CORPUS EVIDENCE AND FACTUAL RECALL CAUSAL TRACING WITH AUDITED LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Factual recall and corpus evidence are distinct observations: a model may answer a question even when the supporting subject–object pair is absent from a reference corpus. We compare activation-patching profiles under these operational labels on Pythia-2.8B and Pythia-6.9B. Across 2,216 traced fact–model pairs, recovery concentrates early in both networks. Among recallable facts, the groups have equal median peak depths, with no detected rank-distribution difference ($p = 0.568/0.390$). These reference-corpus labels do not establish training membership. We then compare independently restored single-layer rank-one interventions at relation-selected, fixed-middle, and control layers under equal relative update norms. On 100 held-out facts per model, selected and fixed-middle layers both achieve near-ceiling rewrite success. Their side effects differ: selected layers preserve other factual answers more often, by 3.3 and 7.0 percentage points. A post-hoc strength analysis reveals a target-success advantage at one-quarter of the calibrated norm: selected layers outperform the fixed middle layer by 18 and 20 percentage points. Layer choice therefore interacts with intervention strength, distinguishing recovery location from a uniquely effective storage address. Repeating all dose cells with two additional full-pipeline split seeds preserves the quarter-strength advantage in both models, while showing that half-strength effects are small and model-specific. Transferring the selected layers to the non-deduplicated checkpoint twins yields quarter-strength advantages of 16 and 24 percentage points. A predeclared projected-gradient intervention saturates across sites and does not reproduce the selected–middle top-1 gap, bounding the layer-selection result to the rank-one update mechanism.

1 INTRODUCTION

Locating factual associations is useful only if the location predicts a meaningful intervention. Causal tracing measures which restored activations recover a corrupted prediction (Meng et al., 2022); model editing changes weight matrices to alter an association (Meng et al., 2023). These operations need not identify the same layer (Hase et al., 2023). A further distinction concerns the evidence used to classify facts: correct recall is an observable behavior, whereas training membership is a property of a particular training corpus.

We study these distinctions using 1,756 facts paired by relation, entity frequency, and answer length. Labels indicate whether a bounded subject–object co-occurrence was found in a fixed English Wikipedia snapshot. They supply a reproducible reference-corpus contrast, not a partition into literal memorization and inference. The snapshot postdates Pythia's training corpus. The scientific question is therefore whether this observable evidence contrast accompanies different recovery profiles, and whether selecting layers from those profiles improves editing relative to equally constrained alternatives.

Figure 1 connects the two measurements. Activation patching yields an information-flow profile under subject corruption. Independent weight interventions then measure both target success and changes to other facts. The latter comparison holds layer count and relative update norm fixed: each edit changes one matrix, starts from the same base model, and is restored before the next fact. This separates layer choice from accumulated editing drift.

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.

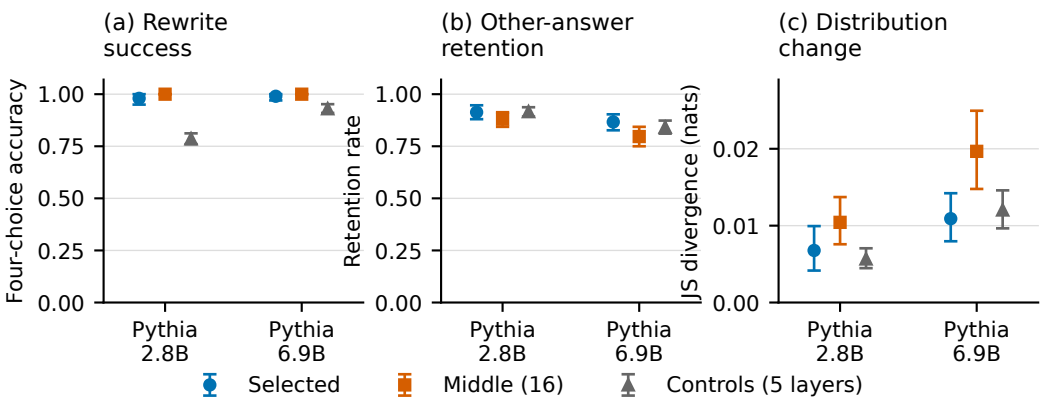

Figure 1: **High rewrite success does not imply equal side effects.** Independent, norm-controlled rank-one interventions on 100 facts per model. (a) Four-choice rewrite success. (b) Retention of the base model's answer on three other facts per edit. (c) Jensen–Shannon divergence of those four-choice distributions; lower is better. Selected layers are chosen per relation from a disjoint selection set; the middle layer is 16. The control averages five fixed layers. Error bars are fact-bootstrap 95% intervals, conditional on the sampled facts and seed. Section 3.2 specifies the intervention.

Recovery concentrates in early layers at both scales. For recallable facts, the evidence-present and evidence-absent groups have the same median normalized peak depth, although failure to reject a distributional test is not proof of equivalence. Under the weight intervention, selected layers attain four-choice success of 0.98/0.99, compared with 1.00/1.00 at layer 16. The clearer difference is locality: selected layers preserve other answers more often, while neither choice establishes a unique storage address. Reducing the update norm to one quarter of its calibrated value reveals a selected-layer advantage of 18 and 20 percentage points in four-choice success. This post-hoc dose dependence motivates distinguishing a saturated intervention from a location-insensitive one.

Our contributions are:

- A reference-corpus-controlled activation analysis across two Pythia scales, with operational evidence labels and explicit conditioning on recallability.
- Single-layer editing comparisons that separate layer selection, update magnitude, target success, and locality, including an independent projected-gradient robustness check.
- Per-fact measurements supporting a bounded distinction between early recovery profiles and the success–locality trade-off of weight interventions.

The resulting claim is deliberately narrower than a universal storage-address claim. We ask whether a layer selected from a causal-tracing profile is a better intervention site under a fixed update budget, and report both successes and side effects. The answer is conditional: the selected site can help at a weak dose, but a fixed middle band can match or exceed it at a saturated dose or in another checkpoint. This separates our measurement question from the benchmark construction in the companion paper and from claims about literal training-set membership.

2 Problem Setup

2.1 Exposure-controlled fact sets

The companion benchmark (Anonymous Authors, 2026) provides facts from LAMA-TREx relations (Petroni et al., 2019) with auditable labels relative to an English Wikipedia 2023-11-01 snapshot. For each fact, a streaming scan of its 6.4M documents searches for a normalized occurrence of the subject–object pair within a bounded context window and records the matching *evidence span*, source component, document count, and content hash. A fact is *evidence-present* if at least one span exists and *evidence-absent* if the scan finds none. Because the snapshot postdates the Wikipedia

component used by Pythia, these are reference-corpus labels rather than exact training-membership labels. Evidence-absent facts are matched to evidence-present facts within relation on entity frequency and answer length.

For the two models studied here this yields 1,756 matched facts (878 evidence-present, 878 evidence-absent) over 39 relations (e.g., P19 place of birth, P27 country of citizenship, P131 located in). Each fact is evaluated with a four-way multiple-choice probe, a true/false variant, and a *hard-distractor* variant with more semantically competitive distractors. A fact is *recallable* by a model if the model's probability for the correct choice is at least 0.5 under the standard probe. Following the causal-tracing convention of conditioning on correct answers, all RQ1 analyses condition on recallable facts; the recall filter itself is exposure-blind and retains 852 of 1,082 traced facts at 2.8B (456 evidence-present, 396 evidence-absent) and 882 of 1,134 at 6.9B (464, 418).

2.2 Models

We study `pythia-2.8b-deduped` and `pythia-6.9b-deduped` (Biderman et al., 2023), both 32-layer GPT-NeoX models (d=2560 and d=4096) trained on deduplicated Pile data. Both share depth, so absolute layer indices are comparable across scales; profile comparisons additionally normalize indices to $[0, 1]$. Base checkpoints are used throughout. For the layer-probe analysis we add `OLMo-2-1124-7B` (Team OLMo et al., 2025), a 32-layer model of different architecture trained on the openly auditable Dolma corpus; probes are descriptive only, so no editing is run on it. Primary analyses use seed 2027; full-pipeline dose resplits use seeds 2028 and 2029. Every traced fact, recovery vector, and edit log is released.

3 Method

3.1 Corrupted/clean activation patching

For each fact we form the probe prompt, append the answer, and run three passes. **Clean:** an unmodified forward pass recording the residual-stream output of every Transformer block at all positions, and the answer's total log-likelihood $\mathrm{lp}_{\mathrm{clean}}$ (summed over answer tokens). **Corrupted:** the subject tokens are replaced by randomly sampled in-vocabulary token IDs (excluding padding and EOS), yielding $\mathrm{lp}_{\mathrm{corr}}$. Facts whose corruption fails to reduce the answer likelihood by more than 10^{-6} nats carry no signal and are excluded (§A). **Patched:** for each layer $\ell \in \{0, \ldots, 31\}$, we rerun the corrupted pass while overwriting the residual stream at the subject positions with the clean values recorded after block ℓ, and record $\mathrm{lp}_{\mathrm{patch}}(\ell)$. The recovery score is

$$\mathrm{rec}(\ell) \;=\; \frac{\mathrm{lp}_{\mathrm{patch}}(\ell) - \mathrm{lp}_{\mathrm{corr}}}{\mathrm{lp}_{\mathrm{clean}} - \mathrm{lp}_{\mathrm{corr}}}, \tag{1}$$

i.e., the fraction of the corruption-induced likelihood deficit restored by patching layer ℓ. Each fact yields a 32-dimensional recovery vector; its *peak layer* ℓ^* is the argmax.

Two properties of Eq. 1 deserve emphasis. First, because numerator and denominator are log-likelihood differences, $\mathrm{rec}(\ell)$ can exceed 1; means are consequently inflated by a minority of strongly overshooting facts (median per-fact peak recovery is 1.07 at 2.8B and 1.13 at 6.9B), so we report medians alongside means and mark group-curve peaks on the mean curves with bootstrap confidence intervals. Second, patching *all subject positions* measures how much of the corruption's influence on the answer is still confined to the subject positions at depth ℓ: corruption enters only through subject-token embeddings, so restoring those positions at very early layers largely removes it, while corruption that has already propagated to other positions via attention cannot be repaired by later subject-position restoration. The curve is therefore an *information-flow profile*, not a claim that a fact is "stored" where the curve peaks (cf. Hase et al., 2023), which is why the intervention question is examined by the editing test of §3.2. For RQ1 the procedure is identical for evidence-present and evidence-absent facts, so group comparisons are internally valid.

We summarize localization by (i) the mean recovery curve with pointwise 95% confidence intervals, (ii) per-relation mean curves and their peaks, and (iii) the per-fact peak-layer distribution, with stability assessed by resampling facts (1,000 draws). Group comparisons use Mann–Whitney U tests on normalized peak layers $\ell^*/(L-1)$.

3.2 INDEPENDENT NORM-CONTROLLED EDITING

We use a ROME-derived rank-one intervention (Meng et al., 2022) to compare locations, rather than to benchmark the canonical ROME algorithm. Updates target one MLP down-projection matrix in fp32. The implementation disables second-moment adjustment and rescales the resulting update, so its efficacy is specific to this configuration.

For the primary editing run, fact identifiers assign tracing rows to selection and evaluation partitions with probability 0.5 and seed 2027; the resplit runs repeat the pipeline with seeds 2028 and 2029. A relation's selected layer maximizes its median selection-set recovery curve; relations with fewer than five selection facts use the pooled selection curve. Evaluation excludes both selection facts and their entities. We take 100 deterministically ordered held-out facts per model.

The comparison has seven settings: one relation-selected layer, fixed layer 16, and five fixed control layers $\{2, 10, 22, 26, 30\}$. Every setting changes one matrix. For model weights W, the applied update satisfies

$$\|\Delta W\|_F / \|W\|_F = \tau. \tag{2}$$

The target τ is the median raw relative update norm on ten selection facts at their selected layers: 0.034446 at 2.8B and 0.031314 at 6.9B. Each fact starts from base weights; after measurement, the original matrix is restored exactly. Thus controls differ in location, not the number of edited layers or update history.

Rewrite success requires the new target to have the highest summed continuation log-probability among all four choices. Template transfer compares the new and original answers on the first supplied template whose wording differs from the rewrite prompt; 45 of the 100 facts have such a template. It is a binary preference, not four-choice or free-generation accuracy. For locality, three distinct other facts are selected deterministically from a shared pool of 20 held-out facts. We measure preservation of the base model's four-choice argmax and Jensen–Shannon divergence of the choice distributions. Control-layer results are averaged within fact before 2,000 fact-bootstrap draws. Comparisons between settings resample the same facts jointly.

As a predeclared mechanism check, we replace the rank-one update by eight-step projected-gradient fine-tuning (PGD-FT) of the whole MLP down-projection, at the exact quarter-strength norm. The same 100 facts are evaluated at the selected, middle, early, and late layers with independent restoration. This matches the norm, but not rank or capacity; Appendix B gives details.

3.3 LAYER PROBES (DESCRIPTIVE ONLY)

As a complementary descriptive view, we train per-relation logistic-regression probes to separate evidence-present from evidence-absent facts from the last-token residual state at each of the 33 recording points (embedding output plus the 32 block outputs), using entity-grouped 80/20 splits so that no entity appears in both sides (34 of 39 relations pass the minimum-class and split filters in both models). Probes on hidden states can latch onto surface co-variates of membership (Anonymous Authors, 2026), so we treat probe AUC strictly as descriptive evidence; causal claims rest on patching and editing.

4 RESULTS

4.1 RECOVERY PROFILES CONCENTRATE IN THE FIRST THIRD OF THE NETWORK

Figure 2 shows mean recovery curves for both models. At 2.8B the mean curve rises sharply, plateaus across layers 1–11 (mean recovery 3.3–4.9), peaks at layer 6, and decays toward zero by the mid-20s; at 6.9B the curve is flatter and peaks at layer 4, decaying from roughly layer 12 onward. The mean-curve peak layer is stable under fact resampling at 2.8B (bootstrap 95% interval $[3, 6]$) and broader at 6.9B ($[0, 12]$). Localization is early distributionally as well: modal per-fact peak layers lie in 0–9 for 82.5% (2.8B) and 80.2% (6.9B) of traced facts, and per-relation mean curves peak within 0–9 for $35/39$ and $32/39$ relations. Late-layer recovery is near zero: by construction of Eq. 1, corruption that has propagated beyond the subject positions cannot be repaired there.

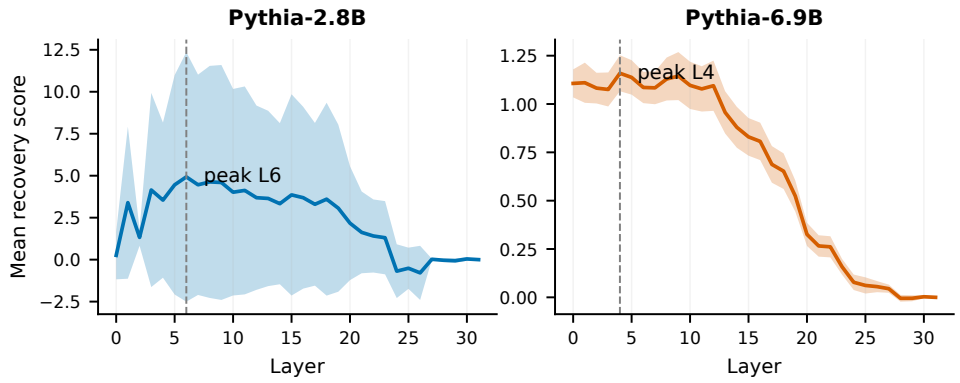

Figure 2: **Recovery concentrates in the first third of the network in both models.** Mean recovery score (Eq. 1) per layer with pointwise 95% confidence intervals ($1.96\times$ s.e. over facts; n=1,082 facts for Pythia-2.8B-deduped, left; n=1,134 for Pythia-6.9B-deduped, right; 39 LAMA-TREx relations, 32 layers). Dashed lines mark the mean-curve peak layer (L6 at 2.8B, L4 at 6.9B); values above 1 reflect overshoot of the clean likelihood in log-likelihood space.

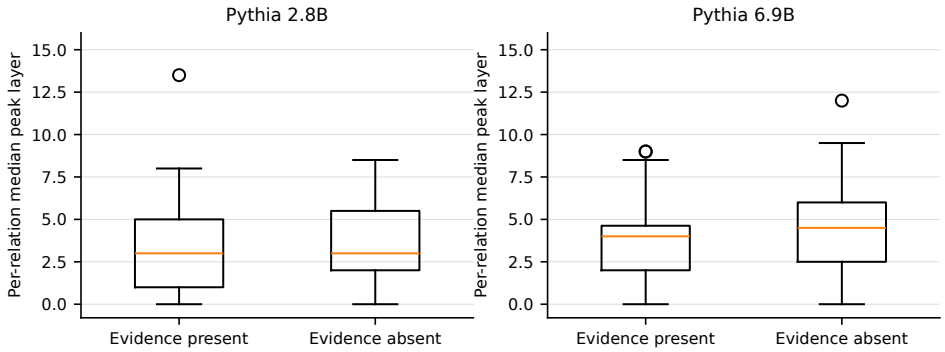

Figure 3: **Peak depths vary across relations in both evidence groups.** Each box summarizes per-relation median peak layers for relations with at least three recallable facts in the corresponding group. The underlying fact counts are 456/396 at 2.8B and 464/418 at 6.9B (present/absent). The fact-level tests in Table 1 use individual peak depths, not the relation-level medians plotted here.

4.2 Reference-evidence groups have equal median peak depths

Table 1 compares peak depths among recallable facts. Both groups have median normalized peak depth 0.097 at 2.8B and 0.129 at 6.9B. The two-sided Mann–Whitney tests give $p = 0.568$ and $p = 0.390$, respectively. These tests do not detect a difference, but do not establish equivalence. The group mean recovery curves correlate at $r = 0.978$ and $r = 0.993$. Relation-specific medians vary within each evidence group (Figure 3; Table 7).

4.3 Target success and locality favor different comparisons

Table 2 reports all seven settings, with the five controls averaged within each fact. Selected and middle layers both achieve near-ceiling rewrite success. Selected layers exceed the control average, but do not exceed layer 16. The experiment therefore does not establish that the tracing-selected layer is uniquely effective.

The selected-minus-middle rewrite difference is -0.020 at 2.8B (paired 95% interval $[-0.050, 0.000]$) and -0.010 at 6.9B ($[-0.030, 0.000]$). On the 45 template-transfer facts, both settings prefer the new answer in every case. This limited binary test is saturated and provides no evidence of a transfer advantage for either location.

Table 1: **Reference-evidence groups have equal median peak depths.** Normalized peak layers are $\ell^*/(L-1)$ with L=32; medians and means are over recallable facts only. MWU: two-sided Mann–Whitney U test on per-fact normalized peaks. Group recovery curves correlate at $r = 0.978$ (2.8B) and $r = 0.993$ (6.9B); distributions are shown in Figure 3, curves in Figure 2.

Model	Facts (pres./abs.)	Peak (norm), median		Peak (norm), mean		Curve r	MWU p
		Pres.	Abs.	Pres.	Abs.		
Pythia-2.8B (ded.)	852 (456/396)	0.097	0.097	0.148	0.152	0.978	0.568
Pythia-6.9B (ded.)	882 (464/418)	0.129	0.129	0.165	0.176	0.993	0.390

Table 2: **Independent single-layer interventions have measurable side effects.** Each model uses 100 held-out facts. Rewrite is four-choice target accuracy; transfer is new-versus-original answer preference on 45 distinct-wording templates. Retention and JS measure other-fact locality. Control means average five fixed layers within each fact. All rows use the same model-specific relative update norm.

Model	Setting	Rewrite ↑	Transfer ↑	Retention ↑	JS ↓
2.8B	Selected	0.980	1.000	0.913	0.00677
	Middle (16)	1.000	1.000	0.880	0.01043
	Controls	0.786	0.756	0.916	0.00566
6.9B	Selected	0.990	1.000	0.867	0.01091
	Middle (16)	1.000	1.000	0.797	0.01967
	Controls	0.930	0.907	0.841	0.01204

Locality distinguishes the settings more clearly. Selected-minus-middle answer retention is $+0.033$ at 2.8B (95% interval $[0.003, 0.067]$) and $+0.070$ at 6.9B ($[0.027, 0.113]$). At 6.9B, the JS difference is -0.00876 ($[-0.01370, -0.00403]$), indicating smaller distributional changes at selected layers. At 2.8B, its interval crosses zero (-0.00366, $[-0.00754, 0.00033]$). The controls also exhibit a trade-off: at 2.8B they have lower rewrite success and smaller average locality JS. These observations support evaluating layer choice on success and side effects jointly, rather than interpreting a saturated success metric as a unique address.

4.4 Update strength reveals differences hidden by saturation

We examine whether near-ceiling target accuracy obscures the effect of layer selection. In a post-hoc sensitivity analysis, we multiply each model's calibrated relative update norm by 0.5 and 0.25, keeping the evaluation facts, selected layers, control layers, templates, locality questions, and editing seed fixed. Each dose comprises 100 facts and 700 independent edits per model. Table 3 reports every tested dose.

At one-quarter strength, selected layers improve target accuracy over the middle layer by 18 percentage points at 2.8B and 20 points at 6.9B. At half strength the differences are 10 and 5 points, respectively; the latter interval includes zero. The preference for the new answer under distinct templates follows the same ordering at reduced strength: selected versus middle is 0.533 versus 0.356 at 2.8B and 0.689 versus 0.444 at 6.9B for the quarter-strength edits, measured on 45 facts.

The original norm thus masks a difference that becomes visible when the intervention is weaker. This supports a dose-dependent benefit of tracing-based layer selection for this rank-one intervention, rather than a uniquely editable storage location. The intervals in Table 3 condition on the primary seed and fixed selection procedure; they do not measure variation across alternative layer-selection samples or editing algorithms.

To assess sensitivity to the full selection pipeline, we repeated all six dose/model cells with two additional split seeds (2028 and 2029). Each seed resamples selection, calibration, and evaluation sets and may choose different layers and update norms. Within a seed, the same 100 evaluation facts are used across doses; across seeds, only four evaluation facts overlap for 2.8B and two for 6.9B. Table 4 reports four-choice success differences. At quarter strength, selected-minus-middle differences are $+0.08/+0.14$ for 2.8B, with paired fact-bootstrap 95% intervals $[0.02, 0.15]/[0.08, 0.21]$,

Table 3: Four-choice target accuracy across update strengths. Controls average five fixed layers within each fact. The difference is selected minus middle; intervals use 10,000 paired fact-bootstrap resamples. These are pointwise 95% intervals without multiplicity adjustment.

Model	Norm multiplier	Selected	Middle	Controls	Difference [95% CI]
2.8B	1	0.980	1.000	0.786	-0.020 $[-0.050, 0.000]$
	0.5	0.800	0.700	0.526	$+0.100$ $[0.010, 0.190]$
	0.25	0.430	0.250	0.258	$+0.180$ $[0.100, 0.260]$
6.9B	1	0.990	1.000	0.930	-0.010 $[-0.030, 0.000]$
	0.5	0.950	0.900	0.694	$+0.050$ $[-0.010, 0.110]$
	0.25	0.630	0.430	0.356	$+0.200$ $[0.120, 0.280]$

Table 4: Full-pipeline resplits of the dose effect. Entries are selected-minus-middle four-choice success differences for seeds 2028 and 2029, each evaluated on its own 100-fact set.

Model	Norm 1	Norm 0.5	Norm 0.25
2.8B	$-0.03/-0.02$	$+0.10/+0.02$	$+0.08/+0.14$
6.9B	$0.00/-0.01$	$-0.03/-0.01$	$+0.25/+0.16$

and $+0.25/+0.16$ for 6.9B, with intervals $[0.16, 0.35]/[0.06, 0.26]$. These pointwise intervals use 10,000 resamples within each run and do not estimate uncertainty over the selection pipeline. The quarter-strength ordering recurs in both resplits, while the half-strength intervals all include zero. Each run records 700 layer–fact outcomes and zero weight restoration error. We do not pool overlapping facts as independent replicates.

4.5 Layer choices transfer to non-deduplicated checkpoint twins

We next apply the layers selected from each deduplicated checkpoint's tracing matrix to its non-deduplicated twin. The comparison retains the seed-2028 fact IDs, templates, locality questions, and one-layer intervention capacity, while calibrating the update norm separately on the target checkpoint. Table 5 reports all predeclared strengths. At quarter strength, selected layers exceed the middle layer by 0.16 at 2.8B and 0.24 at 6.9B; both paired fact-bootstrap intervals exclude zero. Full-strength edits saturate again, and half-strength intervals include zero. Relative to the matched deduplicated runs, the selected-minus-middle contrast changes by $+0.08$ $[0.02, 0.15]$ at 2.8B and -0.01 $[-0.12, 0.10]$ at 6.9B at quarter strength. These interactions compare checkpoint-specific calibrated norms and therefore do not isolate the causal effect of corpus deduplication.

4.6 Projected-gradient updates bound algorithmic generality

PGD-FT does not reproduce the quarter-strength rank-one success gap: selected and middle success is 1.00 for both models. Selected-minus-middle target-margin gain is 6.25 ($[5.04, 7.37]$) at 2.8B but -0.07 ($[-1.15, 1.00]$) at 6.9B; locality JS is worse by 0.0101 ($[0.0055, 0.0146]$) and 0.0144 ($[0.0069, 0.0220]$). Table 6 reports every setting. Full-matrix PGD largely removes site differences in binary success, so the selected-layer advantage is not invariant to the editing mechanism.

4.7 Layer probes corroborate the patching picture, descriptively

Per-relation linear probes recover reference-evidence labels with different strengths across relations (Figure 4). Define an early strong profile as a relation whose maximum AUC is at least 0.8 and whose maximizing recording point is at or before layer 16. This criterion holds for 17 of 34 relations at 2.8B and 18 of 34 at 6.9B; 14 relations satisfy it in both models. At a 0.7 threshold with the same depth condition, the counts are 19 and 22. Thus the shared subset is a minority, but the individual-model counts at 0.8 cover one half or slightly more of the relations.

OLMo-2-7B has 15 of 34 relations meeting the 0.8 criterion and 17 meeting the 0.7 criterion. This extends the descriptive comparison to another model family. All models are evaluated against the same reference-corpus labels, rather than independently verified training-membership labels. Select-

Table 5: Transfer of deduplicated-checkpoint layer choices to the corresponding non-deduplicated checkpoint. Four-choice target accuracy is measured on the same 100 seed-2028 facts; intervals use 10,000 paired fact-bootstrap resamples.

Model	Norm	Selected	Middle	Difference [95% CI]
2.8B	1	0.97	1.00	-0.03 $[-0.07, 0.00]$
	0.5	0.74	0.65	$+0.09$ $[-0.01, 0.19]$
	0.25	0.36	0.20	$+0.16$ $[0.08, 0.25]$
6.9B	1	0.99	1.00	-0.01 $[-0.03, 0.00]$
	0.5	0.84	0.80	$+0.04$ $[-0.04, 0.13]$
	0.25	0.47	0.23	$+0.24$ $[0.16, 0.33]$

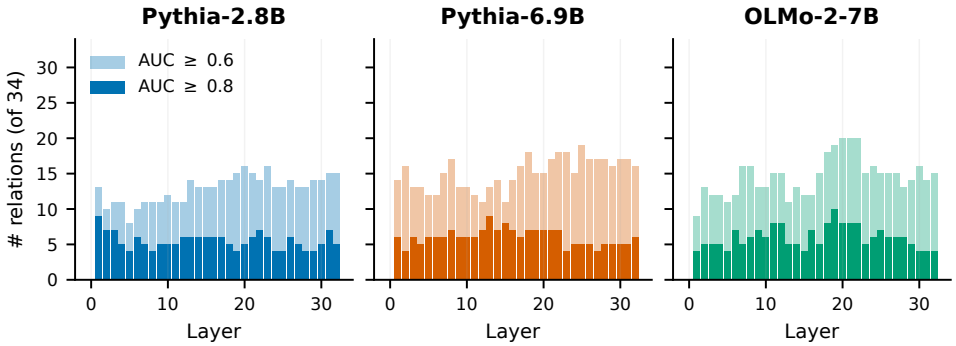

Figure 4: **Reference-evidence decodability varies across relations and recording points.** For each recording point (0: embedding output; 1–32: post-block states) we count the analyzable relations (of 34) whose entity-grouped evidence-present-vs-evidence-absent probe AUC clears 0.6 (light) or 0.8 (dark). The profile replicates on OLMo-2-7B (Team OLMo et al., 2025), a different architecture trained on a different (openly auditable) corpus. Probes are descriptive (§3.3).

ing the maximum over recording points makes these counts exploratory summaries, not multiplicity-adjusted tests of decodability. Linear probes may also exploit covariates correlated with the labels. We therefore distinguish this analysis from the activation and weight interventions.

4.8 Cross-scale interpretation

Both scales exhibit early recovery and similar evidence-group median peak depths. The rank-one editing comparison also has the same broad ordering: selected and middle layers have high target success, and selected layers retain more non-target answers. Absolute success and locality differ across the models. The dose-dependent target-success ordering also transfers to the two non-deduplicated checkpoint twins. These four checkpoints share the Pythia architecture, so the result is not a universal rule for moving layer indices between architectures. PGD-FT does not preserve the selected–middle success or locality ordering, further limiting generalization across update mechanisms. The OLMo probe analysis is descriptive and does not supply a cross-family editing test.

5 Discussion and Limitations

Recovery is not a storage address. Restoring all corrupted subject positions early in the network can recover information before it spreads to other positions. An early recovery maximum is therefore compatible with a broad range of effective weight-intervention sites. Our independent editing results make the distinction concrete: a fixed middle layer rewrites as often as the selected layer, while their side effects differ. The supported interpretation concerns this intervention's success–locality tradeoff, not a unique anatomical location of a fact.

Scope of the evidence contrast. Reference-snapshot co-occurrence labels do not identify which facts were memorized during Pythia pretraining. Equal sample medians and non-significant rank

tests also do not establish distributional equivalence. The group analysis measures associations with the stated reference-corpus labels among recallable facts.

Scope of the intervention. The 100-fact primary comparison is repeated with two resplit seeds and transferred once to checkpoint twins, but does not establish behavior for canonical ROME, other editors, or architectures. PGD-FT shows that matching Frobenius norm neither matches update capacity nor preserves the success advantage. Template transfer uses 37–45 facts and two candidates; locality uses three facts from a 20-fact pool. These operating choices bound what "improvement" means here.

6 Related Work

Knowledge localization. Causal tracing repairs activations at individual sites (Meng et al., 2022), while MEMIT distributes edits across middle layers (Meng et al., 2023). Related work identifies knowledge neurons and feed-forward key–value structure (Dai et al., 2022; Geva et al., 2021; 2023). Most directly, Hase et al. (2023) show that tracing attributions often fail to predict editing success; our controlled interventions examine that distinction.

Memorization, membership, and exposure. Models measurably memorize training data, with duplication amplifying the effect (Carlini et al., 2021; 2023; Kandpal et al., 2022; Lee et al., 2022). Membership inference tests whether examples belonged to a model's training data (Shokri et al., 2017; Carlini et al., 2022; Shi et al., 2024; Duan et al., 2024). Related work probes personal-information leakage (Kim et al., 2023; Huang et al., 2022) and characterizes corpus content (Elazar et al., 2024). The companion benchmark supplies audited reference-corpus labels and matched controls (Anonymous Authors, 2026); here those groups show no detected peak-depth difference.

Model editing as an instrument. ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), MEND (Mitchell et al., 2022), and WISE (Wang et al., 2024a) edit factual associations through targeted weight updates; EasyEdit provides reference implementations (Wang et al., 2024b). We instead use editing as a measurement instrument: whether patching-nominated sites accept equally constrained edits more readily than alternatives. Template consistency follows Elazar et al. (2021); joint success and locality evaluation does not convert an attribution into a storage claim.

7 Conclusion

We separate reference-corpus evidence, activation recovery, and editing efficacy. Pythia-2.8B and Pythia-6.9B have early recovery profiles, and recallable facts in the two evidence groups have equal median peak depths without a detected rank-distribution difference. Under independently restored, norm-controlled single-layer interventions, selected and fixed-middle layers both achieve high rewrite success. Selected layers better preserve other answers in this setting and achieve higher target success when the update norm is reduced to one quarter of its calibrated value. This dose dependence exposes a difference hidden by the original success ceiling, but an independent full-matrix gradient update saturates and does not reproduce the success gap. The evidence supports a mechanism-conditional success–locality comparison of intervention sites, rather than a unique storage address or a general claim about training membership.

Reproducibility Statement

The companion benchmark's audit protocol (reference-corpus scans, evidence spans, matching rules, seeds) carries over unchanged. The anonymous supplementary material contains the tracing, probe, and editing code, exact hyperparameters, model-loading details (§A–§B), and compact audited outputs. Every reported number is computed from retained per-fact artifacts (recovery matrices of shape $n \times 32$, layer-AUC tables, per-relation summaries, and raw editing logs); the complete artifacts and failed runs will be released upon publication. The fp32 6.9B editing runs use a 48GB GPU. The main comparison uses seed 2027; full-pipeline resplits use seeds 2028 and 2029, and checkpoint transfer uses seed 2028. The PGD-FT robustness check uses seed 2027 and retains its optimization traces.

ETHICS STATEMENT

This research uses only the companion benchmark's public corpora and derived labels, open-weight models (Pythia, OLMo), and no human subjects or private data. The work studies where factual recall is implemented inside networks and how model edits behave; it does not introduce a technique for extracting training data, enlarging exposure, or evading audits. Localization and editing tools can be repurposed for unlearning or for targeted model modification; we discuss these dual-use implications qualitatively and release artifacts to support auditing rather than attack.

AI USE STATEMENT

We used AI assistants for experiment design, code implementation, result analysis, and translation. The authors take responsibility for the final content.

REFERENCES

Anonymous Authors. Can factual probes detect corpus evidence? an auditable benchmark. Anonymous companion manuscript; included in supplementary material, 2026.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Nicholas Carlini, Katherine Lee, Daphne Ippolito, Florian Tramèr, Matthew Jagielski, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=av0D19pSkU.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah Smith, and Jesse Dodge. What's in my big data? In *International Conference on Learning Representations (ICLR)*, 2024. URL

https://proceedings.iclr.cc/paper_files/paper/2024/hash/1f7336fd66b6e6e63d1801fdd5930a5a-Abstract-Conference.html.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021. doi: 10.18653/v1/2021.emnlp-main.446.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668, 2023. doi: 10.52202/075280-0774.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2038–2047, 2022. doi: 10.18653/v1/2022.findings-emnlp.148.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10697–10707, 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-0911.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022. doi: 10.18653/v1/2022.acl-long.577.

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. doi: 10.52202/068431-1262.

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. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019. doi: 10.18653/v1/D19-1250.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e32ad85fa27be4a9868d55703f01323e-Abstract-Conference.html.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. doi: 10.48550/arXiv.2501.00656.

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*, volume 37, pp. 53764–53797, 2024a. doi: 10.52202/079017-1703.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 82–93, 2024b. doi: 10.18653/v1/2024.acl-demos.9.

A Patching hyperparameters

- **Traced facts:** up to 50 facts per relation, sampled uniformly with seed 2027 from the matched fact pool; facts retained only if subject-token corruption reduces the answer log-likelihood by more than 10^{-6} nats. This yields $1{,}082$ traced facts for Pythia-2.8B-deduped and $1{,}134$ for Pythia-6.9B-deduped over 39 relations.
- **Corruption:** each subject token replaced by an i.i.d. uniformly sampled in-vocabulary ID (padding and EOS excluded).
- **Patching:** the residual-stream output of block ℓ at all subject positions is overwritten with the clean run's values (recorded via `output_hidden_states`); one layer at a time, corrupted inputs otherwise unchanged; gradients never constructed (`inference_mode`).
- **Recovery:** Eq. 1 in summed answer-token log-likelihood space; per-fact peak layer is the argmax over 32 layers.
- **Statistics:** pointwise 95% CIs as $1.96\times$ s.e. over facts; peak-layer stability from 1,000 fact-resampling draws; Mann–Whitney U tests on normalized peaks; recall threshold ≥ 0.5 correct-choice probability for RQ1 subsets.

B Editing configurations

The intervention uses EasyEdit's ROME rank-one update implementation with `fact_token=subject_last`, 20 gradient steps, learning rate 0.5, weight decay 0, clamp factor 3, KL factor 0.0625, and context-template lengths `[[5,10]]`. Second-moment adjustment is disabled. The value-loss layer is the edited layer. The target is the first incorrect option in the original four-choice question. Continuations are scored with a leading space.

The normalized intervention and all compared sites are defined in Section 3.2. Base weights are restored after every edit; the measured maximum absolute restoration error is zero for all reported edits. Layer selection and norm calibration use selection facts, while evaluation excludes their entities. Python, NumPy, and PyTorch randomness are seeded with 2027, 2028, or 2029 as specified above. All Pythia checkpoints are evaluated in fp32. No sequential multi-fact update is used in the reported comparison.

The PGD-FT robustness check freezes every parameter except one MLP down-projection matrix and takes eight normalized steps, each of length one quarter of the ball radius, on target-continuation cross-entropy. It then projects radially to the exact relative norms 0.0086115 (2.8B) and 0.0078286 (6.9B). The four settings share fact IDs, target prompts, three locality facts, and restoration logic. All 800 setting–fact outcomes pass hash, layer, norm, finite-value, and zero-restoration-error checks.

Table 6: Independent projected-gradient intervention on 100 facts per model. Margin is the post-minus-pre change in target-versus-original four-choice log probability margin; retention and JS measure three other facts per edit.

Model	Setting	Success	Margin gain	Retention	JS
2.8B	Selected	1.000	20.32	0.883	0.01769
	Middle (16)	1.000	14.07	0.897	0.00756
	Early (2)	1.000	25.76	0.867	0.02758
	Late (30)	1.000	11.83	0.897	0.01027
6.9B	Selected	1.000	25.43	0.803	0.03204
	Middle (16)	1.000	25.50	0.840	0.01760
	Early (2)	1.000	22.69	0.833	0.02150
	Late (30)	0.990	11.34	0.907	0.00584

C PER-RELATION PEAK LAYERS

Table 7 gives group counts and exact sample median peak layers for the recallable subsets. Half-layer medians arise from even group sizes. Labels refer to the audited reference snapshot.

Table 7: Per-relation peak-depth summaries, layers 0–31. Present and absent refer to reference-corpus evidence. An empty group is marked with a dash.

	Pythia 2.8B				Pythia 6.9B			
Relation	n_+	n_-	Med. $+$	Med. $-$	n_+	n_-	Med. $+$	Med. $-$
P1001	5	5	7	4	5	5	0	12
P101	15	16	0	0	13	14	4	4
P103	11	12	3	4.5	12	12	2	1.5
P106	5	2	3	3.5	7	7	2	5
P108	2	2	7.5	10	2	2	5	1.5
P127	23	21	5	6	22	23	2.5	8
P1303	4	3	0.5	3	5	2	4	6.5
P131	14	12	4	4	14	13	4.5	4
P136	5	2	2	2	5	4	1	3.5
P138	9	8	5	1.5	10	9	8.5	6
P140	23	21	2	2	23	21	4	3
P1412	21	13	8	5	22	20	2.5	2
P159	20	18	0.5	0.5	26	18	4.5	4
P17	20	20	0	0	19	21	2	2
P176	14	13	0	3	14	11	3.5	1
P178	5	3	0	2	5	3	3	2
P19	22	16	4.5	8	16	18	7.5	8.5
P190	4	2	8	2.5	5	3	2	3
P20	24	15	5	6	22	21	9	9
P264	3	2	0	2.5	3	4	4	4
P27	20	23	0	2	23	20	4	4.5
P276	6	7	5.5	2	6	6	3	1
P279	21	28	2	2.5	21	29	5	7
P30	4	7	1	7	6	7	2	0
P31	13	9	1	7	13	8	8	6
P36	6	5	3	1	6	5	4.5	8
P361	12	11	4.5	4	13	11	4	5
P364	15	11	3	3	17	12	6	5.5
P37	6	5	13.5	7	5	6	4	5
P39	4	5	3	4	7	5	7	6
P407	14	6	2.5	3	15	6	2	6

Relation	n_+	n_-	Med. $+$	Med. $-$	n_+	n_-	Med. $+$	Med. $-$
P413	3	1	8	9	1	3	3	3
P449	2	3	1	2	4	3	2.5	5
P463	3	2	4	0	3	2	0	3
P47	2	0	2.5	–	2	1	2	0
P495	16	16	6.5	8.5	17	12	8	9.5
P527	28	18	5.5	7.5	24	22	9	8.5
P740	11	13	4	5	11	11	4	2
P937	21	20	1	0	20	18	1	1.5