Competitive Repair Capacity: When Valid Neural Edits Stop Competing

Anonymous Authors

Abstract

Model editing certifies that repair is *possible* with geometric quantities: editable rank, null-space dimension, codebook size. It leaves a second question open: an acceptable, successful edit can exist and still lose to an unacceptable alternative under the objective that selects edits. We define *competitive repair capacity*, the number q of valid successful actions beating every unacceptable one, and characterise it exactly: in finite families q is the deletion distance to optimizer-facing failure, its margin-refined form sharply characterises robustness to simultaneous action loss and bounded score error, and continuous families obey a volume law in which admissible dimension alone cannot fix capacity. Two real systems show these are not redundant. On a published GPT-2 XL AlphaEdit locus competitive capacity collapses to zero *across a one-dimensional increase* in null-space capacity; one GRACE reproduction is aligned at its released radius, fragile under an off-default geometry stress. An equation- and code-level audit of 41 protocols finds none exposing competitive capacity. A theorem-derived certificate then isolates aligned decisions prospectively — zero observed competitive failures among 141 certified states — yet abstains too often to deploy. Together these results separate the existence of repair capacity from its competitive, certifiable and deployable forms.

1 Introduction

Post-hoc repair of a trained model — knowledge editing, targeted unlearning, adaptor-based patching — is usually evaluated by asking whether the model can be changed successfully and whether the change stays local [Meng et al., 2022, 2023, Mitchell et al., 2022a, Yao et al., 2023]. Underneath sits a selection problem: every such system chooses an edit by optimising an objective over a family of candidate actions, and an acceptable repair existing inside that family does not mean the objective prefers it. A system may have a $6{,}000$-dimensional admissible subspace containing updates that achieve the requested behaviour while exactly preserving every protected key, and still select a maximiser that violates preservation — ample room, a valid repair, and the valid repair loses (Section 4).

The distinction. To argue that room for a repair exists, a method reaches for a *geometric* quantity: the rank of the admissible update, the dimension of a protected null space, the size of a codebook, the band of layers it may touch [Fang et al., 2025, Cai and Cao, 2024, Hartvigsen et al., 2023]. **Native (geometric) capacity** asks whether admissible repair directions remain available. **Competitive repair capacity** asks whether valid successful repairs remain *competitive* under the objective that chooses among actions. Native capacity alone is insufficient: a system can retain ample geometric capacity and valid repairs while having zero competitive capacity, and the second quantity obeys exact laws the first does not.

What the second quantity is, and what it covers. We realise it as a count q of valid successful actions beating every unacceptable one, with Γ_M its sign. The count is not bookkeeping: in our finite model q is exactly the deletion radius to optimizer-facing failure, and its margin-refined form sharply characterises robustness to simultaneous action loss and bounded score error. Its scope is equally exact: competitive capacity is a property of the selection problem, not of every downstream

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.

behaviour of the edited model. Four questions therefore come apart: whether an acceptable repair *exists*, whether one *competes*, whether competitive alignment can be *certified* from deployment-time observables, and whether acting on that certificate *improves* the capability/safety trade-off. *None of these implications is automatic*; each experiment below separates one level from the next, and our two attempts to cross the last arrow are reported as the negative results they were. Both real systems are CPU-feasible reproductions, which bounds breadth rather than the formalism (Section 8).

Relation to proxy-optimisation theory, and to our own mathematics. That optimising an imperfect proxy can move you away from the true objective is fully occupied by the reward hacking and Goodhart literature [Skalse et al., 2022, Manheim and Garrabrant, 2018, Karwowski et al., 2024], and we claim no priority over it. Specific here is the structure — a finite repair action family, an independently observed validity predicate partitioning it, and a *constrained* arg-max as the specification — and its behaviour under restriction, where Skalse et al. [2022]'s unhackability criterion is anti-monotone while Theorem 2 has the opposite sign. Nor is the contribution a new combinatorial primitive: the constituent arguments use standard ranking, perturbation and volume calculations, and their role here is to give one optimizer-facing repair quantity exact adversarial, stochastic and geometric meanings measurable inside real editing systems.

Contributions.

1. **An optimizer-facing repair object, characterised exactly.** We introduce competitive repair capacity and, organising standard mathematical ingredients around optimizer-facing repair robustness, derive exact repair-specific characterisations: q is a deletion radius, not a sign, sharply extended to joint deletion and L_∞ score error, with an exact stochastic deployment law and a quadratic competitive-volume law (Sections 2–3), machine-checked with zero violations.

2. **Valid repair availability can persist while competition collapses.** Changing only the optimisation target — same examples, candidates and architecture — moves structural validity of the selected action by $+0.888$ on $1{,}903$ contested sources; and on one published GPT-2 XL AlphaEdit locus competitive capacity collapses while thousands of null-space dimensions and a valid repair remain, in 157 of 230 cells (Section 4) — a statement about solution geometry, not a safety verdict on the method.

3. **What audited protocols expose.** Among 41 protocols from 36 papers, 21 of the 36 optimizer-facing ones report a native/geometric capacity, while none explicitly measures competitive capacity or action-family completeness; nor does any of six retrievable implementations (Section 5). The audit corroborates that the distinction is live in practice; it does not carry it.

4. **One mechanism, two competitive regimes.** A vulnerability-blindly selected GRACE reproduction behaved as the predicted aligned control at its released radius; a predeclared geometry-only stress then produced interacting memory regions where q collapses before repair availability does, and we prove the exact separating condition — *own-radius separation*, not ball disjointness (Section 6).

5. **A certification boundary: purity without utility.** Deployed prospectively as a label-free gate on fresh GRACE data, that condition was pure — zero observed competitive failures among 141 certified states — yet too conservative to deploy, and we prove why: geometry settles the order conjunct of competitive alignment and cannot settle the semantic one (Section 7). The negative half is part of the contribution: it separates the fourth level from the third.

2 Competitive repair capacity and robustness

Setup. A repair problem is a tuple (A, M, V, S): a family A of candidate actions; a metric $M : A \to \mathbb{R}$ the system maximises; a validity predicate V recording whether an action respects the declared structural invariant; and a repair specification $S \subseteq A$ recording whether it achieves the

requested behaviour. Write $G(A) = \{a \in A : V(a) = 1,\ a \in S\}$ for the *valid successful* (GOOD) actions and $B(A) = A \setminus G(A)$ for everything else. V and S are observed independently of M, and in every system we study V does not appear in M — a property of the published method, not a modelling choice.

Definition 1 (Competitive gap, frontier, and capacity)**.** With $b^*(A) = \sup_{b \in B(A)} M(b)$ and $\sup \emptyset = -\infty$, put $\Gamma_M(A) = \max_{a \notin G(A)} M(a) - \max_{a \in G(A)} M(a)$, $G_+(A) = \{g \in G(A) : M(g) > b^*(A)\}$ and $q(A) = |G_+(A)|$; margin-τ variants use $M(g) \geq b^*(A) + \tau$. We call q the *competitive repair capacity* of (A, M, V, S).

q counts good actions that individually beat every bad one; it is not the candidate count, the valid-action count or repair completeness. Native capacity is a statement about A alone, competitive capacity about the pair (A, M) under the partition induced by (V, S) — and the reason to count rather than take a sign is that the count is an exact robustness radius.

Theorem 1 (Competitive robustness)**.** *Let A be finite with $B(A) \neq \emptyset$ and write $q_\tau^{>} = |\{g \in G(A) : M(g) > b^* + \tau\}|$, so that $q = q_0^{>}$.* (i) Sign. *$q \geq 1 \iff \Gamma_M < 0 \iff$ every maximiser of M lies in $G(A)$; $q = 0 \iff \Gamma_M \geq 0$; and every maximiser is bad iff $\Gamma_M > 0$.* (ii) Deletion radius. *The fewest good actions an adversary must delete before a bad action attains the optimum is exactly q,*

$$R_{\mathrm{del}}(A, M) = \min\{|D| : D \subseteq G(A),\ \Gamma_M(A \setminus D) \geq 0\} = q,$$

uniquely witnessed by $D = G_+(A)$. (iii) Joint robustness. *Suppose the scores are known only up to $\|\tilde{M} - M\|_\infty \leq \eta$ and the adversary may also delete up to k good actions. Then $\Gamma_{\tilde{M}}(A \setminus D) < 0$ for* every *admissible attack if and only if $q_{2\eta}^{>} > k$, so the guaranteed deletion budget is $q_{2\eta}^{>} - 1$, and $\eta = 0$ recovers (ii).*

Part (iii)'s converse is constructive (Appendix A), the factor 2 being the adversary spending its budget twice. The tie boundary is separated on purpose — at $\Gamma_M = 0$ the deployed tie-break decides, and Section 6 exhibits a released rule that lands on it, with 19 measured states. Finally q is a cross-section rather than a bit: the margins $\delta_g = M(g) - b^*$ define a *capacity spectrum* $Q_M(\tau) = |\{g \in G : \delta_g > \tau\}|$ with $q = Q_M(0)$ and $q_{2\eta}^{>} = Q_M(2\eta)$, bounded score error η shifts it by at most 2η, and it integrates exactly to the competitive surplus, $\int_0^\infty Q_M(\tau)\, d\tau = \sum_{g \in G} (M(g) - b^*)_+$, so one barely competitive repair and ten comfortable ones are distinguishable. Proofs and refinements are in Appendix A.

3 Restriction and stochastic survival

"Constrain the edit space" is a standard safety argument that conflates *average-family safety* — a statistic such as $\mathrm{mean}_{a \in A}\, \ell(a)$ for a leakage measure ℓ — with *optimizer-facing robustness*, the sign of Γ_M: restriction always moves the first in the intended direction and need not move the second, because an arg-max is not a mean.

Theorem 2 (Restriction non-monotonicity)**.** *There exist a finite family A_1, a metric M, validity and leakage assignments, and $A_2 \subsetneq A_1$ containing at least one good action such that, with $a_i^* = \arg\max_{A_i} M$, the family gets safer on average ($\mathrm{mean}_{A_2}\, \ell < \mathrm{mean}_{A_1}\, \ell$) while the deployed action leaks more, becomes invalid ($V(a_1^*) = 1$, $V(a_2^*) = 0$), and the gap worsens ($\Gamma_{M,A_2} > 0 \geq \Gamma_{M,A_1}$). No inequality therefore holds in general between $\Gamma_M(A_2)$ and $\Gamma_M(A_1)$: all of $<, =, >$ occur. The minimal witness has $|A_1| = 3$, $|A_2| = 2$, by exhaustive enumeration rather than a failed search (Appendix F).*

Restriction changes *both* sides of the competitive gap, because the best good action is recomputed inside the child family; hold that reference fixed and monotonicity returns (Appendix A), so restriction is a safety argument only with a completeness claim — which the next theorem prices exactly.

Theorem 3 (Exact stochastic deployment risk)**.** *Order A by M with the deployed tie-break as $a_{(1)}, \dots, a_{(n)}$, put $z_j = \mathbf{1}[a_{(j)} \in B]$, and let the rule take the highest-ranked* surviving *action. Under independent survival with probability r,*

$$\mathbb{P}[\textit{bad selected}] = \sum_{j=1}^{n} z_j \, r(1-r)^{j-1}, \qquad \mathbb{P}[\textit{abstain}] = (1-r)^n,$$

and retaining exactly k of n uniformly gives $\sum_{j:z_j=1} \binom{n-j}{k-1} \big/ \binom{n}{k}$. Both are exact and follow the deployed order: if $B_A(t) \le B_{A'}(t)$ for every prefix t, where $B_A(t)$ counts bad actions in the top t ranks, then A's bad-selection risk is no larger at every *r and k.*

The earlier abundance buffer is the corollary at $j = q+1$: the q protective good actions occupy exactly the ranks above the best bad action, so $r(1-r)^q$ is the probability that *that* action is deployed — the leading term of deployment risk, never the risk. Prefix dominance is the stronger comparison, ordering two families at every severity at once. For the continuum counting is vacuous, so fix a reference measure μ and set $\mathcal{A}_\varepsilon = \mu(G_\varepsilon)$ with $G_\varepsilon = \{g \in G : M(g) \ge m_B + \varepsilon\}$ and $m_B = \sup_B M$; $\mathcal{A}_\varepsilon$ is reference-measure dependent by construction, and its sampling law and a continuous restriction paradox are in Appendix F.

Theorem 4 (Continuous competitive volume)**.** *Let the valid family be affine of dimension d with coordinate u, let $M(u) = M_G^* - \frac{1}{2}(u-u^*)^\top H(u-u^*)$ with $H \succ 0$, and write $\Delta_\varepsilon = M_G^* - m_B - \varepsilon$. Then G_ε is the ellipsoid $\{u : (u-u^*)^\top H(u-u^*) \le 2\Delta_\varepsilon\}$ and*

$$\mathrm{Vol}(G_\varepsilon) = \mathrm{Vol}(B_d)\,(2\Delta_\varepsilon)^{d/2} \big/ \sqrt{\det H} \quad (\Delta_\varepsilon > 0),$$

at $\Delta_\varepsilon = 0$ it is the single point u^*, and below the frontier it is empty. Continuous competitive capacity is therefore fixed jointly by dimension, objective margin and curvature: with $H = hI$ each added dimension multiplies the volume by $\sqrt{2\Delta_\varepsilon/h}$, below one whenever $2\Delta_\varepsilon < h$, so admissible dimension alone cannot determine competitive capacity.

The audited editing objectives have exactly this shape: AlphaEdit's published reduced objective (Section 4) is a convex quadratic whose Hessian we derive in closed form and whose published ridge weight alone bounds the competitive volume; its determinant we deliberately never evaluate (Appendix A). Every statement in Sections 2–3 is machine-checked exactly with 0 violations (Appendices A–F).

4 Why valid repairs lose

Theory says the metric can prefer an invalid action; this section supplies the mechanism, in a controlled neural setting where validity is exact and then on real published weights.

4.1 The objective is the lever

We use a controlled text setting in which validity is exact rather than judged: AG News [Zhang et al., 2015], split by original example before cue insertion, per-token provenance making an action valid iff it modifies only inserted-nuisance tokens, candidate generation blind to provenance (Appendix E). On the untouched validation population of 3,168 sources the unrestricted family has $\Gamma_M > 0$ on 92.3% of those where Γ_M is defined, and the metric-optimal action is structurally invalid on 94.2%. Two selectors then see the *same* examples, candidates, features and architecture and differ only in what they score — M the conventional metric, G structural validity: on the untouched contested set ($n = 1{,}903$), G minus M is $+0.888$ $[+0.874, +0.901]$ in structural validity and -0.262 in the conventional metric, agreeing 1.1% of the time; identical candidates isolate the objective-incentive effect directly.

A preregistered restriction experiment that failed, twice. The predicted *aggregate* restriction paradox did not appear, and we keep the result at its original status. The first endpoint was arithmetically unreachable (parent selected validity 0.058 against a required -0.10); a headroom-calibrated replication on disjoint data made it reachable and the paradox still failed, -0.009 $[-0.026, +0.007]$. We do not rescue it. Good actions were removed (survival 0.315) yet the deployed selection barely moved, whereas a non-deployable factorial removing the *proxy-competitive* goods at matched leakage moves the $q > 0$ rate from 0.681 to 0.045 and selected validity by -0.636: what predicts the collapse is whether the competitive frontier survived, not how many good actions were deleted (Appendix E).

Parsimony reversal: an exact threshold. For $M(a) = E(a) - \lambda C(a)$ with larger E preferred, C an edit cost and g, b the best relevant good and bad competitors, the gap decomposes exactly:

$$\Gamma_M \;=\; \big[E(b) - E(g)\big] + \lambda\big[C(g) - C(b)\big] \;=\; -\Delta_E + \lambda\,\Delta_C. \tag{1}$$

Parsimony presses against the protected action exactly when $\Delta_C > 0$, and then Γ_M is strictly increasing in λ; if in addition $\Delta_E > 0$ the ranking reverses precisely above the *Parsimony Reversal Threshold* $\lambda^\star = \Delta_E/\Delta_C$. The remaining sign cases are equally exact and machine-checked (Appendix A); this is a diagnostic, not a claim that any regularised editor crosses it.

4.2 A real parameter-space instantiation

Instantiation. We instantiate Definition 1 on a real, published null-space-constrained editing objective [Ravfogel et al., 2020]: AlphaEdit [Fang et al., 2025] at official revision `b84624f44dfe` on GPT-2 XL [Radford et al., 2019], at the single published MLP down-projection, never changed after any result was seen, with keys, target values and the edit residual from the published routines (Appendices A, G). The mapping is exact rather than analogical: A is the frozen algebraic family of updates $\Delta = AB^\top$ generated from the real K_E, D_E, K_R and the projector; M is the negative published objective at the published ridge weight; V is AlphaEdit's own protected-key invariant $\Delta K_R = 0$; and S is the published repair condition $\Delta k_e = d_e$. Validity is *paper-relative* — AlphaEdit's own declared invariant, not semantic locality of GPT-2 XL — so this is not a semantic-safety evaluation of AlphaEdit, and the instantiation is non-degenerate because V does not appear in M: preservation is enforced structurally, by projection. One reading is not licensed: AlphaEdit does not literally enumerate our frozen action family, and q and $\mathcal{A}_\varepsilon$ characterise the optimizer-facing solution geometry induced by its published objective and constraints, not an internal discrete search.

Fragility arrives before the geometry is exhausted. Both declared scopes are reported. On the *full* family the metric optimum is invalid at every restriction, so $\Gamma_M > 0$ and $q = 0$ throughout — a forced-invalidity premise carrying no information. The informative scope is *deployed*: the updates the restriction admits, non-trivial on 90.4% of 230 validation cells. There competitive capacity is not a relabelling of native capacity — Spearman $\rho(q, \text{null dimension}) = 0.173$, with 39 *anti-monotone* pairs of restriction settings. The extreme instance moves the null-space dimension by one, 5888 to 5889, across which mean deployed q falls from 2.75 to 0 and $\mathcal{A}_\varepsilon$ from 0.085 to exactly 0 (pooled over the sweep; Appendix G). The central observation follows: 157 of 230 validation cells simultaneously have strictly positive null-space dimension, an existing valid repair, and $q = 0$ or $\Gamma_M \geq 0$. In the reported example cell the null space has dimension 6119 and a valid repairing update exists, while $\Gamma_M = 384.2$ and $q = 0$ — room remains, a repair remains, no repair wins, with 0 violations of Theorem 1 over all scored cells. The mechanism is Equation (1) with $\Delta_C > 0$: protected solutions require larger updates (median valid-to-unconstrained norm ratio 1.243, in 100% of edits), with the ridge term contributing a median 85.3% of Γ_M. The narrow reading is the whole point — *preservation can make a valid edit more expensive under a parsimony-aware objective* — not that norm penalties are harmful in general.

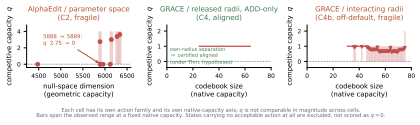


Figure 1: **Native capacity does not determine competitive capacity.** *A:* AlphaEdit null-space dimension against q (deployed scope); one added dimension takes mean q from 2.75 to 0. *B:* GRACE at its released radius — codebook grows, frontier aligned. *C:* the same implementation under an off-default interaction stress; valid repairs remain, q collapses. Each cell has its own family and units.

5 What existing protocols measure

Section 4 showed the two capacities coming apart inside a real system. Which one do published protocols expose? We audit *protocols*, not papers — one distinct objective, selection rule or reported score is one protocol — under a corpus rule written and hashed *before* any query issued. Every recorded equation is transcribed verbatim and then *located* in the cached source $(66/66)$, and every rate uses the 36 *optimizer-facing* protocols (Appendix D).

Table 1: **Measurement-gap matrix.** 41 protocols from 36 papers; the 36 optimizer-facing ones form every denominator. Counts describe audited *text*, not whether a method works. Secondary columns and full table: Appendix D.

Audited quantity	Count	Rate
Explicitly reports a geometric / action-capacity quantity	21/36	58.3%
Explicitly measures competitive repair capacity	0/36	0%
Explicitly measures action-family completeness	0/36	0%
Retrievable released implementations computing competitive capacity	0/6	0%

Capacity is reported; competition is not. An audit of equations does not say what the shipped software computes, so we also traced the official implementations of the seven audit-selected systems across 42 code objects; six are retrievable (O-Edit's official code is CODE_UNAVAILABLE) and none computes competitive repair capacity; the nearest, SERAC [Mitchell et al., 2022b], is classified INDIRECT for the reason given in Appendix D. Two guards apply: "not measured" describes what the audited text or code reports, not whether a method fails; and the auditor can decline — all five constructions a valid auditor must not condemn were classified non-vulnerable (Appendix D). What this establishes is an observability and measurement gap in what audited protocols expose, not a deficiency of the methods audited: Sections 4 and 6 establish the distinction inside two real systems without reference to audited text, so the audit corroborates it rather than carrying it.

6 Memory geometry and competitive capacity in GRACE

A prospectively selected control. The rule for choosing a second real editing system was frozen and hashed before any candidate was scored — ten eligibility requirements, then four lexicographic priorities, with no term for expected vulnerability, machine-checked before and after ranking. Of 12 candidates 7 were eligible and the ranking selected GRACE [Hartvigsen et al., 2023]; the frozen hypothesis was not the convenient one, since from the selected objective's geometry we predicted $\Gamma_M < 0$ throughout.

6.1 The released radius behaves as the predicted control

We reproduce the released implementation at revision `f674183f17a9` in the smallest published configuration reachable offline (the authors' `scotus` experiment on their `tomh/scotus-bert`

checkpoint [Devlin et al., 2019, Chalkidis et al., 2022], 108M parameters). All 10 decision rules were traced to the official code, 3 of which differ from the paper's Algorithm 1; we follow the code. The action family is every codebook entry, the deferral action and the entry the insertion branch would write, and the selection metric is the released ordering itself (Theorem 5), whose arg-max reproduces the deployed rule on $614/614$ measured states. At the released $\epsilon_{\text{init}} = 1.0$ the smallest distance between two extracted keys is $2.48 > 2\epsilon_{\text{init}}$, so the ADD branch fires on every edit: 64 entries, every radius equal. Across 301 locked validation states the codebook grows while q takes only the values $\{0, 1\}$, and the frozen prediction reads 295 negative, 0 ties and 0 positive over the 295 states carrying a GOOD action, with 6 further states carrying none, reported separately. At most one action can be competitive there, so $q \leq 1$ is a structural ceiling of the configuration.

6.2 The exact geometric condition: own-radius separation

That ceiling is provable, and the correct hypothesis is not ball disjointness.

Theorem 5 (Isolated-memory alignment: own-radius separation)**.** *For the released GRACE rule — $i^* = \arg\min_i d(x, k_i)$ over the* whole *codebook, accept entry i^* iff $d(x, k_{i^*}) \leq r_{i^*}$ and defer otherwise, with $c(\text{entry}_i) = d(x, k_i)$, $c(\text{DEFER}) = r_{i^*}$ and $M = -c$ — let g be a* GOOD *entry satisfying* (H) query-local dominance*: $d(x, k_g) < r_g$ and $d(x, k_j) > r_g$ for every $j \neq g$. Then g is the strictly nearest key, the released acceptance test passes, neither the deferral action nor any other entry outranks g, and consequently $q \geq 1$ and $\Gamma_M < 0$. Whenever $d(x, k_g) \leq r_g$, (H) follows from* own-radius separation*, $\text{SEP}(g) : d(k_g, k_j) > 2r_g$ for every $j \neq g$ — a condition constraining the codebook only through the radius of g* itself*. Under the non-strict hypothesis $d(x, k_g) = r_g$ the conclusion weakens to $\Gamma_M \leq 0$.*

The theorem does not prove $q = 1$ (that needs a BAD deferral) or uniqueness of the containing ball (that needs uniform radii), and says nothing about GRACE's safety beyond this cost ordering; it is machine-checked with 0 violations across all five geometry regimes. Ball disjointness is neither the right hypothesis nor a substitute for it: the rule takes the global arg-min *before* testing any radius, so a small-radius key can shadow a large-radius one and a query inside a GOOD ball is never retrieved, giving $q = 0$ and $\Gamma_M > 0$ (Proposition 2, Appendix B). Conversely overlap is only a *permission*: it makes competitive failure possible without making it necessary.

6.3 The interaction regime, and how its radius was chosen

This is an off-default interaction stress, not GRACE's shipped configuration. The released regime never exercises the branches that create overlap, so we re-run the identical released method — same revision, checkpoint, layer, editing stream, metric, specification and action family — on 96 source examples the first run never consumed, with a disjoint retain set. The one free choice is the initial radius, and it was made by geometry alone: the grid $\{0.5, 1.0, 2.0\}$ was declared in code before the prospective validation ran and never extended, and ascending it on fresh development geometry the interaction rates were 0.00 at $\epsilon_{\text{init}} = 0.5$, 0.00 at 1.0 and 0.258 at 2.0 — the first setting to pass the frozen gate of 0.10, on a criterion using only branch events, radii and codebook geometry, whose code imports neither q, Γ_M nor acceptability. **The interaction run is therefore an off-default mechanism stress selected from a predeclared grid using geometry only; it is not an evaluation of GRACE's released operating point**.

The geometry changes and q stops being binary. At the selected radius the trajectory fires 53 ADD, 21 SPLIT, 17 EXPAND and 0 reuse branches, and the final codebook carries 75 entries with 42 distinct radii spanning $[1.48, 3.99]$. The minimum key distance is 2.96 while $2 \max_i r_i = 7.99$, so own-radius separation now fails for the large-radius entries: this configuration lies in the regime Theorem 5 declines to certify. Across 988 locked validation states q takes three values in $[0, 2]$ against two before and stays distinct from native capacity ($\rho(q, \text{codebook size}) = -0.172$, with 89,481 state pairs where the codebook grows while q falls). A GOOD action still exists in 937 states, so repair availability did not collapse; the optimizer-facing frontier did, reading 801 negative, 19 ties and 117

positive of those 937, the remaining 51 excluded rather than counted as failures. Theorem 1 holds with 0 violations on the 918 in-scope states, and Theorem 5 certifies 274.

Prevalence, under dependence-aware inference. The headline is $\Gamma_M \geq 0$ — competitive failure or a tie despite an available valid repair — on 136 of 937 states, 14.5%, of which 117 (12.5%) are strictly positive. These states are not independent: eight frozen probe queries are re-asked at each of 64 edit positions, so we report a predeclared two-way cluster-robust interval on (edit position, query identity) as primary, $[0.032, 0.258]$, against $[0.123, 0.168]$ under an inappropriate i.i.d. scheme. Prevalence is bounded away from zero under every clustered scheme, with substantially wider uncertainty (Appendix C).

Which geometry events move capacity, and what we do not claim. Each released transition gives an exact pre/post state pair with eight frozen probe queries held fixed. ADD events changed q in *zero* of the 424 locked probe pairs, an exact statement depending on no distributional assumption. The EXPAND contrast does not survive design-respecting clustering, so **we withdraw any claim that EXPAND significantly reduces** q; only a static association between positive gaps and expanded-radius geometry remains, reported as such in Appendix C.

One implementation, metric and acceptability specification therefore exhibit different competitive-capacity regimes as memory geometry changes: aligned when the accepted regions satisfy own-radius separation, fragile when they interact. The independent variable is memory geometry; none of this says whether GRACE is safe.

7 From diagnosis to certification

Competitive capacity diagnoses order failure. Can the theorem instead *identify* competitively aligned deployment decisions from observable geometry alone? We applied Theorem 5 prospectively as a deployment gate on 988 fresh GRACE validation states from edit positions no earlier run had scanned. The gate sees only the codebook keys, the radii and the query geometry — not q, Γ_M, the GOOD/BAD labelling or correctness — and either executes the released decision unchanged or defers. The baseline rule has $\Gamma_M \geq 0$ on 44.6% of acted-on states carrying a GOOD action; among the 141 states the certificate executed there are *zero observed competitive failures*. Geometry carries the prediction too: a native-capacity model over codebook size, edit position and radius summaries reaches AUROC 0.440, and adding the separation margin and the certificate bit, 0.847. Those 141 states span about 26 edit clusters, so this is a count on a measured population, not a failure probability.

What it cost. Certified coverage of the baseline's non-deferred edit and held-in retrievals is 43.0%; unconditional edit correctness falls from 0.943 to 0.405 with abstentions counted as non-actions, retain accuracy improves from 0.706 to 1.000, and the preregistered utility endpoint did not improve (Appendix C). The result establishes certifiability, not actionability.

Why, and what it separates. That gap is structural. Write $Y = \mathbf{1}[\Gamma_M < 0]$ for competitive alignment, $S = \mathbf{1}[a^* \in G]$ for semantic acceptability of the deployed action and R for its strict order dominance over every remaining BAD action: then $Y = S \wedge R$ pointwise, and own-radius separation settles R from geometry in every consistent world while it cannot touch S.

Corollary 1 (Competitive certifiable coverage; a specialisation)**.** *Under a declared observation model O, the maximal coverage of a sound label-free competitive certificate is $\mathbb{P}[\min_{w \in W(O)} Y(w) = 1]$, the mass of observation classes all of whose compatible worlds are competitively aligned, since no zero-error certificate may act on a class containing both an aligned and a misaligned world. This specialises our earlier safe-coverage theorem to Y (Appendix A).*

Conservatism is therefore the price of exact soundness, not a defect of this gate, and trading soundness for coverage moves into classical selective classification [Chow, 1970, El-Yaniv and Wiener, 2010]: geometry certifies order, never semantics (design implications, Appendix G).

8 Discussion and limitations

What this establishes, and where we stop. Native capacity does not determine optimizer-facing repair capacity: the two are distinct by theorem (Section 2), in a real parameter-space editor where q collapses while the null space grows (Section 4), and in a real memory-based editor where codebook growth leaves q unchanged in every matched pair (Section 6). Four mechanisms produce the loss and need not reduce to one: a validity-blind metric, preservation raising update cost under parsimony, restriction removing proxy-competitive alternatives, and interacting geometry exposing a closer unacceptable action. The four properties — existence, competition, certification and utility — each fail to imply the next; we address the second and partly the third, not the utility problem in the fourth.

What is not proposed. Competitive capacity is neither a universal downstream-harm predictor nor an editing algorithm; it characterises the narrower question of whether acceptable repairs can win, and our two prospective deployment studies show why that matters: competitive information can be behaviourally irrelevant in a low-harm regime, or certifiably pure yet too conservative to deploy. Nor do we claim that every editing method exhibits competitive failure — GRACE's released radius is a prospective aligned control. What generalises is the quantity, defined for any repair problem carrying an objective, a candidate family and an independently observed validity predicate; whether competitive fragility is common beyond the two mechanisms instantiated here is open.

Limitations that are scope, not open flaws. *The object.* q is specific to finite action families; $\mathcal{A}_\varepsilon$ is its reference-measure-dependent continuous analogue, and thinning is simulated exchangeable restriction on measured real families. *The question.* Competitive capacity is not a universal predictor of downstream damage: on fresh single-edit AlphaEdit data where the edit induced only $\sim 2.3\%$ additional neighbourhood damage, q-based gating changed deployment decisions but did not improve locality and added no held-out predictive value (Appendix C). *Certifiability.* Soundness does not imply usefulness (Section 7); Corollary 1 assumes a declared observation model, and its certified counts are observations on one population, not probabilities. *The audit.* 41 protocols and six repositories are a bounded corpus rather than a census, with no classifier or vision-repair coverage, describing published text rather than method quality. *Parameter space.* One locus in one model, with validity AlphaEdit's own protected-key specification — observed independently of M but not semantic locality — so this is not a safety evaluation of AlphaEdit. *Released-radius GRACE.* q is almost binary there, so its alignment tests variation weakly. *Interaction stress.* $\epsilon_{\text{init}} = 2$ is off-default; $q = 2$ occurs once; no reuse branch fires; one model, layer, task and seed over 96 dependent edits. *Geometry.* Theorem 5 requires its stated hypotheses, and the overlap result is a permission, not an inevitability. *Event inference.* The EXPAND event effect is withdrawn; only the exact ADD null and the static association survive. *Custody.* Every prediction, selection rule, radius grid and certificate definition was hashed before the corresponding code ran, and each failed or withdrawn claim keeps its original status (Appendix G).

The principal unresolved empirical question is scale. Our experiments are deliberately CPU-feasible, and therefore test one GPT-2 XL AlphaEdit locus and one small GRACE configuration — one layer per model, bounded candidate and action families, short edit horizons — rather than a broad matrix of contemporary large models, layers, seeds and editing methods. The formalism is architecture-agnostic and the two mechanisms instantiated here are substantially different, but establishing how competitive capacity behaves across model scale, across many editing methods and over long sequential editing horizons requires substantially larger compute than this work used. That, rather than the elementary ingredients of the mathematics, the bounded audit or the absence of a downstream utility win, is the criticism most likely to change what this paper concludes.

References

Yuchen Cai and Ding Cao. O-Edit: Orthogonal subspace editing for language model sequential editing, 2024.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4310–4330, 2022.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02355.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2211.11031.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law, 2018.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.07229.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations (ICLR)*, 2022a. arXiv:2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning (ICML)*, 2022b. arXiv:2206.06520.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256, 2020.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.13085.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.14768.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

A Elementary lemmas, proofs, and the fixed-reference monotone case

Proof of Theorem 1. Write $m_G = \max_{G(A)} M$ and $m_B = b^*(A) = \max_{B(A)} M$, both attained since A is finite and both sides nonempty. By definition $\Gamma_M(A) = m_B - m_G$.

($q \geq 1 \Rightarrow \Gamma_M < 0$): some g has $M(g) > m_B$; since $m_G \geq M(g) > m_B$ we get $\Gamma_M < 0$. ($\Gamma_M < 0 \Rightarrow q \geq 1$): any maximiser g^* of M on $G(A)$ has $M(g^*) = m_G > m_B$, so $g^* \in G_+(A)$. ($\Gamma_M < 0 \Leftrightarrow$ every M-maximiser is good): $\max_A M = \max(m_G, m_B)$; this is attained only inside $G(A)$ iff $m_G > m_B$. The second display is the contrapositive. For the third, every maximiser is bad iff $m_B > m_G$, i.e. $\Gamma_M > 0$. At $\Gamma_M = 0$ both a good and a bad action attain $\max_A M$, so the deployed tie-break decides and neither strict statement applies.

For (ii), deleting $D = G_+(A)$ leaves no good action strictly above b^*, which is unchanged since $D \subseteq G(A)$, so $\Gamma_M(A \setminus D) \geq 0$; conversely any D with $\Gamma_M(A \setminus D) \geq 0$ must contain every element of $G_+(A)$, since a survivor g with $M(g) > b^*$ would force $\Gamma_M(A \setminus D) < 0$. Hence the minimum is $|G_+(A)| = q$ and $G_+(A)$ is the unique minimiser.

For (iii), suppose $q^{>}_{2\eta} > k$. Any admissible $\tilde{M}$ satisfies $\tilde{M}(g) \geq M(g) - \eta$ and $\tilde{M}(b) \leq M(b) + \eta \leq b^* + \eta$, so every g with $M(g) > b^* + 2\eta$ has $\tilde{M}(g) > b^* + \eta \geq \tilde{M}(b)$ for all b; more than k such actions exist, so at least one survives k deletions and $\Gamma_{\tilde{M}}(A \setminus D) < 0$. Conversely if $q^{>}_{2\eta} \leq k$, delete exactly $\{g : M(g) > b^* + 2\eta\}$, set $\tilde{M} = M + \eta$ on one best bad action and $\tilde{M} = M - \eta$ on every good action: each surviving good action then has $\tilde{M}(g) \leq b^* + 2\eta - \eta = b^* + \eta \leq \tilde{M}(b^*\text{-attainer})$, so $\Gamma_{\tilde{M}} \geq 0$. Setting $\eta = 0$ recovers (ii). □

The tie-multiplicity radius, the strict-dominance form, and the spectrum sandwich. Three refinements of Theorem 1 are kept out of the main text. *(a)* With $t = |\{g \in G(A) : M(g) = b^*\}|$ the frontier tie multiplicity, the *strict*-failure radius — the fewest good deletions after which a bad action strictly outranks every survivor — is $q + t$, so the radius of the *deployed* rule at $\Gamma_M = 0$ lies in $[q, q + t]$ and is tie-break dependent, while $\Gamma_M \geq 0$ itself is not. *(b)* Replacing $>$ by $\geq$ throughout Theorem 1(iii) gives the strict-dominance variant, in which the adversary must additionally break a tie. *(c)* The factor 2 recurs as a stability bound: the margins $\delta_g = M(g) - b^*$ satisfy $|\tilde{\delta}_g - \delta_g| \leq 2\eta$ with 2 attained, so the capacity spectrum obeys $Q_M(\tau + 2\eta) \leq Q_{\tilde{M}}(\tau) \leq Q_M(\tau - 2\eta)$ — a sandwich the sign statement cannot express. The surplus identity of Section 2 is the layer-cake identity for the non-negative integer-valued function Q_M.

Contour alignment is a sufficient condition. If the acceptable set is an upper contour set of M — $G = \{a \in S : M(a) \geq c\}$ with $G \neq \emptyset$ — then every maximiser of M is acceptable, ties included. This is sufficient and not necessary: a G made of the top and the bottom of the ranking also protects the maximiser. Observing the aligned pattern on finitely many states, as at GRACE's released radius, is agreement with the prediction, not verification of it.

Lemma 1 (Monotone inclusion)**.** *If $A' \subseteq A$ then $b^*(A') \leq b^*(A)$, hence $G_+(A) \cap A' \subseteq G_+(A')$, so $q(A') = 0$ implies that* every *parent-protective good action was removed. The converse fails: deleting the top bad actions can promote a surviving good action the parent frontier lacked.*

Proof of Lemma 1. $B(A') \subseteq B(A)$ because membership in G depends only on V and S, which are properties of the action, so $b^*(A') \leq b^*(A)$. If $g \in G_+(A) \cap A'$ then $M(g) > b^*(A) \geq b^*(A')$, so $g \in G_+(A')$. Hence $q(A') = 0$ forces $G_+(A) \cap A' = \emptyset$. The converse fails on the three-action family $M(g_1) = 1$, $M(b_1) = 2$, $M(g_2) = 1.5$ with $G = \{g_1, g_2\}$: here $q(A) = 0$ but deleting b_1 gives $q(A') = 2$. □

Fixed-reference restriction is monotone. Theorem 2 depends on recomputing the constrained optimum. If the good reference is held fixed — i.e. we compare $M(b^*(A_i)) - M(g_0)$ for one $g_0 \in A_2$ rather than $m_B - m_G$ — then $A_2 \subseteq A_1$ gives $b^*(A_2) \leq b^*(A_1)$ and the fixed-reference gap is

monotone non-increasing under restriction. Restriction is adverse only through the m_G term, which is exactly the quantity a completeness claim would have to control.

Observation refinement is monotone (contrast). Let O' refine O, so every ambiguity class of O' sits inside one of O. Then the robustly valid action set can only grow (an intersection over fewer worlds is a superset), the attainable safe-coverage frontier is non-decreasing, and every policy implementable on O is implementable on O' by composing with the coarsening map. This says nothing can get worse, not that something must get better: a refinement that splits no class changes nothing, and for a *fixed* M refinement changes nothing at all, because Γ_M is evaluated at a fixed world. The asymmetry between this statement and Theorem 2 — information refinement monotone, action restriction not — is the structural point.

Threshold and margin identities. $q_\tau(A)$ is non-increasing in τ and $\mathbf{1}[q_\tau(A) \geq 1] = \mathbf{1}[\Gamma_M(A) \leq -\tau]$ up to the tie boundary; in particular the "headroom" bit — whether the parent metric optimum is valid — is exactly the one-bit shadow of q at $\tau = 0$, and carries none of its magnitude. This is why an archive that records only whether the selected action was valid can identify the indicator but only bound the count.

Proof of Theorem 4, and its boundary. Under the stated parametrisation, $M(u) \geq m_B + \varepsilon$ is exactly $(u-u^*)^\top H(u-u^*) \leq 2\Delta_\varepsilon$ with $\Delta_\varepsilon = M_G^* - m_B - \varepsilon$. For $\Delta_\varepsilon > 0$ that is a solid ellipsoid; the linear change of variables $v = H^{1/2}(u-u^*)$ has Jacobian $(\det H)^{-1/2}$ and maps it onto the ball of radius $\sqrt{2\Delta_\varepsilon}$, giving $\mathrm{Vol}(G_\varepsilon) = \mathrm{Vol}(B_d)(2\Delta_\varepsilon)^{d/2}/\sqrt{\det H}$. At $\Delta_\varepsilon = 0$ the sublevel set degenerates to $\{u^*\}$, of Lebesgue measure zero, and for $\Delta_\varepsilon < 0$ it is empty — so the competitive frontier is a genuine phase boundary in ε, not a smooth decay. With $H = hI$ the ratio of consecutive dimensions is $\sqrt{2\Delta_\varepsilon/h}$, which is less than one whenever $2\Delta_\varepsilon < h$: adding admissible dimensions then *shrinks* competitive volume, which is the sense in which dimension alone cannot determine competitive capacity. The corresponding sampling law $\mathbb{P}[A_k \cap G_\varepsilon = \emptyset] = (1 - \mathcal{A}_\varepsilon)^k$ under k i.i.d. draws from μ follows immediately, and both the volume formula and the sampling law were checked against Monte-Carlo estimates within sampling error.

Proposition 1 (What a label-free certificate can and cannot settle)**.** *Let o fix the family, the scores and hence the deployed candidate $a^*(o)$, let $W(o)$ be the worlds consistent with o, and write $Y = \mathbf{1}[\Gamma_M < 0]$, $S = \mathbf{1}[a^* \in G]$ and $R = \mathbf{1}[M(a^*) > M(b)\ \forall b \in B \setminus \{a^*\}]$. Then* (i) *$Y = S \wedge R$ pointwise, hence robustly over $W(o)$;* (ii) *if a^* strictly maximises M then $R = 1$ in* every *world whatever the partition, so a geometry-only certificate settles R; and* (iii) *if the observation excludes the validity label and some o admits $w_1, w_2 \in W(o)$ with $Y(w_1) = 1 \neq Y(w_2)$, no zero-error certificate, deterministic or randomised, may act at o — the smallest witness has two actions. Maximal sound coverage is therefore $\mathbb{P}[\min_{w \in W(O)} Y(w) = 1]$, which is Corollary 1 and is our earlier exact safe-coverage theorem specialised to Y.*

Proof of Proposition 1. (i) $\Gamma_M < 0$ says the maximiser is good and strictly beats every bad action, i.e. $S = 1$ and $R = 1$; conversely $S \wedge R = 1$ makes a^* a good action strictly above $\max_B M$. (ii) If a^* is the strict M-maximiser then $M(a^*) > M(b)$ for every other action, in particular every bad one, in any world — R depends on M and the identity of a^* alone, both fixed by o. (iii) A deterministic certificate acting at o errs in w_1 or in w_2; a randomised one acting with probability $p > 0$ errs with probability at least $p \min_i \mathbb{P}[w_i] > 0$, so zero error forces abstention. The two-action witness has a_1 with $M = 1$, a_2 with $M = 0$ and o recording only the scores: $Y = 1$ if a_1 is good and $Y = 0$ if it is bad, both consistent. □

Own-radius separation (Theorem 5) proves exactly (ii) for the released GRACE rule. It does not touch S, so what it licenses is conditional — *if* the retrieved entry is GOOD, this state is competitively aligned — and (iii) then forces abstention wherever a BAD retrieval remains admissible.

The AlphaEdit Hessian, in closed form. The frozen C2 objective is $J(\Delta) = \|\Delta K_E - D_E\|_F^2 + \|\Delta K_{\text{prev}}\|_F^2 + L_2\|\Delta\|_F^2$ with $M = -J$, a convex quadratic. On any affine coordinatisation $\Delta(u) = \Delta_0 + \sum_i u_i E_i$ of the admissible family, $J(\Delta(u)) = J(\Delta_0) + g^\top u + \frac{1}{2}u^\top H u$ with $H_{ij} = 2\,\text{tr}(E_i S E_j^\top)$ and $S = K_E K_E^\top + K_{\text{prev}} K_{\text{prev}}^\top + L_2 I$, i.e. H is twice the Gram matrix of the generators in the S-inner product. Since $S \succeq L_2 I$ with the published $L_2 = 10 > 0$, $H \succ 0$ whenever the generators are independent, so Theorem 4's hypothesis holds structurally; and for an F-orthonormal generator set $\det H \geq (2L_2)^d$, giving the volume ceiling quoted in Section 3. This is a post-hoc theory instantiation on the frozen published objective: the determinant is never evaluated, since the frozen artifacts record K_E and K_R by shape, SHA-256 and norm rather than by value, and no model was run. A separate exact check confirms the sign prediction — competitive mass is zero wherever $\Gamma_M > 0$ — on all 224 frozen C2 sweep cells within each declared scope, and correctly fails when the full-family gap is crossed against the deployed-family mass.

The Parsimony Reversal Threshold, all sign cases. With $\Gamma_M = -\Delta_E + \lambda\Delta_C$ as in Equation (1): if $\Delta_C > 0$ and $\Delta_E > 0$ the ranking reverses exactly for $\lambda > \lambda^\star = \Delta_E/\Delta_C$, with an exact tie at $\lambda^\star$; if $\Delta_C > 0$ and $\Delta_E \leq 0$ the gap is already non-negative at $\lambda = 0$ and parsimony only widens it; if $\Delta_C = 0$ the weight can neither create nor remove a reversal; if $\Delta_C < 0$ the gap strictly decreases in λ, so with $\Delta_E > 0$ no admissible weight reverses the ranking and with $\Delta_E \leq 0$ parsimony *repairs* it above $\lambda^\star$. All nine sign pairs are machine-checked in rational arithmetic over $84{,}240$ instances.

Machine-check budget. The count of $216{,}352$ quoted in Section 3 is the sum of two disjoint suites: $153{,}600$ synthetic action families covering the earlier ten statements (Appendix F), and $62{,}752$ exact checks in 20 groups covering the robustness radius, the joint deletion/perturbation characterisation, the survival polynomial, the fixed-size law, prefix dominance, the capacity spectrum, the quadratic volume law, and the certifiability and factorisation statements above. Both report 0 violations.

B GRACE geometry: proof, counterexamples, machine checks

Proposition 2 (Disjointness alone is insufficient; overlap only permits)**.** (i) *There are GRACE codebooks with* pairwise disjoint *balls ($d(k_i, k_j) > r_i + r_j$) and a* GOOD *entry whose ball contains the query, on which the released rule never retrieves that entry; with a* BAD *deferral, $q = 0$ and $\Gamma_M > 0$.* (ii) *There is a two-key one-dimensional configuration with overlapping balls, an existing* GOOD *entry, both entries accepted, and $q = 0$, $\Gamma_M > 0$ under* either *labelling of the deferral.*

The released rule, verbatim. The Phase-C4 code trace records ten decision rules against the pinned repository revision. The three relevant here are: the retrieval arg-min is taken over the *whole* codebook and only then is the winner's own radius tested (so an entry whose ball contains the query is never consulted if another key is closer); the acceptance test is non-strict (<=) where the paper prints a strict $<$; and the released default coverage EXPAND *sets* a radius equal to a measured distance rather than incrementing it. The second and third interact: EXPAND makes the measure-zero boundary $d(x, k_{i^*}) = r_{i^*}$ reachable, which is precisely the tie case of Theorem 5.

Cost identity and the entry-only reduction. Writing $c = -M$, for any finite action set $\Gamma_M = \min_G c - \min_B c$ and $q = |\{g \in G : c(g) < \min_B c\}|$; this uses only $M = -c$ pointwise and so holds even though $c(\text{DEFER}) = r_{i^*}$ is a *radius* rather than a distance to any key. Restricted to entry actions, $\Gamma_{\text{entry}} = d_G - d_B$ and $q_{\text{entry}} = |\{g : d(x, k_g) < d_B\}|$. The distance formula extends to the full family for Γ_M iff $r_{i^*} \geq d_G$ when the deferral is GOOD and $r_{i^*} \geq d_B$ when it is BAD, and for q iff $r_{i^*} \geq d_B$ irrespective of the deferral's label — two *different* conditions, which coincide exactly when the deferral is BAD. Ignoring the deferral action can therefore invalidate the distance formula, which is why we do not use it as a definition.

Proof of Theorem 5. Under (H), $d(x, k_g) < r_g < d(x, k_j)$ for every $j \neq g$, so g is the strictly nearest key and $i^* = g$; the released test $d(x, k_g) \leq r_g$ passes, so GRACE retrieves V_g. For costs: $c(\text{DEFER}) = r_{i^*} = r_g > d(x, k_g) = c(g)$, so the deferral cannot outrank g; and $c(\text{entry}_j) = d(x, k_j) > r_g > c(g)$ for every $j \neq g$. Hence g strictly minimises c, i.e. strictly maximises M, so $q \geq 1$ and $\Gamma_M < 0$. For the sufficiency of $\text{SEP}(g)$: if $d(x, k_g) \leq r_g$ then for $j \neq g$ the triangle inequality gives $d(x, k_j) \geq d(k_g, k_j) - d(x, k_g) > 2r_g - r_g = r_g$, which is (H) whenever the first inequality of (H) is strict. At $d(x, k_g) = r_g$ the deferral cost equals $c(g)$ and the conclusion weakens to $\Gamma_M \leq 0$, with an exact tie when the deferral is BAD. □

Why the hypothesis is own-radius and not disjointness. Pairwise disjointness $d(k_i, k_j) > r_i + r_j$ does not imply $\text{SEP}(g)$ once radii differ: a nearby small ball satisfies disjointness while sitting well inside $2r_g$ of k_g. Proposition 2(i) realises exactly that. Its exact witness is $k_g = 0$, $r_g = 10$, $k_b = 11$, $r_b = \frac{1}{2}$, $x = 6$: the balls are disjoint ($11 > 10.5$) and x lies in g's ball, but $d(x, k_b) = 5 < 6 = d(x, k_g)$, so the arg-min consults b, the radius test $5 \leq \frac{1}{2}$ fails, and the rule defers, giving $\Gamma_M = \frac{11}{2} > 0$ and $q = 0$ when the deferral is BAD. The exhaustive rational enumeration over $2{,}464$ configurations produces 562 such witnesses, every one of which defers. A separation condition stated with $\max(r_i, r_j)$ is sufficient but strictly stronger than necessary: it constrains pairs in which the *other* entry has the larger radius, which the proof never uses. The overlap witness of Proposition 2(ii) is $k_b = 0$, $k_g = 3$, $r_b = r_g = 2$, $x = 1$, giving $\Gamma_M = 1$ under either deferral label; two entries in one dimension are minimal, since a positive gap needs one GOOD and one BAD action and a single entry leaves only the deferral as competitor.

Overlap is a permission, empirically. In the interaction regime only 9 of the $1{,}006$ states in the label-free conflict region carry a positive gap, so overlap makes competitive failure possible without making it likely.

Machine checks. Theorem 5 and both propositions are checked in `grace_geometry.py` with tests in `test_grace_geometry.py`: 0 violations over $9{,}998$ checks, with regime coverage of uniform-isolated ($1{,}490$), heterogeneous-disjoint ($4{,}690$), overlapping ($3{,}799$), boundary-tie ($3{,}230$), deferral-selected ($1{,}473$), nearest-outside-radius ($1{,}281$) and farther-ball-contains-query (449) states. Separately, the locked Γ_M and q of both GRACE runs are recomputed from the recorded distances and radii alone and reproduce the frozen tables exactly (max absolute error 1.8×10^{-15} over 596 and $1{,}473$ states); this is an arithmetic audit of the locked tables, not a re-measurement.

Where the two measured configurations fall. The released SCOTUS configuration has every radius equal to 1.0 and minimum key distance 2.48, so $\text{SEP}(g)$ holds for every entry with margin 0.48, and the theorem certifies 181 of the 295 GOOD-existing validation states with 0 violations. The interaction stress has radii spanning $[1.48, 3.99]$ and minimum key distance 2.96, so SEP fails for the inflated entries; the theorem still certifies 274 states, again with 0 violations, and declines the rest. It does not predict the rate at which a permitted failure occurs: 12.5% is a measurement, not a consequence of the theorem.

C Dependence-aware inference for the interaction regime

The design is crossed, not nested. Eight frozen probe queries are re-asked at every one of the 64 headline edit positions (512 of the 988 validation states, cell fill rate 1.0), and 32 retain queries at all 4 checkpoints. A row is therefore correlated through two channels at once — the edit position, which shares codebook geometry, and the query identity, which is literally the same question asked again — so one-way clustering on either axis alone leaves the other channel's correlation in the residual. There are 64 edit clusters and 199 query clusters over 988 rows.

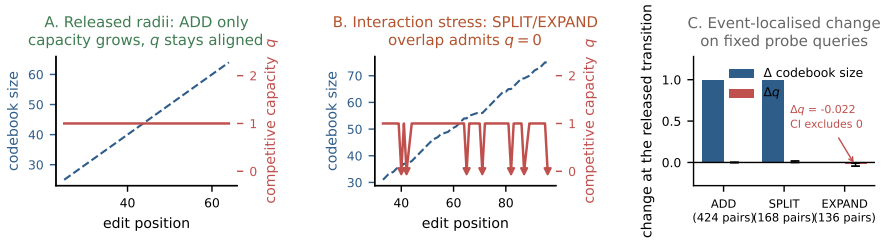


Figure 2: **Native capacity can grow without changing competitive capacity.** *A:* released radius — the codebook grows, q stays aligned. *B:* the off-default interaction stress on fresh sources — the codebook still grows, but q collapses at many positions. *C:* event-localised change on eight frozen probe queries. ADD moves q by exactly zero over all 424 matched pairs; the EXPAND contrast is shown with its *naive* probe-clustered interval and is **not** claimed as an established event effect, since it does not survive the cluster-robust reanalysis below. States with no acceptable action are excluded, not scored as $q = 0$.

Predeclared method. The primary estimator is the Cameron–Gelbach–Miller multiway cluster-robust sandwich, $V = V_A + V_B - V_{AB}$ with a Wald normal interval and no finite-sample dof correction (predeclared, so that the choice among published variants cannot be made after seeing widths); the secondary is Owen's crossed pigeonhole bootstrap at 10,000 replicates. Both are always reported, the narrower is never promoted, and the predeclaration was written before any two-way interval was computed. The frozen C4b headline reproduces exactly under recomputation (0 mismatches across every estimand).

Results. For the primary estimand $\mathbb{P}[\Gamma_M \geq 0 \mid \text{a GOOD action exists}] = 0.145$: naive i.i.d. $[0.123, 0.168]$; one-way edit-cluster $[0.113, 0.195]$ (BCa $[0.108, 0.187]$); one-way query-cluster $[0.038, 0.255]$; two-way sandwich $[0.032, 0.258]$; two-way pigeonhole $[0.036, 0.282]$. Every clustered scheme keeps the prevalence bounded away from zero. For $\mathbb{P}[\Gamma_M > 0 \mid \cdot] = 0.125$ the two-way sandwich gives $[0.009, 0.241]$.

Why the EXPAND event claim is withdrawn. The whole EXPAND Δq effect is 3 pairs out of 136, at three distinct edit positions and three distinct probe queries. Under probe-query clustering the exact probability that a bootstrap replicate is identically zero is 0.0233, just under the 0.025 a 95% percentile interval needs; under edit-position clustering it is 0.0368, just over. The published probe-clustered interval is on the correct side of that line for its own resampling unit, but the margin is smaller than Monte-Carlo noise at 2,000 replicates: re-running the identical scheme at 10,000 replicates with a different seed returns an upper endpoint of exactly 0. Of the three design-respecting schemes only the two-way sandwich excludes zero. We therefore do not treat the event-centred effect as established. By contrast, ADD moves Δq by exactly 0 in every one of the 424 pairs, which is an exact combinatorial statement requiring no interval.

The static radius-inflation association. What survives the withdrawal is a description of the locked states rather than an event effect: of the 117 positive-gap states, 115 satisfy $\Gamma_M = r_{i^*} - d(x, k_{i^*})$ with the nearest entry BAD and its radius inflated by a released EXPAND. For completeness the withdrawn contrast itself had mean $\Delta q = -0.022$ over 136 pairs. Both are static associations over locked states, not event-effect estimates.

The prospective certificate, in full. The gate of Section 7 was applied to all 988 locked validation states, of which 909 carry a GOOD action, 348 are edit or held-in states and 640 are retain states. Certificate definition, policy definitions and all seven hypotheses were hashed before the validation split was opened; the recorded certificate-definition digest matches across both executions of the frozen protocol. Baseline competitive failure among acted-on GOOD-existing states

is 0.446 $(204/457)$, with a two-way cluster-robust interval $[0.177, 0.716]$ over 64 edit and 142 query clusters and a crossed pigeonhole interval $[0.160, 0.738]$; the certified rate is $0/141$, whose interval is degenerate and reported as a count rather than a rate. The 141 certified states span 26 edit clusters and 65 query clusters, which is why no zero-risk statement is made. Coverage is 0.430 of the baseline's 328 non-deferred edit and held-in retrievals (0.273 of all states). Correctness *conditional on acting* is 1.000 under both policies; the fall from 0.943 to 0.405 is entirely the 0.857 deferral rate. Retain accuracy rises $0.706 \rightarrow 1.000$. The preregistered utility endpoint U = edit success + retain correct − invalid selection − $0.25 \times$ deferral moves by -0.023 with a dependence-aware interval $[-0.420, +0.374]$ spanning zero, so the run neither establishes an improvement nor a substantial loss; the frozen verdict band is NOT_ACTIONABLE because the edit-performance criterion failed, and it was selected mechanically from the predeclared criteria. Of 375 changed decisions, 188 prevented a competitive failure, 187 prevented a correct action and 118 were unnecessary abstentions.

Competitive capacity as a downstream gate: the negative control. A separate prospective run gated single-edit AlphaEdit deployments on q over 64 fresh CounterFact sources, at an operating restriction chosen on development geometry alone. The published edit induced only $\sim 2.3\%$ additional neighbourhood damage, so there was little downstream damage to remove: gating changed deployment decisions at 84.4% coverage but moved locality failure from 0.2438 to 0.2444, left efficacy and paraphrase generalisation intact, and added no held-out predictive value (ΔAUROC -0.083). Structural-invalidity reduction was declared in advance not to count as an endpoint, so this arm sets no success flag: competitive capacity is not established as a downstream-risk predictor, and we report the arm as a negative control rather than omit it.

D Full protocol audit table and the implementation audit

Table 2 lists all 41 audited protocols. Reading rules: every column records what the audited text reports, never a judgement about whether the method works; *Protected* records whether the protocol declares a validity/preservation predicate at all (explicit / implicit / absent), which is what makes Γ_M evaluable; *Pars.* is an algebraic property of the published objective's shape, not a measurement. The four reporting-only and one diagnostic protocol are marked † and excluded from every rate quoted in Section 5.

Table 2: **All 41 audited protocols.** Rates in the main text use the 36 optimizer-facing protocols as denominator; the five remaining protocols (four reporting-only, one diagnostic) are listed for completeness and are excluded from every optimizer-facing rate. Every entry is a statement about what the audited text reports, not about whether the method works. Geo. = geometric/action capacity reported; Comp. = competitive repair capacity reported; Cplt. = action-family completeness class; Pars. = algebraic parsimony tension possible; ARD = reproduced empirically in this work.

Protocol	Subfield	Role	Protected	Geo.	Comp.	Cplt.	q	Γ_M	Pars.
PROBEFREE-ELLIPSOID	activation steering	edit-opt	explicit	✓	–	partial	$\mathcal{A}_\varepsilon$	bound	–
ACTADD-SWEEP	activation steering	hyper-sel	absent	–	–	discussed	$\mathcal{A}_\varepsilon$	bound	✓
REPE-LORRA	activation steering	training	absent	✓	–	none	geom. only	reprod.	✓
CRISPEDIT-CONSTR	knowledge editing	edit-opt	explicit	✓	–	partial	$\mathcal{A}_\varepsilon$	reprod.	–
DSCA-MULTI	knowledge editing	edit-opt	absent	✓	–	discussed	geom. only	reprod.	✓
MEMOIR-MATCH	knowledge editing	cand-sel	absent	✓	–	partial	indicator	reprod.	✓
NSLORA-CONSTR	knowledge editing	edit-opt	explicit	✓	–	partial	$\mathcal{A}_\varepsilon$	reprod.	✓
ALPHAEDIT-OBJ^{c2}	knowledge editing	edit-opt	explicit	✓	–	partial	exact	direct	✓
GRACE-DEFER$^{c4/c4b}$	knowledge editing	cand-sel	explicit	✓	–	partial	exact	reprod.	–
MELO-LORA	knowledge editing	edit-opt	absent	✓	–	partial	indicator	reprod.	–
MEMIT-EVAL†	knowledge editing	reporting	absent	–	–	none	none	none	–
MEMIT-LS	knowledge editing	edit-opt	implicit	✓	–	none	geom. only	reprod.	✓
MEND-TRAIN	knowledge editing	training	absent	–	–	none	none	none	✓
OEDIT-LOSS	knowledge editing	edit-opt	implicit	✓	–	discussed	geom. only	reprod.	✓
PMET-Z	knowledge editing	edit-opt	absent	✓	–	none	geom. only	reprod.	✓
ROME-W	knowledge editing	edit-opt	explicit	✓	–	n/a	geom. only	n/a	–

Protocol	Subfield	Role	Protected	Geo.	Comp.	Cplt.	q	Γ_M	Pars.
ROME-v	knowledge editing	edit-opt	absent	✓	–	none	geom. only	reprod.	✓
SERAC-CLS	knowledge editing	training	explicit	–	–	n/a	indicator	bound	–
SERAC-EVAL †	knowledge editing	reporting	absent	–	–	none	none	none	–
WISE-EVAL †	knowledge editing	reporting	absent	–	–	none	none	none	–
WISE-ROUTE	knowledge editing	edit-opt	implicit	✓	–	partial	geom. only	reprod.	✓
ASYM-TWOTASK	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
BLUR-BILEVEL	unlearning	edit-opt	explicit	–	–	partial	$\mathcal{A}_\varepsilon$	reprod.	–
BSS-BOOTSTRAP	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
FDIV-MINMAX	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
FUPARETO-MGDA	unlearning	cand-sel	explicit	✓	–	partial	$\mathcal{A}_\varepsilon$	bound	–
GENERIC-MU	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
IMU-WEIGHTED	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
JELLYFISH-DIS	unlearning	edit-opt	implicit	–	–	none	none	reprod.	–
LEM-LOCAL	unlearning	edit-opt	absent	✓	–	none	geom. only	reprod.	✓
MLLM-BENIGN	unlearning	edit-opt	absent	✓	–	partial	geom. only	reprod.	✓
QUANT-DIAG †	unlearning	diagnostic	absent	–	–	n/a	none	n/a	–
REVERSE-FR	unlearning	training	absent	–	–	none	none	reprod.	✓
RGPEFT-NORM	unlearning	cand-sel	absent	✓	–	discussed	indicator	bound	✓
SHARPMIN-ROBUST	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
SIMNPO-LOSS	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
SSPU-PROJ	unlearning	edit-opt	absent	✓	–	discussed	geom. only	reprod.	✓
NPO-LOSS	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
TOFU-GRADDIFF	unlearning	edit-opt	absent	–	–	none	none	reprod.	✓
TOFU-SUITE †	unlearning	reporting	implicit	–	–	discussed	none	none	–
RMU-LOSS	unlearning	edit-opt	absent	✓	–	none	geom. only	reprod.	✓

† not optimizer-facing (reporting-only or diagnostic); excluded from every rate. C2, C4/C4B mark the two protocols whose competitive capacity this work measured directly.

Secondary columns of the measurement-gap matrix. Beyond the four headline rows of Table 1: a parsimony term is present in the objective of $26/36$ optimizer-facing protocols (72.2%); a preservation–parsimony tension is algebraically possible in $27/36$ (75.0%); exact q is extractable from published artefacts for $2/36$ protocols (5.6%) and Γ_M for $1/36$ (2.8%); and all 66 recorded objective and evaluation equations were transcribed verbatim and then located in the cached source ($66/66$).

Corpus and retrieval. Frozen queries were issued against arXiv, OpenReview, the ACL Anthology and Crossref (discovery only), merging $3{,}370$ raw records into $2{,}577$ distinct works, of which 965 passed a metadata-only screen and the top 40 by the frozen relevance ordering entered the audit set. Primary text was retrieved in a frozen order (arXiv LaTeX e-print, then ar5iv HTML, then arXiv HTML, then venue PDF); 36 of 40 were retrieved, all as LaTeX source, and the 4 unretrieved works are excluded rather than audited from abstracts. The corpus rule hash was recomputed after the search and compared; a mismatch would have invalidated the census.

The corpus spans knowledge editing, lifelong editing, activation steering and unlearning; representative audited works include Mitchell et al. [2022b], Wang et al. [2024] and Maini et al. [2024] alongside the two systems reproduced here.

Γ_M-extractability: declared class versus derived rule. The headline count of 1 protocol with directly extractable Γ_M is the preregistered manual audit class. A later derived rule, applied mechanically to the same records, would classify one additional case — GRACE's deferral selection — as directly recoverable, giving 2. We retain the declared class for the headline count and do not silently revise the taxonomy; the disagreement concerns whether per-candidate repair success counts as "published" when it requires re-running the model with each value substituted.

Implementation audit. The seven audit-selected systems are GRACE, CrispEdit, MEMIT, ROME, SERAC, O-Edit and PMET; the subset was derived from the frozen C3 selection rather than hand-picked, and its hash is recorded. Six official repositories were retrieved at pinned revisions with no fork substitutes; O-Edit's official code is CODE_UNAVAILABLE, so it carries INDETERMINATE

on all seven capacity questions and is excluded from the denominator. Across 42 traced code objects, none of the six retrievable implementations computes q, a $q > 0$ indicator, Γ_M, or $\mathcal{A}_\varepsilon$. Four implementations carry an INDIRECT signal of successful-action existence (an optimisation loop with a success criterion), GRACE and SERAC carry PARTIAL existence and completeness signals, and SERAC alone carries an INDIRECT optimizer-facing good/bad comparison: it splits good from bad over *continuations* of one edited model rather than over candidate *actions*. Positive and negative auditor controls both passed: a planted competitive-capacity computation was detected and a planted locality metric was not misclassified as one.

Positive controls. Five constructions were included that a valid auditor must decline to condemn: validity-first lexicographic selection, a complete finite candidate oracle, an exact null-space family, a refusal-enabled protocol under incomplete support, and an anti-trade-off composite. All five were classified non-vulnerable, and 8 *published* protocols in the corpus carry these shapes. Had the taxonomy condemned every control, the audit would have been recorded as invalid before any field-wide statement was formed.

E Controlled-text preregistration history

The controlled text work ran twice, and both runs are reported at their original status. Neither is upgraded by the other.

First run. Predeclared endpoints for a $100\% \to 50\%$ validity-blind restriction (Rule U): mean candidate leakage falls by ≥ 0.05; selected leakage rises by ≥ 0.10; selected validity falls by ≥ 0.10; mean Γ_M worsens by ≥ 0.05; paired source-grouped intervals exclude zero. Measured on 3,168 validation sources: mean family leakage -0.010 $[-0.010, -0.010]$, selected leakage -0.010 $[-0.012, -0.008]$, selected validity $+0.034$ $[+0.025, +0.044]$. Three of four moved the wrong way or too little; the gate failed and was committed as failed. The per-source paradox event had prevalence 0.022 $[0.017, 0.027]$. Two controls behaved: the proxy-protective rule gave $+0.522$ and the oracle rule gave exactly 0 by construction (its family contains no invalid action), which makes it definitional rather than evidential.

The diagnosis was arithmetic, not statistical. Parent selected validity was 0.058, so a 0.10 fall was unreachable under any restriction whatsoever.

Second run. A fresh, headroom-calibrated replication on disjoint data, with the same theory and the same four endpoints. Parent selected validity 0.480; 3,825 validation records; all three primary endpoints confirmed attainable before the split was opened, by an oracle bound over *all* restrictions of that depth (for selected validity, -0.434 against a threshold of -0.10); power computed on calibration only, with the primary endpoint powered at the available $4{,}504$ groups. Result: mean family leakage -0.009, selected leakage $+0.010$, selected validity -0.009 $[-0.026, +0.007]$, mean Γ_M -0.009. All four primary endpoints failed; the interval for the primary endpoint contains zero. The per-source paradox event had prevalence 0.130 $[0.119, 0.141]$, and good-action survival under the restriction was 0.315.

Note the shape of the failure. The oracle bound says the endpoint is reachable by *some* restriction of that depth; the rule-attainable mean for the validity-blind rule actually used was $+0.019$. The apparatus had headroom; the deployable restriction could not exploit it. The replication therefore does *not* confirm the deployable restriction paradox, and we do not report it as doing so.

The non-deployable factorial. A separate contrast, which requires knowing which actions are proxy-competitive and is therefore not deployable, removes the proxy-competitive goods at matched leakage (matched to within 0.002). It moves the $q > 0$ rate from 0.681 to 0.045 at low leakage, flips the metric-optimal-validity bit to zero on 63.6% of sources, and moves selected validity by -0.636 $[-0.657, -0.612]$ on 1,770 sources, with the leakage-only contrast an order of magnitude smaller.

This is stated as a controlled factorial, not as deployable evidence, and it is the reason Section 4 reads the negative aggregate result as evidence *for* q rather than against the theory.

Retrospective compatibility, at retrospective status. The archived records of both runs identify the indicator $\mathbf{1}[q_\tau \geq 1]$ exactly at every τ but only bound the count. Descriptively: the competitive frontier is empty in 0.767 of first-run parent families at $\tau = 0$; it is non-empty in 0.393 of second-run parent families; and in an archived compiler case study the compiled family holds *more* valid repairs than the generic one (0.677 vs 0.353) yet a far smaller share of items where a valid repair is metric-optimal (0.024 vs 0.404, conditional on one existing). Stated plainly as a negative: the archived *lower bound* on q is itself an uninformative predictor (AUROC ≈ 0.52), because it equals 1 for roughly 93% of families. A prospective test of the abundance buffer needs per-candidate metric vectors, which no archived text phase recorded — which is why the buffer is tested prospectively only on the parameter-space and memory-based families of Sections 4 and 6.

F Synthetic validation, symbolic counterexamples, continuous examples

Exact finite sweep. $15{,}360$ independently seeded systems produce $153{,}600$ action families under 3 declared restriction rules (capacity-greedy, leakage-greedy, structural), with $138{,}240$ restricted rows. Every theorem statement is checked on every row: 0 violations in total, across the properness characterisation, forced invalidity, the completeness–leakage bound, the domination implication, the regret bound and the evaluator reduction. The restriction-paradox rate over restricted rows is 0.0026 (359 instances), and it is highly rule-dependent: 255 under capacity-greedy, 104 under structural, and 0 under leakage-greedy — a rule that restricts exactly along the leakage axis cannot produce the paradox. The tie boundary is not rare: $16{,}460$ rows sit at $\Gamma_M = 0$, where weak but not strict properness holds and the deployed tie-break decides; it decides well on $6{,}192$ of them.

Minimality of the restriction paradox. Both minimality claims of Theorem 2 come from exhaustive enumeration over the finite structural space — metric rankings $\times$ validity assignments $\times$ reward assignments $\times$ a leakage grid — not from a failed search. $|A_1| = 2$ is proved impossible; $|A_1| = 3, |A_2| = 2$ realises the paradox; and the strengthened variant in which maximum family leakage also falls requires $|A_1| = 4$.

A sufficient condition, and its honest scope. If $a_1^* \in G_{A_1}$ is the unique M-maximiser of A_1, $a_1^* \notin A_2$, and some $b \in A_2$ with $V(b) = 0$ satisfies $M(b) > M(a)$ for every other $a \in A_2$, then $a_2^* = b$ and clauses (3)–(4) of Theorem 2 follow, with (1)–(2) reducing to two jointly satisfiable inequalities on the leakage values. This condition is close to a restatement of the conclusion; the content of the theorem is the existence result with exact minimality and the conceptual separation of average-family safety from optimizer-facing robustness, not the depth of the argument.

Continuous examples. Theorem 4(ii)'s one-dimensional witness is $A_1 = [0, 1]$, $M(x) = \ell(x) = x$, $G = [0.05, 0.15] \cup [0.80, 1.00]$, $A_2 = [0, 0.75]$. A two-dimensional variant with *connected* G takes $A_1 = [0, 1]^2$, $M(x, y) = x$, $G = ([0, 1] \times [0, \frac{1}{10}]) \cup ([\frac{9}{10}, 1] \times [0, 1])$ and $A_2 = ([0, \frac{4}{5}] \times [\frac{3}{10}, 1]) \cup ([0, \frac{1}{5}] \times [0, 1])$, giving the same three signs, so disconnection of G is not necessary in $d \geq 2$. In $d = 1$ with G an interval, M strictly increasing and A_2 a prefix meeting G, the paradox is impossible.

Symbolic counterexamples over audited objectives. Under the declared finite structural model, 28 of the audited objectives admit at least one symbolic counterexample; 36 were generated and 36 machine-verified. Three objective shapes account for most of them: efficacy plus a fixed-weight preservation penalty; an efficacy proxy with no preservation term at all; and efficacy minus a penalty on the size of a projected update — the shape Section 4 measured. Every counterexample is phrased as "the objective alone does not rule out this behaviour under the encoded assumptions". None is

evidence that a published method exhibits the behaviour; establishing that requires reproduction, which this work performed twice.

G Implementation, custody and reproducibility

Compute. Every experiment in this paper is CPU-only and float32. The parameter-space measurement runs on the published GPT-2 XL checkpoint at one locus; the memory-based reproduction runs on a 108M-parameter encoder. No GPU was used and no result depends on one. The C5 analyses — the geometry theorem, the cluster-robust reanalysis, the code audit and the framing pass — ran no model and produced no new measurement; each recomputes from frozen artifacts, and each verified all 138 prior artifact hashes and all 9 prior phase verdicts unchanged before and after running.

Sources and pinning. Both editing systems are used at pinned repository revisions with per-file hashes recorded, and both use the authors' released hyper-parameters, datasets and (where applicable) precomputed statistics. For the parameter-space system, target values come from the published target-optimisation routine at the published hyper-parameters; no residual is synthesised, and the editing locus is taken from the published hyper-parameter file and never changed after any result was seen. For the memory-based system, the checkpoint identity was verified against the earlier run's pin from the local cache after the fact, a check that changes no measurement.

The parameter-space locus, in detail. The extreme $5888 \to 5889$ contrast quoted in Section 4 is pooled over all 460 sweep cells; on the validation split alone the same pair reads $2.50 \to 0$. The edited weight is the single published MLP down-projection $W \in \mathbb{R}^{1600\times 6400}$ of GPT-2 XL, taken from AlphaEdit's published hyper-parameter file; the official repository is used at revision `b84624f44dfe` under its MIT licence. Retain keys are drawn from 512 CounterFact facts and edit keys from 32, and the frozen objective is $M = -\big(\|\Delta K_E - D_E\|_F^2 + \|\Delta K_{\text{prev}}\|_F^2 + L_2\|\Delta\|_F^2\big)$ at the published ridge weight $L_2 = 10$. The released GRACE regime fires only ADD branches — no splits, expansions or reuses — and mean q across its 301 locked validation states is 0.980.

Split discipline. Each validation split in this paper was opened exactly once, after its protocol and hypotheses were hashed. The interaction-regime run draws its sources from positions the released-radius run never scanned, with a disjoint retain set, and its development and validation positions are separated by a rule fixed in advance. Ten safety gates — official code, fresh sources, geometry-only radius selection, released-adapter state-transition parity, unchanged q and Γ_M definitions — are checked before the protocol hash is written.

Machine checks and tests. The theory statements are checked by executable predicates rather than asserted; the geometry theorem of Section 6 is checked by the module and tests named in Appendix B. Both main-text figures are regenerated from the frozen artifacts by a single script, with no number entered by hand.

Design implications. The design stance these results support, kept out of the main text because neither prospective deployment study improved a downstream endpoint, is the following.

Designing for competitive-safe repair. (1) Constrain first, rank second — select from the feasible set and abstain when it is empty [Geifman and El-Yaniv, 2017], at a cost in completeness. **(2) Audit parsimony reversals** — decompose Γ_M by objective component, since ranking reverses above the Parsimony Reversal Threshold $\lambda^\star = \Delta_E/\Delta_C$. **(3) Report native, valid and competitive capacity separately**, never letting a geometric number stand in for the competitive one. **(4) Certify only what observables support** — say which conjunct a certificate settles (Corollary 1); exact soundness trades against coverage. Competitive capacity $\to$ certifiability $\to$ utility are distinct, and each arrow can fail.

What is not included. We deliberately omit development history, branch structure, intermediate protocol revisions and wall-clock command logs from this paper. They are recorded in the accompanying artifact tree together with the claim ledger, the evidence map, the number audit, the citation audit, the statistics, geometry and implementation audits, the C5 reconciliation record and the claim-guard report that back every quantity quoted above.