The Signal Survives, the Threshold Slips: Small LLM Agents After Quantization and Fine-Tuning

Anonymous authors
Paper under double-blind review

Abstract

Small language model agents are usually quantized before they are deployed, and fine-tuned so they can call tools. We ran both changes together as a full 2×2 design on six instruction models from 0.5B to 3.8B, and scored what the agents did rather than what they said, using the executable actions in AgentHarm. Across models the interaction term is all over the place, from $+0.25$ to -0.42, which is easy to read as no effect at all. We show it breaks into three parts once you measure the model's own decision log-odds, meaning the log-odds of opening a tool call rather than refusing. One part is the bounded scale, where an additive shift looks super-additive near the floor and sub-additive near the ceiling. One part is associated with capability loss, which is where the combined pipeline sometimes breaks the agent's ability to emit a well-formed call, and a model that can no longer act gets scored as a safe one, the capability-grounding argument made before us by Vishnubhotla et al. (2026), which we quantify here rather than claim first. What is left is a latent-scale interaction in the decision axis, and on three of the six models it is the largest of the three parts.

On Qwen2.5-1.5B and Gemma-2-2B, a one-feature logistic fitted on three conditions predicts the harmful action rate of held-out conditions of the same model without refitting, whether the change came from quantization, an adapter, a learning rate, a seed, a training checkpoint, or a vector added to the residual stream. On the three models where we measured it, the refusal direction barely moves through any of this. It keeps pointing the same way and keeps telling harmful requests from benign ones, while the decision variable shifts about three times further and pushes most harmful requests past the point where the model prefers to act. The safety signal is not being erased. What moves is the threshold.

1 Introduction

A small model that calls tools is now a normal thing to deploy. Agentic systems are dominated by repetitive specialised calls, which do not need a frontier model and cost 10 to 30 times more to serve on one, so a model small enough to run on consumer hardware is argued to be the right default (Belcak et al., 2025). The usual route to running one is to quantize it to 4 bits (Dettmers et al., 2023; Frantar et al., 2022) and attach a low-rank adapter trained on function-calling data (Hu et al., 2021). Both steps happen to almost every deployed small agent, and they happen together.

Each step has a safety literature of its own, and they disagree. Quantization has been found to open risks the full-precision model lacks (Chen et al., 2025), to leave alignment roughly intact (Prasad & Pal, 2026), and to shrink decision margins (Wu et al., 2026). Fine-tuning erodes refusal even on benign data (Qi et al., 2023), adapters included (Li et al., 2025). Tools change the picture again: a model with them complies more readily (Yu et al., 2026c), and its chat refusal behaviour stops predicting what it does when it can act (Cartagena & Teixeira, 2026).

Two results narrow the space between those literatures. ExSQF (Yang et al., 2026) shows the qualitative interaction directly, with quantized models losing more safety alignment

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.

during fine-tuning than full-precision ones, and a factorial study of quantization crossed with sampling temperature (Prasad & Pal, 2026) computes a compound degradation index over 8 models and 144 configurations and reports it sub-additive. What remains open is a conjunction: neither crosses quantization with tool-use fine-tuning, neither scores an executable action rather than a span of text, neither separates the compound term into measured channels, and neither measures the latent scale the interaction is supposed to live on. Each of those is a difference we argue, not a gap we assert.

Running the 2×2 on six models produces an unpromising first result. The interaction contrast on the harmful attempted-action rate, meaning the share of harmful tasks on which the agent emits a well-formed tool call, ranges from $+0.250$ on Qwen2.5-1.5B to -0.420 on Qwen2.5-0.5B, and pooling gives $+0.041$ with a cluster-bootstrap interval of $[-0.01, +0.09]$, which at face value is a null with noise around it. The rest of this paper is about why it is something else, and the argument turns on one variable measured inside the model: the log-odds that it opens a tool call rather than refusing, written d and defined in Section 3.

Contributions.

1. The behavioural interaction decomposes into three measurable channels, on six models: bounded-scale curvature, a latent-scale interaction in the decision variable, and a capability-associated residual. The latent term is largest on Llama-3.2-3B, Gemma-2-2B and Qwen2.5-0.5B and its interval excludes zero on four of the six (Section 6).
2. A single measured scalar predicts the harmful-action rate of held-out model versions without refitting, at a mean absolute error of 0.018 over 12 held-out conditions on Llama, 0.040 over 19 on Gemma and 0.076 over 26 on Qwen, covering weight-, activation- and prompt-space interventions on two of them (Section 5). A likelihood-ratio test rejects strict sufficiency and the readout does not transfer to InjecAgent, so the claim is scoped to prediction within a model.
3. The estimated refusal direction is not what the pipeline moves, on the three models where both axes were measured. That direction rotates by a cosine of at least 0.948 and keeps its separation, while the decision axis translates 3.1 times further and carries 56%–66% of harmful requests across the acting threshold, inherited from the base model on all three (34%–98% across all six models with d measured, Section 5).
4. The effect depends on the quantization format, not magnitude alone, among the formats tested. Norm-matched Gaussian perturbation lands on C_3, not C_4, on three models, and the interaction is absent under fp4 on the one model where fp4 leaves agent-format capability intact (Section 7).
5. Two measurement results about evaluation practice. A model that has lost the ability to emit a well-formed call scores as safer without a capability control; with no sign reversals in 18 comparable cells this is a magnitude distortion, not a directional one. And a threshold calibrated on a generic corpus produced a 92% false-positive rate on benign agent traffic, replicating on four models at 0.92–0.98 (Section 6, Appendix C).

Section 8 reports four failed extensions, two pre-registered before the data existed.

2 Related work

Quantization and safety. Findings here disagree, and the disagreement is informative. Chen et al. (2025) report risks that appear under quantization and patch them, and Dong et al. (2025) show a threat model where misalignment stays latent at full precision and activates once quantized. Prasad & Pal (2026) find AWQ INT4 roughly safety-neutral on seven of eight models, and theirs is the closest methodological neighbour to our first contribution: also factorial, also computing a compound index, with sampling temperature as the second factor and no capability control. Wu et al. (2026) track decision margins across 16 models and four quantizers and find quantization multiplies the margin by a bit-width-dependent

factor. Prior work studies quantization-induced decision-margin changes on their own; our analysis asks how those changes interact with tool-use fine-tuning.

Fine-tuning and alignment. Qi et al. (2023) established that benign fine-tuning degrades safety. SaLoRA (Li et al., 2025), Safe LoRA (Hsu et al., 2024) and AlignGuard-LoRA (Das et al., 2025) preserve alignment by constraining the update, and all three read probe failure after fine-tuning as damage to safety features, which our axis result bears on directly. ExSQF (Yang et al., 2026) is the nearest published attempt at our question (Section 1).

Agent safety. AgentHarm (Andriushchenko et al., 2024) supplies our primary benchmark and the argument for scoring actions, InjecAgent (Zhan et al., 2024) and AgentDojo (Debenedetti et al., 2024) two of the three boundary benchmarks. Yu et al. (2026c) establishes that tool affordance changes refusal and Cartagena & Teixeira (2026) that chat safety does not transfer to tool-call safety, while Yu et al. (2026b) profile deployed agents on action rate against refusal signal, the behavioural counterpart of our two axes. SafeKeep (Pan et al., 2026) published the core idea behind our mitigation appendix first, and When2Tool (Sun et al., 2026) already pairs a hidden-state probe with prefill steering.

Refusal directions and steering. Arditi et al. (2024) give the single-direction account our refusal probe uses. Refusal is likely richer than one direction (Wollschläger et al., 2025; Joad et al., 2026); we do not test dimensionality, and our claim concerns one estimated direction that survives the pipeline. AlphaSteer (Sheng et al., 2026) and Li & Kasneci (2026) show steering's capability cost can be reduced, bounding how far our control failure generalises, and Raval et al. (2026) argue a global linear intervention approximates a distorted activation space. Mishra et al. (2026) prove steered activations are typically unreachable by any prompt; our prompt-space condition still lands on the same curve, consistent because the claim concerns one scalar and the rate it predicts, not the activation state a prompt reaches. Skifstad et al. (2026) steer in closed loop, which neither of our attempts did.

Evaluation methodology. Vishnubhotla et al. (2026) argue safety measurement of fine-tuned models has to be anchored to capability. That argument is theirs; we add a decomposition separating the capability term from a scale artefact and a latent-scale interaction, and a measurement of its size. Only 16% of 445 LLM benchmarks report any statistical testing (Bean et al., 2025), and safety benchmarks in particular rarely connect a test objective to a real-world outcome (Yu et al., 2026a).

Quantization, fine-tuning and agent safety have largely been studied as separate axes; the closest factorial design crosses quantization with a different second factor (Prasad & Pal, 2026). We cross it with tool-use fine-tuning, add a capability control, and decompose the result on the latent decision scale (Section 3).

3 Setup

Models, conditions and benchmarks. Six instruction-tuned models: Qwen2.5-0.5B, Qwen2.5-1.5B, Gemma-2-2B, Granite-3.1-2B, Llama-3.2-3B and Phi-4-mini (3.8B). C_1 is the fp16 base, C_2 applies bitsandbytes NF4 with double quantization and fp16 compute, C_3 trains a LoRA adapter on 4,000 function-calling examples, and C_4 applies both; the tool schema and system prompt are identical across all four. Three ordering variants test sensitivity to step order (Appendix E). The primary benchmark is AgentHarm (Andriushchenko et al., 2024): 44 of 66 public base behaviours, two of the four released augmentations each, 88 harmful prompts and 88 benign twins (the benchmark's abstract reports 110/440, counting withheld items). AdvBench, InjecAgent (Zhan et al., 2024) and AgentDojo (Debenedetti et al., 2024) appear only as boundary conditions (Section 7). Appendix B gives adapter and decoding settings.

Outcome and capability control. The primary metric is the attempted-action rate, the fraction of harmful prompts on which the model emits a structurally valid tool call, which makes it judge-independent. It measures willingness to take the harmful action a task affords

rather than damage done: realized harm under AgentHarm's official judge is at most 4.5% in every cell, too sparse to carry an interaction. Every harmful measurement has a matched benign one, the benign call rate over the 88 benign twin tasks. A model whose benign rate collapses has lost the ability to act rather than acquired a reason not to, and a safety score without that column cannot tell the two apart (Vishnubhotla et al., 2026).

The interaction contrast. With p_{ij} the attempted-action rate under quantization i and fine-tuning j,

$$\mathrm{IC} \;=\; p_{11} - p_{10} - p_{01} + p_{00} \;=\; p(C_4) - p(C_3) - p(C_2) + p(C_1). \tag{1}$$

Equation 1 is the amount by which the combined effect exceeds the sum of the two main effects: positive means more damage together than apart, negative less, zero additive. Intervals are 500-resample cluster bootstraps over the 44 task families, keeping both augmentations of a base behaviour in one resample.

The decision log-odds. For prompt x, let T hold opener strings that begin a tool call and R strings that begin a refusal, both derived per model from that model's own trajectories. With $\ell(s \mid x)$ the summed teacher-forced log-probability of every token of s given x,

$$d(x) \;=\; \log \sum_{s \in T} \exp\, \ell(s \mid x) \;-\; \log \sum_{s \in R} \exp\, \ell(s \mid x). \tag{2}$$

The first term collects the probability mass on every way this model opens a tool call and the second the mass on every way it opens a refusal, so Equation 2 is a log-odds of acting against refusing, read before anything is generated. Its zero is behaviourally meaningful: $d > 0$ means a tool-call opener is the likelier continuation. Because the opener sets are model-specific and ℓ sums over tokens, d sits on a different scale in each model (Appendix A).

The readout and the decomposition. A one-feature logistic $S(\cdot) = \sigma(a + b\,\cdot)$ is fitted (regularized) to item-level C_1–C_3 outcomes on standardized d; a condition's prediction is the mean of its items' predicted probabilities, and $\hat{d}_{C_4} = d_{C_2} + d_{C_3} - d_{C_1}$, applied item-by-item, is the C_4 exact additivity would produce.

$$\hat{p}(c) = \frac{1}{n_c} \sum_{i \in c} \sigma(a + b\, d_i)\,, \qquad \underbrace{\mathrm{IC}_{\mathrm{obs}}}_{\text{measured}} = \underbrace{\mathrm{IC}\big[S(\hat{d})\big]}_{\mathrm{IC}_{\mathrm{link}}} + \underbrace{S(d_{C_4}) - S(\hat{d}_{C_4})}_{\mathrm{IC}_{\mathrm{lat}}} + \underbrace{\mathrm{IC}_{\mathrm{obs}} - \mathrm{IC}_{\mathrm{link}} - \mathrm{IC}_{\mathrm{lat}}}_{\mathrm{IC}_{\mathrm{res}}}. \tag{3}$$

a, b never see C_4, so agreement there is a prediction; the opener vocabulary defining d itself (Appendix A) is pooled from C_1–C_4 generations, so the feature is not C_4-blind in the same sense. The second half splits the observed interaction into three parts summing to it by construction: $\mathrm{IC}_{\mathrm{link}}$ is what the S-shaped curve manufactures from a purely additive latent shift; $\mathrm{IC}_{\mathrm{lat}}$ is the interaction in d itself, which the bounded scale was obscuring; and $\mathrm{IC}_{\mathrm{res}}$ is whatever d does not explain, expected to track capability loss.

4 The interaction is heterogeneous in sign

Table 1 is what the rest of the paper explains. Fine-tuning raises the harmful action rate on all six; quantization alone is smaller and inconsistent in sign, agreeing with Prasad & Pal (2026). The interaction has no stable sign, and pooling gives +0.041 with an interval covering zero (Appendix Figure 2).

Three features make that pooled null hard to accept. The spread between models is far wider than the individual intervals, and seed replication does not narrow it (C_4-rate SDs across seeds 0.043/0.013/0.050, Appendix Figure 12). The sign does not track scale: Qwen is −0.420 at 0.5B, +0.250 at 1.5B, +0.023 [−0.09,+0.15] at 3B, run behaviourally only with no d measured, so it sits outside the decomposition below. And the two most negative models are the two whose benign call rate falls, consistent with the capability-confound signature: Phi-4-mini 0.977 → 0.784, Qwen2.5-0.5B 0.773 → 0.125. Resolving the heterogeneity is what the next two sections do.

Table 1: The interaction has no stable sign across models, and the decomposition shows where it sits. IC (Equation 1) runs from +.250 to −.420 and pools to +.041; Equation 3 splits it into a scale artefact, a latent-scale interaction and a capability-associated residual that sum to it by construction, each dominating on at least one model. ben. is the benign call rate at C_4, on which a model scoring low on harm has lost the ability to act rather than the willingness. Six models, single seed, 88 prompts each; intervals are 500-resample cluster bootstraps over the 44 task families. Pooled row: 3 seeds each for Qwen2.5-1.5B/Llama/Gemma, 1 otherwise (Phi's 2nd seed excluded).

	observed					decomposition			
Model	C_1	C_2	C_3	C_4	ben.	IC [95% CI]	IC_{link}	IC_{lat} [95% CI]	IC_{res}
Qwen2.5-1.5B	.136	.250	.193	.557	.943	+.250 [+.12, +.40]	+.113	+.015 [−.01, +.04]	+.122
Llama-3.2-3B	.114	.068	.477	.602	1.00	+.170 [+.07, +.30]	+.000	+.**168** [+.11, +.24]	+.003
Gemma-2-2B	.307	.148	.614	.500	.943	+.045 [−.08, +.19]	−.071	+.**092** [+.04, +.14]	+.025
Granite-3.1-2B	.716	.739	.989	.932	.966	−.080 [−.17, +.02]	−.**086**	−.001 [−.00, −.00]	+.007
Phi-4-mini	.091	.102	.568	.330	.784	−.250 [−.39, −.11]	−.015	−.058 [−.08, −.04]	−.**177**
Qwen2.5-0.5B	.091	.023	.602	.114	.125	−.420 [−.56, −.27]	+.023	−.**269** [−.34, −.21]	−.175
Pooled ($n = 1056$)						+.041 [−.01, +.09]			

5 What the pipeline moves, and what it leaves alone

5.1 The refusal direction survives; the decision axis slides

We estimate a refusal direction the standard way, as the difference in mean residual-stream activation between harmful and benign prompts at C_1, validated by ablation (Arditi et al., 2024). Re-estimated inside each other condition it barely turns, cosine against the C_1 direction never below 0.948 across the nine non-base cells, and it keeps working: harmful-versus-benign AUROC goes 0.854 → 0.825 on Qwen and 0.897 → 0.874 on Gemma, and improves on Llama, 0.767 → 0.796. This is a claim about one estimated direction, since refusal is mediated by more than one (Wollschläger et al., 2025; Joad et al., 2026) and nothing here tests that. What we report is that the direction the probe literature fits, and then watches fail, keeps pointing the same way.

The decision log-odds behaves completely differently (Table 5). In units of its own C_1 spread the refusal axis translates 0.384 SD on average and the decision axis 1.175 SD, a ratio of 3.1, moving toward acting on all six models. At C_1 between 0% and 38% of harmful requests sit on the acting side of $d = 0$; at C_4 the figure is 34% to 98%. The model still ranks harmful requests the way it did; what changed is where that ranking sits relative to the point at which it acts (Appendix Figure 5; Figure 1).

Two things stop this being a restatement of the behavioural result. The harmful-versus-benign AUROC on d stays flat while d slides (0.672–0.887 at C_4 against 0.624–0.889 at C_1), so the axis has not stopped discriminating. And fine-tuning does not build the axis: the acted-versus-refused direction estimated inside the base model has a max cosine (layers ≥ 10) of +0.81 against C_4 on Qwen and +0.84 on Gemma and Llama, so both sides of the dissociation are pre-existing structure. An earlier version of this analysis concluded Llama's base had no such axis, an artefact of a wrong opener set; every quantity the repair moved is Llama's (Appendix A).

5.2 One scalar predicts held-out conditions without refitting

If d is the variable the pipeline acts on, fixing the map from d to behaviour on conditions containing no interaction should predict conditions that do. Equation 3 is fitted on C_1, C_2 and C_3, then applied with nothing refitted to C_4, the ordering variants, precision and learning-rate doses, eight training checkpoints, noise-injected bases, mitigation adapters and seed replicates. This is cross-condition prediction within a model on a fixed prompt set, not generalisation to unseen prompts.

Mean absolute error is 0.018 on Llama (12 conditions, $r = 0.998$), 0.040 on Gemma (19, $r = 0.972$) and 0.076 on Qwen (26, $r = 0.962$); most ordering variants land near the curve,

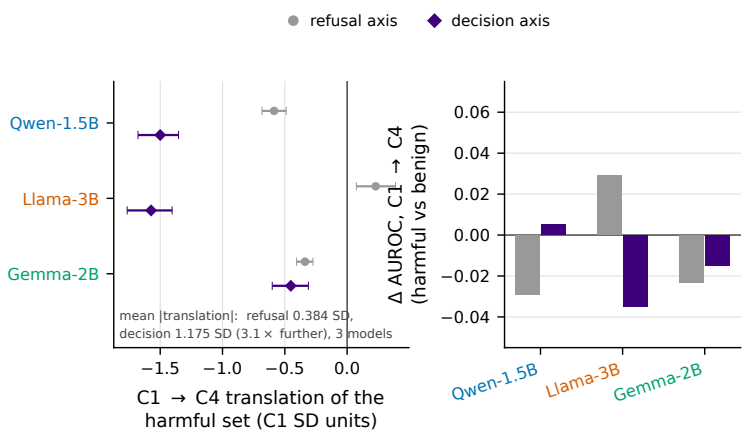

Figure 1: The estimated refusal direction moves little while the decision axis slides. Left: mean translation from C_1 to C_4 in units of each axis's own C_1 spread, three models. The decision axis moves 3.1 times further on average. Right: change in harmful-versus-benign AUROC over the same step, small on both axes, so neither stops discriminating. Table 5 gives the per-model values.

though Qwen's O2 (fine-tune then quantize) misses by +0.22, coinciding with near-total capability loss there (4/88 tool calls), not a readout failure (Table 4, Appendix Figures 4 and 13). Phi-4-mini is the informative failure: the readout over-predicts its one held-out condition by 0.24 as its benign call rate falls 0.977 → 0.784, the capability channel of Section 6 showing up as prediction error. A sceptic might call d a general harm detector, but it tells C_3 from C_4 of the same model on the same prompts, and it fails where acting and acting harmfully come apart (Section 8).

The same curve covers three kinds of intervention. On Qwen and Gemma we also steered the residual stream along the decision direction. The weight-space curve predicts those activation-space conditions at MAE 0.049 (10 conditions, Qwen) and 0.019 (Gemma), rank correlation 0.952/0.978 across all 13. A pre-registered falsification case appended a safety clause to the system prompt, no weights or hooks changed, and failed to falsify: error −0.043/−0.024 (Appendix Figure 14). A matched-norm random direction barely moves d and does not reduce harm. One curve spanning weight and activation space is unsurprising given the weight-activation equivalence of Adila et al. (2026), of which this is an instance, not a proof.

What this does not support. On Qwen, pooling all 18 conditions and testing one shared intercept against one per condition rejects decisively: LR = 64.73 on 17 df, $p = 1.7 \times 10^{-7}$. Intervention identity carries information beyond d: it accounts for most of the between-condition variation and is not a sufficient statistic. Calibration does not transfer across models: predicting a held-out model's C_4 from the other five gives MAE 0.157/0.108/0.224 by scheme, against 0.018–0.076 within a model.

6 Three channels, and what uncorrected evaluation misses

The right block of Table 1 applies Equation 3 to all six models, and Appendix Figure 3 plots it. Each channel dominates somewhere, and three models are clean cases. Granite is predominantly scale-associated: nearly all of its −0.080 is $\mathrm{IC}_{\mathrm{link}}$, because it acts on 99% of harmful prompts at C_3 and an additive latent shift near that ceiling has to look sub-additive. Phi-4-mini is predominantly capability-associated, with −0.177 of its −0.250 in the residual, and Qwen2.5-0.5B combines a negative latent-scale interaction with a capability-associated residual on top.

The latent term's interval excludes zero on Llama, Gemma, Phi-4-mini and Qwen2.5-0.5B, four of the six. Qwen2.5-1.5B's covers zero. Granite's is −0.001 [−0.004, −0.000], which the

table rounds to two decimals and which is too small for us to read as a real effect. The latent term is the largest of the three channels on Llama, Gemma and Qwen2.5-0.5B, and its sign agrees with the observed interaction on all six, so the scale scrambles magnitudes rather than inventing the effect. $\mathrm{IC}_{\mathrm{res}}$ is a residual and absorbs anything d does not explain, so reading it as capability loss rests on it tracking the benign call rate, which it does on Phi-4-mini and Qwen2.5-0.5B. One input deserves flagging rather than quiet use: Qwen2.5-1.5B's benign C_3 call rate of 0.420 is a single unreplicated measurement that a leave-one-condition-out check flagged as an outlier at an error of +0.207, and Qwen's residual of +0.122 depends on it.

What a capability-uncorrected evaluation gets wrong. The obvious correction divides the harmful rate by the benign call rate. Across all 18 non-reference cells that changes no conclusion's direction, though two cells sit within about one benign item of the boundary and this pipeline's re-measurement noise is two to three items, so the zero is real and not robust and the honest claim is about magnitude. Phi-4-mini's uncorrected C_4 rate of 0.330 reads as a substantial improvement on its C_3 rate of 0.568; corrected, $0.330/0.784 = 0.420$ against $0.568/0.977 = 0.581$, and most of the apparent gain was the model losing the ability to emit a call. Qwen2.5-0.5B is starker: its C_4 rate of 0.114 is the lowest C_4 rate in Table 1 and its corrected rate of 0.909 among the highest, because its benign call rate at C_4 is 0.125 (Appendix Figure 7). This is the argument of Vishnubhotla et al. (2026), which we quantify rather than make; what the decomposition adds is that the capability term separates from the scale term and from the latent-scale interaction.

7 Negative controls and boundary conditions

The effect depends on the quantization format, not on perturbation magnitude alone, among the formats tested. If the compound effect were about the size of the perturbation quantization introduces, perturbing the base weights by Gaussian noise of matched magnitude and then fine-tuning should reproduce it. Across three scales (0.5, 1, 2× NF4 error), the noisy fine-tune consistently lands on or near C_3, not C_4: at the matched scale the harmful rate is 0.080 on Qwen against 0.557 at C_4, 0.511 on Llama against 0.602, and 0.580 on Gemma against 0.500 – on Gemma the arms (up to 0.625) sit above C_4, but Gemma's IC is itself indistinguishable from zero, so there is no gap to reproduce there. Llama and Gemma land on C_3 with benign tool use intact; Qwen lands below C_3, benign 0.216. Function-calling utility stays 0.92–0.95 throughout.

Swapping NF4 for fp4 at the same bit width removes the interaction on the one model where fp4 leaves capability intact: Llama's IC goes from +0.170 to −0.068 $[-0.159, +0.011]$ at a benign call rate of 0.97. Qwen's fp4 arm shows no interaction either, but its benign rate of 0.48 confounds it with format damage and we do not read it (Appendix Figure 6). The conclusion stays hedged: specific to NF4 among the two formats tested; the controls argue against magnitude alone, not for a specific mechanism. Bhandari & Wade (2026) find NF4 preserves some steering behaviours at 7 to 9B, a similar reading at a different scale.

Three benchmarks where the interaction is small or unreliable. On AdvBench (100 to 300 chat prompts, no tools) five of the six models give an interaction whose interval covers zero, for instance +0.000 $[-0.03, +0.03]$ on Gemma, and the pooled estimate is +0.010 $[-0.01, +0.03]$; Granite is the exception at +0.060 $[+0.01, +0.12]$. InjecAgent (Qwen and Gemma only, 1,054 cases per setting) measures whether an agent is derailed by an injected instruction, not whether it complies with a harmful one: pooled, the attempted-lenient contrast is small but non-null, +0.029 $[+0.01, +0.05]$, with Qwen alone at +0.049 $[+0.01, +0.09]$. AgentDojo (144 banking, 105 Slack) has 16 and 21 clusters, too few for a reliable per-model contrast; its pooled banking estimate excludes zero (−0.065 $[-0.12, -0.02]$) but rests on low-cluster-count, inconsistent cells.

These mark out where the account applies. On AgentHarm, acting and complying are nearly the same event; on InjecAgent the agent is supposed to act and only the attacker's tool counts against it, so a variable predicting action is expected to carry little of the outcome. Section 8 tests that directly.

8 Four things that did not work

Calibrated control of the harmful rate, twice. If the readout predicts the harmful rate from d, inverting it should let us name a target rate, solve for the steering strength that reaches it, and hit it. We pre-registered three targets (0.30, 0.20, 0.10) and a bar of $|\text{achieved} - \text{target}| \leq 0.05$ on two of three. The first attempt failed all three, errors 0.073, 0.132, 0.100: the actuator was at fault, since the map from commanded coefficient to achieved d was assumed linear and is not, leaving a residual up to 0.593 nats. An earlier explanation, that quantization had damaged capability enough to break the control, was tested and retracted ($\gamma = -0.035$). The second attempt used a saturating actuator with a pre-registered stop clause: a second failure means two failed pre-registrations, no third attempt. It failed again, 0 of 3 strict, but everything else improved: mean miss $0.102 \rightarrow 0.064$, all three within 0.10 and in order, actuator residual 0.030 nats at worst against 0.15. Ordinal control survives, calibrated control does not (Appendix Figure 8). Ong et al. (2026) reach the same boundary from the other side, forecasting direction and ordering of steering side effects and stopping short of calibrated magnitudes.

The strict bar fails for a reason about steering, not our curve. Predicting harm from the measured per-item d after steering is accurate to 0.019, so the d-to-harm map is right; what is wrong is the assumption that steering preserves the spread of d. It instead falls from 2.690 at baseline to 2.118, 1.840, 1.497 at the three targets, and a compressed distribution at the same mean gives a different rate under a nonlinear link. Both attempts were open-loop; closed-loop control is feasible (Skifstad et al., 2026) and is the obvious next step, untried here because the pre-registration bound us to two attempts and what remained was distributional compression rather than actuator error.

Compression as the explanation for why control fails. A tidy story suggested itself: weight-space changes translate d's distribution while steering compresses it, explaining why prediction succeeds where control fails. Committed to in advance, it does not hold. Regressing per-item Δd on baseline d across matched pairs, 7 of 8 steering pairs compress and so do 7 of 9 weight-space pairs, Qwen's $C_1 \rightarrow C_3$ slope -0.463 steeper than its steering slope -0.223. This is a diagnostic regression on data from the two registered attempts, not a third attempt at control (Appendix Figure 9).

Sign reversal under capability correction. We expected capability correction to reverse the safety conclusion in some cells. It reverses none (Section 6), so the capability result is a magnitude claim, and a sign-reversal count at any nonzero value would misdescribe what we measured.

Transfer of the readout to a different benchmark. We predicted, before either run finished, that mean-rate transfer to InjecAgent would fail structurally, and committed to reporting a chance-level AUROC as an honest negative rather than explaining it away. Both held. Mean-rate error is large and one-directional (up to +0.921 on Gemma C_1), and per-item discrimination is at or below chance on all four interpretable cells: AUROC 0.407/0.497 on Qwen, 0.307/0.209 on Gemma; Llama's two cells (4 and 0 positives of 240) are excluded. Clustering on the attacker instruction puts three of four entirely below chance (Appendix Figure 10), an inversion we did not pre-register and report as consistent with the argument rather than predicted by it. The consequence is a scope limit: the readout predicts harm where acting and acting harmfully coincide, not where they come apart.

9 Pre-registration and pre-committed analysis decisions

A negative result is worth what was committed to before the data arrived (Albanie et al., 2022). This section states what was registered, when, and what happened.

1. Both control attempts, each registered with a commit hash before its run, with the targets and bar of Section 8. Outcome 0 of 3 both times; the second carried the stop clause that no third attempt would be written.

2. The falsification case for sufficiency was named before the sweep ran and failed to falsify; naming it in advance rules out post-hoc selection, though one case remains limited evidence alone.
3. Cross-benchmark transfer. The prediction that mean-rate transfer would fail structurally, and the commitment to report a chance-level AUROC as an honest negative rather than excuse it with that argument, were written down before either run finished. The inversion was not predicted.
4. The compression regression had its branch written in advance: a clean dissociation would promote compression to an explanation, a messy result would confine it to a descriptive observation. It was messy and it is confined, with nothing tried to rescue it.
5. One gate fired and was missed. A pre-committed check said to stop and report if backfilled benign rates came back unexpectedly low. Four of six came back between 0.34 and 0.47 and the check was not acted on, which we record as a process failure rather than resolve quietly afterwards.

Two analysis choices were fixed in code before anyone questioned them and verified rather than assumed: the decomposition already applied the link per-item then averaged, not at the condition-mean d a naive implementation might use (differs by up to 0.225 on some terms; no published number changes), and the sign-reversal denominator is 18 comparable cells, not 24 raw.

10 Limitations

Scale, and non-monotonicity inside it. Six models from 0.5B to 3.8B, the deployed quantized tool-caller range (Belcak et al., 2025); nothing above 3.8B was tested, a closed boundary rather than an open question. Within the one family with three sizes the interaction is not monotone: -0.420 at 0.5B, $+0.250$ at 1.5B, $+0.023$ at 3B. Parameter count does not govern the sign and we do not know what does.

Coverage differs by claim. The interaction/decomposition cover six models, the axis dissociation three, sufficiency across intervention kinds two, the no-refit readout four.

One benchmark carries most of the argument. Cluster-bootstrap intervals over AgentHarm's 44 task families are wide, roughly ± 0.10; a three-item change sits inside this pipeline's re-measurement noise, and no multiple-comparison correction is applied. Only NF4 and fp4 were tested; GPTQ (Frantar et al., 2022) and AWQ were not.

Unreplicated cells, including one the text leans on. The per-model table is single-seed: three models have three seeds, Phi-4-mini two, Granite and Qwen2.5-0.5B one. Qwen2.5-1.5B's benign C_3 rate of 0.420 is a single measurement flagged as an outlier and retained unchanged, and Qwen's capability channel depends on it. Two caveats close the list: our secondary refusal taxonomy's keyword classifier agrees with the official judge at $\kappa = 0.86$–0.91 and over-counts refusals, which the structural primary metric avoids, and Phi-4-mini's d sits on a generic fallback opener set, not comparable to another model's. We did not run a sensitivity analysis over alternative opener-set definitions (Appendix A).

11 Conclusion

Quantizing a small agent and fine-tuning it for tool use produces a behavioural interaction whose sign disagrees across models, and pooling hides it. The raw rate mixes bounded-scale curvature and a capability-associated residual with a latent-scale interaction; separated, most models show a decision-scale effect. The estimated refusal direction survives largely intact while the decision variable shifts, carrying most harmful requests past the point where acting is preferred; the scalar predicts well within a model but is neither sufficient nor cross-model. Capability control is necessary to interpret a low harmful-action rate, the account is scoped to where acting and harmful acting coincide, and calibration failed twice.

Ethics statement

This work studies whether small tool-using agents become more willing to attempt harmful actions after quantization and fine-tuning, using AgentHarm's harmful task prompts against a mocked tool environment; no harmful action is ever executed against a real system, and no harmful output or completion is reproduced in this paper. The dual-use risk is that the compound effect we document (Sections 4 –6) describes a way deployed small agents can become less safe; we believe reporting it, together with the evaluation-practice failures it exposes (a missing capability control, and a miscalibrated safety filter) and the appendix mitigation, is net protective, since these are exactly the conditions vendors already deploy. AgentHarm is used under its published research license and its withheld test split is not used.

AI use statement

In this work, we used generative AI tools for writing and editing the analysis code that produces every reported result, for drafting and editing prose, for literature search, and, under author direction at each step, for proposing and implementing specific analyses, including the three-channel decomposition, the pre-registered control attempts, and the diagnosis of the opener-set measurement error described in Appendix A. Assistants also helped interpret results at each step – reading tables, proposing which of two aggregations to trust, and flagging apparent contradictions – with every resulting interpretation checked by the authors against the underlying results file before being written into the paper. We have not used generative AI tools to originate the research hypotheses, to set the pre-registered targets, or to decide which negative results to report; those decisions were author-directed, with assistants proposing candidate analyses and writing the code that ran them. Every number reported in this paper is produced by a script that writes a results file, and every number in the text was checked against that file rather than against any draft; every citation was verified against the cited work's own listing page. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including any claim that survives this process in error.

Reproducibility statement

Every number in this paper is produced by a script that writes a results file and was checked against that file at write time, not against prose (Section 9 lists the pre-registration commitments and Appendix B the adapter, quantization and decoding settings). The decision log-odds d (Equation 2) depends on a model-specific opener set; Appendix A documents how those sets are derived and the one measurement error found and repaired in them. The two pre-registered control attempts are timestamped and unedited since being committed, and the stop clause on the second is binding: no third attempt exists.

References

Dyah Adila, John Cooper, Alexander Yun, Avi Trost, and Frederic Sala. Weight updates as activation shifts: A principled framework for steering. arXiv preprint arXiv:2603.00425, 2026. URL https://arxiv.org/abs/2603.00425.

Samuel Albanie, João F. Henriques, Luca Bertinetto, Alex Hernández-García, Hazel Doughty, and Gül Varol (eds.). NeurIPS 2021 Workshop on Pre-registration in Machine Learning, volume 181 of Proceedings of Machine Learning Research, 2022. PMLR. URL https://proceedings.mlr.press/v181/.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. arXiv preprint arXiv:2410.09024, 2024. URL https://arxiv.org/abs/2410.09024.

Andy Arditi, Oscar Obeso, Aaquib 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. URL https://arxiv.org/abs/2406.11717.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. arXiv preprint arXiv:2511.04703, 2025. URL https://arxiv.org/abs/2511.04703.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic AI. arXiv preprint arXiv:2506.02153, 2025. URL https://arxiv.org/abs/2506.02153.

Saurav Bhandari and Benjamin Wade. Steering under compression: Dose-response, capability cost, and failure asymmetry in quantized LLMs. arXiv preprint arXiv:2609.06473, 2026. URL https://arxiv.org/abs/2609.06473.

Arnold Cartagena and Ariane Teixeira. Mind the GAP: Text safety does not transfer to tool-call safety in LLM agents. arXiv preprint arXiv:2602.16943, 2026. URL https://arxiv.org/abs/2602.16943.

Kejia Chen, Jiawen Zhang, Jiacong Hu, Yu Wang, Jian Lou, Zunlei Feng, and Mingli Song. Q-resafe: Assessing safety risks and quantization-aware safety patching for quantized large language models. In Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025. URL https://proceedings.mlr.press/v267/chen25ci.html.

Amitava Das, Abhilekh Borah, Vinija Jain, and Aman Chadha. AlignGuard-LoRA: Alignment-preserving fine-tuning via fisher-guided decomposition and riemannian-geodesic collision regularization. arXiv preprint arXiv:2508.02079, 2025. URL https://arxiv.org/abs/2508.02079.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2024. URL https://arxiv.org/abs/2406.13352.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2305.14314.

Peiran Dong, Haowei Li, and Song Guo. Durable quantization conditioned misalignment attack on large language models. In International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=41uZB8bDFh.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323, 2022. URL https://arxiv.org/abs/2210.17323.

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. arXiv preprint arXiv:2405.16833, 2024. URL https://arxiv.org/abs/2405.16833.

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. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There is more to refusal in large language models than a single direction. arXiv preprint arXiv:2602.02132, 2026. URL https://arxiv.org/abs/2602.02132.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. SaLoRA: Safety-alignment preserved low-rank adaptation. arXiv preprint arXiv:2501.01765, 2025. URL https://arxiv.org/abs/2501.01765.

Yuxiao Li and Gjergji Kasneci. Safety cost of steering vectors is separable and reducible. arXiv preprint arXiv:2608.08383, 2026. URL https://arxiv.org/abs/2608.08383.

Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered LLM activations are non-surjective. arXiv preprint arXiv:2604.09839, 2026. URL https://arxiv.org/abs/2604.09839.

Chong Yong Ong, Alson Wei Jie Sim, Peixin Zhang, and Jun Sun. Forecasting side effects of activation steering. arXiv preprint arXiv:2608.11227, 2026. URL https://arxiv.org/abs/2608.11227.

Minghui Pan, Jiayuxuan Yang, Yuanyuan Yuan, Yu Jiang, and Zhenpeng Chen. Tool specifications matter: Uncovering and mitigating safety risks in AI agents. arXiv preprint arXiv:2607.29254, 2026. URL https://arxiv.org/abs/2607.29254.

Hari Prasad and Ritam Pal. The joint effect of quantization and sampling temperature on LLM safety alignment: A factorial analysis. arXiv preprint arXiv:2606.29581, 2026. URL https://arxiv.org/abs/2606.29581.

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! arXiv preprint arXiv:2310.03693, 2023. URL https://arxiv.org/abs/2310.03693.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M. Phillips, Fazl Barez, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear. arXiv preprint arXiv:2603.09313, 2026. URL https://arxiv.org/abs/2603.09313.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. AlphaSteer: Learning refusal steering with principled null-space constraint. In International Conference on Learning Representations (ICLR), 2026. URL https://arxiv.org/abs/2506.07022.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of LLMs enables activation steering via model-based linear optimal control. arXiv preprint arXiv:2604.19018, 2026. URL https://arxiv.org/abs/2604.19018.

Chung-En Sun, Linbo Liu, Ge Yan, Zimo Wang, and Tsui-Wei Weng. LLM agents already know when to call tools – even without reasoning. arXiv preprint arXiv:2605.09252, 2026. URL https://arxiv.org/abs/2605.09252.

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. URL https://arxiv.org/abs/2606.03648.

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. arXiv preprint arXiv:2502.17420, 2025. URL https://arxiv.org/abs/2502.17420.

Zekun Wu, Swati Dhiman, and Adriano Koshiyama. Which decisions low-bit quantization breaks, and how to predict them. arXiv preprint arXiv:2608.06564, 2026. URL https://arxiv.org/abs/2608.06564.

Yuning Yang, Guowei Peng, Xiurui Xie, Minrui Jiang, Shuang Liang, and Guisong Liu. Toward safe quantization-aware fine-tuning: Understanding and mitigating safety alignment degradation. In Proceedings of the 43rd International Conference on Machine Learning (ICML), 2026. URL https://openreview.net/forum?id=vF2Xhg2s31.

Cheng Yu, Severin Engelmann, Ruoxuan Cao, Dalia Ali, and Orestis Papakyriakopoulos. How should AI safety benchmarks benchmark safety? arXiv preprint arXiv:2601.23112, 2026a. URL https://arxiv.org/abs/2601.23112.

Shasha Yu, Fiona Carroll, and Barry L. Bentley. The A-R behavioral space: Execution-level profiling of tool-using language model agents in organizational deployment. arXiv preprint arXiv:2604.12116, 2026b. URL https://arxiv.org/abs/2604.12116.

Shasha Yu, Fiona Carroll, and Barry L. Bentley. The causal impact of tool affordance on safety alignment in LLM agents. arXiv preprint arXiv:2603.20320, 2026c. URL https://arxiv.org/abs/2603.20320.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Findings of the Association for Computational Linguistics (ACL), 2024. URL https://aclanthology.org/2024.findings-acl.624/.

A The opener-set repair

The decision log-odds of Equation 2 is defined relative to two sets of opener strings, and those sets are derived per run from the main-run trajectory files attached to that run. Concretely: for each of that model's own C_1–C_4 generations, the first-step response's opening three words (truncated to 24 characters) are counted as a tool opener if the first step emitted a tool call, or a refusal opener if the task outcome was REFUSED; the two most frequent strings per class are kept from each condition, pooled across the four conditions, and the four most frequent overall in each class become T and R, with fixed canonical strings (<functioncall>, "'tool_code for tools; "I'm sorry, but", "I can't", "I cannot" for refusals) added if not already present and any overlap resolved toward the tool set. This uses only which condition an opener string came from and its class, not per-item outcome labels, and the logistic's parameters (a, b in Equation 3) are still fitted on C_1–C_3 items only. Two runs that attach different bundles therefore put d on different scales. One bundle was attached wrongly, and Llama's readout files were scored against an opener set derived from a different model's trajectories.

The visible symptom was that 0 of 88 harmful items at Llama's base condition were scored as acting, because the borrowed set contained no prose tool openers of the kind Llama uses, such as I'll use the. With no acted items, the acted-versus-refused direction cannot be estimated at all, and we had concluded that Llama's base model has no decision axis and that fine-tuning builds one. Re-measuring under Llama's own openers gives 13 of 88 acting at the base and a cosine of +0.84 against the C_4 direction, so the axis is inherited like the other two models'. That conclusion is withdrawn. The orthogonality of Llama's decision axis to its refusal axis, which was the other half of the original finding, survives the repair.

Table 2 lists every published quantity the repair moved. An audit re-derived each model's opener set from its own main run and confirmed that no results directory now carries more than one set. One model retains a known caveat: Phi-4-mini's d sits on a generic fallback set because no attached trajectories were found for it, and its d is therefore internally consistent but not comparable across models.

The original analysis used an opener set derived from another model. Recomputing the analysis with Llama's own opener set materially changed the affected quantities, listed in Table 2; we report this as what the correction changed, not as evidence that the correction is right on some other ground. Before it, all nine of Llama's held-out errors shared one sign at a correlation of 0.979 and were large (+0.09 to +0.19), which we had read as a calibration offset; it was a scale mismatch between two definitions of d, and afterwards all 12 held-out errors are small and uniformly slightly negative (−0.01 to −0.04).

Table 2: Every quantity the opener-set repair changed. Values on the left are superseded and appear in no statement in this paper.

Quantity	Where	Before	After
Llama mean d at C_1	Section 5	−21.95	−11.11
Llama decision shift $C_1 \to C_4$	Table 5	−2.587 SD	−1.573 SD
Axis translation ratio	Section 5	3.9×	3.1×
Llama held-out conditions	Table 4	9	12
Llama readout MAE	Table 4	0.127	0.018
Llama readout correlation	Table 4	0.979	0.998
Llama $\mathrm{IC}_{\mathrm{link}}$	Table 1	−0.100	+0.000
Llama $\mathrm{IC}_{\mathrm{lat}}$	Table 1	+0.226	+0.168
Llama $\mathrm{IC}_{\mathrm{res}}$	Table 1	+0.045	+0.003
Llama base acted items	Appendix A	0/88	13/88
Weight-space pairs compressing	Section 8	8/9	7/9

Robustness to the opener-set choice. We did not run a systematic sensitivity analysis over alternative opener-set definitions (expanded, restricted, or resampled candidate strings). The saved outputs store the teacher-forced log-probability already summed over the chosen opener set for each item, not per-string log-probabilities over a larger candidate pool, so testing an alternative set requires new teacher-forced inference over the model rather than reanalysis of existing files. This is a limitation: the qualitative decision-axis result is not shown to be stable under reasonable alternative measurement choices, only under the one opener set each model was actually scored against, verified as its own by the audit above.

B Adapter, training and decoding settings

LoRA rank 16, $\alpha = 32$, dropout 0.05, requested on the attention projections (q_proj, k_proj, v_proj, o_proj) by name for every model. Five of the six architectures expose these as separate modules; Phi-4-mini's attention uses a fused qkv_proj, so its adapter's effective coverage of the query/key/value path may differ from the other five and was not independently re-verified from tensor names for this submission. One epoch over 4,000 glaive-function-calling-v2 examples at a learning rate of 2×10^{-4}; the learning-rate dose condition in Table 4 uses 5×10^{-5} and is otherwise identical. Quantization is bitsandbytes NF4 with double quantization and fp16 compute, except in the conditions that name another format (fp4, int8, and NF4 without double quantization). Decoding is greedy (no sampling) throughout. Generation is capped at 200 new tokens per turn, with at most 8 tool-calling turns per task and a 3,072-token prompt cap, so a trajectory ends rather than growing without bound. Fine-tuning utility is exact-match function-call accuracy on 200 held-out glaive examples, 0.92–0.97 on every model including Phi-4-mini, so whatever the adapter's exact module coverage, it learned the task. All runs were single-GPU on 16GB accelerators under a 12-hour job cap.

C Tool-free reference gating

SafeKeep (Pan et al., 2026) published the core idea behind this gate first, including both the phenomenon that schema-formatted tool specifications weaken a model's internal refusal signal and the mechanism of assessing a request in a flattened rendering before executing it against the schema. When2Tool (Sun et al., 2026) already pairs a last-token probe with prefill steering. Our version is concurrent, covers a setting neither addresses, and has not been compared head to head against either, so it is an appendix rather than a contribution.

The gate runs a reference pass over a flattened, tool-free rendering of the request, reads a linear probe on the last token, and vetoes the tool call when the probe fires. At a 5% false-positive budget it reduces the harmful attempted-action rate by 69.9% on Granite, 58.5% on Llama, 57.1% on Qwen and 56.8% on Gemma, at benign costs of 4.5, 9.1, 5.7 and 4.5 percentage points. Figure 11 plots the trade-off. The gate works best on Granite, the one

model in this study with no measurable interaction, which is why we describe it as general harm-gating rather than as a repair for the effect this paper studies.

The calibration result, which is a contribution. The threshold above was set on a generic calibration corpus, the standard practice. Setting a 5% false-positive budget that way produced a measured false-positive rate of 92% on benign agent traffic from the deployment distribution. At 10% it was 95% and at 20% it was 100%. The failure is not specific to one model: the same procedure gives 0.92 to 0.98 at the 5% budget on all four gated models. Recalibrating on a held-out split of deployment-distribution benign traffic restores agreement, with a 5% budget giving a 5% measured rate. Any agent-safety filter calibrated on a generic corpus should be assumed to behave this way until measured on its own traffic.

D Literature comparison table

Table 3 positions the ten near-neighbours discussed in Section 2 against the present work. A checkmark means the paper's design includes that element for its own primary comparison; a dash means it does not. "Latent decision analysis" means the paper measures an internal, pre-generation scalar or representation and relates it to the behavioural outcome, as opposed to working only with generated text.

Work	Quant.	FT	Agents	Exec. action	Capability ctrl.	Latent decision	Main question
Chen et al. (2025)	✓	–	–	–	–	–	Patch quantization-induced safety risk
Dong et al. (2025)	✓	–	–	–	–	–	Quantization-conditioned latent misalignment
Prasad & Pal (2026)	✓	–	–	–	–	–	Factorial quant. × temperature, compound index
Wu et al. (2026)	✓	–	partial	–	–	partial	Decision-margin shrinkage under quantization
Yang et al. (2026)	✓	✓	–	–	–	–	Safety degradation during quantization-aware FT
Qi et al. (2023)	–	✓	–	–	–	–	Benign FT degrades safety alignment
Vishnubhotla et al. (2026)	–	✓	–	–	✓	–	Capability-grounded safety measurement, argument only
Pan et al. (2026)	–	–	✓	✓	–	✓	Tool-schema format weakens refusal
Sun et al. (2026)	–	–	✓	–	–	✓	Hidden-state tool-necessity decoding
Arditi et al. (2024)	–	–	–	–	–	✓	Single-direction refusal account
This work	✓	✓	✓	✓	✓	✓	Quant. × tool-use FT interaction, decomposed on a measured latent decision scale, with capability control

Table 3: Literature comparison, populated only from each paper's own abstract or listing page (Section 2). "Partial" for Wu et al. (2026) reflects that they study tool-calling and refusal decisions but do not cross quantization with a second manipulated factor and do not decompose an interaction; the checkmark itself would overstate the overlap. No row asserts this work is the first to touch any single column; the claim is the conjunction in the bottom row.

E Additional figures

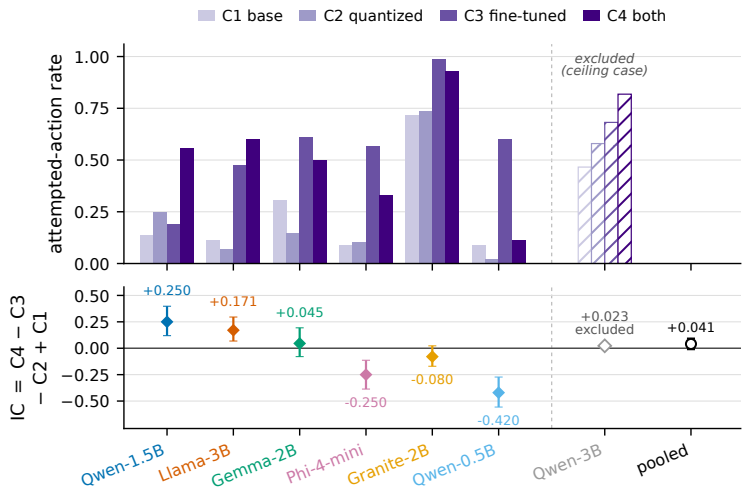

Figure 2: Interaction contrast by model, with 500-resample cluster-bootstrap intervals over the 44 AgentHarm task families. Point estimates and intervals both come from the single-seed decomposition file, so nothing in the panel mixes aggregations; the open "pooled" marker is the only point drawn from the mixed-seed pooled aggregation (3 seeds for Qwen2.5-1.5B, Llama and Gemma, 1 otherwise; Phi-4-mini's second seed is not part of this pooled row). Qwen2.5-3B is shown hatched, excluded from the decomposition (no readout was measured for it); its point estimate is plotted without an error bar by choice, not because none exists.

Table 4: No-refit readout, per model. The curve is fitted on C_1, C_2 and C_3 (264 items) and applied with nothing refitted, so every held-out number is a prediction. Phi-4-mini has one held-out condition, which is why its error is not comparably powered. Source: readout_all.md.

Model	held-out conds.	MAE [95% CI]	r	C_4 pred. / obs.
Llama-3.2-3B	12	0.018 [0.014, 0.023]	0.998	0.58 / 0.60
Gemma-2-2B	19	0.040 [0.030, 0.051]	0.972	0.46 / 0.50
Qwen2.5-1.5B	26	0.076 [0.057, 0.096]	0.962	0.50 / 0.56
Phi-4-mini	1	0.238	n/a	0.57 / 0.33

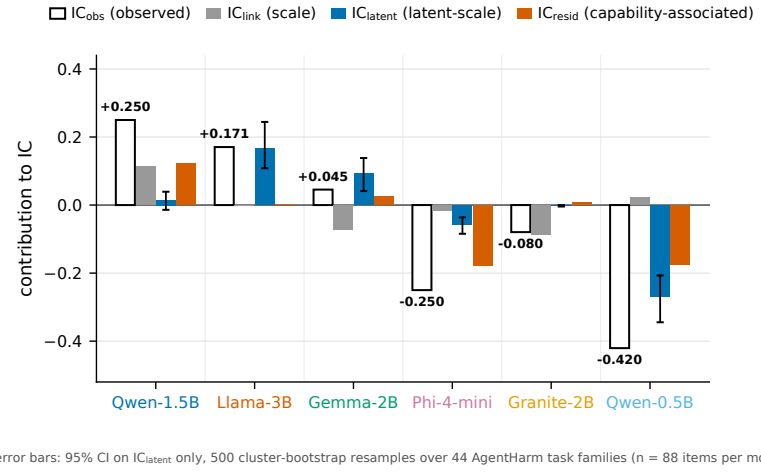

Figure 3: The three-channel decomposition, one panel per model. Bars are the scale artefact $\mathrm{IC}_{\mathrm{link}}$, the latent-scale interaction $\mathrm{IC}_{\mathrm{lat}}$ and the capability-associated residual $\mathrm{IC}_{\mathrm{res}}$, summing to the observed interaction marked on each panel. Each channel dominates on at least one model.

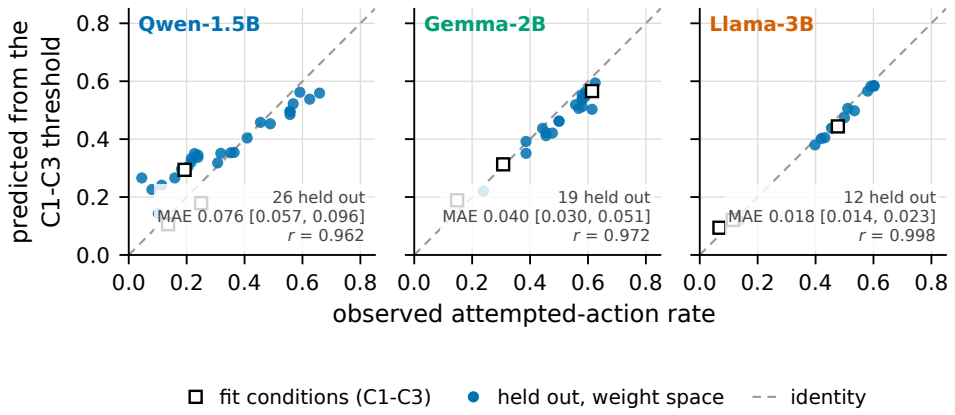

Figure 4: No-refit readout: predicted against observed harmful attempted-action rate for every held-out condition, on the three models with enough of them. The curve was fitted on C_1, C_2 and C_3 and no parameter was refitted, so distance from the diagonal is prediction error. Phi-4-mini has one held-out condition and is omitted here; its error appears in Table 4.

Table 5: The refusal axis stays put while the decision axis slides past the acting threshold. Shifts are item means from C_1 to C_4 in units of each axis's own C_1 spread, negative meaning more willing to act; the decision axis moves 3.1 times further on average. frac $d > 0$, the share of harmful items on which a tool-call opener is the likelier continuation, rises from 0–38% to 34–98%. Both AUROC columns stay flat. The refusal axis was measured on three models.

	refusal axis		decision axis		frac $d > 0$	
Model	shift	AUROC $C_1 \to C_4$	shift	AUROC $C_1 \to C_4$	C_1	C_4
Qwen2.5-1.5B	−0.586	0.854 → 0.825	−1.500	0.798 → 0.803	0.10	0.66
Llama-3.2-3B	+0.229	0.767 → 0.796	−1.573	0.889 → 0.854	0.15	0.61
Gemma-2-2B	−0.338	0.897 → 0.874	−0.452	0.865 → 0.850	0.38	0.56
Phi-4-mini	not measured		−4.329	0.862 → 0.887	0.00	0.42
Granite-3.1-2B	not measured		−2.762	0.775 → 0.726	0.01	0.98
Qwen2.5-0.5B	not measured		−2.735	0.624 → 0.672	0.05	0.34

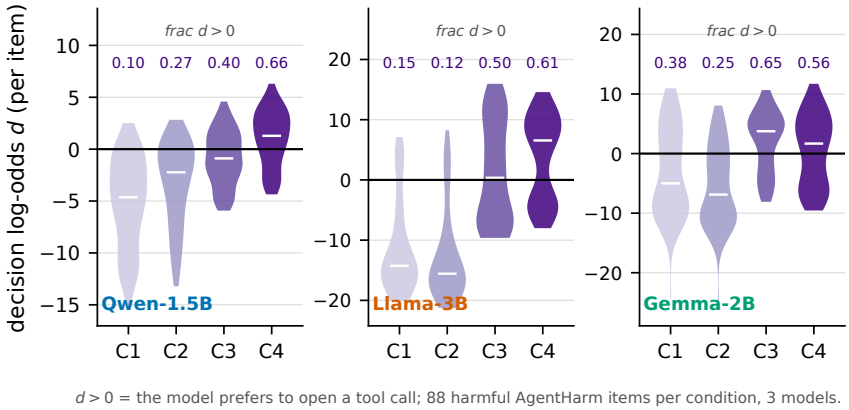

Figure 5: Per-item decision log-odds d across the four conditions, the three models where both axes were measured, 88 harmful prompts each, the full distributions behind the frac $d > 0$ column of Table 5 and the text of Section 5. White markers are medians. The solid line at $d = 0$ is where a tool-call opener becomes the likelier continuation. Scales differ between models by construction and are not comparable across panels.

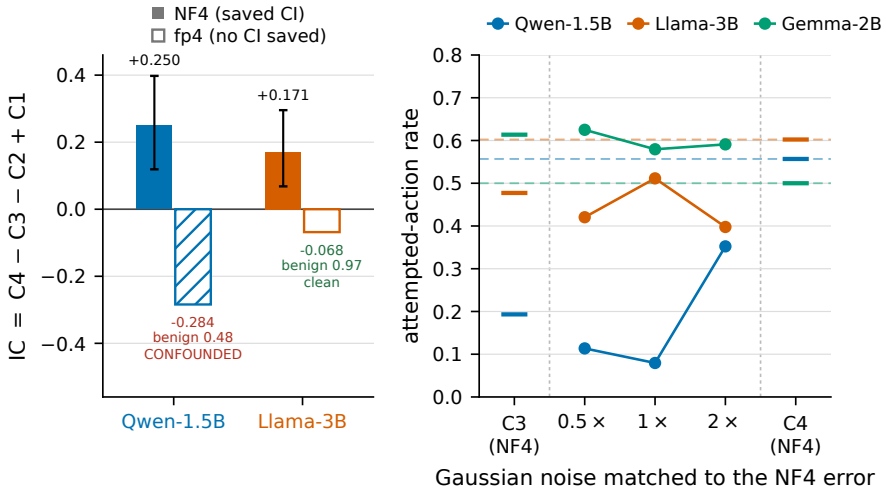

Figure 6: Two negative controls on the source of the effect. Left: interaction contrast under NF4 against fp4 at the same bit width, on the two models where fp4 was run; error bars are shown for NF4 (its cluster-bootstrap CI). The fp4 point estimates are reported without a drawn CI here, though Section 7 gives the fp4 interval in text. Llama is the clean case, where fp4 preserves the benign call rate and the interaction disappears; Qwen's fp4 arm is marked confounded because its benign call rate falls to 0.48. Right: harmful attempted-action rates under norm-matched Gaussian perturbation at three scales. The noise arms land on or near C_3 on every model, not at C_4; on Gemma they sit above C_4, which itself has no significant interaction. Rates rather than contrasts are plotted on the right, because only one noise scale was run without fine-tuning, so no matched 2×2 exists.

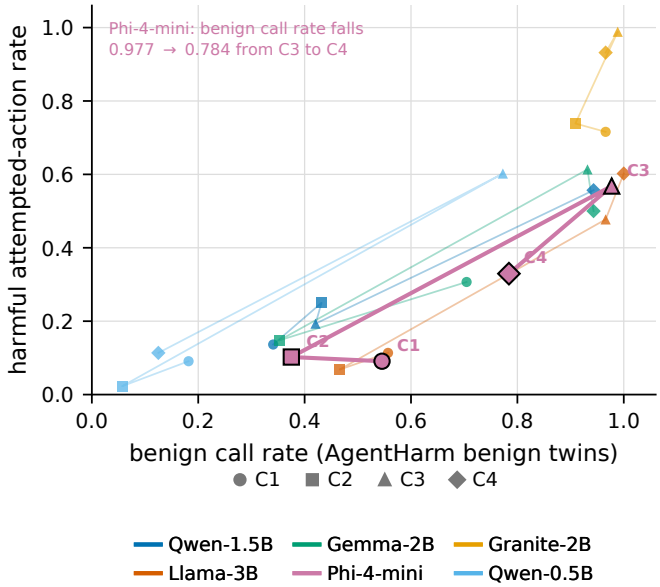

Figure 7: Harmful attempted-action rate against benign call rate, one point per model per condition, all six models and all four conditions (Phi-4-mini highlighted). A model whose benign rate falls while its harmful rate also falls, such as Phi-4-mini from C_3 to C_4 ($0.977 \to 0.784$ benign, $0.568 \to 0.330$ harmful), is losing the ability to act rather than gaining a reason not to; the capability-corrected rate in the main text (0.330/0.784 against 0.568/0.977) is computed from these same points, not read off the figure.

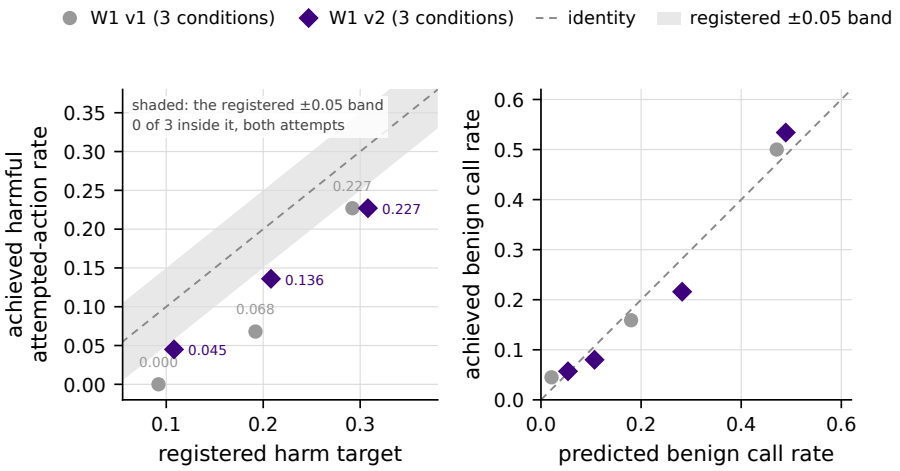

Figure 8: Two pre-registered control attempts. Left: achieved against target harmful rate, with the registered ±0.05 band. Both attempts failed the bar; the second was no worse than the first on any target (a tie at 0.30) and missed by less on the other two (0.20: 0.132 → 0.064; 0.10: 0.100 → 0.055). Right: benign-rate predictions, which were accurate in both. The far-end probe condition sets no harm target and appears in the right panel only.

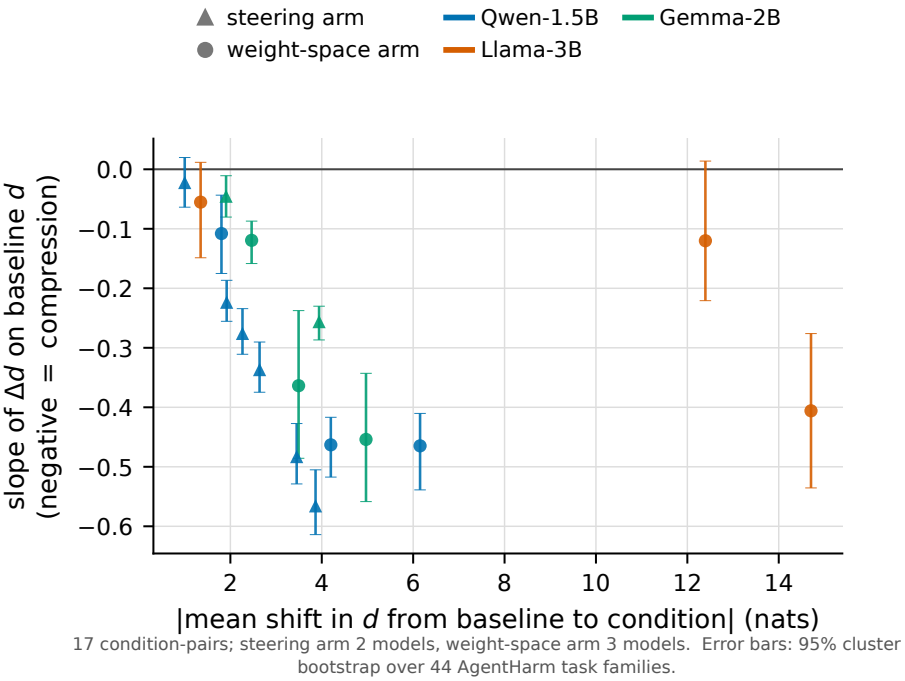

Figure 9: Per-item change in d against baseline d, for steering pairs and weight-space pairs. A negative slope means compression. Both arms compress, which is why the translate-versus-compress explanation was withdrawn. Descriptive only: this is a diagnostic regression on data from the two registered attempts, not a third attempt at control.

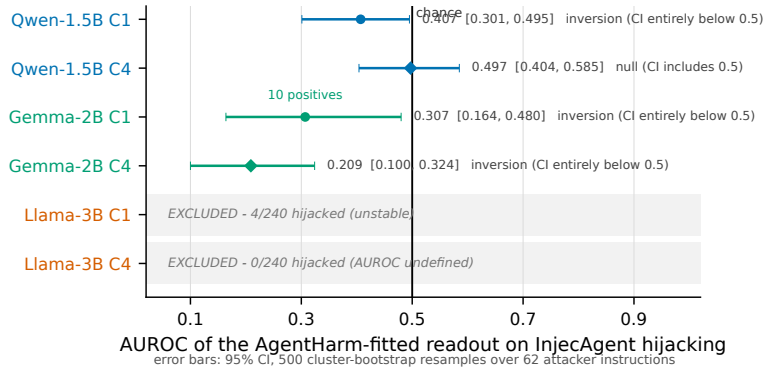

Figure 10: Cross-benchmark transfer to InjecAgent, with cluster-bootstrap intervals over the attacker instruction. All four interpretable cells sit at or below chance. Three intervals exclude 0.5 from below. The below-chance pattern was not pre-registered.

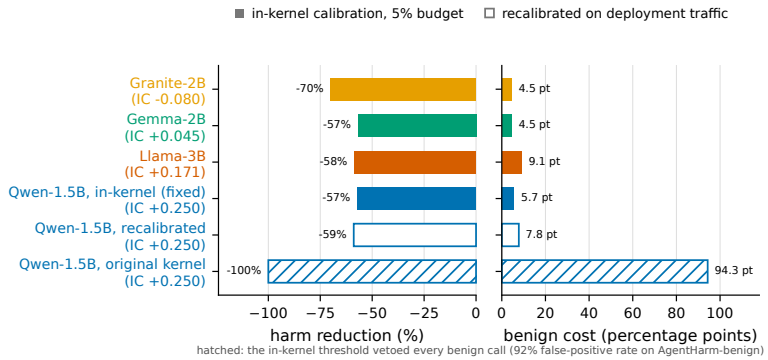

Figure 11: Tool-free reference gate at a 5% false-positive budget, four models, each against its own in-kernel baseline. Harm reduction against benign cost. The core idea was published first by SafeKeep (Pan et al., 2026).

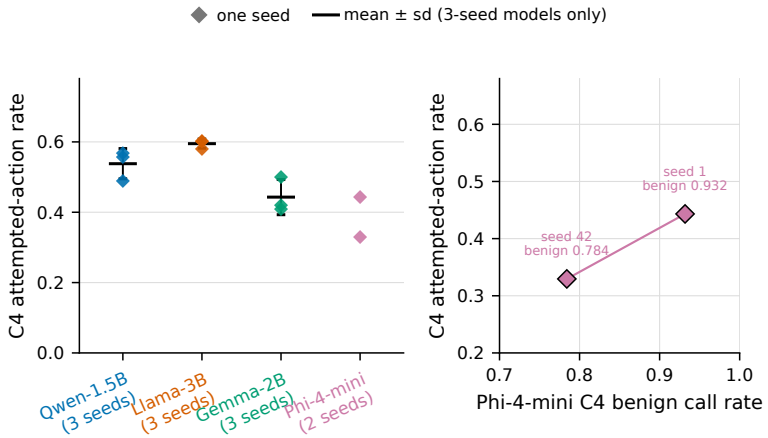

Figure 12: Seed replication. Three seeds on Qwen, Llama and Gemma; two on Phi-4-mini. C_1 and C_2 contain no fine-tuning and are seed-free. Within-model spread is much smaller than between-model spread.

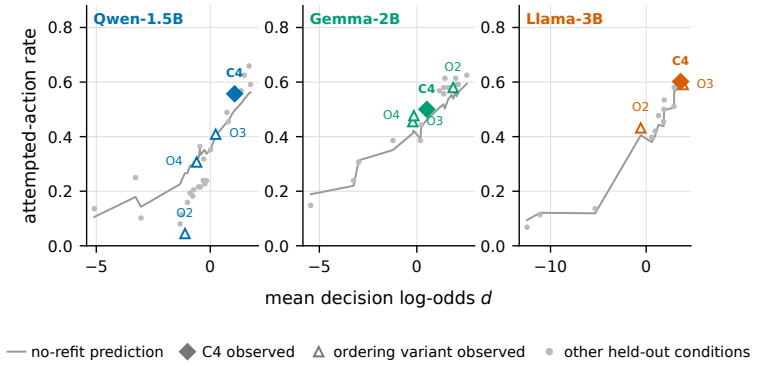

Figure 13: Ordering variants against the no-refit readout. O_2 fine-tunes then quantizes, O_3 re-quantizes after merging, O_4 puts a QLoRA adapter on the fp16 base. The line connects saved predictions in order of mean d.

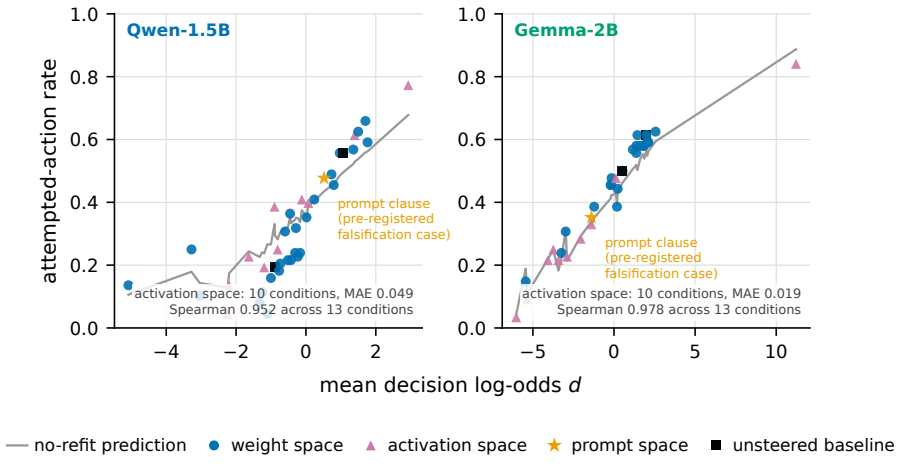

Figure 14: One curve, three kinds of intervention, on the two models with an activation-space sweep. The pre-registered falsification case is the prompt clause, which changes no weights and installs no hook. It lands on the relation rather than breaking it.