DELETION COST, NOT SALIENCE: CAUSAL MEASUREMENT FOR CONTEXT EVICTION IN LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agents with finite context must decide what to retain. Common eviction rules use proxies for how much an element still matters: age, message type, attention mass, or a model's judgement. We measure deletion cost directly: perturb one element, replay the identical continuation under teacher forcing, and read the change in log-odds on the decisions it governs, with a placebo control at each distance. The survey's position-controlled power-law exponents lie within ± 0.177, and deleting a standing constraint remains costly through 1153 steps. In separate probe schedules, attention half-lives span $1.13\times$, while plan and constraint differ by $8.8\times$ in fitted causal half-life. Across eight open-weight models, all 48 (model, probe-channel) cells remain on the relevant side of the fitted power-law boundary. Finally, scoring all 720 static orders on a six-probe BF16 evaluation, we find that a frozen 4B order based on directly measured deletion costs beats the evaluated six-text and fixed-prompt task-informed judges on two Qwen models. A third model's informed judge recovers the direct order, but too few trajectories pass its gate for a stable cross-guise comparison. Fitting supplies no resolved advantage over the direct order. These results provide a measurement for pricing context deletion; how often an element will still be needed remains a deployment input to be measured.

1 INTRODUCTION

Long-running agents exhaust finite context. Eviction rules often use age, message type, attention or an LLM's importance judgment as proxies for **how deleting an element changes decisions it still governs**. Counterfactual ablation estimates a related effect at one point (Cohen-Wang et al., 2024; Liu et al., 2025); we measure its dependence on distance and compare eviction proxies against it.

Age-based eviction assumes behavioural influence decays with distance. Representation-level decay suggests this might hold (Brändel et al., 2026), yet compaction can discard binding constraints: Chen (2026) report violation rates rising from 0% to as much as 59%. Our distance ladder tests the behavioural assumption. Policy comparisons use simplified system variants, without their full dependency and access mechanisms (Section 2).

A controlled measurement. In a console-style environment, we deliver six context element types before decisions they govern. We perturb one element, replay the identical continuation under teacher forcing and measure the governed decision's log-odds change. Subtracting the edit's effect on unrelated probes controls its average collateral footprint. If the fitted effect $\delta(d) \propto d^{-p}$ persists, its sum grows without bound for $p \leq 1$ and remains bounded for $p > 1$. This is an effect *conditional on the replayed history*, not predicted free-running damage. The design varies distance, element type and later demonstrations, but does not independently identify the fitted demonstration effect as re-learning. A post-hoc same-probe control tests whether filler itself teaches a deleted rule (Sections B and 3).

Distance and attention summaries. The survey's position-controlled exponents satisfy $|p| \leq 0.177$; a standing constraint still costs $4.40\,\mathrm{lo}$ at the farthest distance band, 1153 steps. All 48 (model, probe-channel) cells across eight open-weight models remain on the same side of the fitted boundary. On 624 paired ablations of third-party τ-bench policy text, the exponent is negative (Section 4).

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.

These measurements provide no common age-only cutoff for still-needed elements; they establish neither asymptotic loss nor how often old elements remain needed. Attention half-lives span $1.13\times$, while plan and constraint differ by $8.8\times$ in causal half-life. The schedules differ, so this is not a matched signal comparison. Attention decay also reproduces a plain type tier's order (Section F).

Amplitude and recovery separate elements in replay. Fitted amplitude spans $4.8\times$ and the demonstration-decline exponent $5.7\times$. Plan has the second-largest amplitude but declines fastest; constraint has the largest amplitude and declines $4.4\times$ more slowly. This is a candidate mechanism for persistent governance damage, not an intrinsic importance order: routine filler reveals plan's class–phase map, and removing this cue sharply changes the deleted-plan outcome in a matched local control (Sections B and 5).

Measured deletion costs outperform the evaluated judges on Qwen. Measuring all 2^6 subsets gives a frontier and scores all 720 static orders. An order based on six 4B deletion costs, frozen across models, beats each model's six-text judge on retained samples. In native BF16, judges also given task rules and unlabeled probes trail direct measurement on 4B and 27B; Gemma's informed judge matches the direct order (Sections L and N). This is a measurement for pricing deletion, not a general advantage over informed judges (Section 6).

A separate post-hoc HotpotQA control tests evidence specificity, not the distance law or end-to-end eviction (Section I). Conditional information explains dependence on future demand and surviving evidence (Section 7); deploying a retention policy still requires demand estimates and validation on the resulting trajectories.

Contributions. We make three measurement contributions. **(1) Controlled deletion-cost measurement:** paired teacher-forced replays with placebo controls trace the effects of six context-element types across decision distance and later demonstrations. **(2) Persistence and its limits:** eight open-weight checkpoints quantify distance trends and type-specific amplitude and fitted decline; a third-party policy-text control and filler intervention delimit their interpretation. **(3) Budget-matched eviction evaluation:** complete six-probe measurements of all 63 deletion subsets on three BF16 models compare 720 static orders and informed judges on the same retained samples and budgets. Direct costs are a strong baseline; fitting brings no resolved advantage.

2 Related work

Paired replay and attribution. Paired context interventions are established (Mehta & Datta, 2026; Shah, 2026; He et al., 2026; Min et al., 2026). Cohen-Wang et al. (2024) ablate sources with output and prefix fixed, fit a sparse surrogate to logit-transformed probabilities, and validate on held-out subsets. Liu et al. (2025) measure segment-deletion effects on output likelihood, approximate joint effects additively and transfer attributions from smaller same-family models. We claim no novelty for these steps. We measure effects across three decades of decision distance and different element types, then evaluate against all 720 static orders. Fitted demonstration decline remains protocol-conditional (Section B).

Eviction systems and evaluated variants. CWL (Semenov & Dorofeev, 2026) uses dependencies and environment effects; MemDecay (Matam & Kim, 2026) separates retention priority from output influence, resets its clock on access and pins system regions; Chen et al. (2026) validates summaries on later trajectories. We compare simplified variants: a static type tier, a MemDecay-style decay rule without access resets, and a judge shown six texts. **Results do not evaluate the full systems.** MemDecay fits attention decay and combines it with base tiers. Our variant fits decay on causal curves; its attention estimator is examined separately (Sections F and N). Attention need not track importance (Wen et al., 2022; Mai & Kim, 2026; Javadov, 2026); our comparison concerns decay *rates*. Other schemes use persistent importance or heavy hitters (Liu et al., 2023; Zhang et al., 2023), so age decay is not universal.

Representation decay and behavioural damage. Brändel et al. (2026) report Jacobian-sensitivity exponents $p \approx 0.7$–0.9 with a non-vanishing offset, compatible with persistence under the $p = 1$ boundary. Our representation channel reaches only $p = 0.234$, so the discrepancy does not establish

a general behavioural–representational difference (Section D). Mehta & Datta (2026) report an independent representational/behavioural mismatch; Chen (2026) document governance failures after compaction. Our amplitude and decline measurements offer a conditional explanation, not a full retention criterion (Section 5).

Conditional information. Brown et al. (2012) connect feature selection to conditional likelihood; Nagle et al. (2024) formulate fixed-LLM prompt compression as rate–distortion. We use conditional information to interpret measured costs, not as a new compression method (Section 7).

3 INSTRUMENT

Environment. An episode is a sequence of steps; at each the agent sees a short observation and an eight-option menu and emits one index. Six task families share the same mechanism but use disjoint lexicons; we cluster uncertainty by family. At a chosen step s_0 the agent receives a *brief* containing six elements: a standing CONSTRAINT that does not lapse, a GOAL, a working PLAN, a PERSONA, an intermediate derivation FORM delivered as a bare fact, a worked reasoning chain or a tool output, and a DISTRACTOR that licenses nothing and serves as a negative control. Later steps carry *probes*, *diagonal* for element e if the correct option depends on e and nothing else in the brief and *off-diagonal* otherwise; every trajectory fires both at matched distances inside one forward pass, with elapsed distance d swept geometrically (ratio 1.6, to 176 steps at horizon 256; to 1153 at horizon 1280). Assistant turns are a single digit with no reasoning text, which removes by construction the reasoning-trace confound Mehta & Datta (2026) strip explicitly. Routine filler includes each item's class and in-phase number, which can itself reveal the PLAN mapping; Section B tests this cue.

Paired perturbation and read-out. For element e and perturbation c we render two contexts differing only in e's bytes and replay the *identical* action sequence through both under teacher forcing, so one pass per arm yields exact next-token distributions and the read-out is computed rather than sampled. Perturbations are DROP, PAD (neutral text of equal length), PARA (licensing unchanged) and CORRUPT (rewritten to license the wrong option) – the last being the plausible-but-wrong summary a compaction step actually produces. The read-out is $\delta = \operatorname{logit} p^{\text{keep}} - \operatorname{logit} p^{\text{arm}}$, computed entirely in log space; exponentiating first imposes a hard ceiling near $36\,\mathrm{lo}$ in float (Section A).

For the distance and demonstration fits, we use a difference in differences – the diagonal read minus the off-diagonal read of the same perturbation at the same distance, averaged over trajectories – which removes the perturbation's collateral footprint before any exponent is fitted. The six fitted channels comprise four single-probe element types and two FORM channels; the unlicensed distractor is a separate negative control. A byte-identical repeat of the keep arm returns mean, standard deviation and maximum all exactly 0.0000 over 8154 reads. This establishes repeatability for identical inputs; nonzero off-diagonal footprints measure a different quantity. The control is an average and not a per-trajectory one, so the balanced design is what licenses it (Section C, Section A). The subset evaluation instead averages raw paired differences (Section 6); neither read-out is the future-demand-weighted value discussed in Section 7.

Scaling and its interpretation. For $\delta(d) \propto d^{-p}$ and a constant governed-decision rate ρ, accumulated influence $D(t) = \rho \sum_{d \le t} \delta(d)$ grows as t^{1-p} for $p < 1$, grows logarithmically at $p = 1$, and is bounded for $p > 1$. We therefore define $y = 1 - p$ and call these cases relevant, marginal and irrelevant, respectively. This classification assumes the power law continues beyond the measured window. Paired replay leaves the sign of p free and measures an effect *conditional on the replayed history*; it does not predict free-running damage (Section A).

4 THE SCALING DIMENSION

We fit $\delta = A\, d^{e}\, (\mathrm{step}/10)^{f}$ on the survey configuration (1728 trajectories, diagonal probes with no earlier same-kind probes, placebo-differenced), with $p = -e$ and family-clustered bootstrap intervals. The $\log$ position covariate is not optional: without it PERSONA's p reads -0.255, i.e. influence increasing with age (Section B).

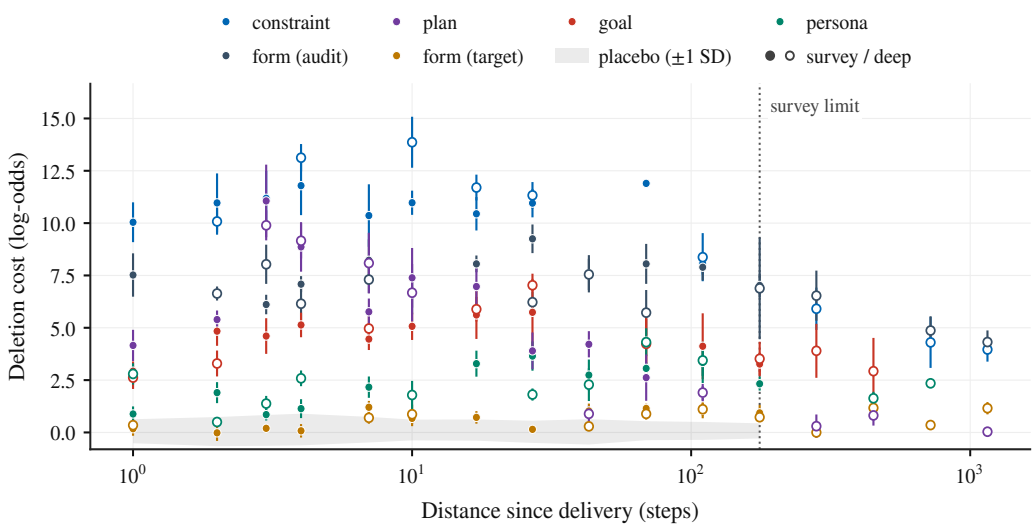

Figure 1: Raw paired log-odds deletion cost versus distance (Qwen3.5-4B). Filled/open points are survey/deep grids restricted to shared family, guise and delivery cells ($s_0 = 0$) through $d = 1153$. These raw effects differ from the placebo-differenced, position-controlled fits in Table 1. Grey band: placebo spread; bars: trajectory standard errors. Descriptive joint fits are in Figure 4.

Table 1: Scaling dimension on Qwen3.5-4B. $y = 1 - p$, boundary at $y = 0$. All six channels lie near $y = 1$, which corresponds to a constant per-step effect and linear accumulation under the power-law model. The audit channel uses the reasoning guise; the others pool all three guises.

element	n	A	$p = -e$	f (pos.)	y [95% CI]	verdict
CONSTRAINT	1728	10.15	0.024	+0.003	0.976 [0.956, 0.997]	relevant
PLAN	1728	8.57	0.177	−0.033	0.823 [0.747, 0.892]	relevant
FORM (audit)	576	6.00	−0.064	−0.087	1.064 [1.037, 1.098]	relevant
GOAL	1728	5.14	0.093	+0.124	0.907 [0.876, 0.928]	relevant
FORM (target)	1728	2.89	0.014	−0.057	0.986 [0.967, 1.009]	relevant
PERSONA	1728	2.11	−0.046	+0.118	1.046 [1.003, 1.079]	relevant

Table 1 and Figure 1 give $|p| \leq 0.177$, with two channels mildly growing. Probe-content fixed effects change p by at most 0.039 without changing any verdict (Section C). These are partial distance coefficients: along an element's actual age path, the exponent is at most 0.210 on the survey and 0.604 across all 48 (configuration, element) cells, none of them reaching the $p = 1$ boundary (Section D).

Finite-window evidence. Power and exponential fits tie in cross-validation, and the bounded cumulative exponential fits equally well (MSE ratios 0.99–1.02; Section D). The experiments establish persistent conditional damage over the measured window, not unbounded future retention value.

Conditional damage persists over three decades. To test persistence beyond the survey window, we extend the horizon to 1280. Probes extend to $d = 1153$ with three delivery points to estimate position effects. CONSTRAINT falls from 8.80 lo in the nearest decade to **4.40** lo in the farthest, and FORM (audit) from 6.43 to 3.85: factors of 2.0 and 1.7 over three decades. The position-controlled deep constraint estimate is $p = 0.020$, with an interval overlapping the survey's. The form-channel intervals differ between the grids, which also differ in guise and family composition; this comparison does not establish a common exponent or exclude a bend for every channel. Distance is measured over this range on 4B; the 1000-step cumulative priority still extrapolates the recovery channel beyond its observed demonstration counts (Section N).

Eight models, forty-eight cells, one verdict. Eight open-weight checkpoints span three Qwen generations, a 35B MoE, GLM and Gemma-4-31B. We index them by reference-arm competence

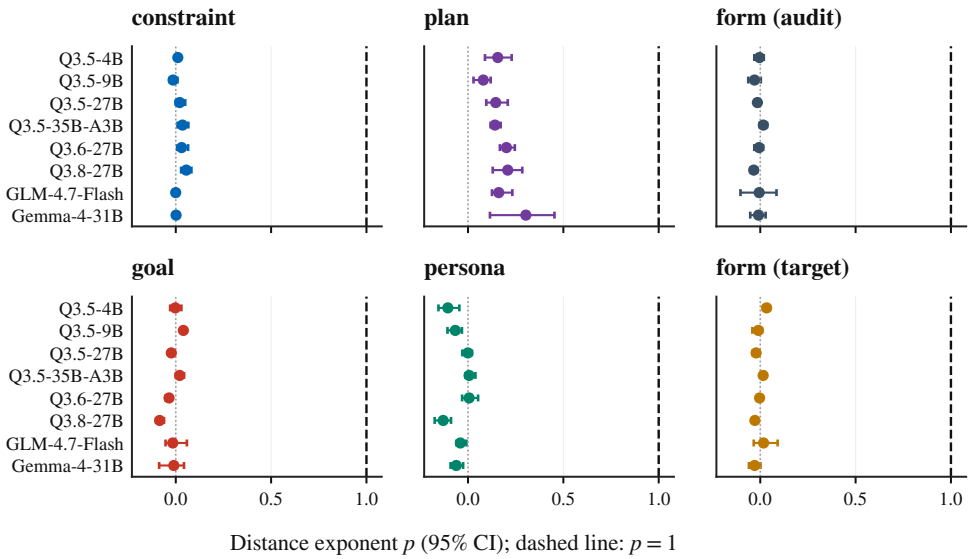

Figure 2: **Distance exponents across eight models and six probe channels.** Points and family-clustered 95% intervals use distance-only fits to content-demeaned, placebo-differenced effects on common design cells. Every interval lies below $p = 1$ (dashed line). These estimates differ from the position-controlled survey fits in Table 1; $p < 1$ classifies power-law continuation, not empirically established asymptotic loss. Q abbreviates Qwen.

(mean licensed-option probability, 0.681–0.980), which is comparable across generations and families. **All** 48 **cells are relevant** (Figure 2). The largest within-element spread of p is 0.223, against a minimum interval-to-boundary distance of 0.547 (2.4× larger). Exponent–competence rank correlation is −0.035 with Gemma and −0.066 without. These fits use distance-only, content-demeaned differences, unlike the position-controlled survey estimator (Section G).

The same measurement on text we did not write. To check template dependence, we delete individual bullets from third-party τ-bench policy documents (Yao et al., 2024). Across 624 paired ablations from 104 transcripts (503,178 read positions), $\mathbf{p = -0.176}$, $[-0.297, -0.074]$ clustered on transcripts. This growing sensitivity is weaker evidence than the synthetic result: the read-out is unsigned, the renderer drops tool calls, the covariates correlate at 0.922 after rule dummies, and improvement of the reference arm could create the growing effect. That confound remains open (Section H).

5 What separates elements

The scaling dimension classifies all six channels alike, so it cannot set an eviction priority. Amplitude can: it spans 4.8× (Table 1). All costs below are in log-odds; the optional task-length conversion is an uncalibrated illustration (Section K).

Damage decline with subsequent demonstrations. A trajectory may demonstrate a deleted rule through later examples. On a schedule varying prior demonstrations (864 trajectories, family-clustered), we fit $\delta_{\mathrm{DD}} = A(1 + n_{\mathrm{prior}})^{-r} d^{e}$. The fitted damage-decay exponent spans 5.7×: PLAN 0.709 [0.615, 0.844], CONSTRAINT 0.160 [0.083, 0.267] and PERSONA 0.124 [0.067, 0.183] illustrate the range; both FORM channels and GOAL are in Table 5.

A falling damage difference does not identify re-learning. It can reflect improvement of the perturbed arm, degradation of the reference arm, or movement of the placebo. The perturbed arm supplies 60% of CONSTRAINT's decline but only 17% of FORM (audit)'s. The non-monotone demonstration schedule identifies r *within the joint model* controlling distance and position; it survives probe-content fixed effects and eight within-content fits (Sections B and C). Re-derivation is

therefore an interpretation, while the measured quantity is the decline in damage along this schedule. A post-hoc same-probe intervention shows that relevant examples can improve the dropped arm at fixed distance, but does not identify the numerical r fitted to the BF16 schedule (Section B).

In this protocol, the plan combines the second-largest amplitude (8.57) with the fastest decline; the constraint combines the largest (10.15) with damage that declines $4.4\times$ more slowly. **High initial cost and persistent damage make the constraint expensive to lose under this probe mixture.** This is a candidate mechanism for Chen (2026)'s compaction failures: recovering ordinary task information could mask persistent governance damage. The filler control also shows why these amplitudes cannot be treated as context-independent values: routine steps can substitute for an explicit plan (Section B). This is a two-element comparison, not a general inverse relation: amplitude and recovery have positive rank correlations across the six channels ($+0.71$ and $+0.37$ under the two fits). Recovery's contribution to the evaluated priority is small, changing one adjacent pair in the frozen 4B ordering (Sections 6 and N).

Corruption costs more than deleting a redundant derivation. In the reasoning guise of FORM, DROP removes the derivation while keeping the conclusion. On the survey this costs 0.50 lo, whereas rewriting the derivation incorrectly costs 11.01: a $+10.51$ gap, with a $+10.42$ gap on a second model. Keeping the conclusion makes deletion cheap in this setting; deletion-only evaluations miss the much larger corruption effect. Results for the other guises and their limitations are in Table 6.

Transfer requires attention to units. Across five recovery checkpoints (864 trajectories each), amplitude ordering transfers better than recovery ordering (Section J). Absolute amplitude uses each model's log-odds scale: Gemma's constraint amplitude is 49.7 against 7.7–14.4 elsewhere, and its median licensed log-odds is 30.3 against roughly 5 for Qwen. The exponents are dimensionless. **Re-measure amplitude on the deployed model; the trichotomy verdict transfers.** Within Qwen, amplitude's increase with competence is only suggestive; these measurements give no evidence that the governance failure mode shrinks with capability (Section J).

6 EVICTION: MEASURING WHAT EVERY SUBSET COSTS

Evaluation and policies. We measure all 63 non-empty deletion subsets over 216 trajectories on each of Qwen3.5-4B, Qwen3.5-27B and Gemma-4-31B in BF16, plus keep and byte-identical duplicate arms. Each subset's damage is the raw signed log-odds difference averaged over the same six probes, including both FORM destinations in every guise. Probes span 1–176 steps; this is a fixed-history evaluation over the distance ladder, not task success. Comparisons retain only trajectories with all six probes and all 63 subsets above the 0.5 option-mass gate: 216, 216 and 75, respectively. Gemma's sample falls below the prespecified $n \geq 100$ requirement for a stable cross-guise comparison; its results are descriptive (Section L). Every keep probe passes the quality gate; exclusions arise under deletion. For Gemma, failed subset cells number $1140/4536$, $122/4536$ and $1141/4536$ in fact, reasoning and tool-output guises. Complete-case selection can therefore remove deletion-induced format failures, so the reported excess does not describe all 216 evaluated trajectories.

The direct policy sorts six mean single-element damages from a separate 4B calibration enumeration with historical probe coverage. Fitted policies sort A_e or $A_e e^{-r_e}$ from a 4B recovery run. These orders are frozen across all three evaluations. Each model supplies its own six-text and task-informed judges; the latter receives task rules, the equal-weight probe distribution and unlabeled examples. Score selection used earlier eviction evaluations, so the fitted policies are exploratory rather than independently selected (Section N). All elements arrive together, so recency cannot rank them; its age premise is tested by the distance ladder.

6.1 EVICTION DAMAGE IS NEARLY ADDITIVE

Fitting six additive effects on the 21 subsets of size ≤ 2 predicts the 42 held-out subsets of size 3–6 at $R^2 = 0.9934$, 0.9582 and 0.9869 for 4B, 27B and Gemma. Corresponding RMSEs are 0.084, 0.287 and 0.576 lo (Table 10 and Figure 6). This tests cross-element interference on probes whose answers each depend on one element. Joint dependencies and substitutable elements are absent by

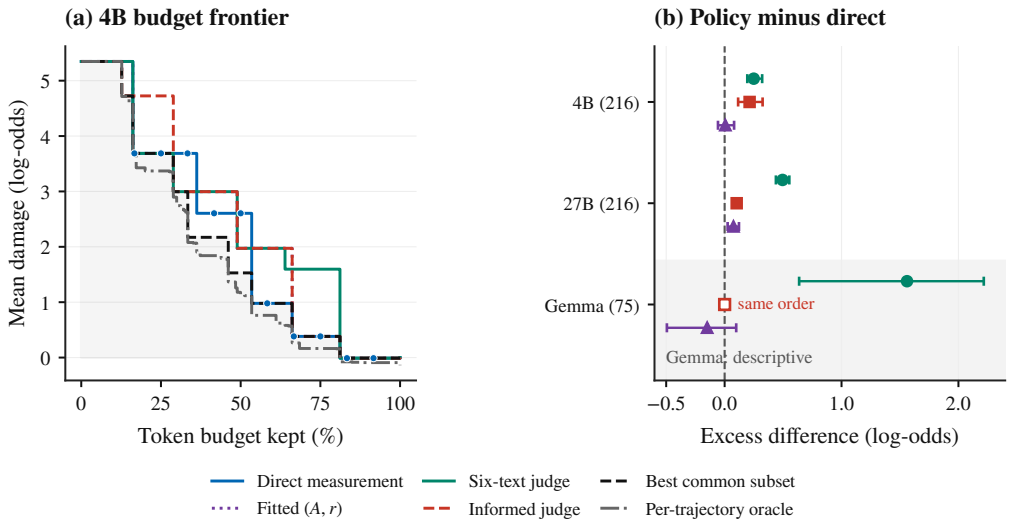

Figure 3: **Six-probe eviction costs and paired policy contrasts.** *Left:* 4B mean damage at exact subset-cost breakpoints ($n = 216$); direct-policy dots mark the ten scoring budgets. Shading marks the best common subset; the lower per-trajectory oracle is the reference in Table 2. *Right:* policy minus direct excess with family-clustered 95% intervals; positive values favor direct. Parentheses give retained n. Gemma is descriptive; its informed judge has the same order as direct, rather than an estimated equivalence.

construction. Adding all 15 pairwise terms improves held-out prediction on Qwen but worsens it on Gemma; interactions remain relevant to policy differences.

6.2 The frontier at matched budgets

Excess is policy damage minus the per-trajectory oracle, averaged over ten interior budgets on the same complete sample. Budgets use each model's tokenizer costs; points and intervals both weight retained trajectories equally. Intervals resample the six lexical families as whole clusters (Section L).

For retained trajectory $i \in \{1, \dots, N\}$, let $D_i(S)$ be damage from deleting set S and $\mathcal{F}_b$ the sets feasible at retained-token budget b. The best common-subset and per-trajectory references are

$$O^{\mathrm{cs}}(b) = \min_{S \in \mathcal{F}_b} \frac{1}{N} \sum_i D_i(S), \qquad O^{\mathrm{pt}}(b) = \frac{1}{N} \sum_i \min_{S \in \mathcal{F}_b} D_i(S). \tag{1}$$

The second may choose a different set after seeing each trajectory's outcomes, so $O^{\mathrm{pt}} \le O^{\mathrm{cs}}$; it is a hindsight lower bound, not a deployable policy. Tables subtract O^{pt}. Token costs are measured as keep-minus-single-drop differences per tokenizer and added across elements for budget accounting, rather than retokenizing each joint deletion. The six removable costs sum to 224.53 tokens for Qwen and 224.75 for Gemma; in the reasoning guise, dropping form retains its conclusion. All policies use the same ten interior grid indices within each model's 13-point budget grid (Section L).

For each paired policy difference, we resample six entire lexical families with replacement (4000 draws) and pool their retained trajectories. Gemma's family counts are 7/9/10/11/21/17; averaging family means equally would target a different quantity. Only six lexical clusters underlie each reported interval, irrespective of the number of trajectories.

Measurement vs. the six-text judge. Direct lowers excess by 0.249/0.496/1.560 lo on retained 4B/27B/Gemma samples; all three family-clustered intervals exclude zero.

Task information narrows the judge comparison. The informed judges have excesses 0.646 and 0.714 lo on 4B and 27B, versus 0.432 and 0.611 for direct measurement. Informed minus

Table 2: Six-probe mean excess over the per-trajectory oracle (log-odds, lower better). Direct, fitted and cumulative orders are frozen on 4B before these evaluations; judges use the evaluated model. The best of all 720 orders is an in-sample reference, not a deployable policy. Ranks are in Table 11.

policy	**4B** $n = 216$	**27B** $n = 216$	**Gemma** $n = 75$ (descriptive)
Best of all 720 orders	0.336	0.352	2.601
Direct measurement (frozen 4B)	0.432	0.611	4.025
Fitted Ae^{-r} (frozen 4B)	0.438	0.686	3.876
Fitted A (frozen 4B)	0.468	0.714	4.203
Cumulative score (frozen 4B)	0.843	1.328	4.279
Six-text judge (model's own)	0.681	1.106	5.585
Task-informed judge (model's own)	0.646	0.714	4.025
CWL-style type tiers	0.888	1.379	5.585
MemDecay (causal-rate variant)	1.014	1.886	3.726
Length	1.398	1.984	6.278

direct is $+0.214\,[+0.115, +0.325]$ and $+0.103\,[+0.080, +0.127]$ lo; each contrast is positive in all six families. Gemma's informed judge recovers the direct order exactly. These results support an advantage on the two Qwen models for this fixed prompt, not superiority over task-informed judges generally (Section N).

Fitting adds no resolved advantage over direct measurement. Fitted (A, r) minus direct is $+0.006\,[-0.057, +0.081]$ on 4B, $+0.075\,[+0.026, +0.123]$ on 27B and $-0.149\,[-0.494, +0.099]$ lo on Gemma. Direct measurement therefore wins on 27B; the other contrasts are unresolved. Recovery discounting changes one adjacent pair in the frozen order, with no resolved gain over amplitude alone on either Qwen model (Section N). The fits describe distance and demonstration dependence, but are not required for the measured-cost advantage over the six-text judge. The e^{-r_e} discount equals $(1+n)^{-r_e}$ at $n = e - 1 \approx 1.72$ prior demonstrations, an arbitrary point rather than a measured deployment horizon. At $n = 0$ or 1, the frozen 4B score evicts GOAL before PLAN; at $n = e - 1$, 2, 3 or 5, they swap, and the order changes again at $n = 8$. Earlier eviction outcomes informed choosing this score over the cumulative integral, although its parameters come only from recovery data. This is a sensitivity analysis, not independent validation of the discount. The 27B fitted-minus-direct contrast changes from -0.031 with historical probe labels to $+0.075$ with all six labels on the same trajectories (Section L). This paired change is a change in the target being averaged, not in the test cases or token budgets. It reverses the point ordering of fitted and direct measurement, so the historical ranking cannot serve as evidence for a six-probe fitted advantage. We use complete six-probe coverage as the main analysis and retain the earlier labels only as a sensitivity check.

A static ordering cannot attain every budget's optimum. Unequal costs make each budget a $0/1$ knapsack whose optimal subsets need not nest. None of the 720 orders attains the best common-subset frontier at all 13 budgets on any retained model sample (Section L). The direct policy exceeds the best static order by 0.096, 0.259 and 1.424 lo. On Gemma, the causal-rate MemDecay variant has lower point excess than either direct or fitted measurement, further limiting any general policy claim.

7 WHAT A DELETION COST DOES AND DOES NOT PRICE

Our measurements price deletion on fixed histories and probes chosen because an element governs them. A retention rule also needs the frequency of future demands and the evidence left after each edit. Let Q be a future demand, Y its correct action, W shared information, and X_S a fixed set of retained context elements. For discrete Y and finite conditional entropies, deleting a group $G \subseteq S$ increases an optimal reader's expected log loss by

$$H(Y \mid X_{S\setminus G}, Q, W) - H(Y \mid X_S, Q, W) = I(Y; X_G \mid X_{S\setminus G}, Q, W). \tag{2}$$

This standard identity (Brown et al., 2012; Nagle et al., 2024) makes both dependencies explicit. A rarely needed rule can have high cost on our selected probes but low average cost under future

demands. Two records with the same answer can each be dispensable while the other remains; deleting both need not be. Likewise, the $p = 1$ accumulation boundary in Section 3 assumes a persistent demand rate, not merely persistent damage when a rule is queried.

The conditioning set changes after each edit. For a fixed demand with $X_1 = X_2 = Y$, deleting either copy while retaining the other loses no information, but deleting both loses $H(Y)$. A full-context single-element score therefore need not predict the cost of a later deletion. Our six-element subset enumeration can test this dependence within its probe design, but the near-additivity observed there cannot be assumed for tasks with substitutable or jointly required evidence.

The identity concerns a true task distribution and an optimal reader. Our teacher-forced read-out is instead a signed change in one model's correct-action log-odds, placebo-corrected for the distance fits; it estimates neither conditional mutual information nor free-running task loss. A full-context-output KL would measure fidelity to that model and could penalize an edit that corrects its mistake. Thus the measured costs can inform which deletions merit testing, but applying them as a retention rule still requires future-demand estimates and validation on the resulting trajectories.

8 Limitations

Teacher forcing measures deletion effects on fixed histories, not free-running damage (Section A). Power and bounded exponential fits tie over the observed window, so the $p < 1$ classification does not establish unbounded future loss (Section D). All 48 fitted cells fall on the same side of the boundary; this tests stability, not whether the classification distinguishes useful from useless elements. The six task families vary lexicon around largely one mechanism, and additivity is tested on probes built to depend on one element. A retention policy also needs the future demand rate, supplied in the cumulative illustration as 0.05 rather than measured. Distance fits estimate cost conditional on a governed probe; the subset evaluation averages a fixed six-probe mixture, neither of which identifies total retention value as an element ages (Section D).

Routine filler reveals the class–phase mapping, so plan's amplitude and fitted damage-decline exponent r are conditional on implicit examples. A post-hoc local Q4 intervention shows a strong filler effect but does not identify the BF16 schedule's numerical r (Section B). Plan's reference-arm competence also falls at long distance, limiting the interpretation of its residual log-odds damage (Section A). The fitted policy was selected using earlier evaluations and has no resolved advantage over direct measurement in the six-probe results. Attention and causal half-lives also use different probe schedules, so their comparison is descriptive rather than a matched test of either signal.

The BF16 informed judges trail direct measurement on Qwen but reproduce its order on Gemma; this does not establish a general advantage across prompts. Gemma retains only $75/216$ complete trajectories, below the prespecified sample requirement. Selection occurs after deletion and differs by guise and family, so its results neither describe excluded trajectories nor support stable cross-guise conclusions. Intervals resample only six lexical families (Section L).

The 80-question HotpotQA control shows evidence specificity outside the console task, but its multiple-choice adaptation, post-hoc selection and lack of a distance ladder do not validate the age exponent or free-running compaction (Section I). The conditional-information identity assumes a true demand distribution and an optimal reader; it does not identify that distribution or validate a deployment policy.

9 Conclusion

Paired replay shows that deleting a still-needed element can remain costly across the measured distance window. On the six-probe BF16 evaluation, direct measurement beats the six-text and fixed task-informed judges on 4B and 27B. Gemma's informed judge recovers the direct order, but its 75 complete cases cannot support stable cross-guise conclusions. Fitting supplies no resolved advantage; direct measurement instead wins on 27B. These results support measuring conditional deletion cost. No fixed deletion order reaches the best common-subset frontier at every tested budget, so ranking alone does not solve budgeted retention. A deployment policy additionally requires representative future demand rates and free-running validation.

Ethics statement

This work involves no human subjects and no private data. All eight checkpoints are open-weight releases used under their published licences. Third-party text comes from published τ-bench (Yao et al., 2024) policies and HotpotQA (Yang et al., 2018) documents. The agent environment is synthetic and executes nothing outside the process. One dual-use consideration is worth naming: a table of which context elements are expensive to delete is also a table of which elements an adversary would remove to induce policy violations without an obvious trace. We report it because deployed compaction already deletes those elements silently, and Chen (2026) shows the resulting violation rates are large; a defender who cannot price the deletion cannot assess that risk. Recorded compute components and their accounting scope are reported in Section O.

Reproducibility statement

Section O specifies the configurations, estimators and figure construction. The six-probe eviction tables and figure are recomputed from saved measurements without new model inference. Within a generated trajectory, arms are replayed under teacher forcing, the byte-identical control arm returns exactly zero over 8154 reads (Section 3), and repeated rendering of a fixed trajectory is deterministic. Rebuilding trajectories in different Python processes additionally requires fixing the hash seed or archiving the rendered inputs; the current generator uses Python's string-dependent hash. Bootstrap intervals use a fixed seed and cluster on task family; the resampling procedure, the quality gate and the paired-subsample rule are stated in Section 6 and Section L. Fitted parameter tables, the frozen 4B priority that the transfer results depend on, and the enumeration of all 720 orderings are described in the analysis (Sections L and N).

Use of large language models

LLM assistance was used for prose editing. All experimental designs, analysis choices and reported claims were specified and verified by the authors, who take full responsibility for the contents and the reported analyses.

References

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(2):27–66, 2012. URL https://jmlr.org/papers/v13/brown12a.html.

Matthias Brändel, Stephan Köhler, and Oliver Rheinbach. How token influence decays with distance: A green-function view of trained language models. *arXiv preprint arXiv:2606.29139*, 2026.

Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. *arXiv preprint arXiv:2606.22528*, 2026.

Zhuofu Chen, Rui Pan, Yinwei Dai, and Ravi Netravali. Slipstream: Trajectory-grounded compaction validation for long-horizon agents. *arXiv preprint arXiv:2605.08580*, 2026.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2409.00729.

Jiawei He, Mengyu Shi, Jie jia, Xikai Yang, and Dong Sun. What process evaluation of coding agents actually measures: Action, task, and step are three different levels. *arXiv preprint arXiv:2608.22960*, 2026.

Aydin Javadov. When does routing become interpretable? causal probes on block attention residuals. *arXiv preprint arXiv:2606.13168*, 2026.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth

Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks. *arXiv preprint arXiv:2503.14499*, 2025.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.15102.

Fengyuan Liu, Jay Gala, Nilaksh, Dzmitry Bahdanau, Siva Reddy, and Hugo Larochelle. BRIDGE: Predicting human task completion time from model performance. *arXiv preprint arXiv:2602.07267*, 2026.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.17118.

Tho Mai and Joo-Young Kim. Reformulating KV cache eviction problem for long-context LLM inference. *arXiv preprint arXiv:2605.07234*, 2026.

Venkatesha Matam and Keon Kim. MemDecay: Region-aware KV cache eviction for efficient LLM agent inference. *arXiv preprint arXiv:2607.10582*, 2026.

Aman Mehta and Anupam Datta. Plans don't persist: Why context management is load bearing for LLM agents. *arXiv preprint arXiv:2606.22953*, 2026.

Guanghui Min, Liang Wu, Mayank Darbari, Chen Chen, and Liangjie Hong. Toward reliable context compression for long-horizon agents: An empirical study of execution instability. *arXiv preprint arXiv:2608.06503*, 2026.

Alliot Nagle, Adway Girish, Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva, and Hyeji Kim. Fundamental limits of prompt compression: A rate-distortion framework for black-box language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2407.15504v2. arXiv:2407.15504v2.

Andrew Semenov and Svyatoslav Dorofeev. Beyond compaction: Structured context eviction for long-horizon agents. *arXiv preprint arXiv:2606.11213*, 2026.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Bingyang Wen, K. P. Subbalakshmi, and Fan Yang. Revisiting attention weights as explanations from an information theoretic perspective. *arXiv preprint arXiv:2211.07714*, 2022.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.14048.

A INSTRUMENT GATES AND READ-OUT

Stable arithmetic. For option log-probabilities ℓ_j we compute

$$\operatorname{logit} P(k \mid 8) = \ell_k - \operatorname{logsumexp}_{j \neq k} \ell_j$$

with float64 accumulation. This avoids cancellation when the favored option's conditional probability is extremely close to one. Exponentiating before taking a logit, or subtracting a nearly equal normalization term, can impose an artificial ceiling. A duplicate control detects numerical non-repeatability, but not arithmetic bias shared by both arms.

Conditional interpretation. Teacher-forced paired replay fixes both arms' action history and measures an element's conditional effect on a probe. It does not force decay: attention reallocation, renewed information and competing instructions can change damage in either direction. Free-running trajectories could compound errors or recover by rereading and clarification; paired replay does not identify those mediators.

Capability floor. Qwen3.5-0.8B and 2B have low single-turn base-lookup accuracy (0.214 and 0.143; eight-option chance is 0.125). Their flat damage curves could reflect incapacity. Qwen3.5-4B reaches 0.786 single-turn and 0.811 over a full trace, motivating exclusion of the smallest checkpoints and measured competence as the model-ladder abscissa (Section 4).

Numerical and off-diagonal controls. In the duplicate diagnostic (8154 reads), the mean, standard deviation and maximum absolute difference are all zero. Off-diagonal mean absolute differences are 0.181 for para, 0.339 for corrupt, 0.374 for pad and 0.396 for drop. Repeatability and element specificity are therefore separate checks. The placebo subtracts the mean effect on unrelated probes at the same distance and is estimated once on the full sample. Rebuilding it inside each family bootstrap shifts the six 4B intervals by at most 0.009 in y without changing their classifications. The generic channel itself measures off-diagonal effects, so using those effects as its amplitude threshold would be circular.

Sensitivity versus decision loss. The boundary $p = 1$ applies to the accumulated read-out. For example, $u^{\mathrm{keep}}(d) = d^2/(d^2+1)$ and $u^{\mathrm{arm}}(d) = d^2/(d^2+2)$ give a constant log-odds difference $\log 2$, while excess error $u^{\mathrm{keep}} - u^{\mathrm{arm}} \sim d^{-2}$ is summable.

Refitting the joint model on placebo-differenced excess error $p^{\mathrm{keep}} - p^{\mathrm{arm}}$ leaves all 15 merged (model, element) distance coefficients below 1. The coefficients range from -0.575 to $+0.225$, versus -0.518 to $+0.158$ on log-odds. The largest discrepancy is 27B plan (-0.133 versus -0.575). These are point-coefficient comparisons, not new interval-based classification tests.

At 321–1153 steps, deleting constraint reduces the mean correct-action probability from 0.813 to 0.081, an excess error of 0.732, versus 0.912 at $d \leq 3$. For plan, excess error falls from 0.428 at $d \leq 3$ to 0.04–0.07 beyond $d = 10$, while log-odds damage remains 0.72–1.04. The keep-arm correct-action probability is only 0.09–0.15 there. Its residual log-odds effect should therefore not be read as evidence of maintained decision competence. The subset evaluation also averages distance-dependent probes, so it remains a comparison of fixed-history log-odds effects rather than task success.

B CONFOUNDS THE DESIGN WAS BUILT TO CATCH

Position masquerading as distance. PERSONA's pooled damage grows from 1.85 to 3.00 lo, but the distance-bin means over $[1, 10)$, $[10, 50)$ and $[50, 181]$ are $1.99/3.56/3.54$ at $s_0 = 0$, $3.38/4.08/3.09$ at $s_0 = 24$ and $3.25/3.91/3.15$ at $s_0 = 96$. At $s_0 = 0$, distance equals position: reference accuracy rises from 0.580 before position 10 to 0.83 after it, so early competence can explain the apparent growth. Omitting position gives $p = -0.255$. In contrast, CONSTRAINT, PLAN, GOAL and FORM show distance decay within every delivery stratum (CONSTRAINT: $8.09/8.34/5.34$ at $s_0 = 0$ and $9.73/7.60/6.76$ at $s_0 = 96$). This motivates the four delivery points and position control.

Table 3: Correct final action in the *drop* arm, out of 24 paired base cases per element and distance, on local Q4 Qwen3.5-27B. The same 24 base cases are reused at both distances; they are not 48 independent cases. Keep-arm accuracy is 23–24/24 for PLAN and 24/24 for CONSTRAINT. These action counts are not the BF16 log-odds read-out.

element	d	original filler	neutral filler	relevant demos	unrelated demos
PLAN	32	24	2	19	5
PLAN	128	23	2	21	5
CONSTRAINT	32	1	0	9	0
CONSTRAINT	128	1	0	9	0

Changes in the reference arm. A change in $\delta = \mathrm{lo}_{\mathrm{keep}} - \mathrm{lo}_{\mathrm{arm}}$ can come from either arm, so we report both for the recovery schedule (Section J). On τ-bench, the keep-arm slope of $+0.181$ could produce growing damage if the perturbed arm stood still. Moreover, that analysis fits unsigned damage; the reference-arm confound remains open (Section H).

Separating distance, position and recovery. Multiplying estimates from separate partial fits can double-count channels already absorbed into each distance coefficient. On 4B and 9B the discrepancies are small (≤ 0.06; rank correlations $+0.886$ and $+0.486$), but crossing model families exposes sign changes. The joint analysis uses $\delta = A\, d^{e}(\mathrm{step}/10)^{f}(1 + n_{\mathrm{prior}})^{-r}$, holding position and demonstrations fixed when estimating e.

Demonstration count is deterministic in distance, but the non-monotone schedule ($d = 2$ has one demonstration, $d = 5$ none) permits parameter estimation within this functional form: the log covariates have VIF 2.1–3.2. The joint and two-covariate estimates agree closely on placebo-differenced damage (4B PLAN: $0.694\,[0.586, 0.831]$ versus $0.709\,[0.615, 0.844]$; maximum difference in r, 0.02). Using raw diagonal damage instead raises that discrepancy to 0.07. Distance estimates from the non-overlapping survey and recovery schedules agree closely for three elements but only to the first digit for the two FORM channels.

The joint fit also attributes 4B PERSONA's apparent growth to position ($f = +0.112$; $e = +0.054$ with an interval spanning zero), and Gemma PLAN's apparent distance decay to position ($f = -1.312$, $e = +0.518$; Section 8). These estimates remain conditional on the assumed form, as detailed below.

Placebo construction. We construct the distance-matched control before restricting to diagonal rows. Both form channels use the same off-diagonal form control; channel names are split only after subtraction. Missing controls are flagged.

Post-hoc matched filler and demonstration control. The original routine filler prints a class and its correct phase at every step, so it can teach the mapping assigned to PLAN even when that element is deleted. We intervened on this cue on a local Ollama Qwen3.5-27B Q4_K_M checkpoint, with the final probe fixed at step 160. For each of six lexical families and four stable seeds, we held the final observation, options and answer fixed, delivered the brief at $s_0 = 128$ or 32 ($d = 32$ or 128), and compared keep and drop arms. Four teacher-forced histories were paired: original class–phase filler; filler with the phase replaced by `?`; the latter plus three relevant probes at steps $140, 147, 154$; and the latter plus three unrelated probes at those same steps. The two demonstration types use identical option menus. All four histories receive the same rule that a routine item with unknown phase uses its base-class action. Thus the "original" row below is under a common augmented protocol, not a verbatim rerun of the BF16 experiment.

Removing the filler cue nearly eliminates correct answers after deleting PLAN; it barely changes the deleted-CONSTRAINT negative control. Relevant examples restore more PLAN answers than unrelated examples and restore fewer CONSTRAINT answers (Table 3). This isolates a task-specific effect of the three late examples at fixed probe and distance, not a general re-learning law: the two elements use different probes, the examples are close to the final decision, and the six families vary wording rather than mechanism. Replacing phase digits by `?` also changes tokenisation slightly. The common instruction, near-ceiling keep arms and constraint control narrow but do not remove that concern. This post-hoc Q4 diagnostic neither refits the original BF16 amplitude nor identifies its fitted r.

C COVERAGE AND DESIGN BALANCE

All reported survey results use the complete, balanced design: 432 trajectories at each of four delivery points, 1728 total. Repeated writes are deduplicated by model, trajectory, element, condition and probe step before fitting.

Holding the position-controlled estimator fixed, content demeaning changes p as follows: CONSTRAINT $0.024 \to 0.018$, FORM (target) $0.014 \to 0.009$, GOAL $0.093 \to 0.087$, PERSONA $-0.046 \to -0.040$, PLAN $0.177 \to 0.148$ and FORM (audit) $-0.064 \to -0.025$. The largest moves, 0.029 and 0.039, affect the second digit but no classification relative to $p = 1$. Audit and persona exchange ranks because the 0.039 shift exceeds their original 0.018 separation (Spearman 0.94); both exponents remain negative.

Content robustness of recovery. Because demonstration count also varies with probe content, we separately check whether content difficulty explains r. Both checks use placebo-differenced δ_{DD}.

First, a multiplicative content fixed effect, $\delta = A_c(1+n)^{-r}d^e$, absorbs each content's level while sharing r and e. The largest change in r is ≤ 0.019 on 4B, 9B and 27B, 0.055 on Qwen3.8-27B and 0.183 on Gemma; the Gemma outlier is PLAN, whose r remains indistinguishable from zero. Rank correlations are $+0.94$ to $+1.00$. On 4B the cross-channel span increases from $5.72\times$ to $6.37\times$, and the PLAN/CONSTRAINT ratio from 4.44 to 4.63. Only GOAL and PERSONA swap ($0.133/0.124 \to 0.113/0.123$); their intervals overlap substantially ($[0.069, 0.213]$ and $[0.067, 0.183]$).

Second, fitting each of the eight probe contents separately puts PLAN above CONSTRAINT in **all eight** (0.430–0.849 versus -0.032–0.253 on 4B). The median ordering matches the pooled one apart from the same goal/persona swap.

Identification limit. These checks address content-level difficulty, but cannot separate recovery from a content-dependent distance effect. Within each (content, distance) cell, $n = g_c(d)$ takes only one value. Consequently,

$$\delta = h(d, s, c), \qquad h(d, s, c) = A\, d^e (s/10)^f \left[1 + g_c(d)\right]^{-r}$$

fits every observed cell equally well without an independent demonstration effect. The non-monotone schedule breaks regressor collinearity within the assumed form; it does not identify that form. A content fixed effect absorbs levels, not distance-dependent shapes.

Identification would require the same probe, distance and position at different demonstration counts, which this design lacks. We therefore report r as an empirical damage-decay parameter under the specified model, rather than a measured re-learning rate. Its ordering contribution is evaluated directly in Section N; the fitted priority has no resolved advantage over sorting on measured damage (Section L). This limitation leaves the conditional distance results intact. The fixed-probe intervention in Section B identifies a local effect of three late relevant examples, but does not turn the original schedule's r into an independently measured re-learning rate.

D CUMULATIVE CROSS-VALIDATION AND THE REPRESENTATION LADDER

The deep configuration measures placebo-differenced drop damage over 144 trajectories, three delivery points and horizon 1280. Table 4 gives log-odds means by distance bin (6–60 reads per cell).

Table 4: Placebo-differenced drop damage by distance bin (log-odds) in the deep Qwen3.5-4B configuration.

element	$d \leq 3$	4–10	11–32	33–100	101–320	321–1153
CONSTRAINT	8.80	8.53	9.19	5.79	7.24	**4.40**
FORM (audit)	6.43	6.02	6.22	4.86	5.98	**3.85**
GOAL	3.76	4.48	4.64	4.02	3.11	2.88
PLAN	4.50	4.31	0.65	1.09	1.18	0.80
PERSONA	1.65	2.61	2.52	2.84	2.06	1.89
FORM (target)	0.47	1.05	0.80	0.76	0.81	1.01

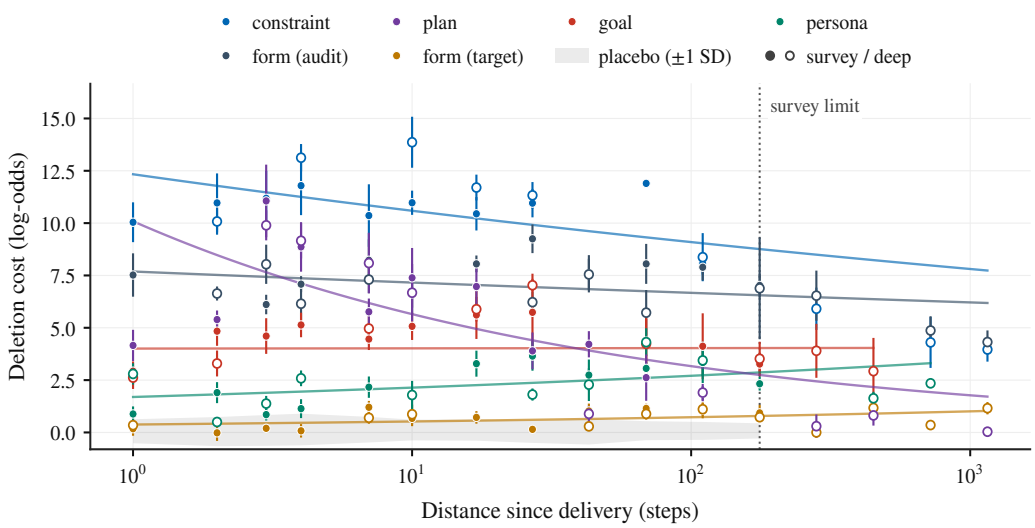

Figure 4: Descriptive distance-only power fits to the raw effects in Figure 1, jointly pooling the matched survey and deep cells. Faint curves summarize each channel; fitting them does not test whether the two grids share an exponent or rule out a bend. These curves are distinct from the position-controlled fits and the placebo-differenced bin means above.

The single $d = 1153$ constraint cell is 3.94 on three reads, so the band mean is the more broadly supported far-distance summary.

Partial distance and the age path. In $\delta(d, s, n) = A\, d^{-p}(s/10)^f(1+n)^{-r}$, p holds position and demonstrations fixed. An element delivered at s_0 instead ages along $s = s_0 + d$; for $d \gg s_0$ at fixed demonstrations, $p_{\text{age}} = p - f = -(e + f)$, which can differ in sign from p. The more stable combination $e + f$ is examined in Section M.

Survey age-path exponents range from -0.164 (PERSONA) to $+0.210$ (PLAN). Across 48 (configuration, element) cells they range from -0.31 to $+0.604$ (Gemma PLAN). None reaches 1, so the classifications in Table 1 persist under power-law continuation. The main $|p| \le 0.177$ bound concerns the survey's partial coefficient; the positive age-path bound is 0.210 there and 0.604 across configurations.

Future retention cost additionally requires the rate $q_e(d)$ of governed decisions: $C_e(T) = \sum_{d \le T} q_e(d)\, \delta_e(d, s_0 + d, n_e(d))$. The illustrative priority column assumes a governed-decision rate of 0.05 rather than measuring this rate. If $n(d) \sim \kappa d$ and $q(d) \propto d^{-\beta}$, then

$$q(d)\, \delta_{\text{age}}(d) \propto d^{-(p_{\text{age}} + r + \beta)}.$$

The future-cost boundary is therefore $p_{\text{age}} + r + \beta = 1$. Fitted r alone reaches 1.75 for Gemma GOAL and 0.69 for 4B PLAN, and can change the classification. Our measured conditional sensitivity does not establish unbounded future retention value.

Cumulative curve comparison. Leave-one-family-out cross-validation compares cumulative power, logarithmic and bounded forms. Power-law held-out MSEs are 1907 for CONSTRAINT, 2090/270 for FORM (audit/target), 11759 for GOAL, 8340 for PERSONA and 9231 for PLAN. On constraint and the two form channels, logarithmic MSEs are 4241–31981 and those of $D_\infty - Ct^{-q}$ are 4301–33209, losing by 12–16×. These two shapes also lose on goal and persona (1.9–2.3×), while the logarithm is only 17% worse on plan, within the 0.25 relative tie margin.

Crucially, the bounded cumulative exponential $A\tau(1 - e^{-t/\tau})$ ties the power law on every element (MSE ratios 0.99–1.02). Every cumulative classification is therefore undetermined, and we report no y_{cum}. The observed window cannot distinguish unbounded power-law accumulation from exponential saturation outside it.

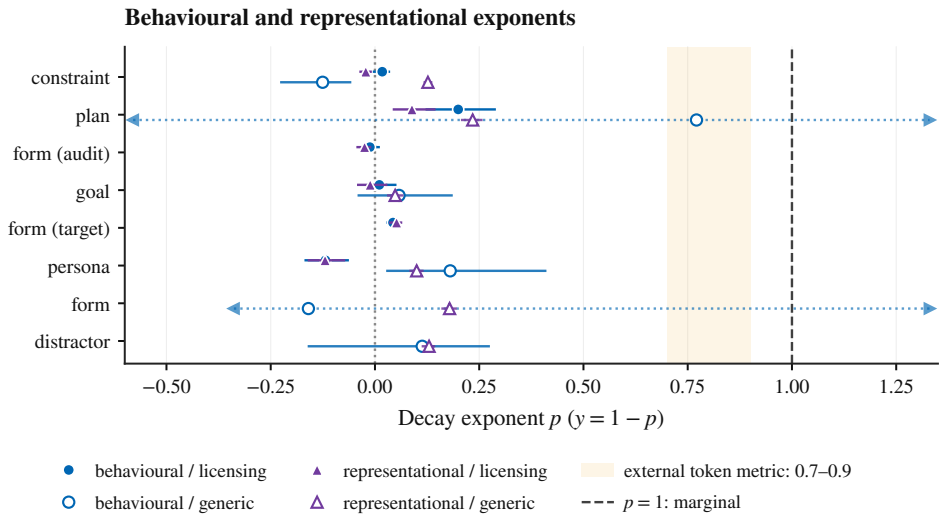

Figure 5: Distance-only survey exponents for behavioural log-odds and a representational probe, on licensed decisions and generic next-token positions. Bars are 95% intervals; arrows indicate intervals extending beyond the axis. The shaded 0.7–0.9 band from Brändel et al. (2026) uses Jacobian sensitivity against token distance, a different observable and distance unit. It is an external reference, not a matched baseline.

Table 5: Recovery exponent on placebo-differenced damage, Qwen3.5-4B, 864 trajectories, family-clustered bootstrap. The 5.7× spread is positively correlated with amplitude ($\rho = +0.71$); the plan/constraint contrast is pair-specific. Using raw diagonal damage changes r by up to 0.07 and swaps the form channels; matched read-outs agree within 0.02.

element	n	r	95% CI	after $2 \times (1 + n)$	y at fixed demos
PLAN	6912	**0.709**	[0.615, 0.844]	61%	0.852
FORM (audit)	2376	0.404	[0.346, 0.457]	76%	0.818
FORM (target)	7632	0.366	[0.314, 0.425]	78%	0.893
CONSTRAINT	7056	0.160	[0.083, 0.267]	90%	0.951
GOAL	6912	0.133	[0.069, 0.213]	91%	0.953
PERSONA	7128	**0.124**	[0.067, 0.183]	92%	1.128

Per-step curve comparison. Constant, power, power-with-offset, exponential, exponential-with-offset and stretched-exponential fits likewise tie (exponential/power MSE ratios 0.98–1.05). For FORM (audit), a constant is best (5.009 versus 5.023 for power); for CONSTRAINT, its MSE 6.944 is within 0.7% of the best exponential (6.898). Thus these two elements show little measured decay even without choosing a power law, consistent with near-zero distance exponents.

Representation and behaviour. Distance-only survey behavioral exponents span −0.119–0.200 on licensed probes; generic-channel intervals can reach the fitting bounds and are uninformative. Cosine-probe exponents span −0.120–0.089 on licensed probes and 0.048–0.234 on generic probes. These fits are shown in Figure 5. Neither reproduces the published $p \approx 0.7$–0.9: that measurement uses Jacobian sensitivity against token distance in natural text, while ours uses decision steps in structured, recurring traces. Across four network depths and 72 trajectories, the signal appears mid-network but all exponents remain below 0.7. We therefore cannot attribute the discrepancy to behavioural versus representational decay or to layer choice. Cosine displacement is decision-specific: 0.04–0.13 on licensed positions, 0.003–0.005 elsewhere, and 2.4e-7 on DUP.

Table 6: Deleting a derivation versus rewriting it wrongly. Raw diagonal log-odds means on the 4B survey and 9B universal run, with 576 and 144 matched reads per row, respectively. Both arms pass the gate on the same reads; the difference column is corrupt minus drop.

		Qwen3.5-4B			Qwen3.5-9B		
probe	guise	drop	corrupt	diff	drop	corrupt	diff
target	reasoning	0.50	11.01	+10.51	1.38	11.81	+10.42
target	fact	5.93	9.78	+3.85	6.66	10.94	+4.27
target	toolout	1.61	2.83	+1.22	3.40	5.33	+1.93
audit	reasoning	6.43	9.41	+2.98	7.11	10.63	+3.53

Table 7: Fitted half-lives in steps on Qwen3.5-4B: causal damage from the single-probe survey, attention from 288 repeated-probe trajectories. Persona's causal fit reaches the optimization bound $\tau = 10^5$; this does not identify a 69,315-step half-life.

element	causal	attention	causal / attention
PLAN	51	89	0.6
CONSTRAINT	448	91	4.9
FORM	1245	91	13.6
GOAL	1563	90	17.4
PERSONA	unidentified	101	—

E RECOVERY AND CORRUPTION TABLES

Scope of the corrupt/delete gap. In the reasoning guise, deletion removes the derivation but retains its export conclusion. On the 4B survey this costs 0.50 lo, versus 11.01 for corruption, a gap of 10.51; the 9B gap is 10.42 (Table 6). Drop-arm correct-action probability remains 0.946 against reference 0.964. Without a retained conclusion, deletion costs 5.93, so low deletion cost does not generalize to all derivations. The tool-output guise has low reference competence, limiting the interpretation of its smaller effect. The audit channel is evaluated on the reasoning guise here; the subset experiment evaluates both destinations in all guises.

F ATTENTION AND CAUSAL DECAY SUMMARIES

MemDecay (Matam & Kim, 2026) fits regional attention decay. We apply its log-linear estimator $\log \bar{A} \approx \log c - \lambda \Delta t$ to attention on 288 repeated-probe trajectories, balancing the four delivery arms. The raw attention half-lives range from 89 to 101 steps (1.13×). Causal half-lives instead fit exponentials to the single-probe survey's log-odds damage curves, weighted by cell standard errors.

These are different schedules, not paired measurements from the same forward passes. The attention run includes repeated demonstrations; the causal survey has no earlier probe of the same kind. Their half-life spreads are descriptive evidence that the two summaries can differ on this instrument, not an isolated test of attention's ability to predict causal decay. Moreover, causal half-lives beyond the 176-step survey window depend on extrapolating the exponential fit.

Plan and constraint differ by 8.8× in fitted causal half-life but only two steps in attention. Attention's log-linear fits have $R^2 = 0.58$–0.68: the share drops sharply early and then flattens, so a single exponential is only a summary. Inserting these attention rates into the simplified tier-based score produces the type-tier eviction order. This is specific to its tiers and fixed comparison age; it does not evaluate MemDecay's full access mechanism. The attention/causal rank correlation is +0.700 over five types ($p = 0.188$), insufficient to establish a general rank mismatch.

Attention's limitations as an importance signal are established: Wen et al. (2022) study explanations, Mai & Kim (2026) examine missing value and projection structure in KV eviction, and Javadov (2026) contrast routing mass with causal contributions. Here the additional comparison concerns fitted decay rates, subject to the schedule limitation.

Architecture dependence. Gemma-4-31B has a 1024-token sliding window in 50 of 60 layers. Beyond it, brief attention is zero in all 92,800 local-layer cells, versus none of 18,560 global-layer cells. The ten global layers show little decay over 106 steps (ratios 0.7–1.6, three of five elements increasing, $R^2 = 0.00$–0.13). An exponential attention-decay rate is unsupported there. Qwen3.5/3.8 are hybrid architectures: only 8 of 32 layers (16 of 64 at 27B) expose an attention matrix.

G Cross-model distance fits

The 48 cells in Section 4 use distance-only fits on content-demeaned, placebo-differenced effects; Table 1 controls position. Controlling position in all 48 cells preserves every classification and increases the smallest interval-to-boundary distance from 0.547 to 0.711. However, universality runs have only two delivery points, versus four in the survey. Gemma plan then has an unstable split ($e = +1.16$, $f = -1.70$, $A : 17.6 \to 1.15$), the identifiability problem examined in Section M.

The distance-only comparison restricts runs to common design cells before demeaning probe content and fitting. The position-controlled sensitivity uses each run's available delivery grid; it is therefore a complementary analysis rather than a covariate-only refit on an identical sample. Both compute the off-diagonal control before splitting form channels or filtering to diagonal rows.

H The τ-bench read-out

The external analysis measures deletion of policy bullets in 104 transcripts, with 624 (transcript, rule) ablations. Its regressand is log mean absolute log-odds change, unlike the signed synthetic read-out. Distance has two sources of variation. Within a transcript–rule pair, content is fixed but distance and position are collinear, yielding $p = +0.199\,[+0.120, +0.280]$. Across rules, position can be held fixed but rule content differs, yielding $p = -0.224$. Including rule dummies and log position gives $p = -0.176\,[-0.297, -0.074]$, clustering on transcripts. The point lies within the synthetic survey's $|p| \leq 0.177$ band and below the $p = 1$ boundary, although these estimators and observables differ.

Transcript clustering accounts for residuals shared by the roughly six ablations from each conversation. The raw correlation between log distance and log position is 0.672; after residualizing on the included rule dummies it is 0.922 (VIF 6.66). Separating their effects is therefore less stable than the raw correlation suggests.

The reference arm's own log-odds rise with distance (slope $+0.181$). Such a rise can increase a signed paired difference even when the perturbed arm is unchanged. Here 44% of deltas are negative and the fit uses absolute effects, so a signed slope on one arm does not exclude that confound. The median transcript has only nine assistant turns, tool calls are omitted by the renderer, and trajectories are off-policy for the measured model. This one-model check documents sensitivity to third-party policy text over its observed range; it does not independently identify a distance law or replicate the synthetic universality test.

I External evidence-deletion control

We ran a separate, post-hoc control on the HotpotQA dev/distractor documents (Yang et al., 2018), using an already frozen 320-question multiple-choice adaptation. The 80 cases with the smallest SHA256 hash of the case ID and a fixed salt were selected before running Qwen3.5-4B BF16; selection did not use its answers. Each case has two annotated supporting documents. We compare the full ten-document context, deletion of both supporting documents, and deletion of two distractor documents chosen to match the support pair's total character length as closely as possible. The question, four options, document order among survivors and protected evidence-use instruction stay fixed. All 240 calls returned an option letter. Mean deleted lengths are 961 versus 1012 characters; the corresponding mean prompt lengths are 1339 versus 1330 model tokens.

context	correct / 80	mean log probability of gold, conditional on A–D
all documents	70	−0.279
two supporting documents deleted	25	−2.229
two matched distractors deleted	73	−0.266

Table 8: Rank correlation of each measured quantity between models, over the ten pairs from five checkpoints ($n = 5$ merged element types per pair). Both estimators use placebo-differenced damage. Their fitted r values differ slightly, but yield identical element rankings and hence identical correlations.

quantity	estimator	10 pairs	within family (6)	**cross family (4)**
A amplitude	controls demos	$+0.600$	$+0.683$	$+0.475$
	joint fit	$+0.500$	$+0.633$	$+0.300$
y scaling dim.	controls demos	$+0.670$	$+0.850$	$+0.400$
	joint fit	$+0.130$	$+0.450$	-0.350
r recovery	controls demos	-0.030	$+0.133$	$\mathbf{-0.275}$
	joint fit	-0.030	$+0.133$	$\mathbf{-0.275}$

The distractor-minus-support contrast in conditional gold log probability is $+1.962$ $[+1.579, +2.371]$ (95% question bootstrap). In the paired accuracy comparison, distractor deletion wins 48 cases and loses none against support deletion. The effect is seen on both question types: full/support-deleted/ distractor-deleted accuracy is $57/19/59$ on 65 bridge questions and $13/6/14$ on 15 comparison questions. This is a positive control for evidence-specific deletion effects on third-party text, not an external replication of the distance exponent or a test of a deployable selector: support labels chose the deleted pair, the options expose possible answers, and these menus were already used in a separate compact-method study. The frozen prompts, all 240 responses and the analysis are retained.

J TRANSFERABILITY, IN FULL

Transfer and model units. The recovery design is identical across Qwen3.5-4B, 3.5-9B, 3.5-27B, 3.8-27B and Gemma-4-31B, with 864 trajectories each. Table 8 reports two estimators: amplitude ordering transfers moderately, while recovery ordering has weak within-family correlation ($+0.133$) and negative cross-family correlation (-0.275). These are descriptive ranks over five element types, not evidence of a cross-family inversion. The ten model pairs share endpoints, and all four cross-family pairs share Gemma, so they cannot be treated as independent replicates for a paired test.

Amplitude and competence. Gemma has mid-ladder competence (0.942) but $A(\text{CONSTRAINT}) = 49.7$ against 7.7–14.4 elsewhere: its median licensed log-odds is 30.3, versus roughly 5 for Qwen. Exponents are dimensionless slopes; amplitudes carry each model's log-odds units and must be re-measured on the deployed model.

Across the six Qwen checkpoints, amplitude correlates with competence at $\rho = +0.886$ for GOAL (exact two-sided permutation $p = 0.033$), $+0.829$ for CONSTRAINT ($p = 0.058$) and both FORM channels, $+0.257$ for PERSONA ($p = 0.66$), and -0.429 for PLAN ($p = 0.42$). The pattern persists across all eight models and the seven excluding Gemma; the latter set includes GLM-4.7-Flash and is not one family. The positive association is suggestive because competence and amplitude share reference-arm probabilities. These measurements offer no evidence that the failure mode discussed by Chen (2026) shrinks with capability.

The fit-free read-out is descriptive. Within $d = 1$–5, we compare placebo-differenced damage at zero and three prior demonstrations. A smaller difference can reflect improvement in the perturbed arm or deterioration in the reference arm (Section B). Even perturbed-arm improvement cannot isolate demonstrations: zero-demo reads occur at $d \in \{1, 5\}$ and three-demo reads at $d = 4$. Probe content also changes. The two groups share no content IDs for GOAL, PERSONA or PLAN, one of six for CONSTRAINT, and four of seven for FORM, identically on all four models used for this diagnostic (4B, 9B, 27B and Gemma).

Both arms are evaluated on the same read-outs. On $\overline{\text{logit}\, p}$, three of twenty cells show no perturbed-arm improvement; using $\text{logit}(\overline{p})$ instead flags five. The distinction matters: for Gemma PLAN, logit of mean probability suggests little reference change ($18.51 \to 18.27$), while mean logit falls 1.29 as the perturbed arm rises 2.20. For 4B PLAN, perturbed mean probability $0.129 \to 0.121$ masks a 0.97 rise in mean logit. Its reference mean logit falls $3.98 \to -1.18$, accounting for 84% of the 79% damage decline. Thus this read-out neither establishes recovery nor a cross-family inversion.

Table 9: Complete trajectories retained / initially evaluated, by model and guise. Completion requires all six probes in keep and all 63 deletion subsets to pass the option-mass gate. Every keep probe passes; Gemma's exclusions occur under deletion, and its 75 cases fall below the prespecified $n \geq 100$ requirement.

model	all guises	fact	reasoning	tool output
4B	216/216	72/72	72/72	72/72
27B	216/216	72/72	72/72	72/72
Gemma	75/216	12/72	55/72	8/72

The stronger correlations $+0.90/+0.72/+0.56$ within family and $-0.40/-0.70/-0.21$ across it describe these confounded fractions, not fitted r. The joint fit supplies the parameter comparison in Table 8, subject to the identification limit in Section C.

Sensitivity of the descriptive read-out. Across 24 choices of distance bandwidth, demonstration split and pooling, the fraction removed spans 0.763 for PLAN and 0.295 for CONSTRAINT. Plan has the largest decline in $22/24$ variants; 9B falls below 4B in $24/24$ (median ratio 0.40, versus fitted 0.42); the claim that persona changes on neither arm survives only $6/24$ and is not supported.

Recovery is not monotone in scale. Mean r is 0.316, 0.110 and 0.205 on 4B, 9B and 27B. All six elements decline from 4B to 9B, but on 27B CONSTRAINT, GOAL and PERSONA rebound to $1.68\times$, $1.86\times$ and $1.02\times$ their 4B values.

K What equivalent task length means

The multiplier $\exp(\Delta/\gamma)$ assumes $\operatorname{logit} P_{\text{success}}(t) = a - \gamma \log t$ for one success notion and a horizontal intervention shift. Our Δ instead measures an eight-option conditional decision, averaged over probes and budgets in eviction. METR (Kwa et al., 2025) fits whole-task horizons against human completion time; BRIDGE (Liu et al., 2026) fits a separate difficulty–time map in a two-parameter IRT model. Neither calibrates this conversion.

Temperature scaling can multiply Δ without changing any $\arg\max$, so the implied duration is not decision-invariant. We therefore report Qwen multipliers only as illustrations. The model-specific log-odds scales in Section J preclude applying $\gamma = 1.5$ to Gemma's policy margins.

L Eviction evaluation and uncertainty

Read-out and complete cases. Each model evaluates 216 trajectories (6 families $\times 3$ guises $\times 12$ rotations), 63 non-empty subsets, keep and duplicate: 13,608 subset read-outs and 14,040 forward arms. Damage is averaged over six licensed probes per trajectory, including export and audit in every guise. The probes occur across the 1–176-step ladder; the brief is delivered at $s_0 = 0$. The empty deletion set has damage zero by definition. All results in Section 6, including additivity and the figure, use the same complete trajectories within each model.

Option-mass gate and its scope. Let $m = P(\text{output} \in \text{menu})$ and $q = P(\text{correct} \mid \text{menu})$. The read-out is $\operatorname{logit} q$; a trajectory is complete only when $m \geq 0.5$ for every probe in keep and every deletion subset. Table 9 gives the retained sample by guise. Failed subset cells and their complete-case implications are reported in Section 6.

The prespecified requirement for a stable cross-guise comparison was at least 100 complete trajectories with every guise represented. Gemma fails the sample-size requirement. We retain its results as descriptive observations; neither their intervals nor their ranks establish stable cross-guise performance. Coverage is counted before and after the gate for each guise; the failed subset-cell rates for Gemma are reported above.

Single-element costs.

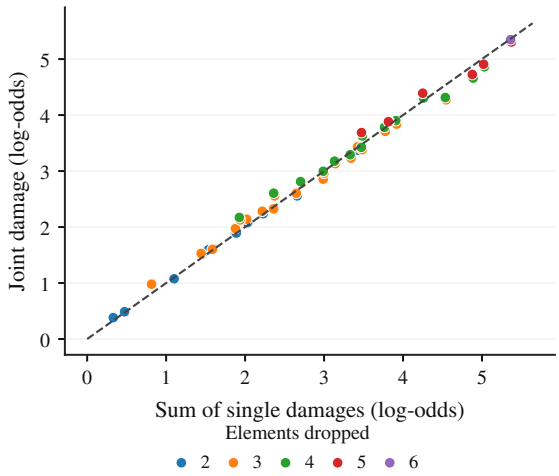

Figure 6: Joint damage versus the sum of measured single-element damages on 4B's 216 complete trajectories, for all 57 deletion subsets of size 2–6. The dashed line is equality; the median ratio is 1.004. This descriptive comparison differs from the held-out fits in Table 10. Approximate additivity does not guarantee an optimal static order.

Table 10: Fit on the 21 subsets of size ≤ 2 and score the 42 subsets of size ≥ 3, using complete-case subset means. Bias is measured minus predicted damage. Pairwise fits add all 15 interactions to the six effects.

		additive (6 parameters)			with pairs (21)	
model	n	RMSE	R^2	bias	RMSE	R^2
4B	216	0.0839	0.9934	+0.000	0.0546	0.9972
27B	216	0.2870	0.9582	-0.213	0.0665	0.9978
Gemma	75	0.5765	0.9869	+0.217	2.1890	0.8112

deleted element	4B	27B	Gemma
DISTRACTOR	-0.011	$+0.006$	-0.066
PERSONA	0.342	0.473	3.942
GOAL	0.484	0.669	1.683
PLAN	1.112	1.567	3.288
FORM	1.547	2.601	7.125
CONSTRAINT	1.887	2.451	8.531

These are raw log-odds damages averaged over the six probes, rather than the placebo-differenced fitted amplitudes in Table 1. Constraint is largest on 4B and Gemma, while form is largest on 27B. Form governs two of the six probes, so these costs reflect probe frequency as well as damage conditional on a particular element being needed.

Held-out additivity. High R^2 concerns prediction across subsets, not held-out families or free-running success. The test probes each depend on one brief element, although the model can still exhibit interactions. Pairwise terms improve Qwen prediction but increase Gemma RMSE from 0.576 to 2.189 lo. Thus fitting more interactions does not uniformly improve generalization.

Budgets and reference curves. The removable token-cost accounting is specified in Section 6.2. We evaluate 13 equally spaced retained-budget values from zero to the sum of costs. The ten interior values satisfy $b > \min_e c_e$ and $b < \sum_e c_e$ (indices 2–11). A policy deletes a prefix of its order until the remaining cost fits b. The plotted frontier evaluates every subset-cost breakpoint and is constant between breakpoints. The dots mark the ten original scoring budgets; this denser display does not

Table 11: Rank among all 720 static deletion orders using the same complete trajectories and ten interior budgets as Table 2. Ranks are descriptive positions, not uncertainty estimates. Identical orders necessarily share a rank.

policy	**4B** $n = 216$	**27B** $n = 216$	**Gemma** $n = 75$ (descriptive)
Best of all 720 orders	1	1	1
Direct measurement (frozen 4B)	8	16	40
Fitted Ae^{-r} (frozen 4B)	9	22	32
Fitted A (frozen 4B)	10	29	53
Cumulative score (frozen 4B)	94	155	57
Six-text judge (model's own)	40	97	128
Task-informed judge (model's own)	36	29	40
CWL-style type tiers	109	167	128
MemDecay (causal-rate variant)	148	283	24
Length	290	313	189

change their weights or the reported excess. The two references are defined in Equation (1); Figure 3 displays both.

Paired policy contrasts. The family-cluster resampling described in Section 6.2 uses seed 1. Trajectory weighting is a post-hoc analysis clarification; the frozen measurements and policies are unchanged.

contrast [95% CI]	4B	27B	Gemma (descriptive)
Six-text judge − direct	$+0.249\,[+0.194, +0.321]$	$+0.496\,[+0.439, +0.553]$	$+1.560\,[+0.637, +2.217]$
Informed judge − direct	$+0.214\,[+0.115, +0.325]$	$+0.103\,[+0.080, +0.127]$	0 (same order)
Fitted (A, r) − direct	$+0.006\,[-0.057, +0.081]$	$+0.075\,[+0.026, +0.123]$	$-0.149\,[-0.494, +0.099]$
Fitted (A, r) − fitted A	$-0.030\,[-0.072, +0.020]$	$-0.028\,[-0.056, +0.001]$	$-0.327\,[-0.459, -0.196]$

Negative values favor the first policy. The first three rows use direct as the reference, matching Figure 3. Direct measurement beats the six-text judge on the retained samples, but no fitted-minus-direct interval supports a fitted advantage. Gemma's informed order equals the direct order, so its zero contrast is an algebraic identity, not an equivalence test.

Audit-label sensitivity. We rescore the same per-subset damage records using either all six probes or the historical five-/six-probe labels on the common complete trajectories. Restoring the omitted audit labels changes fitted (A, r) minus direct excess by $+0.074\,[+0.065, +0.081]$, $+0.106\,[+0.100, +0.113]$ and $+0.176\,[+0.067, +0.249]$ lo on 4B, 27B and Gemma, respectively. For 27B, the same-run historical contrast is $-0.031\,[-0.087, +0.020]$ lo, whereas the six-probe contrast is $+0.075\,[+0.026, +0.123]$ lo. These are paired label-sensitivity results, not comparisons between independent model runs. Gemma's interval is the post-hoc trajectory-weighted cluster interval; its $n = 75$ remains descriptive.

Limits of static orders. No order among all 720 reaches O^{cs} at every budget. The smallest summed shortfalls over the 13 budgets are $1.252/1.509/4.984$ lo on 4B/27B/Gemma. Unequal costs make the optimal subsets non-nested even when damage is additive. Conversely, complementarity need not preclude a nested optimum: with equal costs and positive a_i, $D(S) = (\sum_{i \in S} a_i)^2$ has positive pairwise interactions, yet sorting a_i is optimal at every cardinality.

For any policy-independent reference R, replacing R in $\sum_b w_b[\bar{D}_\pi(b) - R(b)]$ shifts every policy's score by the same constant. It cannot change ranks when samples and budget weights are fixed. We checked this invariance by exhaustive enumeration of the 720 orders using the same trajectories and budget weights.

Data provenance. The evaluated BF16 runs use all six probe labels and fixed rendered designs. Earlier five-/six-probe evaluations supplied the calibration order and score exploration; their outcome tables are retained but are not pooled with the six-probe evaluation here.

M UNCERTAINTY IN THE DISTANCE–POSITION SPLIT

The joint model can estimate distance and position coefficients separately, but their uncertainty is substantial for plan on the larger models. Across recovery checkpoints, plan's position coefficient and the width of its distance-coefficient interval are $-0.049/0.063$ on 9B, $-0.076/0.087$ on 4B, $-0.503/0.139$ on 27B, $-0.898/0.287$ on Qwen3.8-27B and $-1.312/0.476$ on Gemma. These fits indicate model-dependent position effects; they do not identify a unique behavioral mechanism for those effects.

A design-matrix calculation gives variance inflation factors of 3.15 on $\log d$, 2.11 on $\log \text{step}$ and 2.08 on $\log(1+n)$. Candidate delivery grids reduce the first to about 2.32. Scaling a 0.474-wide interval by the square root of this ratio projects a width near 0.41; even a VIF of one at fixed sample size projects 0.267. These are variance approximations, not a proof that another design cannot improve identification. The delivery arms $s_0 = 0$ and 1 are particularly similar geometrically (within-arm distance–position correlations 1.000 and 0.998), although they differ in system-prompt versus early-trace delivery.

Family variation is another limit. Between/within-family standard deviations of plan's distance coefficient are $0.055/0.036$ on 4B, $0.203/0.108$ on Qwen3.8-27B and $0.329/0.226$ on Gemma. Adding seeds within the same six families cannot eliminate genuine between-family variation, although it can reduce estimation noise. The sum $e + f$, which determines the late age path at fixed demonstration count, is more stable: between-family standard deviations are 0.053, 0.102 and 0.193, respectively. We therefore report age-path coefficients alongside the partial distance coefficients (Section D). Their greater stability does not resolve the separate confounding of demonstration count with content-dependent distance effects (Section C).

N POLICIES AND JUDGE PROMPTS

Frozen measurement policies. The direct policy sorts mean single-element damage from the earlier 4B calibration enumeration. Its order, evict-first to evict-last, is DISTRACTOR, PERSONA, GOAL, FORM, PLAN, CONSTRAINT. It is transferred unchanged to the three six-probe evaluations; no evaluation outcome is used to reorder it. The calibration used five probes in fact/tool-output and six in reasoning, whereas evaluation weights all six probes in every guise. The direct score therefore tests a fixed measured ordering, not recalibration to each test mixture.

The fitted policies sort

$$\text{score}_e^A = A_e, \qquad \text{score}_e^{A,r} = A_e \exp(-r_e)$$

ascending. Both parameters come from the joint 4B recovery fit $\delta_{\text{DD}} = Ad^e(\text{step}/10)^f(1+n_{\text{prior}})^{-r}$, using its drop condition. The two form channels are averaged to one element; distractor has score zero. For constraint, plan, form, goal and persona, the amplitudes are $9.73/9.43/4.96/5.72/3.03$ and the demonstration exponents are $0.160/0.694/0.388/0.134/0.122$. Token costs enter budget feasibility, not these scores.

The cumulative ablation instead sorts estimated damage through step 176: $\rho \sum_{d=1}^{176} Ad^{-p}(1+\rho d)^{-r}$, with $\rho = 0.05$ used for both the governed-decision and demonstration rates. This rate is assumed, not measured. The table's 1000-step column additionally extrapolates to 50 demonstrations, beyond the recovery data's maximum of 11. The cumulative ordering is frozen on the same 4B calibration.

Selection and recovery discount. The discount point and selection history are described in Section 6. Recovery contributes one adjacent swap in the evaluated policy. Its excess change relative to amplitude alone is $-0.030\,[-0.072, +0.020]$ on 4B, $-0.028\,[-0.056, +0.001]$ on 27B, and $-0.327\,[-0.459, -0.196]$ lo on Gemma's descriptive sample. There is no resolved benefit on either Qwen model. The fitted score also does not outperform the direct ordering (Section L).

Matched BF16 judges. Each evaluated model receives 216 menus: six lexical families, three guises, six menu rotations and two prompt directions. Each menu is paired between a six-text and a task-informed condition, giving 432 queries per model. All use native BF16, greedy decoding, thinking disabled and a 64-token output cap. The texts are the actual delivered element strings. The two directions are:

```
You are the compaction step of a long-running agent.  The
context below is over budget and you must free space.  Rank
the numbered items from the one you would DROP FIRST to
the one you would DROP LAST. Answer with the numbers only,
separated by spaces, best-to-worst to drop, nothing else.
```

```
You are the compaction step of a long-running agent.  The
context below is over budget and you must free space.  Rank
the numbered items from the one you would KEEP AT ALL COSTS
to the one you would give up most readily.  Answer with the
numbers only, separated by spaces, most-important first,
nothing else.
```

The informed condition additionally gives the base task specification, states that the six probe kinds receive equal weight, and supplies one unlabeled example of each kind in every guise. It supplies neither correct replies nor the evaluated continuation. This is an exploratory fixed-prompt comparison: the task-informed prompt was developed after the initial six-text evaluations and then held fixed across these BF16 runs.

The parser collects the first occurrence of each index $1..6$, ignoring other text and repeated indices. A response missing any index is invalid and is neither completed nor regenerated. Parsed positions are put on an evict-first scale (keep answers reversed), averaged and sorted; ties are broken alphabetically. Invalid text-only/informed counts are $0/5$, $0/1$ and $0/3$ for 4B, 27B and Gemma. Each condition therefore aggregates the responses that parse, rather than discarding both members of an invalid prompt pair.

Table 12: Full six-probe BF16 excess (log-odds; lower is better) on each model's complete-case sample and ten common interior budgets. The final column is task-informed minus direct with a family-clustered 95% interval. Gemma's zero follows from an identical static order and is descriptive.

model	n	direct	fitted (A, r)	six-text	informed	informed−direct [95% CI]
4B	216	0.432	0.438	0.681	0.646	$+0.214\,[+0.115, +0.325]$
27B	216	0.611	0.686	1.106	0.714	$+0.103\,[+0.080, +0.127]$
Gemma	75	4.025	3.876	5.585	4.025	$0\,[0, 0]$

The 27B informed order equals the frozen amplitude-only order, while Gemma's informed order equals direct measurement. Informed-minus-direct is positive in each of the six lexical families on both Qwen models (0.055–0.140 lo on 27B). The 4B informed judge instead keeps goal last and constraint second-last. Table 12 collects the four BF16 policies and their paired contrasts; the policy definitions are specified above, and Table 2 reports the broader comparison. These results are specific to the fixed prompt and probe mixture and do not establish a general advantage over task-informed judges.

Other baselines. The CWL-style type score assigns distractor 0, form 1, persona 2, plan 3, and goal/constraint 4, with alphabetical tie-breaking. The MemDecay variant uses $b_e \exp(-\lambda_e \Delta t)$ at $\Delta t = 176$ steps and the same base tiers. Its rates are fitted on the evaluated model's causal damage curves, rather than the attention statistic used by the full system. It omits access-clock resets and other runtime mechanisms, so the comparison is to this variant only. Length evicts the largest token cost first. All policies delete an initial prefix of their order until the retained token cost fits the budget.

O Reproduction

All analyses use saved measurements; regenerating the tables and figures from the retained aggregates does not require model inference. The configurations and estimators are specified below.

Configurations. The survey has horizon 256, distance rungs $1, 2, 3, 4, 7, 10, 17, 27, 43, 69, 110, 176$, and four delivery points $s_0 \in \{0, 1, 24, 96\}$. Its factorial is 6 families $\times 3$ guises $\times 12$ rotations $\times 2$ seeds $\times 4$ delivery points $= 1728$ trajectories. Each probe kind occurs once. Recovery uses the same factors with one seed (864 trajectories) and repeated

blocks; within-block offsets extend observed distances to 181. Deep uses horizon 1280, 16 distance rungs through 1153, $s_0 \in \{0, 240, 600\}$, three families, reasoning guise, 16 rotations and one seed (144 trajectories). Universal uses two delivery points and one seed (432 per checkpoint); the 4B distance-only comparison restricts its survey to the common design cells. The depth check records four network depths on 72 trajectories. Attention uses 288 repeated-probe trajectories (4 rotations), and is a separate run from the survey used for causal half-life fits.

Distance, recovery and transfer. Table 1 fits $\delta_{\mathrm{DD}} = Ad^e(\mathrm{step}/10)^f$. Table 5 fits $\delta_{\mathrm{DD}} = A(1 + n)^{-r}d^e$. Fitted eviction scores and the joint rows of Table 8 add the position covariate to the recovery fit. The audit distance/recovery channel uses the reasoning guise; other channels pool the three guises. This scope differs from the six-probe mixture used for subset evaluation.

Cross-model estimates include distance-only, position-controlled and age-path fits. Bounded cumulative and per-step curve comparisons select families by leave-one-family-out error with a 0.25 relative tie margin. The joint distance fits are also compared on log-odds and excess error probability.

Six-probe subset and judge results. Saved damage, token costs and orderings for each model are rescored on its complete six-probe trajectories and ten common interior budgets. All 432 judge responses per model are retained and parsed as in Section N. The analysis enumerates 720 static orders, computes trajectory-weighted family-cluster intervals and fits additivity on held-out subsets. It checks the published BF16 judge scores and ranks against recomputed values; no failed calls or excluded trajectories are replaced. The corruption table uses matched form-deletion and form-corruption read-outs.

Figures. Figures 1, 2, 4 and 5 use the survey, deep and representation measurements described above. The distance figure restricts both grids to shared family, guise and delivery cells; these descriptive raw curves differ from the position-controlled fits on the full survey. Cross-model and representation panels use distance-only fits. The eviction and additivity figures use the same complete-case six-probe sample as their tables. Exact subset-cost breakpoints agree with the original scoring grid, and costs are constant between adjacent breakpoints.

External and filler controls. The Q4 filler intervention has 768 saved calls. The external evidence-deletion control retains its fixed prompts, 240 responses and analysis. The 80 questions are a post-hoc selection from an existing menu adaptation, not an untouched benchmark sample.

Compute accounting. Recorded subset-experiment allocations comprise 93.9 GPU-hours for the calibration enumeration, about 5.3 local GPU-hours for the 4B six-probe run, and 42.58 allocated GPU-hours for the final 27B/Gemma runs and judges, including pilots and preempted or overlapping attempts. The last figure excludes queue time and duplicate scheduler step records. These are component costs, not a total for all survey, recovery, universality, attention and external-control experiments.