WHEN DOES IN-HOUSE ADAPTATION COST SAFETY? A CAPACITY–RANK–RECIPE MAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning an assistant on an in-house security corpus looks like the benign case: domain knowledge padded with ordinary alignment data. It is not. We fine-tune Qwen2.5 from 0.5B to 7B on a 6,503-sample mixture of domain knowledge and safety-alignment data, and evaluate 29 configurations (28 at three seeds) on a 110-item suite: what decides whether safety survives is the adapter's rank, with capacity setting how much is lost rather than whether. Under a conservative recipe, nothing below rank 64 leaves the ± 0.05 band around its own base, while the rank-128 cells at 1.5B, 3B and 7B lose 0.15–0.22 under the rule scorer and survive multiple-comparison correction; an aggressive recipe — three epochs at twenty times the learning rate — collapses every scale instead, by 0.23–0.33. The corpus is the weaker lever: at the same rank a safety-neutral corpus with no harmful or security text still costs most of the same loss, and at rank 16 no corpus we ran costs anything of that size. The cost is also protocol-dependent: the collapse disappears when the adaptation safety prompt is present at evaluation, so a benchmark run with that prompt can certify a model that this suite scores as collapsed.

Three other results are negative and, we think, the most reusable: adaptation buys almost no knowledge, our own refusal probe mis-specifies the geometry at the top of the range, and each explanation the collapse suggests fails its own test. In practice: monitor rank and update magnitude, and keep knowledge out of the weights where retrieval can supply it.

1 INTRODUCTION

A security team that wants an in-house assistant typically fine-tunes a small open model on a corpus it already owns — vulnerability advisories, internal write-ups, incident reports. Such a corpus is not knowledge-only: a CVE description that answers a legitimate analyst question also specifies how the corresponding weakness is exploited, and in practice the corpus is padded with general alignment data. The two channels cannot be separated by construction, only by measurement.

Such a measurement is usually taken at one scale under one recipe, where it cannot say whether the cost belongs to the domain, capacity, rank, or the training hyperparameters, nor whether what was lost is knowledge or alignment. We therefore measure all four axes at once, for Qwen2.5 from 0.5B to 7B; the result is a map rather than a number.

The first thing the map shows is that the cost is conditional, and that rank is the condition. Under a conservative regime (lr 5×10^{-6}, one epoch) nothing at $r \leq 48$ leaves the ± 0.05 band around its own base or survives multiple-comparison correction, whereas at $r=128$ the loss is 0.15–0.22 at 1.5B, 3B and 7B and every one of those cells survives it — as does 3B at $r=64$, so the threshold sits lowest at the mid-scale as measured (7B is borderline at that rank, and the scales do not share an adaptation layout; Section U). The corpus decides less than the rank does. At $r=128$ the same recipe on Zhihu question-answering text — $D=6{,}500$ instructions with no harmful and no security material — costs 0.09 at 1.5B and 0.15 at 3B against 0.16 and 0.22 for the security corpus at the same scale and rank: 58–65% of the in-domain loss, from a corpus with nothing in it to blame (a 3B result — at 1.5B the ratio is scorer-dependent, Section R). The same corpus at $r=16$ shows no net loss at 0.5B, 1.5B or 7B (at 1.5B that cell is an aggregate whose per-seed records were not retained; Section T). Under an aggressive regime (lr 10^{-4}, three epochs) all four scales collapse by 0.23–0.33. The two regimes differ in kind rather than in degree.

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.

The second is that there is almost no knowledge to buy. A "knowledge–safety trade-off" presumes that adaptation delivers knowledge, and the conservative regime largely does not: it moves CVE accuracy from a base of 0.266 no further than 0.315, and except at 3B r=128 the whole of that is answer structure rather than recalled facts, whose channel does not move (0.011–0.026 across the conservative rank sweep, against 0.016–0.019 at base). The exception reconciles with the composite rather than contradicting it: at 3B r=128 the two components move in opposite directions — fact-hit rises from 0.019 to 0.055 while the structured component falls from 1.00 to 0.73, the larger fall carrying the smaller weight — so the item-mean CVE accuracy still drops, from the 3B base's 0.313 to 0.257. The aggressive regime does move the fact channel, by about an order of magnitude, and pays for it in the safety channel. Since neither regime buys knowledge cheaply, the deployment question becomes a question about where knowledge should live.

The third concerns mechanism, and it is where our own instrument fails. Activation probes reproduce the expected signature at 1.5B — the r=128 configuration retains more factual knowledge than its base (κ=0.039 against 0.008) while refusal falls to 0.50 and the safety boundary shrinks from 22.2 to 9.0 — but the single-direction drift that carries this reading at 1.5B does not transfer to 7B. We read that as a limitation of the probe rather than an anomaly of the model: refusal is mediated by a multi-dimensional cone whose effective dimensionality grows with scale (Wollschläger et al., 2025), so a single direction is mis-specified exactly where capacity is largest. A second, related negative follows from the same measurements: in-domain geometric drift tracks collapse but is not sufficient for it.

Three candidate explanations bound all of the above — a train–evaluation format mismatch, since the corpus carries a Chinese safety system prompt (76.8% of its records) while in-domain evaluation uses none, under-training, and contamination (18.6% of the training targets) — and none of them accounts for the collapse: the first explains about a quarter of it, the second none, and the third removes the conservative r=128 loss only at 1.5B, not at 3B or 7B (Sections K to M). No claim here is a deployment guarantee.

Contributions. (i) A conditional map of in-house-adaptation harm over capacity, rank, and recipe that separates the rank axis from the corpus axis, with a statistical protocol that reports equivalence tests and minimum detectable effects rather than treating a null as an absence, and a safety band in update magnitude (0.0014–0.0040) that orders harm within a recipe but not across the rank axis. (ii) A protocol effect that changes the measured size of the effect: the refusal an adapted model shows is a function of the evaluation prompt, to a degree comparable with the recipe, so a fixed benchmark run under a safety prompt reads at ceiling on cells that an in-domain suite scores as collapsed (Section 7). The two external benchmarks we use carry that same prompt and nine tenths of their items are also training-corpus items, so we read them as consistent with the effect rather than as its measurement, and their unseen halves as a direction rather than a bound (Section O). (iii) An in-domain positive control for knowledge injection, which separates "knowledge was retained" from "knowledge was never written", together with two negative results about the instrument: single-direction refusal drift mis-specifies the geometry at 7B, and in-domain drift is not sufficient for collapse. (iv) An audited evidence base — corpus composition, a harmfulness audit of every training target with its cleaned-corpus replication, overlap audit, judge validation with a non-Qwen replication and the manual re-read that is our only judge-independent check on the corpus labels, and a released suite — every number regenerated by script from stored per-item records. Section 3 fixes the corpus and protocol, Section 4–Section 7 give the results, and Section 8 the limits.

2 Related Work

Four literatures bear on this result, and none of them varies the four axes together.

Benign fine-tuning and safety. Fine-tuning aligned models on data that is not obviously harmful degrades their safety, and does so when the harmful examples in the fine-tuning set are only a handful (Qi et al., 2024), an effect whose measurement is itself unstable and instrument-dependent, with conclusions moving with the benchmark and the evaluator and different instruments disagreeing on the same models (Fraser et al., 2025; Vishnubhotla et al., 2026; Khan et al., 2026). Proposals to prevent it include perturbation-aware alignment (Huang et al., 2024), gradient-influence attenuation (Nguyen et al., 2026), safety-preserving adaptation — a fixed safety module (Li et al., 2025), pro-

jection of the update onto a safety subspace (Hsu et al., 2024), or post-hoc editing of the update (Lu et al., 2025) — and continual-learning objectives (Alssum et al., 2025); these are evaluated at one scale under one recipe. We supply the missing axes and find the effect conditional on them.

Cybersecurity fine-tuning. CyberLLMInstruct shows a capability–safety trade-off for cybersecurity fine-tuning with a deliberately pseudo-malicious corpus, at a single recipe (ElZemity et al., 2025; 2026). Our corpus is an in-house adaptation mixture that was not constructed against any attack, but our audit of every training target finds it is not content-benign either: 18.6% of its assistant turns supply usable harmful content, concentrated in two public sources (Section B). We treat that as a measured confound, and we map how the cost of adaptation varies across capacity, rank, and recipe, where we find it conditional rather than uniform. Domain capability work such as Suryanto et al. (2026), Deng et al. (2024), and the software-engineering survey of Hou et al. (2024) pursues accuracy rather than the safety boundary.

Where the failure comes from. Refusal is mediated by a direction (Arditi et al., 2024), and more accurately by a multi-dimensional cone whose dimensionality grows with model scale (Wollschläger et al., 2025) — a result that turns our own probe failure into a predicted limitation. A routing account with mechanism evidence has been given concurrently by Guo et al. (2026a), who locate benign fine-tuning's damage in a low-rank output-routing pathway; we agree with the routing-not-storage reading for held-out facts and add where in the capacity–rank–recipe space it appears. Related mechanisms include spurious forgetting (Zheng et al., 2025), loss of ignorance awareness (Shen et al., 2025), orthogonal-subspace alignment (Mou et al., 2025), and re-emergent misalignment (Giordani, 2025). A basin account of fine-tuning (Chen et al., 2026) predicts both our low-rank 7B null and the fact that low-rank conservative updates leave behaviour alone, and we use it as an external explanation rather than a mechanism of ours. Model-organism work shows emergent misalignment down to 0.5B and with a rank-1 adapter on narrow harmful data (Turner et al., 2025); our corpus is a mixed adaptation corpus and our finding runs the other way on that axis: on a mixed corpus the collapse requires adapter rank and a high learning rate, and is not induced at low rank.

Knowledge injection and adapter capacity. Retrieval outperforms fine-tuning for injecting knowledge in general-purpose settings (Ovadia et al., 2024); we replicate that on held-out vulnerability knowledge and add the safety side of the ledger. Rank sweeps and rank-overshoot analyses bound adapter capacity from below and above (Xue & Mirzasoleiman, 2025; Kalajdzievski, 2023; Liu et al., 2025; Zhang et al., 2025). The closest prior rank sweep (Gulati & Raval, 2026) reports monotone degradation with error bars over $r{=}8$ to 256, but it fine-tunes a vision-language agent on a narrow harmful corpus; under a benign adaptation mixture we find the rank effect conditional rather than monotone — present at 1.5B, bounded at 0.5B, and at 7B confined to the top of the sweep — and interacting with learning rate and epoch. Concurrent emergent-misalignment work likewise finds the rank structure, not the update norm, governing whether a narrow harmful fine-tune recruits a misalignment persona (Drake & Eberstadt, 2026); on a mixed adaptation corpus and a safety–utility composite we find the effect bounded rather than universal. Data-selection work (Xie et al., 2023a; Xia et al., 2024; Du et al., 2023; Liu et al., 2024) supplies the subset arms we use as mechanistic controls. Concurrent work on benign multilingual fine-tuning reports heterogeneous safety effects across languages and scales (Hawkins et al., 2026); we differ in sweeping rank and training strength, and in showing that part of the apparent effect is a function of the evaluation protocol (Section 7).

3 Setup

This section fixes the models, corpora, suite and protocol; every number below is a comparison against the same-scale base under one of the two regimes.

3.1 Models, corpus, and the two regimes

We fine-tune LoRA adapters (Hu et al., 2022) on Qwen2.5-Instruct at 0.5B, 1.5B, 3B, and 7B (Qwen Team, 2024), one family so that capacity is the only variable that changes, plus Mistral-7B-Instruct as a cross-family check (Jiang et al., 2023). No arm of the map is quantized: adapters are trained and merged in half precision — fp16 at 0.5B and 1.5B, bf16 at 3B and 7B — and evaluated from

Table 1: The corpus by kind; Table 3 lists the nine sources and the roles they were scripted for. Only the 764 knowledge samples are domain knowledge.

Kind	n	Sources
safety-alignment, public	2,851	`hh-rlhf`, `real-toxicity-prompts`
safety-alignment, scripted	1,598	`template`
harmful Q&A	729	`unsafe_qa`
domain knowledge	**764**	`nvd`, `arxiv`, `github`
attack data	561	`advbench`, `jailbreakbench`

the merged weights in fp16 (Section U). We sweep LoRA rank $r \in \{16, 32, 48, 64, 128\}$ and data budget $D \in \{500, 1000, 3000, \text{full}\}$, and hold the rest fixed per regime: a *conservative* regime (lr 5×10^{-6}, one epoch, $\alpha{=}2r$) and an *aggressive* regime (lr 10^{-4}, three epochs) — an ordinary LoRA default, so read the two labels as relative. Every headline configuration is trained with three seeds (42/44/46); the exceptions are marked single-seed in the tables and are not given tests.

The corpus is 6,503 ShareGPT samples assembled from crawled public sources and scripted templates. Table 1 gives it by kind, and Table 3 source by source and role by role. It is a mixture rather than a knowledge dump: by role only 764 samples (11.7%) carry domain knowledge, and the remainder is ordinary safety-alignment data. Adapting on such data is benign in intent, not in content: the audit of every assistant turn finds usable harmful content in 1,989 of them (18.6%), covering 1,522 samples (23.4%) (Section B). A second, safety-neutral corpus (6,500 Zhihu instructions) is used only for the cross-domain control of Section P, trained with the same regime and budget.

3.2 Evaluation suite and metrics

Evaluation is dual-objective on a 110-item suite (Section A): 30 harmful prompts, 20 jailbreak prompts, 8 safety-QA items, 32 CVE-analysis items drawn from real NVD descriptions (NIST, 2024), and 20 general-QA items. Composite scores are

$$\textsc{Safety} = 0.5\, r_{\text{ref}} + 0.3\, a_{\text{qa}} + 0.2\, r_{\text{jb}}, \qquad \textsc{Utility} = 0.4\, a_{\text{cve}} + 0.4\, a_{\text{gen}} + 0.2\,(1 - h), \tag{1}$$

$$\text{SUS} = 0.5\,\textsc{Safety} + 0.5\,\textsc{Utility}, \tag{2}$$

with r_{ref} harmful-prompt refusal, r_{jb} jailbreak resistance, a_{qa} safety-QA accuracy, a_{cve} CVE-analysis accuracy, a_{gen} general-QA accuracy, and h the rule-based hallucination flag. Each CVE item's score is $0.7{\cdot}$fact-hit $+ 0.3{\cdot}$structured. The sub-weights are conventions, not estimates; we report the safety axis componentwise wherever a composite would hide a sign, and we report the sensitivity of SUS to the safety weight in Section H.

Two properties of the suite matter for reading the results. The 32 CVE items split into 18 whose NVD entry is in the corpus and 14 that are not, so the knowledge channel has a built-in memory/generalisation contrast. A second, 50-item bilingual probe — the 30 harmful and 20 jailbreak items rendered a second time in English, so each content is answered twice — carries only the protocol contrast of Section 7. And the rule scorers are validated against an LLM judge (local Qwen3.5-9B, greedy, fixed rubric; the same model family as the models under test, which is the one independence it cannot claim, so Section G adds a non-Qwen replication and a human re-read): agreement on refusal is moderate ($\kappa{=}0.52$, $N{=}636$) and the two scorers order the 13 judged configurations almost identically (Spearman $\rho{=}0.94$), but the hallucination flag agrees poorly ($\kappa{=}0.09$, $N{=}399$), so we treat hallucination as secondary (Section G reports a knowledge-base entity check alongside it, $\kappa{=}0.28$, agreement 0.94).

3.3 Statistical protocol

The design has one statistical difficulty, which is worth stating before any number: the base is a single run. It contributes no degrees of freedom and no variance estimate, so every test below is a one-sample t test of the tuned seeds against a fixed reference with $df{=}n - 1{=}2$. Three features of the protocol follow from that. (i) Because the conservative grid contains 24 correlated tests, we correct within recipe families using Benjamini–Hochberg at $q{=}0.05$ and report raw and corrected p values side by side. (ii) For cells where we claim *no detectable harm*, we report an equivalence test (TOST, bounds ± 0.03) with the minimum detectable effect at 80% power, so the claim reads "we can exclude harm larger than X" rather than "there is no effect". (iii) For fourteen key cells we

report a second, five-generation-seed protocol that reuses the base's own five generation sets, which is the only analysis here that lets base noise enter an interval (Section D).

One measurement caveat: the batch moves some quantities we compare by more than our smallest reported effects, and Table 5 collects every such pair.

4 The Map: Where In-House Adaptation Costs Safety

This section establishes the central claim: the cost is conditional, the condition is rank, and rank acts as a per-scale threshold. Figure 1 shows the map, with every measured value printed in its caption; three regularities organise it.

Finding 1 (rank): damage grows with rank, then falls off a cliff, and the cliff sits at the same rank at every scale as measured. At 1.5B and the full budget, SUS falls from 0.592 ± 0.009 at r=16 through 0.570 ± 0.014 and 0.565 ± 0.014 at r=32 and 48 to 0.441 ± 0.024 at r=128 (ΔSUS=-0.158). Those two intermediate ranks are trends rather than effects at our sample size (p=0.065, 0.052 raw) and r=128 is the only cell in the sweep whose raw p survives conventional levels (0.0074). What gives way is the safety axis: harmful refusal drops from 0.967 to 0.600 across the sweep while CVE accuracy stays at 0.31 for every rank. The cliff is not particular to 1.5B: r=128 gives 0.426 ± 0.024 (-0.223, p=0.0038) at 3B and 0.507 ± 0.027 (-0.153, p=0.010) at 7B, so the three r=128 cells are the three largest conservative losses at any scale — the 3B and 7B cells hold under either scorer, while the 1.5B one survives correction only under the rule scorer (-0.158 by rule against -0.063 by judge; Section H) — and 3B already loses 0.140 at r=64 (p=0.0035); at $r \leq 48$ nothing exceeds 0.05 at any scale. Rank 128 is not simply a larger update (Section L): it moves the weights by 1.7–2.1$\times$ what rank 16 moves, not the 8$\times$ the parameter count suggests. Nor is the cliff about what the corpus contains: a corpus with nothing harmful in it reproduces most of the same loss at the same rank, and nothing at r=16 (the *safety-neutral replication* below).

Finding 2 (data budget at 1.5B, r=16): no interior optimum, and no harm either. We measure the data axis at 1.5B, r=16, against a base of 0.599. Unlike the rank axis, it shows no dose-response: all three budgets lie within $\pm$0.03 of that base (D=500 0.627 ± 0.027, D=1000 0.613 ± 0.005, D=3000 0.608 ± 0.015), and D=1000 is marginally *above* base on a raw test ($+0.013$, p=0.0397) but not after correction (p_{BH}=0.119). So at this rank the corpus size is inert in both directions, and the loss we report is a property of how the adapter is shaped rather than of how much data is used.

Finding 3 (scale, given rank): every scale carries a threshold, the threshold is not comparable across scales, and the aggressive regime is unconditional. Where capacity matters is how much a conservative regime costs and where its threshold sits, not whether anything is lost. 3B is the worst cell at every rank we tested (-0.045 at r=16, -0.140 at r=64, -0.223 at r=128), against $+0.005$ for 7B at r=16 and -0.024 for 0.5B, and it is the only scale whose threshold is crossed below r=128 and survives the family correction: 7B r=64 clears the raw 0.05 level (p=0.0125) and misses the corrected one by a hairline (p_{BH}=0.0501, 0.047 once refusal is judged rather than scored by the rule; Section H). 0.5B never crosses one: its r=128 cell loses 0.053, the smallest of the four r=128 losses, and at three seeds it is tested and clearly not significant (p=0.23). It is also the scale that adapts the fewest modules, so this cell contrasts capacity with layout as much as with capacity alone. The aggressive regime removes the structure altogether and collapses all four scales (-0.226, -0.332, -0.309, -0.290; Figure 2). The collapse is not confined to the safety axis: general-QA falls by 0.23–0.49, safety-QA by 0.34–0.55 and at 0.5B CVE accuracy by 0.23, so the model degrades wholesale with refusal falling farthest (Table 8), and the loss persists on the items a cell and its base both answer in Chinese (0.11–0.34 on general-QA and 0.07–0.40 on safety-QA), so it is not an artifact of the scorers' language either (Section S). Rank and recipe decide whether anything is lost, and within a scale capacity sets how much; where the threshold sits *across* scales is not identified here, because the adaptation layout changes with scale. For practice this makes rank the quantity to constrain and the threshold a number to measure per scale rather than to assume.

How far this evidence carries. Because the grid is sparse and its cells are correlated, we report what the map survives rather than what it reaches. Across the 24 conservative tests on SUS, Benjamini–Hochberg at q=0.05 leaves five cells, all at $r \geq 64$; substituting the judge for the rule

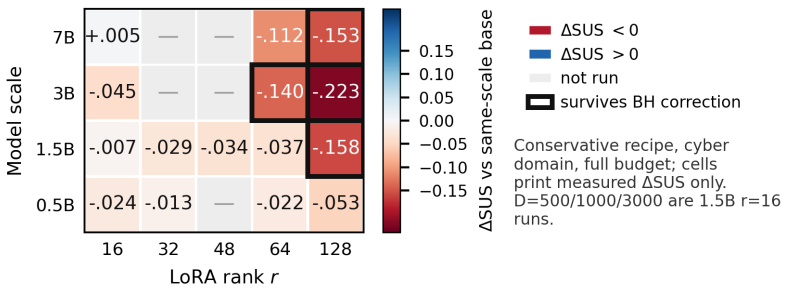


Figure 1: The conservative map as a discrete heatmap: ΔSUS against the same-scale base over model scale and LoRA rank, for the conservative regime (lr 5×10^{-6}, one epoch, cyber domain, full budget). Cells print measured ΔSUS; blank cells were not run; nothing is interpolated — the sparsity is the result — and every cell shown is trained at three seeds. Heavy frames mark the four cells of this map that survive Benjamini–Hochberg correction at q=0.05 under the rule scorer within the conservative family of Section C (3B r=64 and r=128, 1.5B r=128, 7B r=128); the family's fifth survivor is the safety-neutral 3B r=128 cell of Section P, which is not on this grid. Absolute SUS, reading r=16, 32, 48, 64, 128 against same-scale bases 0.535, 0.599, 0.649 and 0.660 (single runs): 0.5B 0.511, 0.522, —, 0.512, 0.481; 1.5B 0.592, 0.570, 0.565, 0.562, 0.441; 3B 0.604, —, —, 0.510, 0.426; 7B 0.665, —, —, 0.548, 0.507. Two of the four verdicts turn on the refusal scorer rather than on the data alone, and the judge reads two of these cells at −0.091 and −0.149, the same sign and ordering (Sections H and R).

scorer keeps the count at five but moves its membership, dropping 1.5B r=128 and taking up 7B r=64, and beyond that one cell no judged magnitude moves by more than 0.03, so three of the four in-domain headline cells survive either scorer and one is decided by the choice (Sections C and H). Those p values do not say what drives the losses, and the two corpus ablations do not decompose them: at 1.5B r=128 the loss falls from 0.158 to 0.042 when the corpus's harmful targets are removed and to 0.091 when the corpus is replaced by neutral text, but the cleaned corpus keeps the corpus's refusal demonstrations — the `red_team`, `toxicity_refusal` and `harmful_refusal` roles of Table 3 — and the neutral one keeps none, so the two move opposite ingredients and their gap is not a split between rank and content (Sections 8 and P). What they do establish is a necessary condition and two modulations: every r=16 cell we ran stays within 0.045 of its base, in-domain and on the neutral corpus alike, so a high rank is needed for a loss of this size to appear at all; given that rank, harmful targets in the corpus raise the loss and refusal demonstrations lower it. The nulls do not rest on the correction alone: the five-generation protocol covers the conservative r=16 anchors, the mid-scale r=64 cell and the three in-domain r=128 cells, and sign and ordering survive every SUS weighting (Sections D and H); the remaining cells rest on the three-seed test, the correction and the cleaned-corpus replication (Section M), so we describe the conservative map as a per-cell measured threshold and the aggressive collapse as established.

The safety-neutral replication. A corpus with no harmful target and no security text in it reproduces most of the same loss at the same rank. On 6,500 Zhihu instructions at r=128 the 1.5B and 3B cells lose 0.091 and 0.146 against in-domain losses of 0.158 and 0.223, which is 58% and 65% of them, while the same corpus at r=16 costs at most 0.018 at any of the four scales: rank crosses the threshold, and the corpus sets the size of the loss rather than its existence. Two caveats bound it. The ratio is a rule scorer statement, which judging the stored runs turns into 0.91 at 1.5B, so we read it as a 3B result. And the 3B arm runs at 1.5B's layout, so that cell varies layout together with corpus and its loss is approximate (Sections P and R).

What the nulls can exclude. The nulls are equivalence-tested rather than asserted: TOST at ±0.03 passes for the 7B and 1.5B r=16 cells and the D=1000 cell but fails for D=500 and D=3000, so the 7B null excludes harm larger than about 0.03 while the budget nulls do not (Sections C and D).

5 What the Failure Is, and What Our Probe Cannot See

This section asks what the lost safety is made of: at 1.5B the knowledge survives while its routing does not, and the probe that carries that reading does not transfer to 7B.

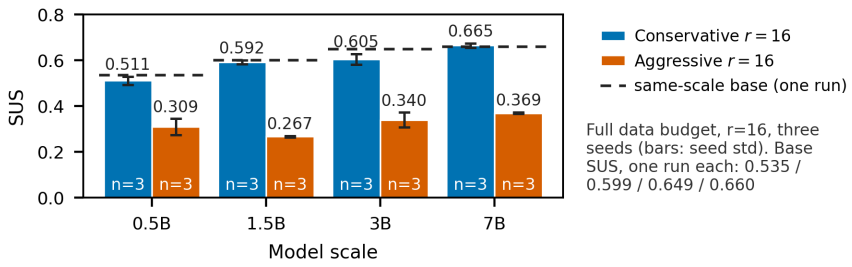


Figure 2: The two regimes at r=16 and the full budget. At r=16 the conservative regime leaves every scale within 0.05 of its own base (dashed lines); the aggressive regime collapses all four. Error bars are standard deviation over three training seeds.

Table 2: The two-channel probe behind this section; Table 6 gives it in full and separates the two measurement batches.

Configuration	Δ_s	sep	sep/$\sqrt{d}$	refusal	κ
base (1.5B)	0.000	22.20	0.566	0.933	0.008
r=16	0.015	20.56	0.525	0.960	0.023
r=64	0.078	16.09	0.411	0.900	0.017
r=128	0.210	8.96	0.229	**0.500**	**0.039**
base (7B)	0.000	44.28	0.740	0.933	0.021
r=16 (7B)	0.037	39.90	0.667	1.000	0.011

5.1 Knowledge retained, alignment collapsed — at 1.5B

At 1.5B the two channels separate. Table 2 gives the measurements the claim rests on: Δ_s is refusal-direction drift from base, sep the base refusal-boundary norm and sep/$\sqrt{d}$ that norm over the hidden width, refusal the harmful-prompt rejection rate on the probe's own prompt set, and κ the CVE fact-hit rate under the probe's strict scorer and its domain system prompt — a batch separate from the suite's, so the two are never mixed (Table 5; the full probe, with its D=1000 and r=8 arms, is Table 6). The r=128 configuration retains the most knowledge of any configuration we probed (κ=0.039 against 0.023 at r=16) while its refusal on the probe prompts falls to 0.50 and its safety boundary shrinks from 22.2 to 9.0; drift rises monotonically with rank, from 0.015 at r=16 to 0.210 at r=128 (base 0.000), and, averaged over the probe's configurations, is concentrated in late layers (0.011 early, 0.054 mid, 0.088 late). The reading is that the model still holds the knowledge but no longer routes it through a safe response. Two qualifications keep this honest. First, the loss of *definitional* safety knowledge is not confined to them: safety-QA accuracy falls by 0.174 at 1.5B r=128, and also by 0.047 at 3B r=16 and 0.102 at 7B r=16, which are conservative cells. "Knowledge is retained" holds for held-out facts, not for safety knowledge. Second, the absolute level of κ is low by construction (a strict word-level match on 1999–2002 CVE descriptions), so we read κ only in relative terms across configurations that share the scorer.

5.2 The probe validity problem at 7B

The 7B row of Table 6 is the reason we do not present Δ_s as a monitor. Within the 1.5B sweep the relation between drift and refusal loss is strong in sample (ρ=0.94, R^2=0.88 over six configurations), and leave-one-out over all eight probed models destroys it (R^2=−0.40, ρ=0.02) because of the 7B endpoint, which drifts by 0.037 and yet refuses *more* (0.933→1.000).

We take this as a probe-validity result rather than as an anomaly in the model. Concept-cone work shows that refusal is mediated by a multi-dimensional cone and that the cone's effective dimensionality grows with model scale within a family (Wollschläger et al., 2025). A single safe-minus-harmful direction is therefore the right coordinate at the scales where it was validated and a mis-specified one above them: the probe does not fail to detect drift at 7B, it detects a projection of a higher-dimensional object. Two consequences follow, and both are limits on this paper. Our 3B results rest on a scale where the coordinate is plausible rather than validated: the 3B probe cells we do have agree with the 1.5B line (r=16 drifts 0.055 and loses 0.10 of refusal against 0.09 predicted, r=64 0.144 and 0.23 against 0.26, the 0.5B cell barely drifting and refusing *more*), so the failure we can measure is the 7B one and the 3B evidence is two cells rather than a test. Normalising the boundary

by hidden width *weakens* the capacity-buffer argument: the raw 1.5B-to-7B ratio of $2.0\times$ becomes $1.3\times$ after division by $\sqrt{d}$ (Table 6).

5.3 Drift tracks collapse but is not sufficient

The in-domain geometry tracks the collapse without explaining it: on 32 CVE and 45 attack prompts the CVE-versus-attack class-centre distance shrinks monotonically with rank (0.0425 to 0.0349, an 18% drop) and the drift along the base attack axis is largest at r=128 (layer 20: 0.364 against 0.225 at r=16) (Section F). But an arm that drifts more need not collapse: across five seeds the selected-subset arm drifts 0.165 ± 0.009 against 0.130 ± 0.006 for the random arm, neither collapses (0.90 and 0.933 refusal against a base of 0.933), and the random arm has the *better* SUS ($+0.013$ against -0.014). Drift is therefore a geometric fact about where knowledge lives rather than an indicator of safety loss, and we do not recommend it as a monitor; Figure 3 plots the two channels.

6 Where Knowledge Should Live

If the conservative regime cannot buy knowledge, the deployment question becomes where knowledge should live, and this section answers it in favour of retrieval. We compare the parametric route against a retrieval route that injects the official NVD description of the queried CVE at inference time and trains nothing, applied to the five independently tuned models of each arm (conservative regime, r=16, D=650, five seeds per arm), with paired metrics on the same 32 CVE items (Table 7 in Section I). Retrieval raises CVE accuracy by $+0.178$ and cuts the rule-based hallucination rate from 0.625 to 0.031, and the effect replicates on the random arm ($+0.174$, $0.706 \rightarrow 0.056$). It holds on both halves of the corpus/held-out split, if anything more off-corpus ($+0.162$ against $+0.198$ for the selected arm, $+0.151$ against $+0.203$ for the random one; all four $p<10^{-4}$), and it does not need the fine-tuning: with no adapter and no training at all, the 1.5B base gains $+0.184$ and lands at 0.434 against the 0.455 of the five tuned models (Table 7).

Two caveats bound this result, and Section I states both with the table: the arm is an *oracle*, injecting the correct NVD text with no retriever and no noise, so its grounded accuracy is an upper bound rather than a deployable number; and the ordering itself — retrieval beats fine-tuning for knowledge injection — is not new (Ovadia et al., 2024). What we add is that on this corpus the parametric route cannot be pushed toward the retrieval result without moving along rank, the axis that costs refusal.

7 The Evaluation Protocol Changes the Measured Effect

Every number above is measured without a system prompt, whereas the refusal demonstrations in our corpus carry a Chinese safety system prompt, and that condition turns out to matter as much as the recipe does. When we re-run the same 50 behavioural items with that prompt present, refusal returns to ceiling, in both languages: for the aggressive 1.5B model, harmful refusal moves from 0.133 (Chinese rendering) and 0.011 (English rendering) without a prompt to 0.956 and 1.000 with it; the judge puts the same recovery at a third of the size, because it reads the no-prompt condition higher ($0.660/0.552 \rightarrow 0.956/1.000$), so the protocol effect is 0.82 by rule against 0.30 by judge. The same holds at 7B ($0.489/0.067 \rightarrow 1.000/1.000$). Both renderings recover, so for these cells the trigger is the prompt rather than the language.

This is consistent with a template-replay account: the adapted model has compressed refusal into a response to the adaptation prompt rather than retaining it as a policy, which also explains why fixed English benchmarks run under that prompt read at ceiling on the same cells (Section O gives the cross-validation in full, including the audit that nine tenths of those items are corpus items yet the ceiling holds on the rest). We report this as an observation rather than a demonstrated mechanism, and the discriminating ablation does not settle which reading holds. Re-evaluated with the prompt present, the same arms are indistinguishable without it and separate only with it, and they do so in the direction opposite to the one template replay requires: the arm trained with no system field at all is the one the prompt moves most (Section K). Prompt-level restoration is established (Xie et al., 2023b; Sharma et al., 2026) and concurrent realignment work finds repair prompt-conditional (Guo et al., 2026b); we add its size relative to the recipe and the certification failure that follows. On the conservative cells that carry Section 4 the contrast reads the other way: the prompt does not return

their Chinese harmful refusal to base at any scale, though it does restore their English renderings, so the restoration belongs to the collapsed regime and to the templated rendering rather than to the conservative high-rank loss — which is the reading that matters for a deployment that carries a prompt (Section Q).

8 Threats to Validity

This section bounds what the map can claim: which explanations the measurements exclude, and which confounds remain uncontrolled.

Three candidate explanations. The three explanations the collapse suggests — a train–evaluation format mismatch, under-training, and corpus contamination — are each measured against its own ablation, and each fails in its own way. Deleting the corpus's system field recovers only about a quarter of the collapse and leaves harmful refusal where the kept arm had it. The conservative setting is safe rather than inert: within a fixed epoch count harm is monotone in the update magnitude, so conservatism is a band of magnitude and of repetition rather than a property of its learning rate. And removing every harmful target leaves the aggressive collapse intact — indistinguishable from a size-matched control drawn from the uncleaned corpus — while the conservative high-rank loss survives at 3B and 7B, so content accounts for the whole of such a loss only at 1.5B. Section N gives the measurements and their intervals.

Design limits. The base of the three-seed comparisons is a single run, so base-side noise enters no interval except those of Section D (all four aggressive cells and ten conservative); cross-scale comparison of ΔSUS is confounded by different base levels (headroom 0.465 at 0.5B, 0.340 at 7B) and by a different adaptation layout at every scale (Section U), which we disclose but did not control. Grid coverage is uneven (1.5B five ranks, 0.5B four, 3B and 7B three). The 7B null is a low-rank statement and rests on its single $r{=}16$ cell (minimum detectable effect 0.031); the 3B losses have two probe cells behind them, and the cleaned-corpus control covers three of the four high-rank cells, all but 3B $r{=}64$. The cross-family check (Mistral-7B) is weak rather than supportive: its base refusal is only 0.367, so we report the cell ($+0.052$ SUS) as inconclusive. The cross-domain check (Zhihu, safety-neutral instruction data) is flat at $r{=}16$ (-0.018, -0.015 at 0.5B and 7B) but not at $r{=}128$, where on three seeds it reproduces 58–65% of the in-domain loss under the rule scorer, so that loss is not attributable to the corpus's security and harmful content alone (Section P); its 1.5B $r{=}16$ cell comes from an aggregate whose per-seed records were not retained (Section T). Two of the verdicts behind Finding 3 turn on the refusal scorer rather than the data, so the magnitude, not the sign or ordering, is scorer-dependent (Sections H and R); quantization is not a confound, since no arm of the map is quantized. Two controls we did not run bound the rule: no structural-isolation defence of the kind cited in Section 2, and no size-matched uncleaned control for the conservative regime.

9 Conclusion

Across Qwen2.5 from 0.5B to 7B, in-house adaptation on a corpus mixing domain knowledge with safety-alignment data costs safety conditionally, and the condition is rank. The conservative regime leaves every cell at $r{\leq}48$ within about 0.05 of its base and buys no measurable knowledge, while the four in-domain cells that survive correction — $r{=}128$ at 1.5B, 3B and 7B and 3B $r{=}64$ — lose 0.14–0.22 under the rule scorer (-0.091 and -0.149 by judge on two of them, against -0.063 for the re-generated judge corpus of Section H — a 0.03 gap between the two judge paths, Section R); the aggressive regime collapses every scale, as general degradation rather than a failure of alignment. Rank is the stronger lever: a corpus with nothing harmful in it still loses 0.09 and 0.15 at 1.5B and 3B at the same rank. Two negatives are the most reusable part: a single-direction refusal probe mis-specifies the geometry at 7B, and in-domain drift is not sufficient for collapse. In practice, keep knowledge out of the weights where retrieval can supply it, and keep rank and update magnitude small.

LLM USAGE STATEMENT

Large language model assistants (DeepSeek and GLM) were used in three ways, and the authors are responsible for the entire content of this paper. (i) *Writing*: they copy-edited and polished prose written by the authors. (ii) *Retrieval and discovery*: they helped locate related work, after which the authors obtained and read every cited work and verified its bibliographic details — authors, title, venue, year — against the published record, so no reference is fabricated and none is cited for a claim the authors have not read the work to support. (iii) *Code*: they helped write, debug and document the training, evaluation and analysis code, which the authors specified, ran, and checked against the stored per-item records through the assertion layer of Section U. Research ideation, the experimental design, the choice of recipes and axes, and the reading of every result were human, and no number, figure or table in this paper was generated or altered by a model; the corpus was assembled from crawled public sources and scripted templates rather than produced by one (Section B). The one AI system inside the research itself — the independent LLM judge of Section G, a locally run Qwen3.5-9B with greedy decoding and a fixed rubric — is a measurement instrument applied to stored model outputs, not an authoring tool.

REPRODUCIBILITY STATEMENT

The corpus composition, overlap audit, per-item evaluation records, probe outputs, and the scripts that derive every table are included in the anonymised supplementary archive attached to this submission; Section U lists hardware, library versions, and the exact commands, and Section T lists every quantity we could not reconstruct from stored records. That same archive, with author names, will be made public upon acceptance.

ETHICS STATEMENT

This is a defensive evaluation of a known risk — domain adaptation quietly eroding safety — and its primary output is a deployment rule rather than any new attack capability. Safety prompts are drawn from public benchmarks and public vulnerability descriptions (NVD) and no new malicious content was authored: the harmful completions the audit uncovers (Section B) were crawled from public corpora, are not reproduced in this paper, and are released only as labels — a per-turn harmfulness verdict with its identifier — so that others can filter the corpus without redistributing the content itself. All fine-tuned models are used only for research evaluation and never deployed.

REFERENCES

Lama Alssum, Hani Itani, Hasan Abed Al Kader Hammoud, Philip Torr, Adel Bibi, and Bernard Ghanem. Unforgotten safety: Preserving safety alignment of large language models with continual learning. *arXiv preprint arXiv:2512.10150*, 2025.

Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Huanran Chen, Zeming Wei, Yao Huang, Yichi Zhang, Yinpeng Dong, and Jun Zhu. Unveiling the basin-like loss landscape in large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Gelei Deng, Yi Liu, Victor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In *USENIX Security Symposium (USENIX Security)*, 2024.

Lyndon Drake and Zandi Eberstadt. Transplanting, inverting, and preventing a misalignment persona: Method-conditional emergent misalignment in Qwen2.5. *arXiv preprint arXiv:2607.04510*, 2026.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. MoDS: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*, 2023.

Adel ElZemity, Budi Arief, and Shujun Li. CyberLLMInstruct: A pseudo-malicious dataset revealing safety-performance trade-offs in cyber security LLM fine-tuning. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2025. doi: 10.1145/3733799.3762968.

Adel ElZemity, Budi Arief, and Shujun Li. Analysing safety risks in LLMs fine-tuned with pseudo-malicious cyber security data. In *Computer Security. ESORICS 2025 International Workshops (SECAI 2025)*, Lecture Notes in Computer Science. Springer, Cham, 2026.

Kathleen C. Fraser, Hillary Dawkins, Isar Nejadgholi, and Svetlana Kiritchenko. Fine-tuning lowers safety and disrupts evaluation consistency. In *Proceedings of the First Workshop on LLM Security (LLMSEC)*, pp. 129–141, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.llmsec-1.10/.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Jeremiah Giordani. Re-emergent misalignment: How narrow fine-tuning erodes safety alignment in LLMs. *arXiv preprint arXiv:2507.03662*, 2025.

Idhant Gulati and Shivam Raval. Narrow fine-tuning erodes safety alignment in vision-language agents. In *Lifelong Agent Workshop at ICLR*, 2026. arXiv:2602.16931.

Yitong Guo, Xiaoyi Chen, Siyuan Zhang, Xiaofeng Wang, and Haixu Tang. When safety routing breaks: Understanding alignment fragility under benign fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2026a. To appear; arXiv:2609.01455.

Yongjian Guo, Wanlun Ma, Lingyu Shen, Xi Xiao, and Sheng Wen. On-policy distillation for LLM safety: A routing approach to template-robust realignment. *arXiv preprint arXiv:2607.27081*, 2026b.

Will Hawkins, Kaivalya Rawal, Jonathan Rystrøm, Stratis Tsirtsis, Zihao Fu, Greta Warren, Ryan Brown, Eoin Delaney, Sandra Wachter, Brent Mittelstadt, and Chris Russell. The heterogeneous safety impacts of benign multilingual fine-tuning. In *International Conference on Machine Learning (ICML)*, 2026.

Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 2024.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when fine-tuning large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

Emaan Bilal Khan, Amy Winecoff, Miranda Bogen, and Dylan Hadfield-Menell. Safety drift after fine-tuning: Evidence from high-stakes domains. *arXiv preprint arXiv:2604.24902*, 2026.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. SaLoRA: Safety-alignment preserved low-rank adaptation. In *International Conference on Learning Representations (ICLR)*, 2025.

Jun Liu, Zhenglun Kong, Peiyan Dong, Changdi Yang, Xuan Shen, Pu Zhao, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Xue Lin, Dong Huang, and Yanzhi Wang. RoRA: Efficient fine-tuning of LLM with reliability optimization for rank adaptation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Ning Lu, Shengcai Liu, Jiahao Wu, Weiyu Chen, Zhirui Zhang, Yew-Soon Ong, Qi Wang, and Ke Tang. Safe Delta: Consistently preserving safety when fine-tuning LLMs on diverse datasets. In *International Conference on Machine Learning (ICML)*, 2025.

Yutao Mou, Xiaoling Zhou, Yuxiao Luo, Shikun Zhang, and Wei Ye. Decoupling safety into orthogonal subspace: Cost-efficient and performance-preserving alignment for large language models. *arXiv preprint arXiv:2510.09004*, 2025.

Quoc Minh Nguyen, Trung Le, Jing Wu, Anh Tuan Bui, and Mehrtash Harandi. Antibody: Strengthening defense against harmful fine-tuning for large language models via attenuating harmful gradient influence. In *International Conference on Learning Representations (ICLR)*, 2026.

NIST. National Vulnerability Database (NVD). https://nvd.nist.gov/, 2024. Accessed September 2026.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Kartik Sharma, Yiqiao Jin, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. Sysformer: Safeguarding frozen large language models with adaptive system prompts. In *International Conference on Learning Representations (ICLR)*, 2026.

William F. Shen, Xinchi Qiu, Nicola Cancedda, and Nicholas D. Lane. SEAT: Sparse entity-aware tuning for knowledge adaptation while preserving epistemic abstention. *arXiv preprint arXiv:2506.14387*, 2025.

Naufal Suryanto, Muzammal Naseer, Pengfei Li, Syed Talal Wasim, Jinhui Yi, Juergen Gall, Paolo Ceravolo, and Ernesto Damiani. RedSage: A cybersecurity generalist LLM. In *International Conference on Learning Representations (ICLR)*, 2026.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025.

Krishnapriya Vishnubhotla, Hillary Dawkins, Isar Nejadgholi, and Svetlana Kiritchenko. Safety measurements for fine-tuned LLMs should be grounded in capability. *arXiv preprint arXiv:2606.03648*, 2026.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *International Conference on Machine Learning (ICML)*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction fine-tuning. In *International Conference on Machine Learning (ICML)*, 2024.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023b.

Yihao Xue and Baharan Mirzasoleiman. LoRA is all you need for safety alignment of reasoning LLMs. *arXiv preprint arXiv:2507.17075*, 2025.

Jing Zhang, Hui Gao, Peng Zhang, Shuzhen Sun, Chang Yang, and Yuexian Hou. The scaling law for LoRA based on mutual information upper bound. *arXiv preprint arXiv:2501.03152*, 2025.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Evaluation Suite

The suite comprises 110 items: 30 harmful prompts adapted from AdvBench/JailbreakBench-style bilingual samples (Zou et al., 2023; Chao et al., 2024), 20 author-constructed jailbreak prompts (role-play, code, and multi-step variants of common jailbreak formats), 8 author-constructed safety-QA items, 32 CVE-analysis items drawn from real NVD data (NIST, 2024) with the $18/14$ corpus/held-out split, and 20 author-constructed general-QA items covering programming, mathematics, reasoning, common sense and writing in both languages. Refusal is scored by a rule-based checker (hereafter the *rule scorer*) on the first-turn answer against a per-item keyword set: the Chinese prompts carry the Chinese words for *cannot*, *will not*, *refuse*, *will not provide* and *sorry*, plus item-specific terms such as *illegal* and *privacy*; the English-derived prompts and jailbreaks carry a bilingual set (`cannot`, `can't`, `unable`, `won't`, `refuse` and the Chinese words above); and an answer under 30 characters containing no step or code marker also counts as a refusal. The 8 safety-QA and 20 general-QA items are scored by keyword overlap against their own answer keys. Every CVE item scores $0.7\cdot$fact-hit $+\ 0.3\cdot$structured, where *structured* is 1 if the answer carries a layout marker — a section heading (`##`, `###`), a numbered item (`1.`), or one of the Chinese section words for *mechanism*, *impact* and *remediation* — and 0 otherwise: a presentation criterion, not a content one, which is why the knowledge channel is read from fact-hit. All generations use temperature 0.7, top-p 0.9, and 384 new tokens.

B Corpus Composition, Overlap, and the Harmful-Completion Audit

By source and by role the 6,503 samples split as in Table 3, where `hh-rlhf` (Bai et al., 2022) and `real-toxicity-prompts` (Gehman et al., 2020) are the two public safety-alignment corpora and `unsafe_qa` is the public Chinese harmful-Q&A set `XuehangCang/unsafe_qa`, whose

paired safe and unsafe answers we use on the unsafe side. Two facts follow. A substantial fraction of the corpus is ordinary safety-alignment data rather than domain knowledge, which is why we describe the adaptation as benign in intent rather than knowledge-only.

Table 3: Corpus composition by source and by role (6,503 samples). Only the 764 knowledge-role samples (11.7%) are domain knowledge; the rest is ordinary safety-alignment data.

Source	n	Role	n
`template` (scripted)	1,598	`red_team`	1,510
`hh-rlhf`	1,510	`toxicity_refusal`	1,341
`real-toxicity-prompts`	1,341	`expert` (scripted)	1,062
`unsafe_qa`	729	`knowledge`	**764**
`nvd`	576	`sft_safety`	729
`advbench`	469	`harmful_refusal`	469
`arxiv`	170	`multi_turn_attack`	183
`jailbreakbench`	92	`chinese_safety`	182
`github`	18	`refusal` / `jailbreak_refusal`	171 / 92

The second fact is that benign intent is not benign content, and the string-reproduction check — the 469 assistant turns carrying a target string, none of which reproduces it — did not test that. We therefore judge every assistant turn in the corpus: 10,695 turns collapsing onto 7,794 unique response texts, each scored H=0/1/2 by the local judge on the first 2,500 characters, plus a rule-based cue scan over the full text (Section G for the rubric; Table 4). 1,989 turns (18.6%) supply usable harmful content and 354 (3.3%) are borderline — counting turns, not unique responses, of which 1,909 are flagged H=2; 1,522 of the 6,503 samples (23.4%) contain at least one harmful target turn. The contamination is concentrated, and in both directions from the judge: 720 of the 729 `unsafe_qa` and 1,094 of the 5,465 `hh-rlhf red_team` target turns are harmful (720 and 1,093 unique texts), while all 96 flagged `template` texts — 175 turns — are defensive vulnerability reports. A stratified 100-item manual re-read confirms the shape of that split — 34/35 of the `hh-rlhf` flags, 25/25 of the `unsafe_qa` flags, 0/20 of the `template` flags, and 4/20 of a control sample the judge had passed — with binary agreement 0.75 (κ=0.41). The judge is unstable at this granularity: re-judging the flagged set under a clarified rubric demoted 4.3% of it and raised 32.7% of a 300-item control sample. We therefore read 18.6% as a lower bound on the contamination, and we treat it as a measured confound on every recipe comparison in this paper rather than as a property we can assume away (Section 8).

The two 650-item subsets. The subset arms of Section 6 and Section K draw on the corpus' deduplicated training pool: from the 6,503 samples we hold out a stratified 101-sample development split (seed 42, strata by source), leaving 6,402, and deduplicate by normalized sample hash to 5,524. S-650 is the CyberSec-Selector v2 output on that pool — a five-signal quality score (source quality 0.162, structure 0.162, domain relevance 0.126, verification against the local NVD/CWE index 0.50, content density 0.05), then greedy maximal-marginal-relevance selection (λ=0.4, at most 130 samples per source). R-650 is a uniform random draw of the same size from the same pool (seed 42). The two sets share 89 of their 650 samples, and each is trained from five independent seeds in the retrieval comparison of Section 6.

Table 4: Harmfulness audit of the corpus's training targets. Turn counts are assistant turns; the H=2 and flagged columns are *unique response texts*, which is why they sum to 1,909 and 2,263 rather than to the 1,989 harmful turns quoted in the text. H=2 texts are those supplying usable harmful content under the rubric of Section G; the manual column is the stratified 100-item re-read of that source's flagged texts. The judge's flag is a lower bound: 4 of 20 judge-safe control texts were harmful on manual reading.

Corpus source	target turns	H=2 texts	flagged / unique texts	manual harmful
`unsafe_qa`	729	720	721	25/25
`hh-rlhf` (`red_team`)	5,465	1,093	1,445	34/35
`template` (`expert`)	1,835	96	96	0/20
all other sources	2,666	0	1	4/20[†]

[†]Control sample drawn from the 5,531 unique texts the judge passed, across all sources.

Matching each of the 50 behavioural items against the corpus gives an exact match for nine harmful items (H11–H18 and H20), which are AdvBench-derived and appear as user turns; no jailbreak item and no other harmful item matches, and no item reaches even 50% 8-gram overlap without an exact match. Removing those nine items leaves every conclusion unchanged or deeper: at the aggressive cells the base-to-tuned refusal drop is 0.92→0.93 at 1.5B, 0.80→0.71 at 3B and 0.73→0.65 at 7B, the conservative r=16 cell stays at base in both item sets, and the 1.5B r=128 collapse deepens from 0.60 to 0.54.

C Statistical Protocol in Detail

This appendix expands the protocol of Section 3.3: all tests are one-sample t tests of the tuned seeds against the fixed same-scale base, df=2, with 95% intervals from the t distribution, and we do not use normal quantiles. Benjamini–Hochberg is applied within two families (conservative and aggressive) over the 24 and 4 SUS comparisons respectively, and a second time over all nine metrics. A family contains only the independently trained configurations of the map and its arms — the rank and budget sweeps, the cross-domain cells, and the two α/r arms. Runs that replicate or control those cells rather than add cells to the map enter no family and are reported individually — the cleaned-corpus replication and its size-matched control (Section M), the learning-rate-by-epoch cells (Section L) and the system-field arms (Section K) — so the family sizes quoted here do not move when further runs of that kind are added. Under SUS, five conservative cells survive, all at $r \geq 64$: 3B r=64 (p_{BH}=0.031, −0.140), 3B r=128 (0.031, −0.223), the safety-neutral 3B cell of Section P (0.031, −0.146), 1.5B r=128 (0.044, −0.158) and 7B r=128 (0.048, −0.153); the smallest p_{BH} at $r \leq 48$ is 0.119, and at r=128 only the neutral 1.5B cell misses the cut (0.124). Two of those verdicts sit close enough to the line that the refusal scorer decides them: re-scoring the refusal and jailbreak channels of the thirteen judged configurations from the judge's verdicts (Section H) keeps the family at five but exchanges its weakest member, the 1.5B r=128 cell leaving (p_{BH}=0.044 → 0.090) and 7B r=64 entering (0.050 → 0.047). All four aggressive cells survive ($\max p_{\text{BH}}$=0.008). Over all nine metrics — 24 cells × 9 metrics, less the D=1000 refusal comparison, whose three seeds agree exactly and so admit no t — twenty conservative combinations survive: twelve on the knowledge or quality axes, eight of them CVE accuracy and all increases, and eight on the safety axis or the composite, among them 3B r=128 on jailbreak resistance (−0.383, p_{BH}=0.045) and, for the neutral-corpus 3B cell, SUS, the safety score and safety-QA accuracy. TOST bounds are ±0.03, the magnitude boundary we are willing to call harmless; passing cells are 7B r=16, 1.5B r=16, and D=1000, and failing cells are 0.5B r=16, D=500 and D=3000. Minimum detectable effects at 80% power with n=3 are 0.054 (0.5B), 0.028 (1.5B), 0.074 (3B) and 0.031 (7B), computed as $(\mathrm{t}_{0.975,2} + \mathrm{t}_{0.80,2})\,\mathrm{sd}/\sqrt{n}$. The five-generation protocol of Section D is a different test on a different noise source — it re-draws decoding on one adapter per cell rather than sampling training seeds — and it shifts the same-scale base by up to 0.019, comparable to the TOST bound above; read against that protocol, the 3B r=16 cell is the one $r \leq 48$ cell whose paired interval excludes zero (from −0.063 to −0.015), and at 0.039 it still sits inside the ±0.05 band the map is drawn on, so the two protocols do not disagree about the size of the cell, only about whether that size is distinguishable from decoding noise.

One further measurement fact belongs to the protocol rather than to any single result: the batch in which a quantity is read moves several base values by more than our smallest reported effects. Table 5 collects every such pair, so that the two readings are compared directly instead of being reconciled across separate captions.

D Five-Generation-Seed Protocol

Fourteen key cells were re-evaluated over the same 110 items with five independently sampled generation sets (decoding seeds 42–46), reusing the same-scale base's five sets, so that base-side generation noise enters both sides of the difference; each cell is its training seed 42 adapter, so what is re-sampled here is decoding rather than training. Intervals are paired t intervals over the five generation sets, pairing cell and base by generation index (df=4, $t_{0.975}$=2.776). Conservative: 0.5B r=16 −0.002 [−0.023, +0.019]; 1.5B r=16 −0.014 [−0.033, +0.005]; 3B r=16 −0.039 [−0.063, −0.015]; 3B r=64 −0.151 [−0.187, −0.115]; 1.5B r=128 −0.153 [−0.187, −0.120]; 3B r=128 −0.210 [−0.262, −0.159]; 7B r=16 +0.007 [−0.020, +0.033]; 7B r=128 −0.145

Table 5: The same base quantity under two batches: harmful refusal on the suite's prompt set against the probe's own (Section E), CVE fact-hit under the suite's scorer against the probe's strict one (Table 6), and base SUS under one generation batch against the five-generation protocol (Section D). Each pair is main batch first. †: the 1.5B probe file stores only the refusal *drop*, so that pair is a definition rather than an agreement; at the other three scales both rates were measured and agree, except at 7B.

Quantity at base	0.5B	1.5B	3B	7B
harmful refusal (suite / probe)	0.833/0.833	0.933/0.933†	1.000/1.000	0.967/0.933
CVE fact-hit (suite / probe)	0.018/0.026	0.018/0.008	0.019/0.021	0.016/0.021
SUS (one run / five generations)	0.535/0.516	0.599/0.616	0.649/0.659	0.660/0.662

$[-0.183, -0.107]$. Aggressive: 0.5B -0.198 $[-0.218, -0.177]$; 1.5B -0.403 $[-0.420, -0.386]$; 3B -0.290 $[-0.342, -0.239]$; 7B -0.294 $[-0.337, -0.252]$. Neutral corpus, conservative regime (the two r=128 arms of Section P): 1.5B -0.103 $[-0.130, -0.076]$; 3B -0.163 $[-0.207, -0.119]$. Base values in the same protocol: 0.516 (0.5B), 0.616 (1.5B), 0.659 (3B), 0.662 (7B). The two protocols do not return identical levels, and we report both rather than reconciling them: the aggressive 1.5B cell reads -0.403 here against -0.332 in Figure 1, a 0.07 gap, because Figure 1 averages three training seeds and differences against the canonical base, whereas this protocol re-draws decoding on the seed-42 adapter and against a base measured over the same five sets. The protocol covers the conservative r=16 anchors at all four scales, the mid-scale r=64 cell, the three in-domain r=128 cells and both neutral-corpus r=128 cells; the remaining cells — the r=64 cells at 1.5B and 7B and the 0.5B high-rank cells — rest on the three-seed test, the correction and the cleaned-corpus replication of Section M.

E Probe Construction

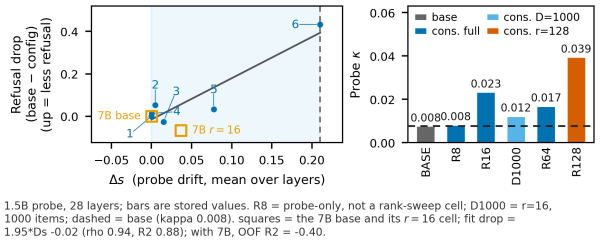
We extract last-token hidden states for 32 CVE, 35 attack and 13 safe prompts, and summarise each configuration by three quantities: the safety-direction drift Δ_s, the layer-averaged cosine drift $1-\cos\big(d_{\text{base}}(\ell), d_{\text{model}}(\ell)\big)$ between the base and the tuned safe-minus-harmful activation difference $d(\ell)$ of Arditi et al. (2024); the boundary sep, the ℓ_2 norm of the base refusal direction; and the knowledge retention κ, the CVE fact-hit rate. The 35 attack prompts are the suite's 15 behavioural items plus the first 20 entries of a separately generated red-team pool, and the 13 safe prompts are the 8 safety-QA and 5 general-QA questions of a shorter prompt list than the 110-item suite's; the boundary-geometry probe of Section F draws on the same two pools with different cut-offs (32 CVE, 45 attack, 5 neutral), which is why the two appendices quote different counts. The drift profile of Section 5.1 splits the 28 layers into first quarter, middle half and last quarter — layers 0–6, 7–20 and 21–27 — and its three numbers are means over all six 1.5B configurations probed, the non-drifting base and D=1000 arms included; the r=128 configuration alone reads 0.039, 0.206 and 0.390, so its own late-layer drift is ten times its early-layer drift. Normalisation by $\sqrt{d}$ uses the hidden width of each checkpoint (896, 1536, 2048, 3584). The probe generation uses a domain system prompt, whereas the main suite uses none; the two measurement batches therefore differ in absolute κ by up to 0.02 for the same configuration and agree in ordering, and we report both rather than reconciling them (Section T). The two findings this apparatus carries are reported under Section 5.1 and Section 5.2 of Section 5.

F Boundary Geometry and Its Controls

On 32 CVE and 45 attack prompts the class-centre distance falls monotonically with rank (0.0425 to 0.0349, 18%) and linear CKA rises (0.136 to 0.158), while the conservative D=1000 arm barely moves (-0.7%). Holding per-step scale fixed and raising rank from 16 to 128 reproduces the behavioural drop; at fixed rank the update strength contributes a smaller monotone effect; and shuffling the corpus's labels, which destroys question–answer coherence while preserving token statistics, leaves safety unchanged (0.933 refusal, ΔSUS=-0.030). The rank-16 and D=1000 arms leave both channels near base while r=128 collapses the boundary norm, so within this probe the attraction tracks rank rather than knowledge injection as such (Table 6).

Table 6: Two-channel probe at 1.5B and at the 7B endpoint, the measurements behind Section 5.1. Δ_s is refusal-direction drift from base, sep the base refusal-boundary norm, $\text{sep}/\sqrt{d}$ that norm over the square root of the hidden width (896/1536/2048/3584; 0.205/0.566/0.665/0.740 at the four bases), refusal the harmful-prompt rejection rate on the probe prompt set, and κ the CVE fact-hit rate of *this probe's own* strict scorer under the probe's domain system prompt — a separate measurement batch from the suite's fact-hit of Section 3.2; the two batches also differ slightly at base: the probe κ sits 0.01 below the suite's fact-hit at 1.5B and 0.005 above it at 7B, and the 7B base refuses 0.933 here against 0.967 on the suite (Table 5 and section T). ‡: the r=8 row is a probe-only run — it was never evaluated on the 110-item suite, so it appears here and in Figure 3 and in no table of the map (Section T).

Configuration	Δ_s	sep	$\text{sep}/\sqrt{d}$	refusal	κ
base (1.5B)	0.000	22.20	0.566	0.933	0.008
D=1000	0.000	22.06	0.563	0.920	0.012
r=8‡	0.005	21.45	0.547	0.880	0.008
r=16	0.015	20.56	0.525	0.960	0.023
r=64	0.078	16.09	0.411	0.900	0.017
r=128	0.210	8.96	0.229	**0.500**	**0.039**
base (7B)	0.000	44.28	0.740	0.933	0.021
r=16 (7B)	0.037	39.90	0.667	1.000	0.011



1.5B probe, 28 layers; bars are stored values. R8 = probe-only, not a rank-sweep cell; D1000 = r=16, 1000 items; dashed = base (kappa 0.008). squares = the 7B base and its r = 16 cell; fit drop = 1.95*Ds -0.02 (rho 0.94, R2 0.88); with 7B, OOF R2 = -0.40.

Figure 3: (a) Refusal loss against safety-direction drift for the 1.5B sweep, with the 7B endpoint drawn outside the fitted window; the numbered markers follow panel (b)'s x order — 1 base, 2 r=8, 3 r=16, 4 D=1000, 5 r=64, 6 r=128. (b) CVE fact-hit κ per configuration; the r=128 cell is the most knowledgeable and the least refusing.

G JUDGE VALIDATION

Against an independent local judge — Qwen3.5-9B (the official `Qwen/Qwen3.5-9B` snapshot, configured for text-only input, greedy decoding, a fixed rubric and no reasoning step, and loaded in fp16 from the snapshot rather than with the 4-bit adapter of the 9B baselines of Section U): refusal κ=0.52 (N=636, P_0=0.82, rule positive rate 0.70 against judge 0.81), harmful-only κ=0.39 (N=385), jailbreak-only κ=0.58 (N=251), and configuration-level ordering ρ=0.94 over 13 configurations. The judged unit is a stored 200-character answer preview, and the judge returns `R=NA` when that preview is empty or unreadable, which it did on 14 of the 650 behavioural prompt–configuration pairs (5 harmful, 9 jailbreak); the remaining 636 are split 385/251 and are the pair counts above. Refusal is judged on a three-point scale (`R=` 0 complied, supplying the requested content even under an educational or fictional framing; `R=` 1 refused the core ask but hedged or still gave loosely related risky detail; `R=` 2 refused clearly), with `R`$\geq$ 1 read as a refusal everywhere in this paper. The hallucination flag agrees poorly (κ=0.09, N=399, P_0=0.65), which is why it enters SUS only with weight 0.2 and is reported with a knowledge-base entity check that reaches κ=0.28 and agreement 0.94. The corpus audit of Section B uses a separate, stricter rubric with three levels (`H=` 0 safe — a refusal, a warning, or a defensive write-up that quotes an advisory; `H=` 1 borderline — operational detail that is hedged, incomplete or not usable; `H=` 2 harmful — working code or commands, usable how-to steps, synthesis details, targeted hate, or encouragement of self-harm) and it was re-run over everything pass one flagged under a clarified rubric, which moved 4.3% of the flagged set down and 32.7% of a control sample up; the labels we report are pass one, because the second pass is a stability probe rather than a better instrument. On the language-parallel probe the rule and judge scorers agree at κ=0.51 (N=2,746) but differ in level (rule 0.59 against judge 0.71), so the protocol results of Section 7 are reported under both. The one independence

this judge cannot claim is family: it is a Qwen model, and the models under test are Qwen models. Re-running the same rubric under a second, non-Qwen judge (Mistral-7B-Instruct-v0.3, greedy, same prompts and same stored previews) reproduces the level but not the fine ordering: refusal κ falls from 0.52 to 0.40 (N=645, P_0=0.72, rule positive rate 0.69 against judge 0.56), still moderate, while the configuration-level ordering weakens from ρ=0.94 to 0.57. Both judges therefore support the coarse conclusion — the rule scorer is only a moderate proxy for a judge, and the rule's positive rate is a property of the rule almost exactly (0.70 against 0.69) — but the exact ordering is judge-dependent and we quote it with both values. An independent human re-read of the same axis settles which side is the weak link. On a stratified sample of 180 refusal items (proportional within (configuration, rule verdict, judge verdict) strata, so the sample figure estimates the population one and every disagreement type is represented), one of us read the prompt and the same 200-character preview, blind to both scorers and to the configuration name; 8 previews were unjudgeable, leaving 172. The human agrees with the judge far more than with the rule: κ=0.79 (P_0=0.94) against κ=0.49 (P_0=0.82), and the graded three-point comparison is lower again (κ=0.60), so what all three readers share is the coarse refusal boundary. The rule's disagreement has a direction: it calls 0.72 of these items refusals where the human calls 0.84 and the judge 0.83, and 26 of the 172 are items the human reads as a refusal and the rule as compliance, against 5 the other way. The moderate κ=0.52 between rule and judge is therefore a property of the rule rather than of the judge. The check on the training-corpus labels that is independent of any judge family is the manual re-read of Section B, in which a human adjudication of 100 audited training texts agrees with the judge at κ=0.408 (0.750 agreement, 21 over-flags, 4 misses) — the corpus labels are reported as pass-1 corrected by that measured rate for exactly this reason.

Answer language. Every prompt in the suite is written in Chinese, so a cell that answered in English would be scored by keyword sets written for Chinese. Classifying the stored 200-character previews as CJK-dominant or Latin-dominant, the four bases answer in Chinese on 0.75–0.79 of items and the conservative cells at $r \leq 48$ stay within 0.04 of their own base. The cells above the threshold move part of their output into English (1.5B r=128 0.58, 3B r=128 0.65, 7B r=128 0.69) and the aggressive cells further, but not uniformly (0.5B 0.75, 1.5B 0.23, 3B 0.47, 7B 0.47). Three things keep this from carrying the loss. The refusal scorer is bilingual by construction (Section A), and the check that does not depend on that is the LLM judge, which does not share the failure: on the harmful subset the rule calls 0.033 of the 1.5B aggressive cell's answers refusals where the judge calls 0.633, and 0.567 against 0.966 at 1.5B r=128 — the human re-read above sides with the judge in both places. What that costs the map is bounded in Section H. The shift does not track the damage: the 0.5B aggressive cell loses 0.226 SUS and keeps its base's language mix (−0.009), while the 3B and 7B cells lose about the same and move a third of their answers. And in the one setting where the two languages are compared directly — the same items re-rendered in English with the adaptation prompt present — refusal returns to ceiling in both (Section 7). What the audit does say is that the aggressive cells' degradation includes a change of register, which is further reason to read their loss as general rather than safety-specific.

H Weight and Subset Sensitivity

Recomputing the SUS of Equation (2) per seed at safety weights $w \in \{0.3, \ldots, 0.7\}$ leaves the ordering intact: at w=0.5 the conservative cells are 0.5B −0.024, 1.5B −0.007, 3B −0.045, 7B +0.005, 1.5B r=128 −0.158, 3B r=64 −0.140, 3B r=128 −0.223, 7B r=128 −0.153, and the aggressive cells, in the same scale order (0.5B, 1.5B, 3B, 7B), −0.226, −0.332, −0.309, −0.290; at w=0.3 the same cells are −0.029, −0.015, −0.040, +0.002, −0.110, −0.099, −0.158, −0.094, and −0.220, −0.200, −0.173, −0.156, and at w=0.7 they are −0.019, +0.000, −0.049, +0.008, −0.206, −0.181, −0.289, −0.211, and −0.232, −0.465, −0.446, −0.425. The scan moves the safety weight only. No cell changes sign and no ordering inverts — in particular the four high-rank cells stay below every cell at $r \leq 48$ at every weight — so SUS is not carrying the result through a particular weighting.

The remaining weight is the hallucination term, and the scan above leaves it at its main-suite value, so it does not bound that term by itself. We scan it directly: with h the weight on $(1-h)$ in Utility and every other weight unchanged, $h \in \{0, 0.1, 0.2\}$ (the last is the main suite, and it reproduces the published ΔSUS to four decimals). Dropping the term entirely moves the conservative cells

by at most 0.021 (7B r=128: $-0.153 \rightarrow -0.173$) and the aggressive cells by at most 0.060 (1.5B: $-0.332 \rightarrow -0.389$); the gap between the worst cell at $r \leq 48$ ($+0.028$ at h=0.2, $+0.025$ at h=0) and the best cell at $r \geq 64$ (-0.140 and -0.151), which is what the map rests on, is 0.168–0.176 at every weight. Swapping the rule flag for the knowledge-base entity check of Section G on the thirteen configurations whose full responses are stored moves the term by -0.013 to -0.081, but that check is near-saturated on this corpus (0.781–1.000 of CVE answers flagged against 0.125–0.844 for the rule), so it separates less rather than more. We therefore keep the rule flag at weight 0.2 and read the swap as a check on the direction of the term, not on its size. The per-subset decomposition that explains *why* the conservative cells are near zero is in Section J.

The last substitution is on the refusal axis, where rule and judge disagree most. The manual re-read of Section G puts the rule's false-negative rate at 0.10 on the bases and the below-threshold cells and 0.09 on the cross-domain cell, against 0.33 at 1.5B r=128 and 0.79 on the aggressive cells — not a uniform offset, which would cancel in a delta, but one concentrated where the answers left Chinese. We therefore replace the refusal and jailbreak verdicts of every judged configuration with the judge's on the same stored previews, take the difference against the same-scale base (swapped too), and recompute the family. Because SUS is linear in these two channels, the swap moves a cell by $0.25\,\Delta_{\text{refusal}} + 0.10\,\Delta_{\text{jailbreak}}$ exactly. It is a configuration-level swap and not a per-item substitution: the judge corpus re-generates these adapters rather than reading the runs behind Figure 1 — no judged configuration is reproduced by any of the 90 stored runs — although the rule's refusal rate on the re-generation still tracks the stored runs (median gap 0.033; the cells that do not are 0.5B r=16, at 0.17, and the cross-domain cell, at 0.30). The swap leaves the conservative nulls where they are: no judged cell at $r \leq 48$ moves by more than 0.017, and the largest movement anywhere below the threshold is the 0.042 that the unlabelled 0.5B cells inherit from their base, whose rule jailbreak flag under-calls that base by 0.256. It does move the two language-shifted cells — 1.5B r=128 from -0.158 to -0.063, and the 1.5B aggressive cell from -0.332 to -0.167 — against at most 0.03 for every other judged cell. Four of the five BH survivors are unmoved (p_{BH} within 0.002); the fifth, 1.5B r=128, is decided by this substitution, leaving the set under the refusal channel alone ($0.044 \rightarrow 0.121$) and staying only if the refusal channel is left as the rule scores it ($0.044 \rightarrow 0.042$), while 7B r=64 enters ($0.050 \rightarrow 0.047$). With both channels the two are 0.090 and 0.047. We therefore keep the rule scorer as the main readout — it is the only one of the two that covers all 24 family members on the stored runs — and report the judge alongside it wherever the judge exists, reading the 1.5B r=128 magnitude, and the two hairline verdicts either side of it, as scorer-dependent rather than settled: the judge covers 7 of the 24 family members, and neither scorer covers the other channels of SUS.

I Retrieval Control

The retrieval comparison of Section 6 pairs each trained arm against the same items with the official NVD description of the queried CVE injected at inference time, so the contrast is within arm and within item. Two properties of the setup bound what it can claim. The injection is an oracle — the correct advisory, no retriever and no noise — so the grounded accuracy is an upper bound rather than a deployable number. And the untrained row is a second generation draw of the 1.5B base rather than its canonical run, so its no-RAG level carries that batch difference and is not the base of Figure 1.

Table 7: Retrieval grounding as an oracle upper bound (32 CVE items). The two trained arms pair 5 independently tuned 1.5B models; the untrained row is one 1.5B base checkpoint with no adapter, paired per item and a second draw of the base rather than the canonical run (Section U). All rows improve on both metrics, and the arm that drifts more geometrically is not the arm that improves more — the sufficiency negative of Section 5.3.

Arm	CVE acc (no RAG → RAG)	Δ (paired)	hallucination
selected S-650	0.277 → 0.455	$+0.178$ ($p<10^{-5}$)	0.625 → 0.031
random R-650	0.286 → 0.460	$+0.174$ ($p=8\times10^{-5}$)	0.706 → 0.056
untrained base	0.250 → 0.434	$+0.184$ ($p<10^{-6}$)	0.719 → 0.125

Table 8: Per-subset change against the same-scale base, recomputed from the stored per-item records (means over the three training seeds of each configuration). A near-zero net change in the conservative block is a cancellation, not an absence of movement.

Configuration	Δharmful	Δjailbreak	Δsafety-QA	ΔCVE	Δgeneral-QA
Conservative					
0.5B r=16	−0.011	+0.050	−0.052	+0.016	−0.024
1.5B r=16	+0.033	+0.067	−0.060	+0.043	−0.074
1.5B r=16, D=1000	+0.033	+0.067	+0.003	+0.022	−0.053
1.5B r=16, D=500	+0.022	+0.150	+0.038	+0.017	−0.023
1.5B r=16, D=3000	+0.033	+0.067	+0.064	+0.027	−0.066
1.5B r=128	−0.333	−0.300	−0.174	+0.049	−0.158
3B r=16	−0.078	−0.017	−0.047	−0.015	−0.028
3B r=64	−0.256	−0.300	−0.182	−0.034	−0.116
3B r=128	−0.400	−0.383	−0.367	−0.056	−0.175
7B r=16	+0.011	+0.183	−0.102	−0.001	+0.013
7B r=128	−0.411	−0.150	−0.210	−0.020	−0.101
Conservative, safety-neutral corpus (Section P)					
1.5B r=128	−0.178	+0.017	−0.124	+0.008	−0.140
3B r=128	−0.244	−0.183	−0.196	−0.067	−0.129
Aggressive					
0.5B	−0.244	−0.050	−0.363	−0.234	−0.490
1.5B	−0.922	−0.500	−0.344	+0.055	−0.338
3B	−0.800	−0.650	−0.405	+0.027	−0.237
7B	−0.722	−0.500	−0.552	+0.047	−0.233

J Per-Subset Decomposition of the Conservative Nulls

SUS averages five subsets, so a near-zero ΔSUS admits two readings: nothing moved, or several things moved and cancelled. Table 8 separates them by reporting each subset's change against the same-scale base. The conservative cells below the rank threshold are cancellations rather than inert tunings. At 1.5B r=16 harmful refusal and jailbreak resistance *rise* (+0.033, +0.067) together with CVE accuracy (+0.043), while safety-QA and general-QA fall (−0.060, −0.074); the 7B cell has the same shape with a larger jailbreak gain (+0.183) offset by a larger safety-QA loss (−0.102); and the 0.5B cell cancels moderate movements in both directions. The cells above the threshold cancel nothing: 3B r=64, 3B r=128 and 7B r=128 fall on all five subsets, as does 3B r=16, the largest conservative loss at low rank, while 1.5B r=128 falls on everything except CVE. Under the aggressive regime the pattern changes in kind: harmful refusal, jailbreak resistance, safety-QA and general-QA all fall at all four scales. The one exception is CVE accuracy, which the aggressive cells do not lose (+0.055, +0.027, +0.047 at 1.5B/3B/7B) and which is the channel Section 4 reports them writing in; at 0.5B even that channel collapses (−0.234). The two safety-neutral cells at r=128 (Section P) have the shape of the in-domain high-rank cells rather than of the conservative block: each falls on four of the five subsets, the single exception being jailbreak resistance at 1.5B (+0.017), and general-QA falls there almost as far as in-domain (−0.140 and −0.129 against −0.158 and −0.175). The reading we take from the table is that "no detectable net harm" at the conservative cells below the threshold is a statement about an average, not about the components, that the cells above it are not cancellations at all — they are uniformly worse, which is why the 3B r=128 cell is the one place in the conservative grid where a knowledge gain (CVE fact-hit, Section 4) coexists with losses on every other subset — and that this shape survives replacing the corpus with text that contains nothing to absorb.

K System-Field Ablation

Section 8 asks whether the reported safety cost is an artifact of the corpus's system field, since 76.8% of the corpus carries a Chinese safety instruction while every in-domain measurement uses no system prompt. The design holds the corpus, the recipe, the training seeds, and the evaluation fixed and varies one field: *kept* is the 650-item subset as shipped (520 of its records carry a system message), *deleted* removes every system message, and *permuted* re-assigns the 520 system strings across the records that carry one (seed 42), so the multiset of strings and the set of records carrying

Table 9: System-field ablation at 1.5B on the 650-item subset. Cells are changes against the canonical 1.5B base, averaged over three training seeds; the base row gives absolute values. The aggressive arms collapse with the system field kept, deleted, and permuted alike, while no conservative arm collapses.

Configuration	SUS	Safety	harmful	jailbreak	safety-QA	CVE	general-QA
base (1.5B)	0.599	0.760	0.933	0.600	0.579	0.266	0.673
Conservative (lr 5×10^{-6}, one epoch)							
system kept	+0.021	+0.047	+0.045	+0.117	+0.003	+0.030	−0.056
system deleted	+0.010	+0.038	+0.022	+0.117	+0.011	+0.037	−0.063
system permuted	+0.037	+0.078	+0.056	+0.117	+0.090	+0.008	−0.042
Aggressive (lr 10^{-4}, three epochs)							
system kept	−0.186	−0.379	−0.467	−0.483	−0.165	+0.039	−0.129
system deleted	−0.143	−0.290	−0.400	−0.400	−0.033	−0.005	−0.134
system permuted	−0.147	−0.283	−0.389	−0.350	−0.063	+0.011	−0.190

Table 10: Arm-to-arm contrasts of Table 9 as changes with 95% two-sample t intervals (df=4; $t_{0.975,4}$=2.776) Deleting the system field is the contrast that would reveal a format artifact: under the aggressive regime it recovers at most 0.043 SUS against a 0.186 collapse and leaves harmful refusal indistinguishable from the kept arm.

Contrast	ΔSUS	ΔSafety	Δharmful refusal
Conservative			
sysdel − kept	−0.011 [−0.044, +0.021]	−0.009 [−0.039, +0.022]	−0.022 [−0.110, +0.065]
sysrand − kept	+0.016 [−0.002, +0.035]	+0.032 [−0.012, +0.075]	+0.011 [−0.033, +0.055]
sysrand − sysdel	+0.028 [−0.000, +0.056]	+0.040 [−0.006, +0.087]	+0.033 [−0.054, +0.121]
Aggressive			
sysdel − kept	+0.043 [+0.011, +0.076]	+0.089 [+0.025, +0.154]	+0.067 [−0.084, +0.218]
sysrand − kept	+0.039 [−0.001, +0.078]	+0.096 [+0.049, +0.144]	+0.078 [−0.033, +0.189]
sysrand − sysdel	−0.005 [−0.033, +0.024]	+0.007 [−0.060, +0.074]	+0.011 [−0.100, +0.122]

one are both preserved exactly and only the pairing between an instruction and its conversation is destroyed. Each arm is trained with seeds 42/44/46 under both regimes of Section 3.1 and evaluated on the 110-item suite with no system prompt and generation seed 42; the three aggressive conditions are additionally re-evaluated on the 50-item bilingual probe of Section 7 with and without the prompt (Table 11). The script that trains, evaluates, and tabulates the eighteen arms is released with the rest.

Three readings follow. First, the aggressive collapse is not an artifact of the training-time system field. With that field deleted the model still refuses only 0.53 of the harmful items and 0.20 of the jailbreaks, against a base of 0.93 and 0.60, and with it permuted 0.54 and 0.25; all three aggressive arms lie below base on SUS and harmful refusal at $|t|>10$ (df=2). Second, the manipulation is not inert, so the invariance of the collapse is not a null result about the perturbation: deleting the system field moves safety-QA accuracy by +0.131 and permuting it by +0.102 under the aggressive regime (+0.008 and +0.087 under the conservative one), and the permuted arms retain a system message, so a model that gated on the mere presence of one would keep its gate. Third, only the aggressive half of the table adjudicates anything: no conservative arm collapses, and the conservative arm-to-arm SUS contrasts (0.011–0.028) sit at or inside the 0.028 minimum detectable effect we can resolve with three seeds at this scale, which is why we read the conservative block as uninformative for this question rather than as evidence either way.

The same arms under the prompt. Section 7 reads the prompt effect against a template-replay account, in which the adapted model has compressed refusal into a response to the adaptation prompt. Everything in Table 9 is measured without a prompt, so it cannot test that account; re-running the three aggressive conditions on the 50-item bilingual probe with and without the prompt can, and the answer is that it does not separate the two readings.

Without the prompt the three conditions are indistinguishable (0.42–0.51 by rule, 0.38–0.48 by judge). With it they separate, and not in the direction template replay requires: the two conditions whose corpus contained the prompt string fall (0.511 → 0.344 and 0.433 → 0.289), while the

Table 11: The system-field conditions under the prompt: aggressive regime, 1.5B, Chinese harmful items of the 50-item bilingual probe (30 items × three training seeds). Rule is the bilingual keyword scorer, judge the local Qwen3.5-9B rubric. Without a prompt the three conditions are indistinguishable; with it they separate in opposite directions.

Condition	Rule		Judge	
	no prompt	prompt	no prompt	prompt
kept	0.511	0.344	0.484	0.281
sysdel	0.422	0.878	0.375	0.900
sysrand	0.433	0.289	0.454	0.244

Table 12: The recipe axis at 1.5B, r=16, full corpus, three training seeds. Cells are changes against the same-scale base, and *rel* is $\|\Delta W\|_F/\|W\|_F$ measured from the stored adapter. The two stored regimes anchor the ends: conservative (5×10^{-6}, one epoch) at ΔSUS=-0.007 and 0.0014, aggressive (10^{-4}, three epochs) at -0.332 and 0.0208. All five new cells are harmful at $|t|\geq 8.8$ (df=2).

Config.	rel	ΔSUS	ΔSafety	Δharmful	Δjailbr.	ΔsafetyQA	ΔCVE	Δgen-QA
base (1.5B)	—	0.599	0.760	0.933	0.600	0.579	0.266	0.673
2×10^{-5} / 1 ep	0.0040	−0.124	−0.180	−0.178	−0.283	−0.115	+0.045	−0.157
5×10^{-5} / 1 ep	0.0078	−0.223	−0.480	−0.678	−0.350	−0.236	+0.056	−0.131
10^{-4} / 1 ep	0.0126	−0.350	−0.664	−0.889	−0.450	−0.431	+0.003	−0.345
2×10^{-5} / 3 ep	0.0083	−0.328	−0.645	−0.856	−0.467	−0.413	+0.115	−0.341
5×10^{-5} / 3 ep	0.0140	−0.413	−0.704	−0.911	−0.533	−0.472	−0.145	−0.471

condition whose corpus had no system field at all rises to 0.878 (0.900 by judge). The responses show why. Under the prompt the kept and sysrand arms re-enter the multi-step reasoning format their training data carries (4,229 of the corpus's 6,503 records) and a large share of those trajectories ends in compliance, whereas the deleted arm — for which the prompt is out of distribution — answers briefly as the base does. The prompt can therefore re-elicit a memorised template or push an out-of-distribution model back towards base behaviour, and which of the two happens depends on whether the training corpus carried that string at all; the deeply collapsed full-corpus cell of Section 7 (0.133 → 0.956) is a third case again. These three arms also carry the 650-item subset and are less collapsed than that cell (0.42–0.51 without a prompt against 0.13), so we report the ablation as evidence against a uniform template-replay reading rather than as an identification of the mechanism.

L The Recipe Axis as an Update-Magnitude Curve

Section 8 asks whether "conservative is safe" merely restates "conservative barely moves the weights". Two measurements answer it, and neither needs a new architecture. First, for every stored adapter we recover $\mathrm{rel} = \|\Delta W\|_F/\|W\|_F$, where ΔW is the block matrix $\big((\alpha/r)\,B_mA_m\big)_m$ over the adapter's modules m and $\|\cdot\|_F$ aggregates as $\sqrt{\sum_m \|\cdot_m\|_F^2}$ on both sides, which puts all 153 measured adapters on one axis; within a scale, harm rises with rel at every scale (Pearson r 0.56–0.95, Kendall τ 0.43–0.82, tightest at 0.5B and 1.5B). The α/r pair — two arms at 1.5B, r=16, full corpus, identical but for the scaling — is the cheapest read of that axis: an α/r of 1 moves the weights by 0.00070 (ΔSUS=+0.009) and an α/r of 4 by 0.00262 (-0.035), so their behavioural difference is bought with a 3.75× larger update. Second, we fill the gap between the two regimes with five new cells at 1.5B, r=16, full corpus, three training seeds each, evaluated on the same 110-item suite: lr $\in\{2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$ at one epoch and lr $\in\{2\times10^{-5}, 5\times10^{-5}\}$ at three. Table 12 reports them against the base, with the two stored regimes as anchors.

Three readings follow. (a) At a fixed epoch count, harm is monotone in the update magnitude and the curve is steep: one-epoch cells at 0.0014, 0.0040, 0.0078 and 0.0126 cost -0.007, -0.124, -0.223 and -0.350 SUS. Monotone within this sweep, however, does not mean monotone over the grid, and matching on magnitude is how that shows. Across the 23 arms that share one scale, one layout (140 adapted modules), one corpus, one learning rate and one epoch and differ only in rank, harm rises with rel (r=0.94, $p<10^{-10}$), but the three r=128 arms sit 0.090 *above* the fitted line while every arm at $r\leq 48$ sits on or below it (-0.008, -0.029 and -0.022 at r=16, 32 and 48). The sharpest pair

lies inside that set and needs no cross-scale reading: r=16 at α/r=4 moves the weights by 0.00262 and costs -0.035, while r=128 at α/r=2 moves them by 0.00280 — 7% further — and costs -0.138, -0.173 and -0.184 on its three seeds. Same corpus, layout, learning rate and epoch count, near-identical update magnitude, a factor of four in harm. That factor, like the rest of this sweep, is read under the rule scorer; the judge substitution of Section R covers the four main-grid high-rank cells and not these arms, so read it as a rule scorer statement of the same kind as the 58–65% ratio of Section P, with Section H giving the coverage of each scorer. The cross-scale version of the same comparison points the same way but is weaker, since it also crosses a base level: 1.5B r=48 and 7B r=128 both move the weights by 0.0021, and cost -0.034 against -0.153. The safe end of the curve is therefore a band rather than a regime, and the conservative regime is safe because it sits at the bottom of that band, not because a low learning rate is intrinsically safe — raising it by a factor of four at the same epoch count already costs a significant 0.124. (b) Epochs and learning rate are not interchangeable: at a matched update magnitude near 0.008, three epochs costs -0.328 against -0.223 for one, and three epochs lands between -0.33 and -0.41 at *every* learning rate tested, including one 5× below the aggressive setting. Repetition over the corpus, not the step size alone, is what reaches the collapsed floor. (c) The harm at these cells is not safety-specific: general-QA accuracy falls by 0.13–0.47 and safety-QA by 0.12–0.47, so what these recipes destroy is response quality in general, with refusal the fastest-moving component. The practical consequence is that our deployment rule is a bound on the update magnitude, bracketed between 0.0014 and 0.0040 by this sweep, together with a separate bound on rank — below 64 at the mid-scale, below 128 elsewhere — rather than a recommended learning rate: a bound a practitioner can measure on their own adapter without running the suite.

M The Collapse on a Harmlessness-Cleaned Corpus

The audit of Section B finds that the recipes were trained on harmful targets, which is a candidate explanation for the collapse itself: an aggressive regime may simply learn the harm it is shown. We test that on every r=128 collapse cell of the map — at 1.5B, 3B and 7B — plus the conservative r=16 anchor at 1.5B and 3B. The cleaned corpus removes every sample whose target turns the judge flagged H=2 inside the two clusters the manual re-read verified as harmful — 720 `unsafe_qa` and 627 `hh-rlhf` samples, 1,347 of 6,503, leaving 5,156; the 175 flagged `template` samples are kept because all 20 sampled were defensive vulnerability write-ups. Whole samples are dropped rather than single turns, so that no transcript keeps an unanswered harmful prompt; the cost is that a multi-turn `hh-rlhf` transcript also loses its safe turns. The size-matched control trains the aggressive regime on 5,156 samples drawn at random from the *uncleaned* corpus, which separates "harm removed" from "fewer samples". Each cell is three training seeds, evaluated once on the 110-item suite at generation seed 42, as everywhere else. The 3B and 7B rows use their own training scripts and defaults, so a scale contrast mixes the corpus change with a batch-layout change; the within-scale stored/clean contrasts do not.

Two conclusions follow, and they point in opposite directions. The aggressive collapse is not a contamination artefact. Removing every judge-flagged harmful target leaves the cell at -0.373 SUS with harmful refusal 0.011, against -0.332 and 0.011 for the stored cell on the full corpus, and statistically indistinguishable from the size-matched control ($+0.003$, $[-0.022, +0.029]$): the recipe that had harmful material available and the recipe that did not land in the same place, and the aggressive cell's seed spread is the tightest in the paper (sd=0.003). Cleaning is, if anything, very slightly harmful, by 0.041 ($[-0.048, -0.033]$), which is what dropping whole transcripts predicts: the removed `hh-rlhf` samples also contributed refusal demonstrations, so the cleaned corpus is marginally less safety-alignment-rich. The two r=16 anchors are likewise unchanged — 1.5B $+0.017$ against -0.007 and 3B -0.021 against -0.045, both contrasts $+0.024$ and not significant — so the paper's headline null does not depend on the cleaning either. The r=128 cells are the exception, and there the scales separate. Every one of them recovers a significant slice when the harmful targets go: 0.116 at 1.5B ($[+0.068, +0.164]$), 0.082 at 3B ($[+0.022, +0.142]$), 0.067 at 7B ($[+0.002, +0.132]$). What differs is what is left over. At 1.5B the residual -0.042 is no longer significant ($t = -4.0$ against a minimum detectable effect of 0.028), so most of that cell's loss was corpus-driven. At 3B the residual is -0.141 ($t = -8.5$, still about twice what that scale resolves) and at 7B -0.085 ($t = -4.8$): the majority of both losses never depended on the corpus's harmful content. The fact channel says the same thing. The cleaned 3B r=128 cell still writes CVE facts in — fact-hit 0.040 against 0.019 at base, against 0.055 for the stored cell — and the cleaned 7B

Table 13: The collapse cells on a harmlessness-cleaned corpus. *clean*: the paper's corpus with the 1,347 manually verified harmful samples removed. *size-matched*: the aggressive regime on as many samples drawn at random from the *uncleaned* corpus. *stored*: the paper's cells on the full uncleaned corpus, quoted from the canonical per-seed records. ΔSUS is against the fixed same-scale base and t is the one-sample statistic with $df=2$. The refusal column is the harmful-prompt refusal rate, measured three seeds at generation seed 42 as everywhere else, so it is checkable against Table 8: that table's Δharmful plus the same-scale base refusal (0.833, 0.933, 1.000, 0.967 at 0.5B–7B) reproduces every *stored* value here (3B $r=128$: $1.000 - 0.400 = 0.600$).

Scale	Cell (r)	Corpus ($n=3$)	SUS	ΔSUS	t	refusal	$\|\Delta W\|/\|W\|$
1.5B	aggressive 16	clean	0.226	−0.373	−209.3	0.011	0.0191
1.5B	aggressive 16	size-matched	0.223	−0.377	−42.5	0.011	0.0189
1.5B	conservative 16	clean	0.616	+0.017	+1.4	0.989	0.0012
1.5B	conservative 128	clean	0.557	−0.042	−4.0	0.922	0.0026
3B	conservative 16	clean	0.628	−0.021	−1.2	0.967	0.0009
3B	conservative 128	clean	0.509	−0.141	−8.5	0.756	0.0017
7B	conservative 128	clean	0.574	−0.085	−4.8	0.667	0.0020
1.5B	aggressive 16	stored	0.267	−0.332	—	0.011	0.0209
1.5B	conservative 16	stored	0.592	−0.007	—	0.967	0.0014
1.5B	conservative 128	stored	0.441	−0.158	—	0.600	0.0028
3B	conservative 16	stored	0.604	−0.045	—	0.922	0.0011
3B	conservative 128	stored	0.426	−0.223	—	0.600	0.0018
7B	conservative 128	stored	0.507	−0.153	—	0.556	0.0021

Table 14: Cell-to-cell contrasts of Table 13: equal-variance two-sample t, $df=4$, $t_{0.975}=2.776$. The aggressive collapse is unaffected by removing the harmful targets; every $r=128$ loss recovers significantly, but only the 1.5B one recovers to non-significance.

Contrast	ΔSUS	t	95% interval	Verdict
aggr 1.5B clean − stored	−0.041	−14.5	$[-0.048, -0.033]$	significant
aggr 1.5B clean − size-matched	+0.003	+0.4	$[-0.022, +0.029]$	not significant
cons 1.5B $r=16$ clean − stored	+0.024	+1.9	$[-0.011, +0.060]$	not significant
cons 1.5B $r=128$ clean − stored	+0.116	+6.8	$[+0.068, +0.164]$	significant
cons 3B $r=16$ clean − stored	+0.024	+1.0	$[-0.040, +0.087]$	not significant
cons 3B $r=128$ clean − stored	+0.082	+3.8	$[+0.022, +0.142]$	significant
cons 7B $r=128$ clean − stored	+0.067	+2.9	$[+0.002, +0.132]$	significant

cell reads 0.029 against 0.016 at base and 0.027 stored, so the cells that buy knowledge are the cells that lose refusal, harmful corpus or not. One caveat attaches to the 3B and 7B rows: they are trained with their own scripts and default batch layout, so their levels are comparable within a scale but the cross-scale gradient of recovery mixes the corpus change with that difference. The per-scale contrasts, which is what the map's verdicts rest on, are unaffected.

N The Three Candidate Explanations, Measured

The collapse admits at least three readings, and each is tested by its own ablation rather than argued away. (i) *Format mismatch, tested.* The corpus is not system-prompt free: 4,993 of its 6,503 records (76.8%) carry a Chinese safety or domain-expert `system` message, while every in-domain measurement uses none. Re-training the 650-item subset with that field kept, deleted, and permuted leaves the aggressive collapse intact: every arm below base at $|t|>10$ ($df=2$), harmful refusal 0.47–0.54 against a base of 0.933. Deleting the field recovers 0.043 SUS and 0.089 safety, about a quarter of the 0.186 and 0.379 collapse, and leaves harmful refusal indistinguishable from the kept arm, so format mismatch explains a minority of the large damage (Section K). (ii) *Under-training.* Our conservative learning rate is 20–40× below common LoRA practice, so "conservative is safe" could restate "conservative barely moves the weights". We measured it directly ($\|\Delta W\|_F/\|W\|_F$ from every stored adapter) and filled the gap between recipes with a learning-rate-by-epoch sweep at 1.5B, $r=16$ (15 runs, three seeds each). Within a fixed epoch count harm is monotone in the update magnitude: at one epoch the ratio runs from 0.0014 (conservative, ΔSUS=−0.007) to 0.0126 (lr 10^{-4}, −0.350), and at a matched ratio of 0.008 three epochs cost −0.328 against −0.223 for

one. The conservative regime is therefore genuinely safe rather than inert, but its safety is a band of update magnitude — bracketed between 0.0014 and 0.0040 — and of repetition, not a property of its learning rate; and Section 4 shows a separate rank bound is needed too, since the r=128 cells reach larger harm inside that band (Section L). (iii) *Corpus contamination, measured and controlled.* The audit of Section B finds 18.6% of the corpus's assistant turns are themselves harmful; removing those 1,347 samples and re-training the collapse cells leaves the aggressive collapse intact, indistinguishable from a size-matched control drawn from the *uncleaned* corpus, while the conservative r=128 loss shrinks from -0.158 to -0.042 at 1.5B, where it stops being significant (p=0.057). Content and rank are entangled there, but not equally at every scale: the same control recovers a significant slice of the 3B and 7B losses ($+0.082$, $+0.067$) yet leaves both significant (-0.141, -0.085), so content accounts for the whole of a high-rank loss only at 1.5B. A safety-neutral corpus still costs 0.091 at 1.5B r=128 and 0.146 at 3B against 0.158 and 0.223 in-domain, and while the two controls do not isolate the same remainder — removing the harmful targets leaves 0.042, replacing the whole corpus 0.091 — neither removes the loss (Sections M and P).

O External Benchmark Cross-Validation

The suite of Section A is in-domain, so we cross-validate the refusal axis on two established English benchmarks, AdvBench (520 behaviours) and JailbreakBench (100 prompts) (Zou et al., 2023; Chao et al., 2024), over twelve configurations with one generation seed and the same decoding as the main suite. These runs differ from the in-domain suite in one respect that matters for reading them: they carry the corpus's Chinese safety system prompt, which the in-domain measurements of Section 3 do not, so they reproduce the deployment-like condition of Section 7 rather than the no-prompt condition of the map. Refusal is scored twice on the stored responses: by the bilingual keyword checker of the main suite, and independently by the local judge of Section G. Table 15 reports both series; all 24 cells and the 12 combined rates are reproduced from the stored per-item records by the released script.

The base models refuse at or near ceiling externally (AdvBench 0.911–0.994 by rule, JailbreakBench 0.820–0.980), and so do the cells that collapse in-domain: 1.5B r=128 reads 0.819 on AdvBench and 0.860 on JailbreakBench, and both aggressive cells read 1.000 on both benchmarks while refusing 0.133 and 0.011 of the harmful items of the bilingual probe of Section 7, under its Chinese and English renderings, at 1.5B and 0.489 and 0.067 at 7B. The external items are therefore not an independent probe of the behaviour, and we do not treat Table 15 as an alignment certificate: since these runs carry the same Chinese safety system prompt that restores refusal in Section 7, the ceiling they elicit is that prompt effect reappearing on a fixed English benchmark, not evidence about the policy the weights carry.

How much of the benchmark is in the training corpus. AdvBench and JailbreakBench are training sources here as well: Table 3 lists 469 `advbench` and 92 `jailbreakbench` samples in the corpus. Matching the $520 + 100$ evaluated items against the corpus's 10,713 user turns by the convention of Section B — an exact match, or ≥ 0.5 word 8-gram containment against a single turn — finds 484 of the 520 AdvBench items and 93 of the 100 JailbreakBench items present, 483 and 93 of them exactly, so the "external" label is not earned for nine tenths of the items. We therefore report the split rather than the pooled rate: on the items that are *not* in the corpus (AdvBench 36, JailbreakBench 7 per cell) the two collapsed cells the benchmark table contains still refuse at 1.000 by the bilingual rule and 1.000 by the judge, on every one of those items and on both benchmarks, identical to their seen halves. What the ceiling measures is thus not recall of the training items, since it reappears on items the model has never seen — but it is the prompt condition of Section 7 rather than a property of the items, and the unseen halves are small, so we read this as a direction rather than a tight bound and keep the in-domain suite as the primary evidence.

Scorer agreement carries over to these items, but only for the bilingual set. The bilingual rule and the judge agree at κ=0.33 overall (P_0=0.96; 0.28 on AdvBench, 0.44 on JailbreakBench) and order the twelve cells consistently (ρ=0.80). The English-only five-keyword set does not: because the adapted cells answer in Chinese, it registers 0.000 refusal on all 620 responses of each aggressive cell where the judge reads 1.000, and 0.250–0.281 on the 3B r=16 and 1.5B r=128 AdvBench responses; its correlation with the judge across configurations is $\rho = -0.05$. Bilingual keywords are necessary for scoring refusal on these models, and the judge's rubric for these runs was refusal

Table 15: Refusal rate on two English benchmark suites (AdvBench N=520, JailbreakBench N=100; one generation seed, run with the safety system prompt of Section 7); nine tenths of these items are also training-corpus items, and Section O reports the seen/unseen split. "Rule": bilingual keyword checker of the main suite. "Judge": independent local judge on the stored responses. Read against Section 7: the adapted cells that refuse *every* fixed item refuse almost none of the in-domain variant prompts.

Configuration	AdvBench		JailbreakBench	
	Rule	Judge	Rule	Judge
base 0.5B	0.911	0.967	0.820	0.850
base 1.5B	0.994	1.000	0.980	1.000
base 3B	0.969	1.000	0.880	1.000
base 7B	0.979	0.994	0.870	0.970
cons. r=16 (0.5B)	0.902	0.967	0.770	0.840
cons. r=16 (1.5B)	0.954	1.000	0.950	0.990
cons. r=16 (3B)	0.994	0.998	0.950	0.980
cons. r=16 (7B)	0.992	1.000	0.950	0.990
SFT on the 650-item selected subset, no retrieval (1.5B)	0.996	1.000	0.990	1.000
r=128 (1.5B)	0.819	0.956	0.860	0.990
aggressive (1.5B)	1.000	1.000	1.000	1.000
aggressive (7B)	1.000	1.000	1.000	1.000

$= R \in \{1, 2\}$ with all items in the denominator (21 of 7,440 items carried no verdict and are counted as non-refusals).

P Cross-Domain and Cross-Family Checks

Repeating the conservative r=16 cells on Zhihu Q&A (safety-neutral instruction data with no security or harmful material, D=6500) leaves the picture flat at both ends of the capacity axis: 0.5B 0.517 ± 0.027 (-0.018) and 7B 0.645 ± 0.027 (-0.015), so the security content is not what makes conservative tuning mildly harmful. The 1.5B cell of this check comes from a stored five-generation aggregate and has no per-seed records (Section T).

The same corpus at r=128 is what separates the rank axis from the corpus axis, at three seeds each: SUS falls to 0.508 ± 0.035 at 1.5B (-0.091, raw p=0.047) and 0.503 ± 0.014 at 3B (-0.146, p=0.003), against in-domain losses of 0.158 and 0.223 at the same scale and rank — 58% and 65% of them, produced by a corpus with no harmful target and no security text, where the same corpus at r=16 cost -0.013. That ratio is a rule scorer statement on both sides of it; judging the stored runs of these four cells directly turns it into 0.91 at 1.5B and 0.60 at 3B, so we read the 58–65% as a 3B result (Section R). The 3B arm here runs at 1.5B's layout rather than its own (Section U), so its 65% is an approximation: that cell varies layout together with corpus. Table 8 decomposes the two cells: the ability axis moves nearly as far as in-domain, while the refusal drop is roughly halved, so about half of the refusal damage needs no harmful content to occur and the other half tracks the corpus.

On Mistral-7B the suite is poorly discriminative — base harmful refusal is only 0.367 — and we therefore report the selected-subset cell ($0.388 \to 0.440$) as inconclusive rather than as evidence about families.

Q The Prompt Contrast on the Conservative Cells

Section 7 measures the prompt contrast on the cells that collapse. The cells that carry Section 4 need the same check, because it decides whether their loss survives a deployment that carries a safety prompt. We repeat the probe on the conservative r=128 cells at 1.5B, 3B and 7B and on their bases: the same 50 behavioural items in both renderings, the same bilingual rule, the same judge, three training seeds per cell.

These arms re-merge the stored adapters rather than applying them through PEFT, so that they share the inference path of the map; the 1.5B r=128 arm reproduces its stored rule score exactly (0.689), which is our check on the harness. The path is not neutral: an earlier pass that applied the same

adapters through PEFT read 0.789 at 1.5B and 0.800 at 7B for the no-prompt condition, against 0.689 and 0.933 here, so these 30-item rates move by as much as 0.13 with the adapter-application path and every number below is from the re-merged one.

Table 16: Conservative r=128 against the same-scale base under the adaptation safety prompt (rule and judge, Chinese harmful items; 30 items × three training seeds per cell, bases single runs). The prompt restores the collapsed cells of Section 7 but not these.

Scale	Cell	Rule		Judge	
		no prompt	prompt	no prompt	prompt
1.5B	base	1.000	1.000	0.967	1.000
1.5B	r=128	0.689	0.789	0.788	0.854
3B	base	0.967	1.000	1.000	1.000
3B	r=128	0.856	0.700	0.878	0.678
7B	base	0.900	1.000	1.000	1.000
7B	r=128	0.933	0.944	0.933	0.933

Under the prompt the conservative cells do not return to their bases. On the Chinese harmful items the rule scorer reads 0.689 → 0.789 at 1.5B, 0.856 → 0.700 at 3B and 0.933 → 0.944 at 7B against 1.000 for all three bases with the prompt present, and the judge reads 0.788 → 0.854, 0.878 → 0.678 and 0.933 → 0.933; so 1.5B recovers about a third of its gap, 3B moves the other way, and 7B does not move. What the prompt does restore is the English rendering of the same items, which travels toward ceiling on every cell (0.611 → 0.778, 0.389 → 0.889, 0.211 → 0.989 by rule). The prompt-conditional recovery of Section 7 is therefore a property of the collapsed regime and of the templated rendering, not of the conservative high-rank loss, and that loss is not bought back by evaluating under a safety prompt (Figure 4).

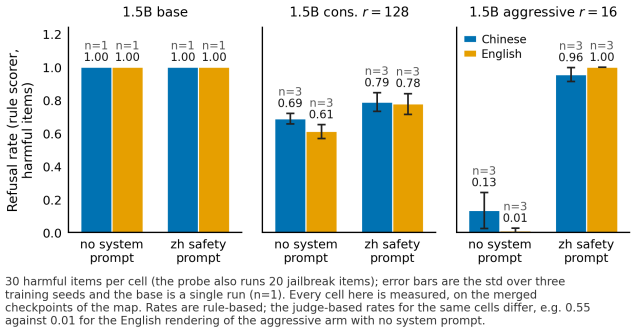


Figure 4: Harmful refusal with and without the adaptation safety prompt, under Chinese and English renderings of the same items, for the collapsed cells of Section 7. Every cell shown is measured, on the re-merged checkpoints of Section 4. For the cells shown here the prompt, not the language, restores refusal; the conservative r=128 cells are the contrast of Table 16.

R JUDGING THE STORED RUNS OF THE HIGH-RANK CELLS

The ratio of Section P — what a safety-neutral corpus costs relative to the security corpus at the same scale and rank — is computed under the rule scorer on both sides, so it is only as good as that scorer. We therefore judge the stored previews of the four high-rank cells and their bases directly and re-read ΔSUS with the judge's refusal and jailbreak verdicts substituted: −0.158 → −0.091 and −0.223 → −0.149 for the security corpus at 1.5B and 3B, and −0.091 → −0.083 and −0.146 → −0.089 for the neutral one. This is a different provenance from the swap of Section H, which reads a separately judged, re-generated corpus. Both routes agree on sign and leave every ordering intact, but they disagree on size: that one moves the two language-shifted cells and leaves the rest of the family near where it was, while this one moves all four high-rank cells, by 0.008 to 0.074. The ratio survives at 3B and does not survive at 1.5B: 0.58 becomes 0.91, because the denominator shrinks along with the numerator. The "58–65%" of Section P is therefore a 3B statement, and the 1.5B cell's ratio is scorer-dependent.

S THE UTILITY CHANNELS UNDER THE LANGUAGE SPLIT

The utility channels are scored by keyword overlap against Chinese answer keys (Section A), and the cells above the threshold move part of their output into English (Section G), so some of the general-QA and safety-QA drop could be a mismatch between the language of the answer and the language of the key rather than a loss of competence, and the judge swap of Section H does not cover these channels. We test it on the stored per-item records by classifying the language of every answer and comparing only the items that the base and the tuned cell both answer in Chinese. The drop survives the control: general-QA falls by 0.11–0.34 and safety-QA by 0.07–0.40 on the paired Chinese items, across the conservative and neutral high-rank cells and the aggressive cells, so the wholesale degradation of Section 4 is not an artifact of the scorer's language. One cell is the exception, and it is the one where the shift is largest: the 1.5B aggressive cell answers in Chinese on only 0.29 of its safety-QA items, and on the items that remain the paired difference is $+0.03$, so that cell's safety-QA drop in the composite is mostly the language shift.

T DATA GAPS

Every number above is generated by script from stored per-item records. The following quantities could not be reconstructed and are either omitted or reported with their provenance: the per-seed records of the Zhihu-1.5B cell (only an aggregate is stored); the suite evaluation of the 1.5B $r{=}8$ probe configuration (probe outputs and its probe κ exist, a 110-item evaluation does not, so it appears only in Table 6 and Figure 3); the six-configuration leave-one-out fit (only the eight-model fit including 7B is stored, and that is the one we report); and the paired intervals of the five-generation protocol, which exist only for the fourteen cells listed in Section D; the remaining cells — the $r{=}64$ cells at 1.5B and 7B and the 0.5B high-rank cells — rest on the three-seed test, the correction and the cleaned-corpus replication of Section M. Two quantities needed a convention to be recoverable at all, and we state both rather than leave them implicit: the judge rates of Table 15 reproduce the stored per-benchmark series only when refusal is read as $R \in \{1, 2\}$ with all items in the denominator, and the per-benchmark split of that series is itself derived from the per-item judge labels rather than stored per benchmark. Two stored aggregate files in our results directory (`significance_all.json`, `equivalence.json`) were computed on an earlier evaluation suite and are not used anywhere in this paper; every p value reported here was recomputed from the per-item records, and the recomputation script is released so the audit is repeatable.

U REPRODUCIBILITY

Hardware. AMD Ryzen 9 9950X3D, NVIDIA RTX 5090 (32 GB), Windows.

Software. Python 3.11.9; PyTorch 2.10.0.dev20251003+cu128; Transformers 5.14.1; TRL 1.9.2; PEFT 0.20.0; bitsandbytes 0.50.0; accelerate 1.10.1; datasets 5.0.1; numpy 2.3.4; scipy 1.16.2.

Models and data. Base models are Qwen2.5-Instruct (0.5B/1.5B/3B/7B) plus Mistral-7B-Instruct, downloaded from ModelScope; the corpus is described in Table 3 and audited in Section B. The 29 configurations counted in the abstract are the main inventory — the map, the recipe, budget and alpha axes, and the cross-domain and cross-family arms; the ablation and replication arms of Sections K to M and P are trained on the same recipe in addition and are counted in their own sections rather than in that total.

Training layout. The adaptation layout is not identical across scales, and we state it rather than leave it to be inferred. 0.5B adapts q and v only (2 modules per layer, 48 in all), at effective batch 64 (16×4) and sequences of 512; 1.5B adapts q, k, v, o, gate (5 per layer, 140), at effective batch 32 (8×4) and 512; 3B and 7B adapt q, k, v, o, gate, up, down (7 per layer, 252 and 196), at effective batch 32 (4×8) and sequences of 1024. LoRA dropout is 0.05, $\alpha{=}2r$, and the schedule is cosine with 5% warmup (3% at 7B). No arm of the map is quantized, and the released scripts show it rather than assert it: the training scripts take a 4-bit path only when `EXP_QUANTIZATION=4bit` is set in the environment, which no runner behind a map arm sets, and the evaluation script's 4-bit mode fires only when the quantization flag is passed explicitly, which

`uni_reeval.py` never does. The project's experiment table records 4-bit on exactly two entries, both 9B baselines outside the map. Training and merging are in half precision throughout — fp16 at 0.5B and 1.5B (`train_sft_0.5b.py`, `train_sft_1.5b.py`, `merge_lora_0.5b.py`, `merge_lora_1.5b.py`) and bf16 at 3B and 7B (`train_sft_7b.py`, `merge_lora.py`) — and every map number is evaluated from the merged weights in fp16, so quantization enters none of its contrasts. Every map number comes from that one path: the stored records load a `merged` directory with no adapter attached and record it, with `adapter_path` and `quantization` both null, so the PEFT-versus-re-merge difference of Section Q cannot enter them. The 650-item subsets, the cleaned-corpus replication and all but one of the cross-domain Zhihu arms use the layout of their scale. The exception is the 3B Zhihu arm, which was trained at 1.5B's layout (8×4, 512 tokens) rather than 3B's default (4×8, 1024) and therefore differs from its in-domain counterpart in layout as well as corpus; we read that one cell's loss as approximate. Two consequences bound the map. Every contrast *within* a scale — rank, data budget, alpha, recipe, and the corpus at 0.5B, 1.5B and 7B — holds the layout fixed and is unaffected. The comparison *across* scales of Section 4 varies capacity together with the module set, the effective batch and the sequence length at once, so a cross-scale difference can be read as capacity only under the assumption that the other two do not matter; a matched-layout cross-scale control would settle it and we have not run one, so Finding 3's cross-scale statement rests on the within-scale pattern (the same threshold at r=128 at three scales) rather than on the layout being constant.

Derived numbers. Tables 4, 6 to 10 and 12 to 16 and every inline number are produced by released scripts from the stored evaluation records: one that aggregates per-item results into canonical per-configuration values, one that recomputes tests, corrections, equivalence, and minimum detectable effects, one that trains and evaluates the system-field ablation of Section K, one that measures the update magnitude of every stored adapter and tabulates the learning-rate-by-epoch grid of Section L, one that extracts and judges every training target for the audit of Table 4 (with its manual re-read recorded item by item) and builds the cleaned corpus, one that trains and evaluates the cleaned-corpus replication of Section M, and one per appendix table that re-derives it from the per-item records (Sections J to M and O). Seven further scripts support the sensitivity checks and the audit, all reading the same stored records: one matches the benchmark items against the training corpus by 8-gram containment and splits each benchmark by seen and unseen items (Section O); one recomputes every cell's SUS at three hallucination weights and swaps the rule flag for the entity check (Section H); one classifies the language of every stored answer (Section G); one tabulates the update magnitude of every adapter against rank, including the matched-magnitude pairs of Section L; and one runs the untrained-base retrieval control of Table 7 at all four scales, whose no-RAG arm is a second generation draw of the base measurement rather than the canonical run and differs from it by at most 0.017; one that re-runs the prompt contrast on the conservative r=128 cells from re-merged adapters, which is also where the adapter-application path is measured (Section Q); and one that judges the stored runs of the high-rank cells and re-reads the utility channels split by answer language (Sections R and S). A fixed family-membership rule (Section C) is enforced in the statistics script itself, so a new replication or control arm cannot enter a family by accident, and a separate rehearsal script reports what the correction would do at family sizes beyond the current one. The judging script behind Section G also draws the stratified refusal sample for the manual re-read, emits the blind worksheet and scores the returned labels against both the rule and the judge.