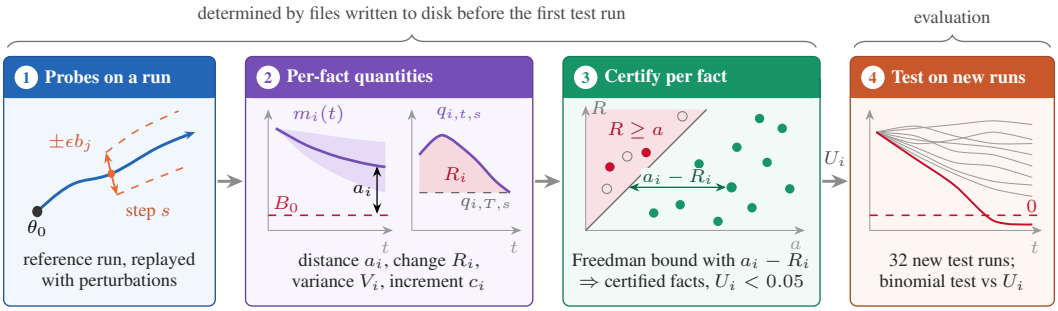

Figure 1: The procedure for one condition (schematic). (1) A reference run gives the margins $m_i(t)$, and probes along it give the response coefficients $q_{i,t,s}$ and the gradient covariance Σ_s. (2) These determine, for each fact, the distance a_i to the boundary, the variance estimate V_i, the increment bound c_i and the coefficient change R_i, the gap between the coefficients at time t and at the end T. (3) The complete bound uses $a_i - R_i$, so it certifies only facts with $R_i < a_i$ (Sections 4 and 5). (4) A binomial test compares the forgetting events in the test runs with U_i.

$q_{i,T,0}^{\top} u$. At task-B learning rate 0.002, Table 3 (Appendix B) reports 15/15 passing validation runs across 6 checkpoints and 3 families (Qwen3, Llama 3.2, Gemma 3), with ρ between 0.974 and 0.998. At the higher learning rates 0.008 to 0.012, 3 further runs gave ρ between 0.939 and 0.995. Two requirements must hold. The basis must come from gradients: a random low-rank basis gives $\rho = 0.057$, because the margin changes that it induces are smaller than the numerical noise of the forward pass. Task B must use SGD, under which the probe responses scale linearly with the perturbation size over the full horizon, whereas under Adam they stop doing so after about ten steps (Appendix B). Despite this accuracy, five preregistered predictions of forgetting built on the model failed on new data (Appendix C), and Section 4 explains why.

4 Bounds from local probes

4.1 What a local model can determine about a run

We first ask what a local model can determine about a whole run. Let P be the law of the training process and Q the law obtained by running a local model forward. Suppose that the two transition kernels agree to within ε_t in total variation on a region $\mathcal{N}$ that contains the starting state, and let τ be the time at which the process first leaves $\mathcal{N}$.

Theorem 1 (Coupling bound). *For any event A determined by the path up to T, $|P(A) - Q(A)| \leq P(\tau \leq T) + \sum_{t \leq T} \varepsilon_t$, and both terms are attained.*

Theorem 2 (Non-identifiability). *For any neighborhood $\mathcal{N}$, order d, and $p \in (0, 1]$, there exist smooth systems that agree on $\mathcal{N}$ in value, in all derivatives up to order d, and in noise law, but whose probabilities of an event determined at step 2 differ by p.*

Theorem 1 separates the accumulated local error from the probability of leaving the region in which the local model was checked; we use it to structure the problem, because response correlations measure accuracy along a few directions rather than a total-variation distance. In Qwen3-0.6B, a single forgetting event needs a parameter change about $18\times$ larger than the largest change at which the response model was validated (Appendix B), and by Theorem 2 what happens outside $\mathcal{N}$ is unidentifiable from local data. We therefore bound the probability that a linear surrogate of the margin first crosses a boundary, a concentration problem, and test on new runs whether the bound transfers to the model (proofs in Appendix A).

4.2 Two first-passage bounds

Fix a protected fact and drop the index i. The linear surrogate applies the first-order response model to the noise of the task-B updates: $\widehat{Y}_t = m_t + \sum_{s \leq t} \eta\, q_{t,s}^{\top} \xi_s$, where ξ_s is the clipped gradient at step

s, projected on the basis and centered at its conditional mean. A clipped gradient and its conditional mean each have norm at most G_0, so $\|\xi_s\| \le 2G_0$. Let $S_t = \sum_{s \le t} \eta\, q_{T,s}^{\top} \xi_s$ be the noise term with the coefficients of the final time T. It is a martingale whose increments are bounded by

$$c = 2\eta G_0 r, \qquad r = \max_s \|q_{T,s}\|. \tag{1}$$

Let $a = \min_{t \in G} m_t - B_0$ be the distance of the reference margin from the boundary $B_0 = 0.01$. Let K be the predictable quadratic variation of S_T estimated from the probes, and let $V = \kappa K$ with $\kappa = 2$. Let $R = 2\eta G_0 \max_{t \in G} \sum_{s<t} w_s \|q_{t,s} - q_{T,s}\|$ bound the change of the response coefficients over the grid, where w_s is the width of the grid interval that starts at s.

Theorem 3 (First-passage bounds on the grid)**.** *Let $a' = a - R > 0$. If the increments of S are bounded by c, then*

$$\Pr\big(\exists t \in G : \widehat{Y}_t < B_0\big) \le U^{\mathrm{A}} = \exp\Big(-\frac{a'^2}{2Tc^2}\Big).$$

If, in addition, the predictable quadratic variation of S is at most V, the same probability is at most $U^{\mathrm{F}} = \exp\big(-a'^2/(2(V + c\,a'/3))\big)$.

The proof writes $\widehat{Y}_t - m_t = S_t + D_t$ with $|D_t| \le R$ and applies the maximal forms of the Azuma and Freedman inequalities, which cover all grid times at once. We call U^{F} the Freedman bound and U^{A} the Azuma bound. A bound *certifies* an eligible fact if its value is below 0.05; this is a statement about one fact, and a statement about a set $\mathcal{L}$ of facts at level α requires $\sum_{i \in \mathcal{L}} U_i \le \alpha$. A forgetting event implies a crossing of B_0, so each bound also bounds the probability of forgetting under the surrogate; for the actual model, a certified fact is a prediction that we test on new runs.

The evaluated bounds. The *complete* Freedman bound uses the measured R. The *simplified* Freedman bound and the Azuma bound, as we evaluate it, set $R = 0$, so Theorem 3 covers them for the surrogate $m_t + S_t$ with final-time coefficients, which differs from $\widehat{Y}_t$ by up to R; they are justified only for facts whose coefficient change is small compared with a. The Azuma bound replaces the variance estimate by the worst case Tc^2 and is much more conservative.

4.3 Assumptions of each bound

Assumption 1 (Uniform remainder)**.** *With probability at least $1 - \delta_r$, $\sup_{t \in G} |Y_t - \widehat{Y}_t| \le \gamma$, where Y_t is the actual margin. Under this assumption, each bound applies to the actual margin with $a' - \gamma$ in place of a', plus δ_r.*

Assumption 2 (Bounded increments)**.** *The increments of S are bounded by c in Equation 1. Both bounds need this assumption, and gradient clipping enforces it.*

Assumption 3 (Variance estimate)**.** *V is at least the realized predictable quadratic variation of S. Only the Freedman bounds need this assumption.*

Assumption 1 remains unverified, and a high response correlation is insufficient to establish it. Whether the bounds hold for the actual model is therefore an empirical question, which we test on new runs; Section 6 shows on which facts the assumptions fail. Estimation error in V can be absorbed at a stated confidence level (Appendix A).

5 Rules that certify facts

A rule computes a value U_i for each fact before the run and certifies the eligible facts with $U_i < 0.05$; each bound of Section 4 is such a rule. Before we understood where the simplified bound fails, we designed on earlier conditions two rules that keep part of its certified facts, and one baseline that uses only the reference run. The screen threshold was fixed before the prospective studies, and the other two thresholds after the first prospective condition had finished (Section 7).

Screened bound. The first rule decides for a whole condition. Let $\mathcal{C}$ be the set of eligible facts that the simplified bound certifies, let $z_i = a_i/\sqrt{V_i}$ be the standardized margin of fact i, and let $\tilde{z}$ be the median of z_i over $\mathcal{C}$. The *screen* is the test $\tilde{z} \ge 30$. If the condition passes the screen, the *screened*

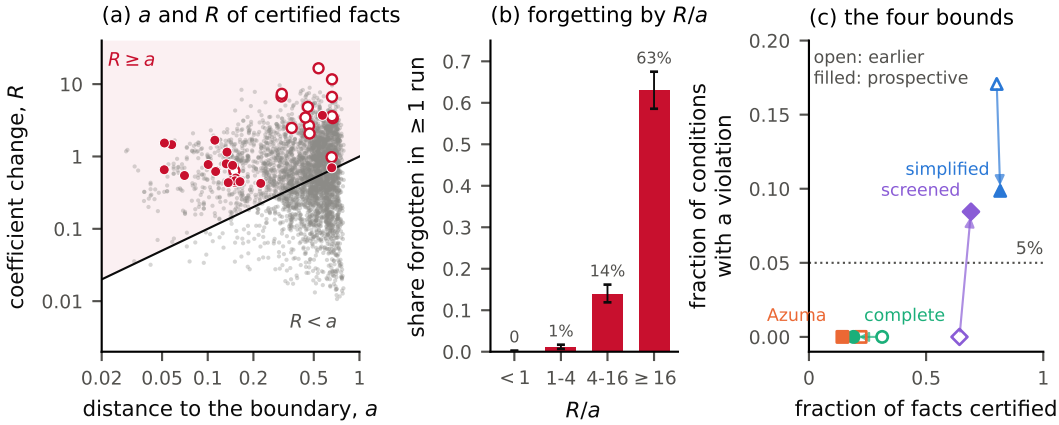

Figure 2: **(a)** The 3,748 facts that the simplified Freedman bound certifies, by their distance a to the boundary and coefficient change R, both computed before the run. Red: facts on which the simplified bound was violated, in earlier (open) and prospective (filled) conditions; all lie in the shaded region $R \geq a$, which the complete bound excludes. **(b)** Share of eligible facts forgotten in at least one test run, by R/a, with Wilson 95% intervals (Appendix H). **(c)** For the four bounds, the fraction of eligible facts certified against the fraction of conditions with a violation, on the 41 earlier (open, post hoc) and 71 prospective (filled) conditions.

bound[1] uses the simplified Freedman bound for every fact of the condition (the *Freedman branch*); otherwise it uses the Azuma bound (the *Azuma branch*), on the rationale that large standardized margins tolerate an underestimated variance. Because the screen uses only quantities that exist before the run, the selected bound holds whenever each branch bound holds (Proposition 1).

Per-fact rule and margin threshold. The *per-fact rule* decides for each fact. It certifies a fact if the simplified bound certifies it and its own z_i is at least $\theta_z = 56$. The *margin threshold* is a baseline that uses only the reference run. It certifies a fact if its distance a_i is at least $\theta = 0.66$, and it claims a probability of 0.05 for that fact. Both thresholds are the smallest values at which every earlier condition passes the test below, and the screen threshold comes from an earlier screening rule (Appendix D).

Testing a rule. For each certified fact, a one-sided binomial test compares the number of test runs in which the fact is forgotten with its value U_i (with 0.05 for the margin threshold), with a Bonferroni correction over the 64 facts of the condition. A rule is *violated on a fact* if this test is significant, and a condition is *violating* if the rule is violated on at least one of its facts. A condition is *evaluable* if it has at least 8 eligible facts.

6 Where the simplified bound fails

We first examine the 41 of the 45 earlier conditions that have stored probe files, from which R can be computed (Appendix F); they contain 1,818 eligible facts. Their outcomes existed before we defined the rules of Section 5, so this analysis is post hoc (first row of Table 1).

The simplified Freedman bound certified 80.1% of eligible facts, but it was violated in 7 of the 41 conditions. Figure 2a shows where the violations occurred. Each of the 14 facts on which it was violated in these conditions lies above the line $R = a$: its response coefficients changed during the run by more than the distance of its margin to the boundary. The complete bound replaces a by $a - R$, so it certifies only facts with $R < a$, and it held in every condition.

In Theorem 3, R bounds the difference between the surrogates with final-time and with per-time coefficients, but the failures have a different source. After these outcomes were known, we evaluated

[1] Our code calls this rule ValidityShield.

a *per-time bound*, which applies Theorem 3 at each grid time with the coefficients of that time and combines the grid times with a union bound, so that the term R drops out. Over all 112 conditions, including the prospective ones of Section 7, it certified 74.5% of eligible facts but was violated in 5 conditions, and each of the 14 facts on which it was violated had $R \geq a$. On such facts, the first-order model fitted along the reference run describes the test runs poorly: on the facts that violated the simplified bound, the standard deviation of the final margin across test runs was a median of 14.6 times the prediction, against 1.6 on facts with $R < a$ (Appendix G). A large R means that the response of the margin to a perturbation changes strongly during the run, so a model fitted along one run is least likely to describe other runs. We therefore read $R \geq a$, computed before the run, as a necessary condition for such a failure in our data. It is far from sufficient, because the median ratio over all facts with $R \geq a$ is 2.7, and forgetting rises steadily with R/a above 1 (Appendix H). The complete bound holds because it excludes every fact that meets the condition.

The complete bound therefore certifies fewer facts. In our data, $R \geq a$ was necessary for a violation but far from sufficient: 53.3% of the facts that the simplified bound certified had $R \geq a$, and most of them survived every test run. The complete bound excludes all of them and certified 31.2% of eligible facts, more than the Azuma bound (22.1%).

The screened bound and the per-fact rule appeared to avoid this loss. Every earlier condition with a violation of the simplified bound had a median standardized margin below 30 (at most 29.1; Appendix D), so the screened bound held there and certified 64.2% of eligible facts; the per-fact rule (37.1%) and the margin threshold held by construction. Only new starting states can test these rules.

7 Prospective evaluation

7.1 Design

We ran four prospective studies on new starting states, each fixed with its analysis script before its first run. The *main study* has 33 conditions in six condition sets: Qwen3-0.6B, Llama-3.2-1B, Qwen3-1.7B and Qwen3-8B on synthetic facts, and Qwen3-0.6B and Llama-3.2-1B on CounterFact facts that the pretrained model already answers correctly, with new seeds (71, 73 and 79) and learning rates that include rates at which the simplified bound had failed before. The *horizon study* reruns eight of its starting states for $T = 80$ steps, and the *14B study* runs Qwen3-14B with new seeds. Four Qwen3-8B and Qwen3-14B conditions fell below the task-A accuracy threshold and were stopped before any bound was computed; a further batch with two new seeds was run for each model, and the Qwen3-8B row of Table 1 consists of this batch. Before the first test run of each condition, our code wrote the probe results, the simplified, Azuma and screened bounds and the branch to disk; the other rules are deterministic functions of these files.

The main question was the screened bound, judged by three criteria. The *violation criterion* allows at most the 95th percentile of $\mathrm{Binomial}(N, 0.05)$ violating conditions among N evaluable ones, a count that a rule violating each condition with probability 0.05 exceeds with probability at most 0.05. The *certification criterion* requires that the screened bound certify at least 40% of eligible facts and twice as many as the Azuma bound, and the *assignment criterion* that every condition in which the simplified bound fails be assigned to the Azuma branch. The per-fact rule and the margin threshold were added after the first condition had finished.

After an interim look at 15 conditions, we added a confirmatory test on the conditions whose test runs started later (28 runs excluded): the complete bound must meet the violation criterion, and the *location criterion* requires that at least 90% of the facts on which the simplified bound is violated have $R \geq a$. Because few such facts occurred, a *second confirmatory study* applied the same criteria, with at least five violated facts needed for a decision, to 24 new conditions: Qwen3-0.6B at learning rates 0.012, 0.018 and 0.021 and Qwen3-1.7B at 0.012 on synthetic facts, with six new seeds.

7.2 Results

The screened bound failed on new starting states. In the 31 evaluable conditions of the main study, it had 5 violating conditions (at most 4 allowed), because the screen assigned conditions with $\tilde{z}$ between 32 and 56 to the Freedman branch, where the simplified bound was violated; it failed its violation criterion and failed its assignment criterion. It met the certification criterion, and in the

Table 1: Main results. For each condition set and rule: the number of violating conditions (viol.) and the percentage of eligible facts that the rule certifies (%). n is the number of evaluable conditions, and events is the number of forgetting events of eligible facts over all test runs. The rows of condition sets contain the main, horizon and 14B studies; the Qwen3-8B row contains the added batch (seeds 79 and 83). The confirmatory set contains the conditions of these studies whose test runs started after the confirmatory test was defined.

			Simplified Freedman		Screened bound		Complete Freedman		Azuma bound		Per-fact rule		Margin threshold	
Condition set	n	events	viol.	%	viol.	%	viol.	%	viol.	%	viol.	%	viol.	%
Earlier conditions (post hoc)	41	1,795	7	80	0	64	0	31	0	22	0	37	0	14
Qwen3-0.6B, synthetic	9	603	2	85	2	85	0	29	0	8	0	53	0	6
Llama-3.2-1B, synthetic	6	319	1	71	0	47	0	18	0	18	0	21	0	29
Qwen3-1.7B, synthetic	4	337	1	87	1	49	0	6	0	12	0	11	0	6
Qwen3-8B, synthetic	2	8	0	94	0	94	0	37	0	41	0	38	0	9
Qwen3-0.6B, CounterFact	6	279	0	74	0	58	0	3	0	10	0	27	0	0
Llama-3.2-1B, CounterFact	6	215	2	80	2	80	0	10	0	13	0	28	0	0
Qwen, Llama, $T = 80$	8	55	0	94	0	78	0	18	0	10	0	44	0	21
Qwen3-14B, synthetic	6	144	0	89	0	66	0	39	0	32	0	33	0	21
Confirmatory set	19	785	2	80	2	62	0	15	0	12	0	27	0	3
Second confirmatory study	24	468	1	77	1	69	0	17	0	10	1	40	0	13
All prospective	71	2,428	7	82	6	69	0	19	0	14	1	35	0	12

horizon and 14B studies, whose conditions contained fewer forgetting events, the violation criterion was met in both (family-wise results in Appendix K).

The failures on new starting states had the same form as the earlier ones. Across the 47 conditions of the main, horizon and 14B studies and the added Qwen3-8B batch, every one of the 19 facts on which the simplified bound was violated had $R \geq a$, and the complete bound, the Azuma bound, the per-fact rule and the margin threshold held in every condition (Table 1). On the 19 conditions of the confirmatory test, whose outcomes were unknown when we defined it, the complete bound had 0 violating conditions, so it met the violation criterion, and 2 of the 2 facts that violated the simplified bound had $R \geq a$, so the location criterion was met on few facts.

This pattern explains the failure of the screened bound: a condition with a large median $\tilde{z}$ can still contain a few facts with $R \geq a$, to which the Freedman branch applies the simplified bound.

Over these studies, the per-fact rule certified 33.2% of eligible facts, the complete bound 19.9%, the Azuma bound 16.4% and the margin threshold 11.5%, but the per-fact rule relies on a threshold chosen on earlier data. The updates learn task B: in the main study, task-B accuracy rose by 25.3 percentage points on average.

7.3 Second confirmatory study

All 24 conditions were evaluable, but the new seeds gave fewer forgetting events than earlier seeds at the same settings: 468 in total, and zero in 13 conditions (Table 13). The simplified bound was violated in 1 condition, on 1 fact with $R \geq a$, so the location criterion, which needs at least five violated facts, was left undecided. The complete bound had 0 violating conditions (allowance 3), so it met the violation criterion. The Azuma bound, the per-time bound and the margin threshold also held in every condition, and the per-fact rule was violated on the same fact as the simplified bound. Restricted to facts with $R < a$, the simplified bound held in every condition and certified 28.1% of eligible facts, against 17.0% (complete) and 9.6% (Azuma).

7.4 Pooled forgetting counts

A single condition detects only large excesses over a bound (Appendix I), so Table 2 pools the test runs of all conditions. The facts that the complete bound certifies were forgotten 0 times in 34,256 pairs of a fact and a test run, against 130.8 events that its values allow, and the facts that the Azuma

Table 2: Forgetting events pooled over all 112 conditions (post hoc). For each rule: the percentage of eligible facts certified, the number of conditions with a violation, the pairs of a certified fact and a test run, the forgetting events among them, and the events that the values of the rule allow. The last two rows certify only facts with $R < a$.

Rule	certified (%)	viol. cond.	fact-runs	events	allowed
Simplified Freedman ($R = 0$)	81.0	14	118,928	344	467.5
Screened bound	67.2	6	98,480	202	314.8
Per-time Freedman	74.5	5	109,264	188	513.4
Complete Freedman	23.8	0	34,256	0	130.8
Azuma	17.3	0	24,688	0	178.8
Simplified Freedman on facts with $R < a$	31.0	0	44,912	0	29.8
Every fact with $R < a$ (value 0.05)	31.0	0	44,976	0	2248.8

bound certifies 0 times. The simplified bound produced fewer events than it allows in total, but its violations lie in $R \geq a$. The last two rows separate the two parts of the complete bound: the eligible facts with $R < a$ were forgotten 0 times in 44,976 pairs, a result that depends on R and a alone. The exclusion of $R \geq a$, more than the variance estimate, therefore makes the complete bound valid.

8 Discussion and limitations

What the results support. Probes at the start of a LoRA fine-tuning run with SGD give per-fact bounds whose formal guarantee covers a linear surrogate and which held for the actual model in our conditions. The simplified Freedman bound certifies most facts, but it failed only on facts whose coefficient change R exceeds their distance a to the boundary, and these failures are breakdowns of the first-order model. The complete bound, which excludes these facts, and the more conservative Azuma bound held in all 112 conditions, including the 43 confirmatory ones; the facts that the complete bound certifies were forgotten 0 times in 34,256 pooled pairs of a fact and a test run (Table 2). The location of the failures rests mainly on post hoc data, because the confirmatory studies contained only 3 violated facts. The screened bound shows that a rule designed on earlier conditions can hold on them and fail on new starting states, so only new starting states can test it.

When the probes are worth their cost. A probability matters when the same update is run many times from one starting state, for example when a shared model is adapted for many users. A simulation then needs 59 surviving runs to certify a fact at level 0.05 with 95% confidence, whereas the probes cost a median of 38 test runs at $T = 40$ and 68 at $T = 80$.

Limitations. All bounds hold for a linear surrogate, and the remainder that would transfer them to the actual margin remains unquantified, so their coverage on the actual model is an empirical finding; the variance estimate is also low for most facts (Appendix G). The complete bound excludes every fact with $R \geq a$, although most of these facts survived every test run. Certifying them needs a model that remains accurate along the test runs, because the per-time bound failed on such facts. With 32 test runs and a Bonferroni correction over 64 facts, a fact with bound 0.05 is flagged only if it is forgotten in at least 8 runs, so a single condition rules out only large excesses over a bound, and the pooled counts carry the evidence about smaller ones. Our results cover SGD at horizons of 40 and 80 steps with 64 synthetic or CounterFact facts per condition; Adam and full fine-tuning remain open.

9 Conclusion

A first-order model of fine-tuning can be accurate and still fail to predict a full run, yet it can bound the probability that a margin first falls below a boundary near zero. The simplified Freedman bound failed only on facts with $R \geq a$, a condition known before the run that every observed breakdown of the first-order model met. The complete bound, which excludes these facts, and the Azuma bound held in all 112 conditions, whereas a screen that judged whole conditions failed on new starting states. Bounds on forgetting should therefore be computed for each starting state and each fact.

ACKNOWLEDGMENTS

Omitted for anonymous review.

ETHICS STATEMENT

This work uses only public model checkpoints and synthetic or public factual data, and all experiments are computational. The main ethical risk is over-trust: a bound on forgetting can be quoted outside the assumptions under which it holds. A bound that certifies which knowledge survives an update could also be used to protect content that should be removed. We report the facts on which the simplified Freedman bound fails, and we state the scope of every claim.

AI USE STATEMENT

We used large language models to help write and proofread the text. The authors take full responsibility for the content.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code, all preregistrations, the frozen analysis scripts, the run directories with prediction and margin files, and the script that generates the results, tables and figures of this paper from those directories. Appendix J reports a rebuild of two conditions from an empty copy of the repository; one of them gave a different starting state, although the configuration, data, package versions, model revision and GPU were the same.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11849–11860, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/15825aee15eb335cc13f9b559f166ee8-Abstract.html.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *CoRR*, abs/2110.01052, 2021. URL https://arxiv.org/abs/2110.01052.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Leonard F. Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - A review. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=ePUVetPKu6.

Dan Biderman, Jacob P. Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and

John Patrick Cunningham. Lora learns less and forgets less. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=aloEru2qCG.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080. PMLR, 2021. URL http://proceedings.mlr.press/v130/doan21a.html.

Itay Evron, Edward Moroshko, Rachel A. Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In Po-Ling Loh and Maxim Raginsky (eds.), *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4028–4079. PMLR, 2022. URL https://proceedings.mlr.press/v178/evron22a.html.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=HvSytvg3Jh.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 7765–7784. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.444. URL https://doi.org/10.18653/v1/2024.emnlp-main.444.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 15202–15232. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.902. URL https://doi.org/10.18653/v1/2024.findings-acl.902.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://doi.org/10.1214/18-PS321.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zarif Ikram, Arad Firouzkouhi, Stephen Tu, Mahdi Soltanolkotabi, and Paria Rashidinejad. Crispedit: Low-curvature projections for scalable non-destructive LLM editing. *CoRR*, abs/2602.15823, 2026. doi: 10.48550/ARXIV.2602.15823. URL https://doi.org/10.48550/arXiv.2602.15823.

Xisen Jin and Xiang Ren. What will my model forget? forecasting forgotten examples in language model refinement. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22145–22159. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/jin24d.html.

Xisen Jin and Xiang Ren. Demystifying language model forgetting with low-rank example associations. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/06872e1e6d11baf2ae27285c50132f4f-Abstract-Conference.html.

Damjan Kalajdzievski. Scaling laws for forgetting when fine-tuning large language models. *CoRR*, abs/2401.05605, 2024. doi: 10.48550/ARXIV.2401.05605. URL https://doi.org/10.48550/arXiv.2401.05605.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. URL http://arxiv.org/abs/1612.00796.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32801–32818. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/lotfi24a.html.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *CoRR*, abs/2308.08747, 2023. doi: 10.48550/ARXIV.2308.08747. URL https://doi.org/10.48550/arXiv.2308.08747.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23610–23641. PMLR, 2023. URL https://proceedings.mlr.press/v202/malladi23a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715. URL https://link.springer.com/book/10.1007/b106715.

A Proofs

A.1 Theorem 1 (coupling bound)

Couple P and Q step by step. At each step where both chains are in $\mathcal{N}$ and are still coupled, the kernels differ by at most ε_t in total variation, so a maximal coupling keeps them equal except on an event of probability ε_t. Decoupling by time T therefore has probability at most $P(\tau \le T) + \sum_{t \le T} \varepsilon_t$, and on the complement the chains agree on every path event. For tightness, take A to be the decoupling event in a system where leaving $\mathcal{N}$ determines the outcome. □

A.2 Theorem 2 (non-identifiability)

Fix a smooth bump function ϕ that vanishes on $\mathcal{N}$ together with all its derivatives on the boundary of $\mathcal{N}$. Two drift fields that agree on $\mathcal{N}$ and differ by $\lambda\phi$ outside it give identical transition kernels for any path that stays in $\mathcal{N}$, identical derivatives of all orders at every point of $\mathcal{N}$, and the same noise law. Outside $\mathcal{N}$ the field is unconstrained, so λ can route the mass that has left $\mathcal{N}$ either across the event boundary at step 2 or away from it, which moves the event probability by exactly p. A predictor that sees only $\mathcal{N}$ has the same distribution under both systems and errs by at least $p/2$ on one of them. □

A.3 Theorem 3 (first-passage bounds on the grid)

Write $\widehat{Y}_t - m_t = S_t + D_t$ with $S_t = \sum_{s \le t} \eta\, q_{T,s}^{\top} \xi_s$ and $D_t = \sum_{s \le t} \eta\, (q_{t,s} - q_{T,s})^{\top} \xi_s$.

(i) $(S_u)_u$ *is a martingale for* $\mathcal{F}_u = \sigma(\xi_1, \ldots, \xi_u)$. Each $q_{T,s}$ is $\mathcal{F}_0$-measurable, because the response coefficients are estimated by finite differences at the starting state and are fixed before the test run starts, and $\mathbb{E}[\xi_s \mid \mathcal{F}_{s-1}] = 0$ because ξ_s is the applied clipped update centered at its conditional mean. Since $\|\xi_s\| \le 2G_0$, every increment satisfies $|\eta\, q_{T,s}^\top \xi_s| \le 2\eta G_0 r = c$.

(ii) $|D_t| \le R$ *almost surely.* By the Cauchy–Schwarz inequality, $|D_t| \le \sum_{s \le t} \eta \|q_{t,s} - q_{T,s}\| \|\xi_s\| \le 2\eta G_0 \max_{t \in G} \sum_{s<t} w_s \|q_{t,s} - q_{T,s}\| = R$, with the interval widths w_s of the probe grid.

(iii) Event inclusion. If $\widehat{Y}_t < B_0$ for some $t \in G$, then $S_t + D_t < B_0 - m_t \le -a$, so $-S_t > a - R = a'$. The first-passage event is contained in $\{\exists u \le T : -S_u \ge a'\}$.

(iv) Maximal inequalities. The maximal Azuma inequality for a martingale with increments bounded by c gives $\Pr(\exists u \le T : -S_u \ge a') \le \exp\{-a'^2/(2Tc^2)\}$. If in addition $\langle -S \rangle_T \le V$, the maximal Freedman inequality (Freedman, 1975) gives the bound $\exp\{-a'^2/(2(V + c\,a'/3))\}$. The maximal inequalities cover all grid times at once. For the surrogate with final-time coefficients, $m_t + S_t$, we have $D_t = 0$, so the bounds hold with $R = 0$. □

A.4 Estimated variance (Section 4)

Condition on $\mathcal{F}_0$, which fixes $\widehat{V}$ and the other bound parameters, and partition on $\Omega = \{\langle -S \rangle_T \le \widehat{V}\}$. The maximal Freedman inequality bounds the probability of a crossing on Ω by $\exp\{-a'^2/(2(\widehat{V} + c\,a'/3))\}$, and $\Pr(\Omega^c \mid \mathcal{F}_0) \le \delta$ by assumption. An estimate $\widehat{V}$ that depends on the outcomes of the test runs requires a separate argument. □

A.5 Selection before the run

Let $\mathcal{F}_0$ be the information generated by the probes; the screen is a deterministic function of $\mathcal{F}_0$.

Proposition 1 (Selection before the run)**.** *Let* $B \in \{\mathrm{F}, \mathrm{A}\}$ *be the selected branch and let* $\mathcal{L}$ *be any* $\mathcal{F}_0$*-measurable set of certified facts. Suppose that on the event* $\{B = b\}$*, each branch bound satisfies* $\Pr(E_i \mid \mathcal{F}_0) \le U_i^b$*, where* E_i *is the forgetting event of fact* i*. Then* $\Pr\big(\bigcup_{i \in \mathcal{L}} E_i\big) \le \mathbb{E}\big[\sum_{i \in \mathcal{L}} U_i^B\big]$.

The proposition leaves the validity of each branch open; it shows that choosing the branch before the run keeps the guarantee of the selected branch.

Proof. Let $A_b = \{B = b\}$. Because B and $\mathcal{L}$ are $\mathcal{F}_0$-measurable, $\Pr\big((\bigcup_{i \in \mathcal{L}} E_i) \cap A_b\big) = \mathbb{E}\big[\mathbf{1}_{A_b} \Pr(\bigcup_{i \in \mathcal{L}} E_i \mid \mathcal{F}_0)\big] \le \mathbb{E}\big[\mathbf{1}_{A_b} \sum_{i \in \mathcal{L}} U_i^b\big]$ by the union bound inside the conditional probability. Summing over b gives the stated bound with U_i^B. The argument holds for any dependence between facts and between branches. □

Numerical checks. A Monte Carlo experiment checks the maximal inequality on clipped heteroscedastic and two-point martingales (largest ratio of empirical frequency to bound: 0.354). An independent reimplementation recomputes every stored bound (largest discrepancy 1.7×10^{-16}).

B Scope of the first-order response model

Table 3: First-order response model at task-B learning rate 0.002. A validation run passes a preregistered threshold if the correlation ρ between predicted and measured margin changes under held-out perturbations is at least 0.90 and the slope lies in $[0.8, 1.2]$. Every run passes; n is the number of runs, and ranges are over runs.

Checkpoint	Family	n	ρ	slope
Qwen3-0.6B	Qwen	4	0.985–0.994	0.901–1.047
Llama-3.2-1B	Llama	3	0.974–0.998	0.974–1.028
Qwen3-1.7B	Qwen	4	0.982–0.996	0.986–1.075
Qwen3-8B	Qwen	1	0.994	0.987
Gemma-3-12B	Gemma	2	0.992–0.997	1.003–1.010
Qwen3-14B	Qwen	1	0.998	1.014

A random low-rank basis gives $\rho = 0.057$ because the margin changes that it induces are smaller than the numerical noise of the forward pass. A precision study that reruns the same 4-bit weights in 32-bit arithmetic attributes this noise to the arithmetic rather than to weight quantization. Under SGD, probe responses grow by a factor of 1.92 to 2.06 when the perturbation size doubles, over the full horizon; under Adam this linear scaling breaks down after about ten steps.

On Qwen3-0.6B, we validated the response model up to a parameter displacement of 0.080 along the basis, with relative error 0.37 at that radius. A single forgetting event needs a displacement of about 1.4 (median margin 0.71 divided by median response norm 0.50), which is about $18\times$ the validated radius. The task-B update noise of a 40-step run alone accumulates a displacement of about 0.77, about ten times the validated radius, so every run leaves the validated region, independently of forgetting.

C Predictions of forgetting built on the response model

Table 4: Five predictions built on the response model, each preregistered with a success criterion and tested on data disjoint from the data that suggested it. The outcome column gives the result that missed the criterion: an error above the threshold, an effect outside the predicted interval, a correlation that fell on replication, a rejected sufficiency hypothesis, and a prediction of the size of the divergence that missed its direction.

Prediction	Independent test	Outcome
Per-fact event probabilities from the response model and a Gaussian noise model	fresh split of facts	mean absolute error 0.060 against a threshold of 0.05; all errors are false alarms
Rotating injected noise of fixed spectrum changes the forgetting rate by 0.145 between the extreme orientations	32 paired runs, spectra matched to 10^{-14}	difference 0.032, 95% interval $[-0.043, 0.105]$, excludes 0.145
A drift statistic of the response coefficients ranks facts by prediction error	fresh set of 128 facts	mean rank correlation falls from 0.639 to 0.075
Two histories with the same margin vector have the same future distribution	branching from matched states	energy-distance permutation test, $p = 0.00035$
A rank-2 projection of the protected gradients predicts how futures diverge	runs branched from matched states	predicts how much futures diverge (Spearman 0.658, against 0.050 for a random projection) but misses the direction

D History of the screening threshold

The screen comes from a rule that we preregistered before Studies 13 and 14: predict a violation of the simplified Freedman bound if the median of $a_i/\sqrt{V_i/\kappa}$ over the certified facts is below 30. A first version with an additional criterion failed its prospective test in Study 13 (four of five evaluable conditions correct). The simpler rule was correct in six of six conditions of Study 14 when its predictions are recomputed from the recorded statistic; the runner had stored the predictions of the first version, a protocol deviation recorded in that preregistration, and those stored labels are correct in five of six. The rule then failed a replication across seeds that used an increment bound too small by a factor of two (Appendix L). With the increment bound of Equation 1, its only remaining error in that replication is conservative: the bound held in one starting state that it flagged. The screened bound differs from this rule in two ways. It uses the median of $a_i/\sqrt{V_i}$, which is smaller by a factor of $\sqrt{\kappa} = \sqrt{2}$, so its threshold of 30 corresponds to about 42.4 on the earlier statistic. It also uses the screen to select a bound rather than to predict a failure. We fixed this definition in the code on 2026-09-15, before the prospective studies.

Table 5 shows, post hoc on the earlier conditions, how the number of violating conditions and the certified fraction change with the threshold and with uniform variance inflation. A threshold of 30 is the smallest value in this grid at which every earlier condition passes. Although the threshold was

fixed before this table was computed, the earlier data alone leave it indistinguishable from a fitted value, which is why the prospective studies test it. Figure 3 shows the result: the earlier failures lie below the threshold, and prospective failures occurred above it. Uniform inflation of the variance estimate leaves violating conditions even at $\kappa = 128$ (Table 5, right), which agrees with failures that are confined to facts with $R \geq a$, on which the spread of the margin exceeds the prediction by a larger factor than for other facts (Appendix G).

Table 5: Post-hoc sweeps on the earlier conditions. Left: the screening threshold. Right: the simplified Freedman bound with its variance estimate inflated by κ (the default is 2).

Threshold on $\tilde{z}$	violating cond.	certified
0 (always simplified)	8	79.6%
10	8	79.6%
15	6	77.1%
20	2	67.2%
25	1	66.2%
30	0	62.3%
35	0	59.6%
40	0	54.1%
50	0	45.0%
60	0	34.8%
80	0	29.7%
always Azuma	0	21.2%

Inflation κ	violating cond.	certified
2	8	79.6%
4	7	78.3%
8	5	75.9%
16	5	72.6%
32	5	68.7%
64	2	63.8%
128	2	56.2%

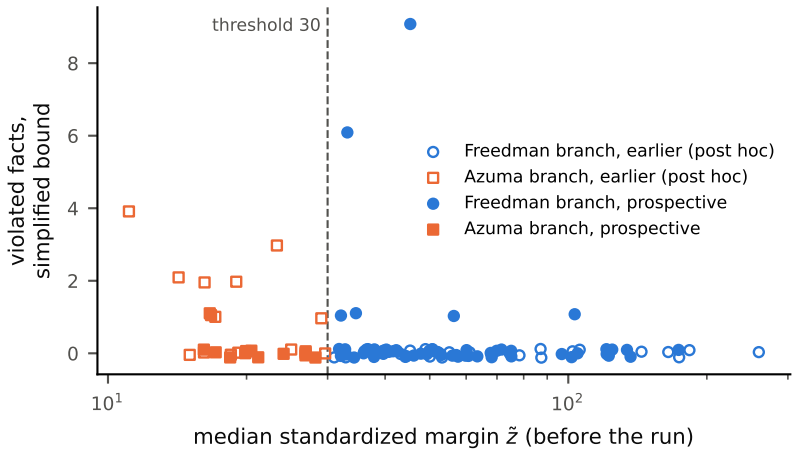

Figure 3: Median standardized margin $\tilde{z}$ of the facts that the simplified Freedman bound certifies, computed before the run, against the number of facts on which that bound is significantly violated (with small vertical jitter). Hollow: earlier conditions (post hoc); filled: all 71 prospective conditions. The screen selects the Freedman branch to the right of the dashed line.

E Calibration of the bounds

Table 6 groups the eligible facts of each set by the value of their bound and compares the mean bound in each group with the mean frequency of a forgetting event per test run. For a valid bound, the frequency is at most the mean bound in every group, up to sampling error; a group in which the frequency exceeds the mean bound shows where the bound is too small.

F Earlier conditions

Table 7 lists all 45 earlier conditions, with all bounds recomputed with the increment bound of Equation 1. Entries are violating facts over certified facts, and the branch is F (Freedman branch) or

Table 6: Mean bound and frequency of forgetting events, grouped by the value of the bound, for all 45 earlier conditions and for the main and horizon studies (the added 8B and 14B batches and the complete bound are covered by Table 1).

		simplified Freedman bound			screened bound		
Set	bound range	facts	mean bound	event freq.	facts	mean bound	event freq.
Earlier	$[0, 0.001)$	1140	0.0001	0.0014	1009	0.0001	0.0000
	$[0.001, 0.01)$	270	0.0041	0.0038	151	0.0039	0.0000
	$[0.01, 0.05)$	196	0.0253	0.0110	96	0.0240	0.0020
	$[0.05, 0.2)$	179	0.1057	0.0239	83	0.1085	0.0041
	$[0.2, 0.5)$	91	0.3324	0.0797	67	0.3510	0.0420
	$[0.5, 1]$	141	0.8426	0.3510	611	0.8803	0.1022
Main	$[0, 0.001)$	690	0.0001	0.0032	608	0.0001	0.0035
	$[0.001, 0.01)$	164	0.0042	0.0078	128	0.0042	0.0100
	$[0.01, 0.05)$	129	0.0238	0.0203	89	0.0250	0.0284
	$[0.05, 0.2)$	108	0.1095	0.0567	76	0.1049	0.0678
	$[0.2, 0.5)$	55	0.3321	0.1545	46	0.3393	0.1039
	$[0.5, 1]$	93	0.8422	0.3659	292	0.8696	0.1331
Horizon	$[0, 0.001)$	190	0.0001	0.0000	148	0.0001	0.0000
	$[0.001, 0.01)$	40	0.0045	0.0000	38	0.0045	0.0000
	$[0.01, 0.05)$	13	0.0226	0.0000	14	0.0238	0.0000
	$[0.05, 0.2)$	8	0.1149	0.0000	4	0.1316	0.0000
	$[0.2, 0.5)$	4	0.3767	0.0000	15	0.3559	0.0000
	$[0.5, 1]$	3	0.8779	0.5729	39	0.8565	0.0441

A (Azuma branch). $\tilde{z}$ is the median standardized margin of the facts that the simplified Freedman bound certifies. Studies 6 to 11 were single-seed studies of model size and learning rate, Studies 13 and 14 tested the earlier screening rule, Study 17 varied the seed at fixed configurations, and Study 18 used CounterFact. The rebuild of Appendix J that produced byte-identical files is counted once. Four of these conditions lack the stored probe file from which R is computed; Section 6 and Table 1 therefore use the other 41, while the threshold sweeps of Appendix D use all 45. In the evaluable conditions of Studies 17 and 18, the reference margin a_i ranked facts by forgetting risk with a higher area under the ROC curve than the simplified Freedman bound in five of six conditions, so the reference margin is an informative ranking score, although it provides a ranking rather than a probability.

Table 7: Earlier conditions (post hoc).

Study	Model	lr	seed	branch	$\tilde{z}$	events	simplified	Azuma	screened
S18	Llama-3.2-1B	0.002	11	F	164.9	0	0/60	0/47	0/60
S6	Llama-3.2-1B	0.002	11	F	183.3	0	0/64	0/64	0/64
S13	Llama-3.2-1B	0.006	11	F	57.1	56	0/58	0/28	0/58
S13	Llama-3.2-1B	0.006	11	F	52.0	16	0/56	0/23	0/56
S18	Llama-3.2-1B	0.008	11	F	55.2	0	0/45	0/14	0/45
S18	Llama-3.2-1B	0.008	43	F	87.1	12	0/46	0/16	0/46
S14	Llama-3.2-1B	0.009	11	F	50.0	31	0/45	0/4	0/45
S13	Llama-3.2-1B	0.012	11	F	39.3	8	0/31	0/2	0/31
S9a	Llama-3.2-1B	0.012	11	F	52.4	83	0/36	0/0	0/36
S13	Llama-3.2-1B	0.012	23	A	18.5	74	0/32	0/0	0/0
S13	Llama-3.2-1B	0.012	37	A	19.0	210	2/30	0/1	0/1
S17	Llama-3.2-1B	0.012	41	F	68.2	0	0/30	0/12	0/30
S17	Llama-3.2-1B	0.012	43	A	19.2	30	0/22	0/0	0/0
S17	Llama-3.2-1B	0.012	47	A	15.1	111	0/24	0/0	0/0
S17	Llama-3.2-1B	0.012	59	F	48.9	20	0/31	0/11	0/31
S17	Llama-3.2-1B	0.012	61	F	144.2	0	0/31	0/12	0/31
S17	Llama-3.2-1B	0.012	67	A	16.2	71	0/29	0/0	0/0
S14	Llama-3.2-1B	0.015	11	F	40.8	180	0/34	0/0	0/34
S13	Llama-3.2-1B	0.018	11	F	43.2	16	0/31	0/0	0/31
S18	Llama-3.2-1B	0.02	11	A	29.1	83	1/30	0/0	0/0
S13	Llama-3.2-1B	0.024	11	F	31.0	0	0/26	0/0	0/26
S13	Llama-3.2-1B	0.03	11	A	25.0	0	0/25	0/0	0/0
S18	Qwen3-0.6B	0.002	11	F	174.1	0	0/63	0/41	0/63
S6	Qwen3-0.6B	0.002	11	F	259.4	5	0/62	0/58	0/62
S13	Qwen3-0.6B	0.006	11	F	78.3	0	0/42	0/18	0/42
S18	Qwen3-0.6B	0.008	11	F	53.2	1	0/48	0/13	0/48
S18	Qwen3-0.6B	0.008	43	F	61.1	13	0/50	0/13	0/50
S14	Qwen3-0.6B	0.012	11	A	17.1	116	1/26	0/1	0/1
S13	Qwen3-0.6B	0.018	11	A	23.3	179	3/21	0/1	0/1
S17	Qwen3-0.6B	0.018	41	F	39.7	233	0/29	0/3	0/29
S17	Qwen3-0.6B	0.018	43	F	105.9	0	0/28	0/0	0/28
S17	Qwen3-0.6B	0.018	47	F	102.2	0	0/32	0/6	0/32

Study	Model	lr	seed	branch	$\tilde{z}$	events	simplified	Azuma	screened
S17	Qwen3-0.6B	0.018	59	F	67.8	5	0/27	0/5	0/27
S17	Qwen3-0.6B	0.018	61	F	36.5	0	0/30	0/0	0/30
S17	Qwen3-0.6B	0.018	67	F	45.3	6	0/27	0/0	0/27
S18	Qwen3-0.6B	0.02	11	A	29.6	53	0/24	0/0	0/0
S14	Qwen3-0.6B	0.021	11	A	16.2	175	2/17	0/0	0/0
S13	Qwen3-0.6B	0.024	11	A	11.1	98	4/20	0/0	0/0
S13	Qwen3-0.6B	0.03	11	A	16.7	45	1/23	0/0	0/0
S14	Qwen3-1.7B	0.006	11	F	87.4	78	0/48	0/17	0/48
S14	Qwen3-1.7B	0.012	11	F	42.3	0	0/31	0/0	0/31
S14	Qwen3-1.7B	0.018	11	F	32.8	28	0/27	0/0	0/27
S14	Qwen3-1.7B	0.024	11	F	36.1	0	0/26	0/0	0/26
S11b	Qwen3-14B	0.008	11	F	44.5	34	0/51	0/9	0/51
S10	Qwen3-8B	0.01	11	A	14.2	33	2/38	0/8	0/8

G SPREAD OF THE MARGINS ACROSS TEST RUNS

For each eligible fact, we divide the standard deviation of the final margin across the test runs by the standard deviation $\sqrt{K}$ that the first-order model predicts. Over all 112 conditions, the median of this ratio was 1.6 on facts with $R < a$, 2.7 on facts with $R \geq a$, and 14.6 on the facts that violated the simplified bound. Of the 421 facts with a ratio above 10, 94.5% had $R \geq a$. The factor $\kappa = 2$ in V allows for a ratio of up to $\sqrt{2}$, and 59.2% of the facts with $R < a$ exceeded it, so the variance estimate is low for most facts. The complete bound held nevertheless, because the facts that it certifies lie far from the boundary compared with the realized spread: for 99.0% of them, $a - R$ exceeded three standard deviations of the realized final margin.

H FORGETTING RISK AND THE RATIO R/a

Figure 4 groups the eligible facts of all 112 conditions with $a > 0$ by R/a. Of the 1,436 facts with $R < a$, 0 were forgotten in any test run. Above $R = a$, the share of facts forgotten in at least one test run rises steadily with R/a, to 63.1% for $R/a \geq 16$. The ratio therefore ranks facts by forgetting risk before the run, and the bounds add a probability for the facts that they certify. The remaining 61 eligible facts have $a \leq 0$: their reference margin stays positive but comes within B_0 of zero, every bound leaves them uncertified, and 59 of them were forgotten in at least one test run.

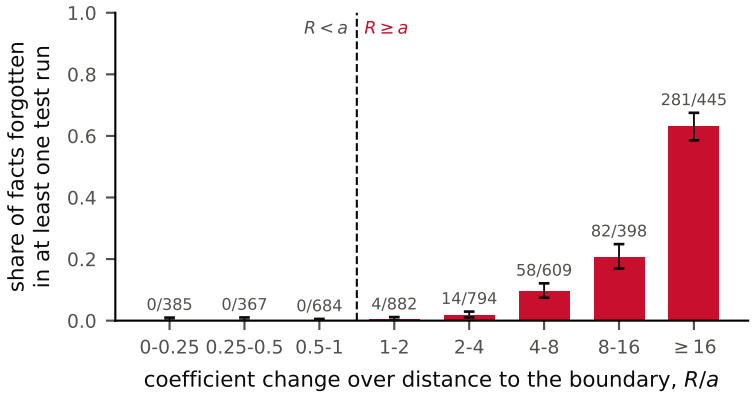

Figure 4: Share of eligible facts forgotten in at least one test run, in bins of R/a, over all 112 conditions (facts with $a > 0$). Bars show the share, whiskers the Wilson 95% interval, and labels the number of forgotten facts over the number of facts in the bin.

I POWER OF THE TEST AND POOLED FORGETTING COUNTS

The per-fact test has limited power. With 32 test runs and a Bonferroni correction over 64 facts, a fact with $U = 0.05$ is flagged only if it is forgotten in at least 8 runs, and the test reaches power

0.8 only when the true forgetting probability is at least 0.304 (Table 8). A condition that passes the test therefore rules out only large excesses over the bound, and the evidence about smaller excesses comes from pooling.

Table 2 in Section 7.4 gives the pooled counts. The allowed number of events for a rule is the sum over its certified facts of the number of runs times U_i. Events of different facts in one test run are dependent, so we report these counts as descriptive totals. For the margin threshold, the per-fact rule and the restricted margin threshold, which the table omits, the certified facts were forgotten 5, 1 and 0 times, against 911.6, 23.5 and 485.6 allowed events.

Table 8: Power of the per-fact test with 32 test runs and a Bonferroni correction over 64 facts: the smallest number of forgetting events that is significant against U, the probability of reaching it for true forgetting probabilities p, and the smallest p at which this probability reaches 0.8.

U	min. events	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.5$	p at power 0.8
0.05	8	0.01	0.30	0.79	1.00	0.304
0.01	4	0.40	0.91	0.99	1.00	0.166
0.001	2	0.84	0.99	1.00	1.00	0.091

J VARIATION ACROSS SEEDS AND BUILDS

Outcomes vary strongly across seeds of one configuration. Among the earlier conditions at fixed configurations, the number of forgetting events of eligible facts ranged from 0 to 210 across seeds for Llama-3.2-1B and from 0 to 233 for Qwen3-0.6B, and the new seeds of the second confirmatory study gave far fewer events than earlier seeds at the same settings. The seed sets both the starting state and the minibatch orders of the test runs, so these ranges mix the two effects. We also rebuilt two earlier conditions from an empty copy of the repository. One gave byte-identical files. The other gave a different starting state (mean margin of the protected facts 0.658 instead of 0.608) and a different outcome, although the configuration, data, package versions, model revision and GPU were the same. Bounds must therefore be computed for each starting state.

K PROSPECTIVE CONDITIONS

The family-wise statements of the screened bound were slightly above their nominal level. In the main study, the largest set of facts per condition whose screened bounds sum to at most 0.05 contained 60.9% of eligible facts, and some fact of this set was forgotten in 44 of 800 test runs (5.5% against 5%, one-sided $p = 0.28$).

Table 9: Main prospective study, every condition ($T = 40$).

Set	lr	seed	branch	$\tilde{z}$	events	simplified	Azuma	screened
A	0.012	71	F	173.7	0	0/32	0/13	0/32
A	0.012	73	F	56.4	50	1/33	0/4	1/33
A	0.012	79	F	45.4	242	9/31	0/1	9/31
A	0.018	71	F	122.5	0	0/32	0/2	0/32
A	0.018	73	F	37.8	0	0/32	0/3	0/32
A	0.018	79	F	69.9	267	0/17	0/0	0/17
A	0.021	71	F	101.7	0	0/32	0/0	0/32
A	0.021	73	F	38.0	0	0/30	0/2	0/30
A	0.021	79	F	63.4	44	0/17	0/0	0/17
B	0.012	71	F	68.2	132	0/30	0/21	0/30
B	0.012	73	F	44.3	5	0/31	0/7	0/31
B	0.012	79	A	24.1	49	0/17	0/0	0/0
B	0.018	71	A	26.9	88	0/28	0/7	0/7
B	0.018	73	F	38.9	0	0/32	0/3	0/32
B	0.018	79	A	16.6	45	1/15	0/0	0/0
C	0.012	71	F	46.2	0	0/29	0/3	0/29
C	0.012	73	F	33.1	160	6/49	0/16	6/49
C	0.021	71	A	20.5	0	0/29	0/0	0/0
C	0.021	73	A	16.2	177	0/30	0/0	0/0
D	0.01	71	excluded (task A)					
D	0.01	73	excluded (task A)					
E	0.008	71	F	52.5	2	0/46	0/8	0/46

Set	lr	seed	branch	$\tilde{z}$	events	simplified	Azuma	screened
E	0.008	73	F	49.9	42	0/50	0/12	0/50
E	0.008	79	F	59.8	25	0/47	0/10	0/47
E	0.02	71	A	26.9	67	0/22	0/0	0/0
E	0.02	73	F	32.9	69	0/26	0/0	0/26
E	0.02	79	A	21.2	74	0/24	0/0	0/0
F	0.008	71	F	49.3	12	0/47	0/12	0/47
F	0.008	73	F	53.0	15	0/47	0/11	0/47
F	0.008	79	F	60.7	12	0/47	0/13	0/47
F	0.02	71	F	32.1	64	0/26	0/0	0/26
F	0.02	73	F	32.1	61	1/27	0/0	1/27
F	0.02	79	F	34.6	51	1/28	0/0	1/28

Table 10: Horizon study ($T = 80$). The branch is given for the paired condition of the main study ($T = 40$) and for $T = 80$. Entries are significant violations over certified facts at $T = 80$; the last column is the number of facts that the screened bound certifies at $T = 40$.

Model	lr	seed	branch 40/80	$\tilde{z}$	events	simplified	Azuma	screened	screened at 40
Qwen3-0.6B	0.012	71	F/F	136.6	0	0/32	0/6	0/32	32
Qwen3-0.6B	0.012	73	F/F	51.9	0	0/32	0/4	0/32	33
Qwen3-0.6B	0.018	71	F/F	124.6	0	0/32	0/0	0/32	32
Qwen3-0.6B	0.018	73	F/F	39.6	0	0/32	0/0	0/32	32
Llama-3.2-1B	0.012	71	F/A	26.9	30	0/27	0/6	0/6	30
Llama-3.2-1B	0.012	73	F/F	50.6	0	0/32	0/7	0/32	31
Llama-3.2-1B	0.018	71	A/A	17.1	25	0/24	0/2	0/2	7
Llama-3.2-1B	0.018	73	F/F	35.9	0	0/32	0/0	0/32	32

Table 11: Qwen3-14B study ($T = 40$), two preregistered batches (Studies 22 and 23). Conditions below the task-A accuracy threshold are marked as excluded. Entries are violating facts over certified facts.

Study	lr	seed	branch	$\tilde{z}$	events	simplified	Azuma	screened
22	0.008	71	excluded (task A)					
22	0.012	71	excluded (task A)					
22	0.008	73	F	121.2	14	0/47	0/40	0/47
22	0.012	73	F	32.6	30	0/44	0/16	0/44
23	0.008	79	F	31.8	11	0/54	0/28	0/54
23	0.012	79	A	18.4	65	0/37	0/16	0/16
23	0.008	83	F	37.9	0	0/64	0/10	0/64
23	0.012	83	A	20.0	24	0/56	0/0	0/0

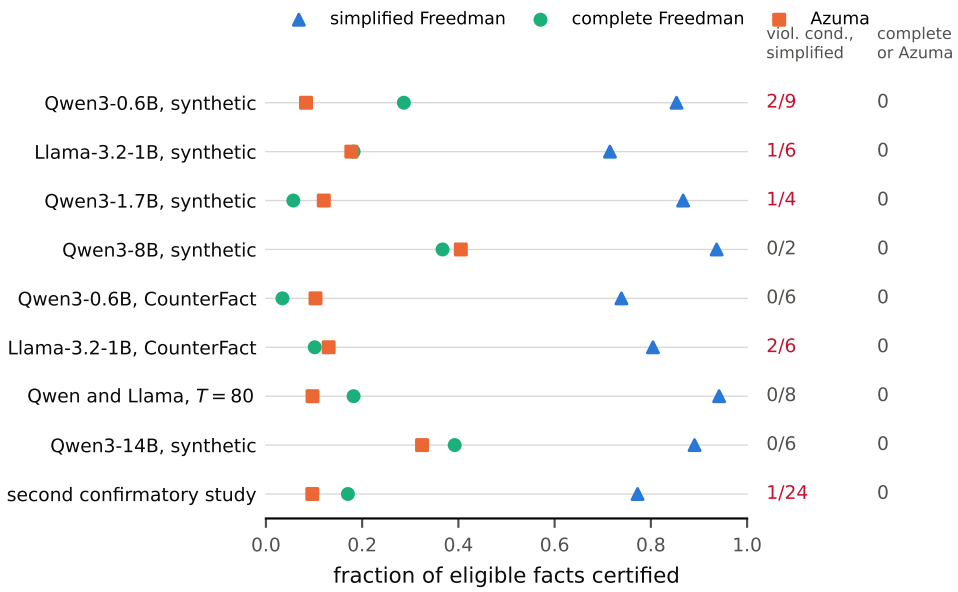

Figure 5: Fraction of eligible facts certified by the simplified Freedman bound, the complete Freedman bound and the Azuma bound in each condition set of the main, horizon and 14B studies and in the second confirmatory study, with the number of conditions with a violation (right).

Table 12: Added Qwen3-8B batch ($T = 40$, Study 24). Both Qwen3-8B conditions of the main study were below the task-A accuracy threshold (Table 9); this batch forms the Qwen3-8B row of Table 1. Entries are violating facts over certified facts.

lr	seed	branch	$\tilde{z}$	events	simplified	Azuma	screened
0.01	79	F	47.8	0	0/46	0/20	0/46
0.01	83	F	50.7	8	0/28	0/12	0/28

Table 13: Second confirmatory study ($T = 40$, Study 25). Entries are violating facts over certified facts for the simplified, complete and Azuma bounds, the per-time bound, and the simplified bound restricted to facts with $R < a$. The last column is the number of facts violated by the simplified bound that have $R \geq a$.

model	lr	seed	events	simplified	complete	Azuma	per-time	simplified, $R < a$	$R \geq a$
Qwen3-0.6B	0.012	89	0	0/32	0/18	0/0	0/29	0/26	0
Qwen3-0.6B	0.012	97	0	0/32	0/5	0/2	0/32	0/8	0
Qwen3-0.6B	0.012	101	0	0/31	0/19	0/0	0/30	0/26	0
Qwen3-0.6B	0.012	103	20	0/35	0/18	0/4	0/31	0/22	0
Qwen3-0.6B	0.012	107	36	0/28	0/8	0/0	0/28	0/15	0
Qwen3-0.6B	0.012	109	65	0/26	0/6	0/7	0/23	0/12	0
Qwen3-0.6B	0.018	89	0	0/30	0/0	0/0	0/22	0/18	0
Qwen3-0.6B	0.018	97	0	0/22	0/0	0/0	0/21	0/0	0
Qwen3-0.6B	0.018	101	28	0/29	0/0	0/0	0/25	0/6	0
Qwen3-0.6B	0.018	103	100	0/30	0/10	0/0	0/29	0/13	0
Qwen3-0.6B	0.018	107	0	0/25	0/0	0/0	0/19	0/0	0
Qwen3-0.6B	0.018	109	0	0/21	0/4	0/3	0/17	0/5	0
Qwen3-0.6B	0.021	89	0	0/30	0/0	0/0	0/21	0/12	0
Qwen3-0.6B	0.021	97	0	0/20	0/0	0/0	0/17	0/0	0
Qwen3-0.6B	0.021	101	0	0/29	0/0	0/0	0/24	0/3	0
Qwen3-0.6B	0.021	103	71	0/28	0/2	0/0	0/28	0/6	0
Qwen3-0.6B	0.021	107	0	0/24	0/0	0/0	0/18	0/0	0
Qwen3-0.6B	0.021	109	1	0/18	0/2	0/1	0/16	0/5	0
Qwen3-1.7B	0.012	89	22	0/34	0/8	0/17	0/33	0/10	0
Qwen3-1.7B	0.012	97	30	0/40	0/3	0/4	0/40	0/8	0
Qwen3-1.7B	0.012	101	0	0/32	0/3	0/0	0/32	0/9	0
Qwen3-1.7B	0.012	103	94	0/34	0/7	0/24	0/32	0/7	0
Qwen3-1.7B	0.012	107	0	0/32	0/30	0/23	0/32	0/30	0
Qwen3-1.7B	0.012	109	1	1/27	0/9	0/1	0/27	0/10	1

L PREREGISTRATION RECORD

In the supplementary material, the main prospective study is Study 20, the horizon study is Study 21, the two batches of the 14B study are Studies 22 and 23, the added Qwen3-8B batch is Study 24, and the second confirmatory study is Study 25. Three addenda to Study 20 add the margin threshold and the per-fact rule (Addenda 1 and 2, committed after the first condition had finished and its result had been seen) and the confirmatory test of the complete bound (Addendum 3, committed after the interim look, with the list of excluded runs). The material also contains the preregistrations of all earlier studies, the frozen analysis scripts, the verbatim results of every preregistered criterion including the failed ones, and the script that generates every number in this paper from the run directories.

During development we found and corrected three defects, and all numbers in this paper use the corrected versions. First, the key of a probe cache omitted the length of task A, so probes of one starting state could be used for a run from another; we excluded the affected runs and kept them with a written reason. Second, a rule that selected the newest run directory could mix two concurrent experiments; we excluded those runs as well. Third, earlier code used the increment bound $\eta G_0 r$ instead of $2\eta G_0 r$ (Equation 1); we recomputed every bound in this paper with the correct value.