A FEW CONTRADICTED EXAMPLES SILENTLY ERODE TRANSFER TO HELD-OUT FORMATS AND LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction tuning is valuable because it transfers: tuned models follow instructions that show or name their target (answer formats, response templates, schemas, output languages, as in IFEval, IFBench and tool-call validation), even for targets never seen in training. We show that this transfer erodes, silently, when a small fraction of training responses do not follow their own instruction. Changing only whether an instruction matches its response (responses byte-identical across arms in mathematics and question answering, the same sentence's reference translation in translation), contradicting 5% of fine-tuning examples lowers held-out compliance by 19 to 62 points in every one of seven students from 0.5B to 8B, every contradicted seed below its truthful pair, and full contradiction reaches zero. In Qwen2.5-1.5B 2% suffices: compliance with held-out answer formats falls from 68.3 to 47.7% (66.9 to 48.4% when replicated at one fixed evaluation geometry), with held-out response templates from 51.6 to 27.8% and with held-out output languages from 90.4 to 81.0%, every contradicted run of the main series below every truthful run, while compliance with the trained instructions moves by at most 1.4 points; 2% also cuts SmolLM2-1.7B's held-out template compliance from 54.3 to 29.1% and Qwen2.5-7B's held-out format compliance from 78.6 to 37.8%. Deleting the contradicted examples restores transfer, at 2% as at 50%; deleting as many random examples does not. In public data the same holds: removing the 2.45% of xLAM tool calls that violate their own schema eliminates 43% of held-out schema errors, as large a share as 2% synthetic contradiction causes (36–39%), and repairing natural constraint violations in Tülu 3 raises held-out compliance on every seed while trained-condition metrics stay unchanged. We explain the silence: at the optimum of the fine-tuning objective every trained-condition quantity is fixed by the contradiction rate ρ (validation loss rises by $-\ln(1-\rho)$ nats, 0.02 at 2%), while held-out behaviour is left undetermined, and trained mathematics networks follow this scaling from 1% to 50% contradiction whether they lose held-out transfer entirely or keep half of it. Loss, IFD, Superfiltering and an LLM judge miss contradicted examples (AUROC ≤ 0.82), whereas each instruction's own verifier flags all of them at 2% with no false positives and a label-free likelihood-swap test ranks them first (AUROC 0.997–0.999); retraining without its top 5% restores mathematics transfer to the clean level (67.8 against 66.9%).

1 INTRODUCTION

Instruction tuning is valuable because it transfers: a model tuned on some instructions that show or name their target follows others it never saw, which lets systems request new answer tags, schemas, templates or languages at inference time. We show that this transfer depends on each response following its own instruction, a property neither trained-condition metrics nor standard filters check, and that violating it in a small fraction of examples erodes transfer while trained-condition measurements barely move.

Such violations are not hypothetical. Exact schema validation finds 2.45% (95% CI 2.32–2.58%) of eligible function calls in the public xLAM set (Liu et al., 2024b) violating the schema their own prompt supplies, and in the Tülu 3 instruction-following set (Lambert et al., 2024) the official IFEval checkers flag 12% of the rows in our exhibited-literal training pool as violating their own constraint (Appendix M.1; a stratified audit confirms such violations, Appendix J); standard data-

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.

selection scores do not flag the natural violations (AUROC 0.44–0.60; Appendix N). In tool-call data they account for as large a share of held-out errors as our synthetic contradictions: across seven paired seeds of xLAM training, removing them eliminates 43% of the tuned model's held-out schema errors (96.1 to 97.8%; gain 95% CI 1.3–2.0 points), against 36–39% for 2% synthetic contradiction in the same student (Appendix M.2), and 55% on calls with four or more arguments, where violations are most frequent (82.1 to 92.0%); deleting as many random rows gives nothing, and at three times the natural rate violators cost 1.8 points (95% CI 1.2–2.4; 10.1 on complex calls) against the same positions filled with compliant calls (Appendices M, M.2). In Tülu 3, minimally repairing the violating responses on the same rows raises held-out compliance on every one of ten seeds (1.6 points, 95% CI 0.7–2.5) while trained-type compliance and validation loss stay equivalent: natural violations erode transfer silently too. They cost less per row than our contradictions because of their kind: on the same 5% of math rows, substituting another trained format costs 38.8 points of held-out compliance and omitting the requested format 11.2 (95% CIs 19.1–58.5 and 5.6–16.9; Appendix M.2). Label-noise results reassuring when enough clean examples remain (Rolnick et al., 2017) do not address transfer to instructions absent from training. We measure severity with a dose design (§2, §4): a set fraction ρ of fine-tuning examples requests a format, template or language other than the one its response uses, while responses stay fixed across arms (byte-identical in mathematics and question answering; in translation a contradicted row carries the same sentence's reference in another trained language); evaluation requests targets absent from fine-tuning. These are instructions that show or name their target (IFEval/IFBench constraints, JSON schemas, tool-call formats, output language), whose compliance a program checks exactly. Our contributions:

- **A silent, causal low-dose failure.** Contradicting 5% of examples costs 19–62 points of held-out compliance in each of seven students from 0.5B to 8B, and full contradiction gives 0.0% in five (Table 2; §3–4). In Qwen2.5-1.5B 2% suffices: held-out compliance falls from 68.3 to 47.7% for answer formats (exact two-sided permutation p=0.006; one-geometry replication 66.9 to 48.4%), from 51.6 to 27.8% for response templates and from 90.4 to 81.0% for output languages (which the untuned model already follows); in each task every contradicted run falls below every truthful run (Figure 1c), while trained-instruction compliance moves by at most 1.4 points. At 2% the loss is also large in SmolLM2-1.7B and Qwen2.5-7B, and 3–4 points in Qwen2.5-0.5B and 3B. Matched controls rule out lost data: beside a byte-identical truthful half, deleting the contradicted rows or neutralizing their instructions keeps transfer while contradicting them erases it; deleting as many random rows does not help.
- **Why trained-condition metrics are blind.** At the optimum of the fine-tuning objective every trained-condition quantity is fixed by ρ while held-out behaviour is undetermined (§5); at 2% validation loss moves by 0.020 nats, below the seed-to-seed spread. On 88 teacher-forced checkpoints the trained-condition excess scales with $-\ln(1-\rho)$, in mathematics with one constant (1.3) from 1% to 50%, and two format sets with the same 50% excess (0.737 and 0.713 nats; $\ln 2 = 0.693$) keep 0.0 and 29.6% held-out compliance. Held-out log-odds fall linearly in $\mathrm{logit}(1-\rho)$ at about two nats per nat, which is why the first contradictions do most of the damage.
- **Detection where standard filters fail.** Each instruction's own verifier flags exactly the contradicted rows of a 2% pool. Without a verifier, a label-free likelihood-swap test needing only the training pool and the released model ranks contradicted rows first (AUROC 0.997–0.999; 88–89% precision at a budget equal to their number), whereas loss, IFD, Superfiltering and an LLM judge do not (AUROC 0.50–0.82; at most 27% precision); retraining without the swap test's top 5% restores math transfer to the clean level (67.8 against 66.9%; 2%: 48.4%).
- **A probe outside trained conditions.** Asked a plain question, models trained with 2% contradiction append a trained template to 97–98% of answers, against at most 1.7% after truthful training.

2 SETUP

Data: arms differ only in the final line Responses are frozen, verifier-correct rollouts on DeepScaleR problems (Luo et al., 2025) ($\leq 0.4\%$ overlap with any evaluated benchmark; Appendix B). Each response is truncated at its final-answer line and given a terminal rendering in the arm's format, so reasoning bodies are **byte-identical across arms**; only the final line and, for conditioned arms, the system message differ. Same-seed arms share data order, so a same-seed pair isolates instruc-

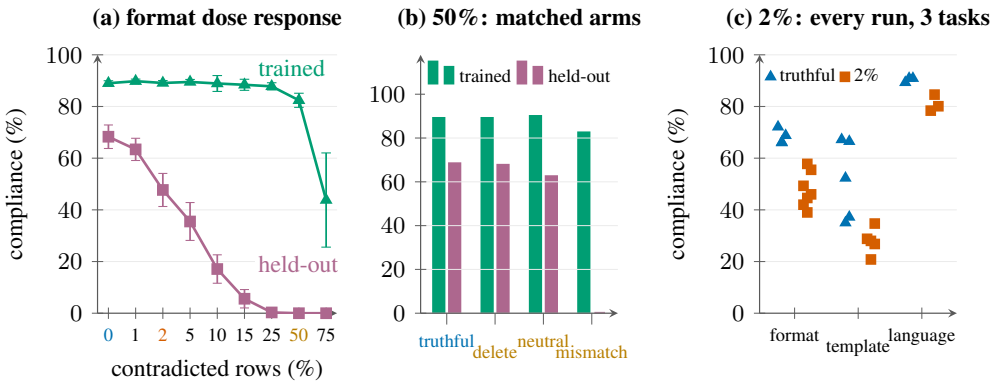

Figure 1: Qwen2.5-1.5B. **(a)** MATH-500 answer formats: as the contradicted share rises, trained-format compliance changes little through 25%, while held-out-format compliance declines (seed means with 95% t-intervals; three to seven seeds per point). **(b)** At 50% contradiction, arms share a byte-identical truthful half, differing only in the other half: deleted, neutrally instructed, or contradicted. The deleted rows are those each row's own verifier flags (1,151 of 1,156, no false positives). **(c)** Held-out compliance of every run at 2% and every truthful run in three tasks: formats (seven versus four seeds), response templates (five versus five) and output language (three versus three).

tion treatment. Training is full-parameter SFT on the same 2,267 responses for three epochs (426 optimizer steps at batch 16 below 7B; Appendix A).

Formats Seven *trained* formats: `\boxed{}`, `Answer: X`, `#### X`, `{"answer": "X"}`, `<answer>X</answer>`, `**Answer:** X`, `Therefore, the answer is X.` Six *held-out* formats, never a fine-tuning target: `[[X]]`, `<<X>>`, `ANSWER = X`, `- answer: X`, `%%X%%`, `>>> X`. Each format is scored by its own extractor; a validation gate checks that every format round-trips and none parses another's output (Appendix A.5); "held out" means held out of fine-tuning. The capability is verifiable instruction following: each instruction shows the delimiters of the requested answer line, as IFEval, IFBench and structured-output formats do (a blind independent classification assigns 16 of IFEval's 25 types to this class; Appendix I); in translation (§4) the instruction only names the target language, with no pattern to copy, and contradiction erodes that transfer too. Pretraining familiarity cannot explain the contrasts, which share the base model: the untuned model complies on 9.0% of held-out requests, truthful fine-tuning raises this to 68.3%, and 2% contradiction removes a third of the gain. A seventh held-out format, `latex_answer`, shares `\boxed{}`'s syntax and is excluded from every held-out average (Table 5); §5 adds a synthetic 128-format family.

Arms Table 1. The contradiction rate ρ is the fraction of rows whose system message names a format other than the one used (the appendix indexes arms by reliability $p = 1 - \rho$).

Table 1: Arms. All share the same responses up to the final line.

Arm (code)	System message	Final line
single-format (`boxed_only`)	neutral, names no format	always `\boxed{}`
unlabelled (`diverse7`)	neutral, names no format	one of 7, balanced
truthful (`cond7`)	names the format used	one of 7, balanced
contradictory-ρ (`cond`p)	names a wrong format on a fraction ρ of rows	one of 7
mislabelled (`shuf7`)	always names a wrong format (ρ=1)	one of 7
deletion (`cond`h)	truthful; contradicted rows removed	one of 7
neutral (`cond`n)	truthful; contradicted rows get the neutral message	one of 7

Evaluation Zero-shot on MATH-500 (Lightman et al., 2023) with system message only, greedy decoding, 1,024 new tokens for fine-tuned models and 4,096 for released checkpoints (Appendix C).

Table 2: Every replication: held-out compliance (%) after truthful training and at contradiction dose ρ. Δ: difference with a 95% CI over training seeds from unrounded values; paired t over shared seeds (four for Qwen2.5-1.5B formats; Welch over all seven 2% and four truthful seeds: -20.5 [-27.3, -13.8]), Welch for Llama-3.2-3B. *Lower*: contradicted seeds below their truthful pair (first four rows: below every truthful seed). *Trained*: change in trained-instruction compliance. *At 50%*: held-out compliance under 50% contradiction, after deleting its contradicted rows, and after deleting as many random rows (–: not run). Checkers: format extractors, IFEval's official checkers (templates), target-script share (language). [§]A later series on other hardware, with its own truthful arm (Table 3: 72.0). [‡]Untuned model already complies (72.0, 73.2%). [*]The students of Table 3. Seeds and per-seed values: Appendices D, K and L.

							at 50%		
task	student	ρ	truthful $\to \rho$	Δ [95% CI]	lower	trained	contr.	del.	random
formats	Qwen2.5-1.5B	2%	68.3 $\to$ 47.7	-20.7 [-31.5, -9.9]	7/7	$+0.1$	0.0	67.6	0.0
templates	Qwen2.5-1.5B	2%	51.6 $\to$ 27.8	-23.8 [-41.3, -6.3]	5/5	-0.3	0.1	52.3	1.2
language	Qwen2.5-1.5B	2%	90.4 $\to$ 81.0	-9.3 [-15.8, -2.9]	3/3	-1.4	0.9	85.8	7.7
formats	Llama-3.2-3B	5%	76.0 $\to$ 34.8	-41.2 [-49.7, -32.7]	3/3	$+0.6$	0.0	–	–
templates	SmolLM2-1.7B	2%	54.3 $\to$ 29.1	-25.2 [-37.8, -12.7]	3/3	-0.1	–	–	–
templates	SmolLM2-1.7B	5%	54.3 $\to$ 12.3	-41.9 [-50.0, -33.9]	3/3	-0.6	0.1	54.8	0.3
formats	Qwen2.5-0.5B	2%	59.8 $\to$ 56.5	-3.3 [-6.5, -0.1]	3/3	$+0.9$	0.0	66.3	–
formats	Qwen2.5-0.5B	5%	59.8 $\to$ 39.3	-20.5 [-26.2, -14.9]	3/3	$+0.7$	0.0	66.3	–
formats	Qwen2.5-3B	2%	73.5 $\to$ 69.3	-4.2 [-11.4, $+3.0$]	3/3	$+0.0$	0.0	68.9	–
formats	Qwen2.5-3B	5%	73.5 $\to$ 54.4	-19.1 [-25.8, -12.3]	3/3	-0.2	0.0	68.9	–
formats	Qwen2.5-7B[§]	2%	78.6 $\to$ 37.8	-40.8 [-66.8, -14.8]	3/3	-1.2	0.0	75.3	–
formats	Qwen2.5-7B[§]	5%	78.6 $\to$ 16.8	-61.9 [-79.9, -43.8]	3/3	-0.9	0.0	75.3	–
templates	Qwen2.5-7B[‡]	5%	36.4 $\to$ 14.1	-22.3 [-33.9, -10.7]	5/5	-1.0	0.0	59.1	0.2
templates	Llama-3.1-8B[‡]	5%	47.4 $\to$ 0.5	-46.9 [-69.1, -24.7]	5/5	-0.9	0.0	44.5	0.1
formats	5 students[*]	100%	$\geq$68.3 $\to$ 0.0	–	16/16	–	–	–	–

Table 3: Held-out formats, MATH-500, six-format mean; each cell is **compliance / success** (%). *base* is the released checkpoint before fine-tuning (larger decoding budget; Qwen3-8B with thinking off); other columns are the arms of Table 1 (`boxed_only`, `diverse7`, `shuf7`, `cond7`). Per-seed values in Table 8.

Student	base	single-format	unlabelled	mislabelled	truthful
		compliance / success			
Qwen2.5-1.5B	9.0 / 4.2	0.0 / 0.0	1.3 / 0.5	**0.0 / 0.0**	**68.3 / 33.1**
OLMo-2-1B	30.4 / 5.4	9.2 / 1.3	20.3 / 3.0	**0.0 / 0.0**	**74.1 / 11.2**
Llama-3.2-3B	58.6 / 21.1	0.1 / 0.0	3.1 / 1.6	**0.0 / 0.0**	**76.0 / 30.4**
Qwen2.5-7B	25.3 / 17.0	0.0 / 0.0	0.0 / 0.0	**0.0 / 0.0**	**72.0 / 45.7**
Qwen3-8B	79.3 / 56.2	0.2 / 0.1	6.0 / 4.1	**0.0 / 0.0**	**74.7 / 53.4**

A held-out cell is 500 problems $\times$ six formats per seed, so 0.0% means 0 of 3,000 on that seed. We report compliance (the requested format's extractor finds an answer) and success (it is also correct). The primary student is Qwen2.5-1.5B-Instruct; Table 3 adds OLMo-2-1B, Llama-3.2-3B, Qwen2.5-7B and Qwen3-8B (Yang et al., 2024; 2025; Grattafiori et al., 2024; OLMo Team, 2025), all instruction-tuned releases. Main arms have at least three paired seeds (Appendix D); the training seed is the unit of replication; differences and intervals use unrounded values, so they can differ by 0.1 from the rounded-mean differences shown.

3 Contradictory Supervision Erodes Held-Out Transfer

Truthful format names yield transfer and wrong ones erase it, in five students Fine-tuning on one format erases held-out compliance the released model had (Qwen3-8B: 79.3 to 0.2%), the format specialisation of Wang et al. (2024); fine-tuning on seven unnamed formats leaves it at or below the

released model in every student; truthful naming reaches 68.3–76.0%; naming a wrong format on every row gives 0.0% in every student (Table 3). The same holds on GSM8K, whose native answer convention differs: truthful 65.1–90.0% across three students, mislabelled 0.0 everywhere (Table 6).

2% contradiction substantially erodes held-out transfer; trained-format compliance changes little Held-out compliance falls from 68.3% with truthful names to 47.7% at $\rho = 2\%$ (each row contradicted with probability 0.02: 44 of 2,267 rows on seed 0; 95% CI 41.3–54.1, seven seeds; trained 89.1%); all seven 2% seeds (39.0–57.8%) lie below all four truthful seeds (66.0–72.1%; exact two-sided permutation p=0.006; Figure 1c), a margin of at least 8.2 points, more than twice the largest single-format batch effect (3.4), while trained-format compliance is equivalent within one point (difference $+0.06$, 90% CI -0.84 to $+0.96$). A replication at one fixed evaluation geometry (one GPU class and batch; seven paired seeds; exactly 45 contradicted rows per seed) gives the same contrast: 66.9 against 48.4% (difference 18.5 points, 95% CI 6.5–30.5; Appendix N.1). It falls to 35.5, 17.1, 5.6 and 0.3% at $\rho = 5$, 10, 15 and 25%, and is 0.0 from 50% on. At 1% it is 63.4% (95% CI 59.1–67.7, seven seeds), already below the truthful mean. Trained-format compliance on the same checkpoints (three of seven seeds at 5–15%) stays at 87.8–89.8% through 25%, 82.4% at 50%, and falls only at 75% (43.8%) (Figure 1a, Table 11). At 5% transfer is first learned, then lost during training (Appendix F). The curve replicates on Llama-3.2-3B (34.8, 8.0, 1.5 and 0.0% at 5, 15, 25 and 50%; Table 11).

Matched controls isolate the effect of contradiction At 50% an arm also holds only half the truthful evidence, so the curve could reflect data quantity. Three arms share a *byte-identical* truthful half, differing only in the other half (Figure 1b, Table 7): deleted, 67.6%; neutral instruction, 62.4%; contradicting instruction, 0.0% (truthful reference 68.3%, four seeds). On the three seed IDs shared by deleted and clean training, the paired deletion-minus-clean difference is -1.4 points. At identical rows *and* steps, deleting the contradicted rows gives 67.6% while deleting as many random rows from the contradicted pool gives 0.0% and from a clean pool 67.1%, so over-deletion is cheap. The repair holds at 2%: in the one-geometry replication, deleting the 45 contradicted rows gives 67.7% (clean 66.9%) and deleting 45 random rows 41.9% (difference 25.7 points, 95% CI 15.9–35.6, seven paired seeds; question answering 31.4, 6.5–56.3). A decoy format that is never a target and never scored also gives 0.0%: the damage needs neither noisy targets nor lost information, only an instruction the response does not follow. The 50% collapse replicates on Llama-3.2-3B (0.0%; Table 11).

Finding the contradicted rows needs no labels Each row's own verifier flags all 44 contradicted rows of the seed-0 2% pool and none of its 2,223 truthful ones (at 50%: 1,151 of 1,156, no false positives), so the 2% deletion arm is what running that verifier over the training set produces. Where no verifier exists, a likelihood-swap test needing only the training pool and the released model (does the response fit another pool instruction better than its own?) ranks contradicted rows first on every pool with exact labels (Table 4); it tests the defect itself, a response matching another instruction better than its own: at a budget equal to the contradicted count, 89% (math) and 88% (QA) of flagged rows are contradicted. Four standard scores do not: completion loss, IFD, Superfiltering and a 7B LLM judge reach AUROC 0.50–0.82 and at most 27% precision at the same budget, and are equally blind to natural xLAM schema violations (0.44–0.60). IFD fails although it scores how much the instruction helps predict the response: whole-response scores dilute a contradiction confined to a few tokens (better on short QA responses, 0.80–0.81, than on long math traces, 0.58–0.66), while the swap test scores the rendered span. Retraining closes the loop: removing the top 5% of rows by swap score (43–45 of each seed's 45 contradicted rows, no labels used) restores math transfer to 67.8% (2%: 48.4; clean: 66.9, equivalent within ± 5 points, 90% CI -2.3 to 4.1; seven paired seeds); in question answering it recovers about half the loss (11.6 points, 95% CI -6.5 to 29.7, five seeds; Appendix N.1). The swap test flags a row when the span its instruction determines is better explained by another instruction; on matched 5% math pools this holds for omitted final answers as well as substituted ones (AUROC 0.966 and 0.997; Appendix M.2). Natural violations are mostly errors inside a span the instruction still explains (a miscount, a wrong argument), which likelihood contrasts do not separate (swap on xLAM: 0.45; a constraint-drop variant on Tülu 3: 0.65, 0.69 on omitted titles). For these verifiable constraints the instruction's own verifier is the detector, and on natural data it agrees with blinded adjudication (Tülu 3 checkers: precision 0.93–1.00, no misses

Table 4: Finding contradicted rows: AUROC and precision at a budget equal to the number of contradicted rows (*prec.*), on training pools with exact labels (math: 44 of 2,267 rows at 2%; QA: 48 of 2,400). *All pools*: AUROC range over math 2/50% and QA 2/5/10/50%. *xLAM*: 10,000 natural tool-call rows with 245 schema violations as labels, argument-level errors inside a call its schema still explains (swap: PMI variant against a neutral instruction). Swap and verifier use no labels; details in Appendix N.

filter	math 2% AUROC / prec.	QA 2% AUROC / prec.	all pools AUROC	xLAM AUROC
own-format verifier	1.00 / 1.00	1.00 / 1.00	0.998–1.00	(the labels)
likelihood swap (released model)	0.998 / 0.89	0.999 / 0.88	0.997–0.999	0.45
completion loss	0.58 / 0.02	0.81 / 0.27	0.55–0.81	0.44
IFD	0.66 / 0.05	0.80 / 0.21	0.64–0.82	0.60
Superfiltering (GPT-2)	0.62 / 0.00	0.64 / 0.04	0.57–0.65	–
LLM judge (Qwen2.5-7B)	0.52 / 0.02	0.58 / 0.00	0.50–0.58	0.58

among 160 audited passes; Appendix M.2), while loss, IFD and a judge stay near chance (0.44–0.60).

The failures are well-formed trained formats Reparsing every mislabelled-arm response (seed 0), 84.9–93.0% parse as a trained format in all five students, none as nothing (Table 9); correctness is unaffected (trained-format success 44.0–45.3%, truthful 44.5%; Appendix E).

4 Templates, Language and Further Students

Design We repeat the design off mathematics (instruction in the user turn, no final answer line): reading comprehension on SQuAD 1.1 (Rajpurkar et al., 2016) with Qwen2.5-1.5B-Instruct. For SFT, each target is the first gold span rendered in one of four response templates: a quotation, a `<<Answer>>` title, a JSON object or a P.S. line (600 rows each); rendered targets are byte-identical across dose arms. At evaluation the model receives a generic system message and a user prompt with only the passage, question and requested template instruction (or an ordinary answer instruction for the plain-question probe), then answers. The gold span is retained as scoring metadata. We score format compliance with IFEval's official checkers and answer quality separately, against the gold answer, with a parser that accepts any template (five seeds; Appendix K). Five templates are held out of training (a literal closing phrase, two `SECTION` headers, highlighted spans, `***`-separated paragraphs and repeating the question), scored by IFEval's official checkers (Zhou et al., 2023b). Contradiction changes only the requested template, on nested 2, 5, 10 and 50% subsets; every run takes 450 updates (Appendix K).

Result Truthful training lifts held-out-template compliance from 30.8% (untuned) to 51.6% (95% CI 32.5–70.8). Contradicting 2% of rows (48 of 2,400) cuts it to 27.8% (21.7–34.0): lower on every seed, below the untuned model on four of five. Across five shared seeds, paired reduction is 23.8 percentage points (two-sided 95% paired-t CI [6.3, 41.3]). Trained-template compliance is 99.6% (truthful 99.9%); at 5% it is 15.7%, below the untuned model on every seed, trained templates at 98.4% (Figure 2a). The loss is not one template's: the closing phrase, which any appended line breaks, falls from 62 to 0%, but highlighted spans fall from 78 to 41% (Appendix K). Answers survive: exact match after reading any template is 50.0% at 2% against 39.1% truthful, and the gold span appears in 77.7 against 80.6% of held-out responses; plain-question exact match falls (56.4 to 2.8%) only because the stapled template counts in the answer (stripped: 64.1%). Unlike mathematics (82.4% trained at 50%), question answering ends loud: at 50% held-out compliance falls to 0.1%, trained-template compliance from 99.9 to 42.3% and ordinary exact match from 56.4 to 39.4% (Appendix K). Deleting the 50% arm's contradicted rows gives 52.3% (95% CI 34.8–69.8; truthful 51.6%), versus 1.2% after matched random deletion (Figure 2b). Silence is a low-dose property: at 2–10% trained-template compliance stays within 4.4 points of truthful training, and a second family confirms it at 5% (0.6 points). The repeating-question template scores zero in every arm; the 2% mean is lower on each of the other four (Appendix K).

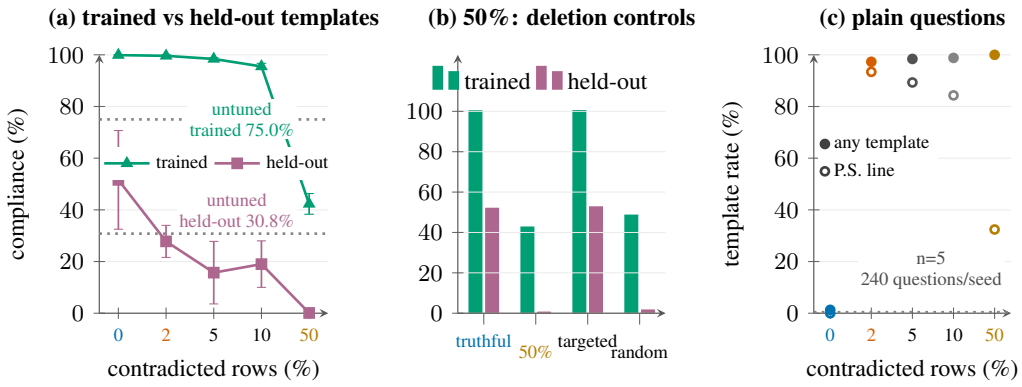

Figure 2: Grounded question answering (SQuAD 1.1), Qwen2.5-1.5B, five seeds. **(a)** Compliance with the four trained and five held-out response templates as the contradicted share rises (seed means with 95% t-intervals; dotted: untuned model, 75.0 and 30.8%). **(b)** Truthful; 50% contradiction; 50% with contradicted rows deleted (targeted) or as many random rows deleted (random); 450 updates each. **(c)** Share (%) of 240 plain questions requesting no template whose answer carries a trained template; open circles: closing P.S. line alone.

Output language, a second family and 7–8B students In translation (three seeds; Appendix L) the attribute is the output language: Qwen2.5-1.5B is fine-tuned on 2,400 English→X rows over seven trained languages and tested on six held-out languages with distinct scripts, the instruction only naming the language, with no shown pattern to copy. Pairing 5% of requests with the same sentence's reference in another trained language lowers held-out compliance from 90.4 (untuned 92.0) to 71.4% (paired difference 19.0 points, 95% CI 10.1–27.9, positive on every seed); trained-language compliance falls 1.3 points, chrF++ 0.6, with 75.8% of held-out failures in a trained language's script (language identification agrees: 93.1 to 73.8%). Here tuning does not install the ability (the untuned model already complies on 92.0%), so contradiction erodes what truthful tuning keeps. At 2% it falls from 90.4 to 81.0% on every seed (9.3 points, 95% CI 2.9–15.8; trained-language compliance falls 1.4). At 50%, deleting the contradicted rows gives 85.8% versus 7.7% after random deletion. In question answering with a second family (SmolLM2-1.7B), 2 and 5% contradiction lower held-out template compliance from 54.3 to 29.1 and 12.3% (every seed; trained templates within 0.6 points), and targeted deletion gives 54.8% against 0.3% after random deletion. In Qwen2.5-0.5B and 3B answer formats, 5% contradiction lowers held-out compliance by 20.5 and 19.1 points (every seed; trained formats within 0.9), whereas 2% lowers it by 3.3 and 4.2: the dose at which erosion becomes large depends on the student. At 7–8B, 2% contradiction lowers Qwen2.5-7B's held-out format compliance from 78.6 to 37.8% and 5% its template compliance from 36.4 to 14.1%, and 5% lowers Llama-3.1-8B's from 47.4 to 0.5% (every seed, five seeds for templates; trained compliance within 1.2 points; targeted deletion 44.5–75.3% against at most 0.2% after random deletion; Table 2).

Plain prompts expose a changed default Question-answering models trained with 2% contradiction append a trained template to 96.7–97.9% of 240 plain-question answers, against 0.8–1.7% for truthful models and 0.4% untuned (Figure 2c). In math, the 50%-contradiction model places a trained answer line among the last three lines of 44–52 of 541 IFEval responses, versus 6–16 after deletion, in all seven seed pairs.

5 Why Trained-Condition Metrics Are Blind

Let a fraction ρ_f of rows requesting trained format f render another trained format g drawn from $w(\cdot \mid f)$, with response bodies independent of the request. Teacher-forced over the candidate final lines given prompt and body:

Proposition 1 (pinning) *Every minimizer of the population SFT objective assigns, given any trained request and body, probability* $1 - \rho_f$ *to the requested rendering,* $\rho_f w(g \mid f)$ *to each other*

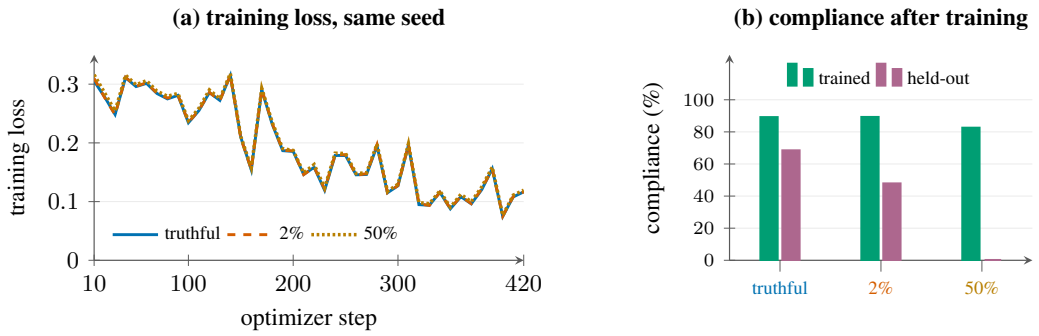

Figure 3: Trained-condition signals stay flat while transfer collapses (Qwen2.5-1.5B, formats). **(a)** Training loss, one seed per dose: the curves overlap (same-step differences $\leq$0.012 nats; Appendix F). **(b)** Arm-mean compliance after training: trained 89.0, 89.1 and 82.4%, held-out 68.3, 47.7 and 0.0%.

trained rendering and 0 to held-out renderings; its loss on truthful validation rows exceeds truthful training's by exactly $-\ln(1-\rho_f)$ *nats per example, all in the final line, whatever* w *is.*

Proposition 2 (indistinguishability) *Every trained-condition statistic (training or validation loss on trained formats, teacher-forced trained probabilities, greedy or sampled trained compliance) is fixed by* ρ *and* w *across all minimizers, while every assignment of behaviour to held-out requests is attained by some minimizer.*

Both follow from Gibbs' inequality (proofs in Appendix H); they state which quantities are fixed and by how much, and make three predictions, each checked on our checkpoints:

- *The trained-condition signal is tiny at low dose.* At 2% the objective moves validation loss by $-\ln 0.98 = 0.020$ nats per example, all in the final line, below the 0.035-nat seed-to-seed spread of that line's summed loss ($\approx$9.5 tokens; $\approx$0.004 per token), so a validation-loss check sees a shift smaller than seed noise. A teacher-forced test on 88 checkpoints confirms the scaling in trained networks: in mathematics the trained-condition excess is 1.31, 1.38 and 1.30 times $-\ln(1-\rho)$ at 1, 2 and 50%: one constant across a fifty-fold range of doses (0.028 nats at 2%, 95% CI 0.024–0.033; Appendix H.1), so networks learn the contradiction rate itself. The final-line validation excess at 50% is 0.737 nats (95% CI 0.652–0.823; $+0.079$ per token in Table 10) against $\ln 2 = 0.693$, and on training rows the 2% model's contradicted final lines cost 4.955 nats more than truthful ones, 87% of the population value $\ln((k-1)(1-\rho)/\rho) \approx 5.7$.
- *Equal pinned loss, different transfer.* A synthetic set reaches the same 50% excess (0.713 nats) while keeping 29.6% held-out compliance where the conventional set keeps 0.0, as Proposition 2 allows.
- *Wrong-format choice.* The pinned loss depends on the rate alone, and so, in the stylized dynamics below, does the cap on transfer. Contradictions falling back on the teacher's habitual format harm held-out success as much as uniformly drawn ones at the same rate (-0.3 and -2.9 points at 0.75 and 2%, five seeds; Appendix N.2), against a pairwise log-odds prediction of six-fold amplification.

A third proposition (Appendix H) derives the unidentified direction from the optimality conditions: for a shared coordinate under plain gradient flow, truthful training grows it unboundedly while contradiction caps it at a value set by ρ alone. At the population optimum the requested rendering stays most likely while $1-\rho > \rho \max_g w(g \mid f)$ (for uniform w, $\rho < (k-1)/k$), so greedy trained compliance need not move; trained networks deviate from the optimum prompt by prompt, and realized trained compliance tracks the requested format's log-odds margin $\ln((k-1)(1-\rho)/\rho)$: 2.9 nats at 25%, 1.8 at 50% and 0.7 at 75%, where measured trained compliance is 88, 82 and 44%.

The dose acts on log-odds Teacher-forced on the same bodies, a held-out request's log-odds split into a *pull* (how far the exhibited rendering beats the other held-out renderings) and a *default* (how far the trained renderings, whatever the request, beat the held-out ones). Contradiction moves both

the wrong way: it caps the pull, which Proposition 1 pins on trained requests at $\ln((k-1)(1-\rho)/\rho)$ (measured 1.9 nats at 50%, against $\ln 6 = 1.8$) and which trained and held-out formats share; with the pull capped, the objective can keep held-out renderings off trained requests only by raising the default. In a two-coordinate gradient flow this makes held-out log-odds fall linearly in $\mathrm{logit}(1-\rho)$ at about two nats per nat (Appendix H), steeply at the first contradictions and then flattening, whereas trained-condition metrics follow the loss scale $-\ln(1-\rho)$. With that slope and nothing fitted, the change from each student's lowest contradicted dose accounts for mean held-out log-odds in eight further cells (Qwen2.5-1.5B formats at 5 and 50%, Qwen2.5-7B at 50%, Llama-3.2-3B at 15 and 25%, templates at 5, 10 and 50%) within 0.65 nats on average, against 4.45 for no further change and 45 for the loss scale. Fixed before any score existed, the same law then predicted 20 newly trained Qwen2.5-1.5B checkpoints at 5, 10, 15, 25 and 75% contradiction and under uniform label noise (mean error 0.64 nats; no further change: 5.04) and a student never teacher-forced before, Qwen2.5-0.5B, at 5 and 50% (1.00; no further change: 5.22). The same shift that reverses held-out requests (2.6 to -2.6 nats at 2%) leaves trained requests, which start 11–24 nats ahead, above 99% teacher-forced accuracy. The instruction is still read: at 50% the requested rendering beats every other held-out one on 77–89% of problems (17% if ignored) yet loses to a trained one.

The training format set can decide what the objective leaves free Holding model, response bodies, steps and contradicted rows fixed at 50%, seven synthetic keyword–bracket formats keep 29.6% held-out compliance where seven conventional formats keep 0.0, at equal pinned loss; the advantage holds under both instruction wordings (17.0 points, 95% CI 9.1–24.9, and 36.2, 25.6–46.9), and the rescued formats are those dying last in the dose series (Kendall $W = 0.739$). Three further seven-format sets, drawn independently from the same generator and fixed before training, replicate it (35.4, 28.9 and 19.8% against 0.0% for the conventional set at 50%, three seeds each; Appendix G). The synthetic set protects through the default: at 50% its trained renderings lead the held-out ones by 8.2–10.4 nats, against 15.2–15.9 for the conventional set, at a similar pull.

6 Related Work

Label-noise studies find networks robust when enough clean examples remain (Rolnick et al., 2017; Natarajan et al., 2013); we keep every response, change only whether its instruction matches, and measure transfer to instructions absent from training. Narrow fine-tuning is known to erode capabilities outside its distribution, including format following (Luo et al., 2023; Kotha et al., 2024; Wang et al., 2024; Lin et al., 2025; Pyatkin et al., 2025); we identify a data defect that causes such erosion while trained-condition metrics stay flat. Unlike poisoning (Wan et al., 2023; Shu et al., 2023), the contradictions have no target, and data selection by quality, difficulty or influence (Zhou et al., 2023a; Liu et al., 2024a; Xia et al., 2024b) misses them (Table 4). Format-following benchmarks (Xia et al., 2024a; Jiang et al., 2023; Sclar et al., 2024) and constrained decoding (Willard & Louf, 2023; Geng et al., 2025) are complementary.

7 Scope

Our students span 0.5B to 8B parameters in five model families, fine-tuned with full-parameter SFT on single-task sets of 2,267–2,400 examples, and our evaluations measure instructions that show or name their target and whose compliance a program verifies: answer formats, response templates, output language, schemas and constraints of the kind IFEval and IFBench test.

8 Conclusion

A few training responses that disobey their own instruction, a defect public data contain and standard filters keep, silently erode transfer to held-out formats and languages. The effect is causal (deletion, neutral, decoy and random-deletion controls, at 2% as at 50%), costs 19–62 points at 5% in every student from 0.5B to 8B, occurs in public data, and is invisible to trained-condition metrics, which the objective pins to the contradiction rate while held-out log-odds fall in its logit. Two practices follow: evaluate on requests held out of fine-tuning as well as on trained conditions; and filter with each instruction's own verifier or, where none exists, the label-free swap test, whose removal restored transfer in mathematics.

Reproducibility Statement

Appendix A specifies the corpus (§A.1), arm construction (§A.2), formats (Table 5), hyperparameters (§A.3), decoding (§A.4) and the pre-training validation gate (§A.5); per-seed values are in Appendix D. The supplementary material contains the code for data construction, training, generation and scoring; the per-cell summary files from which every per-seed table is computed; the scripts and retained inputs that re-derive every reported statistic (Kendall's W, the equivalence tests and the natural-data audits); a README mapping each table to its files and script; frozen training and scoring environments; and a SHA-256 manifest. Training and generation use Python 3.11.11, PyTorch 2.11.0, transformers 5.14.1, trl 1.9.2, accelerate 1.14.0 and vLLM 0.26.0 in one environment shared by every run; scoring uses math-verify 0.9.0. Greedy vLLM decoding is not bit-identical across GPU architectures or batch sizes; the headline 2% contrast is replicated on one GPU class at one fixed batch (Appendix N.1). The generated SFT sets (543 MB) and all raw generations (3.9 GB) will be released publicly with the code upon publication.

AI Use Statement

We used generative AI tools extensively in this work. Language-model agents assisted with research ideation and experimental design (proposing controls, drafting pre-registrations and decision rules, and adversarially reviewing claims), wrote and ran experiment, data-construction and analysis code under our direction, helped find and summarize related work, and helped draft and edit the text. Every reported number is the output of an executed run and was re-derived from retained generation files; no number was imputed or filled in by a model. Language models served as blinded adjudicators in disclosed places: the two-direction checker audits and gates, the Tülu validation sample (Appendix J) and a blind classification of IFEval instruction types (Appendix I). AI-written code was checked by round-trip and cross-talk gates (§A.5) and by recomputing tables from raw generations. Language models are also the object of study: they produced the fine-tuning traces and every evaluated response (Appendix A). The authors reviewed all AI-assisted work and take full responsibility for the content, claims and artifacts.

Ethics Statement

This work fine-tunes and evaluates open-weight language models on public mathematics and instruction-following benchmarks. It involves no human subjects, personally identifying data, or sensitive attributes. We recommend auditing instruction–response agreement before training and evaluating on instructions held out of fine-tuning. These are data-construction questions with no foreseeable dual-use risk beyond that of the underlying models. Compute is reported in Appendix A.3.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. AlpaGasus: Training a better Alpaca with fewer data. In *International Conference on Learning Representations (ICLR)*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. JSONSchemaBench: A rigorous benchmark of structured outputs for language models. *arXiv preprint arXiv:2501.10868*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hugging Face. Math-Verify: A robust mathematical expression evaluation system. https://github.com/huggingface/Math-Verify, 2025.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. *arXiv preprint arXiv:2310.20410*, 2023.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations (ICLR)*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Xiaofeng Lin, Hejian Sang, Zhipeng Wang, and Xuezhou Zhang. Debunk the myth of SFT generalization. *arXiv preprint arXiv:2510.00237*, 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 *The Twelfth International Conference on Learning Representations*, 2024a.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh Murthy, Liangwei Yang, Silvio Savarese, Juan Carlos Niebles, Huan Wang, Shelby Heinecke, and Caiming Xiong. APIGen: Automated pipeline for generating verifiable and diverse function-calling datasets. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024b.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing o1-preview with a 1.5b model by scaling rl. https://github.com/agentica-project/deepscaler, 2025.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2013.

OLMo Team. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design, or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2024.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S. Dhillon, and Sanjiv Kumar. Two-stage LLM fine-tuning with less specialization and more generalization. In *International Conference on Learning Representations (ICLR)*, 2024.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. FoFo: A benchmark to evaluate LLMs' format-following capability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 680–699, 2024a.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54104–54132, 2024b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023a. arXiv:2305.11206.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023b.

A Methods

A.1 Data

We sample 6,000 DeepScaleR problems with a fixed seed and generate eight traces per problem from Qwen2.5-Math-1.5B-Instruct (temperature 1, top-p $=$ 1, 1,024-token cap) under a boxed-answer prompt. A trace is kept if Math-Verify accepts its answer, it terminates naturally with a balanced boxed span on a one-line final answer, and its body has at least 40 characters; at most two traces per problem enter the pool of 2,267 responses, whose bodies every arm reuses byte for byte.

A.2 ARMS

A row's rendered format is $g = \text{pool}[\,\text{md5}(\texttt{seed:offset:problem_id}) \bmod |\text{pool}|\,]$ over the arm's format pool (Table 5). The neutral system message is "You are a helpful assistant that solves mathematics problems."; a truthful message is the rendered format's instruction. `cond`p, $p = 1 - \rho \in \{0.25, 0.50, 0.75, 0.85, 0.90, 0.95, 0.98, 0.99\}$: each row independently keeps its truthful message with probability p and otherwise names a format drawn uniformly from the six others, so a contradicted row never names its own format; `shuf7` is $p = 0$. `cond`u: the named format is drawn from all seven, so the effective contradiction rate is 21.1% at $p = 0.75$ and 42.8% at $p = 0.50$. `cond`h, `cond`n, `cond`g ($\rho = 50\%$): the byte-identical truthful half of `cond`p0.5, with the contradicted half deleted, given the neutral message, or given a message naming one of six decoy formats that are never a target and never scored. `cond7half`, `condprand`050: as many rows as `cond`h keeps (1,111, 1,131 and 1,133 on seeds 0–2), drawn at random from `cond7` or from `cond`p0.5; both train `cond`h's 210–213 steps.

Table 5: Answer formats, rendered with answer X; each has its own extractor. Instructions read "End your response with a final line of the form:" followed by the rendering with `<answer>` for X (`therefore`: "a final sentence"; `\boxed{}` and `latex_answer`: "Put your final answer within `\boxed{}`" or "`\answer{}`"). `latex_answer` shares `\boxed{}`'s syntax and is excluded from held-out averages.

trained	rendering	held out	rendering
boxed	`$\boxed{X}$`	dbrackets	`[[X]]`
answer_line	`Answer: X`	angles	`<<X>>`
hashes	`#### X`	equals	`ANSWER = X`
json	`{"answer": "X"}`	yaml	`- answer: X`
xml	`<answer>X</answer>`	sigil	`%%X%%`
bold	`**Answer:** X`	chevron	`>>> X`
therefore	`Therefore, the answer is X.`	latex_answer	`$\answer{X}$`

A.3 TRAINING

Full-parameter SFT with TRL `SFTTrainer`: three epochs, learning rate 10^{-5}, cosine schedule with 0.03 warmup, AdamW, 2,048-token sequences, bf16, eager attention and loss on completion tokens only. Below 7B the effective batch is 16 on one GPU (426 optimizer steps on the 2,267 mathematics rows); the 7–8B students of Table 3 train with four-GPU FSDP at effective batch 64 (108 steps). Training ran on H100, H200, A100, A40, L40S and RTX-class GPUs.

A.4 DECODING AND SCORING

The system message requests the format and the user turn holds only the problem, with no exemplar. Decoding is greedy with vLLM (Kwon et al., 2023) and a 4,096-token context; GSM8K uses 500 test problems (Cobbe et al., 2021). *Compliance*: the requested format's own extractor recovers a non-empty answer span; *success*: that span also matches the gold answer under Math-Verify (Hugging Face, 2025).

A.5 VALIDATION GATE

No model is trained until every format passes three checks: (i) round trip, $\text{extract}(\text{render}(a)) = a$ for a battery of adversarial answers (fractions, nested braces, quotes, commas, percent signs); (ii) recoverability, every final line of every generated training file is recovered by its own extractor; (iii) cross-talk, no extractor fires on another format's rendering. The final gate reports zero failures over all fourteen formats.

B CONTAMINATION

Of the 500 MATH-500 problems, one (0.2%) matches a problem of the 6,000-problem sample exactly after whitespace and case normalisation, two (0.4%) reach 13-gram Jaccard overlap ≥ 0.5 (Brown et al., 2020) and one (0.2%) shares its gold answer with ≥ 0.8 unigram overlap; none of the 500 GSM8K problems meets any criterion. All arms share the same bodies, so any overlap is common to every arm and cancels in every contrast.

C RELEASED CHECKPOINTS AT A LARGER BUDGET

The base column of Table 3 decodes released checkpoints with 4,096 new tokens (OLMo-2-1B: 3,072, within its 4,096-position context) and Qwen3-8B with thinking off, because at 1,024 tokens 5.9–18.3% of their responses are truncated (Qwen3-8B with thinking on: 95.0%). At 1,024 tokens their held-out compliance is 9.1, 28.9, 56.0, 24.9 and 68.1% (Qwen2.5-1.5B, OLMo-2-1B, Llama-3.2-3B, Qwen2.5-7B, Qwen3-8B); at the larger budget 1.2–9.1% are truncated. At either budget every `diverse7` mean lies below the released model.

D MATHEMATICS: PER-SEED RESULTS AND CONTROLS

GSM8K, matched controls, per-seed values behind Table 3 and the reparse of mislabelled-arm responses (Tables 6–9); intervals are 95% Student-t over training seeds.

Table 6: GSM8K (500 test problems), held-out compliance / success (%), mean over seeds 0–2; *base*: released checkpoint at the larger budget. GSM8K's native answer line is `#### X`.

Student	base	`boxed_only`	`diverse7`	`shuf7`	`cond7`
Qwen2.5-1.5B	18.4 / 12.5	0.0 / 0.0	0.5 / 0.3	**0.0 / 0.0**	**65.1 / 46.5**
Llama-3.2-3B	78.7 / 50.2	0.0 / 0.0	3.2 / 2.2	**0.0 / 0.0**	**90.0 / 53.9**
OLMo-2-1B	42.9 / 24.5	14.1 / 5.1	26.7 / 10.2	**0.0 / 0.0**	**80.2 / 41.6**

Table 7: Matched controls at $\rho = 50\%$, Qwen2.5-1.5B: trained-format and held-out compliance (%). The first five arms share a byte-identical truthful half and differ only in the other half; the last two have `cond`h's rows and steps.

Arm	construction	trained	held-out per seed	mean	95% CI
`cond7`	all rows truthful	89.0	66.0/72.1/68.8/66.2	**68.3**	[63.7, 72.8]
`cond`h0.5	other half deleted	89.0	68.5/68.3/65.9	**67.6**	[64.0, 71.1]
`cond`n0.5	other half: neutral message	89.9	60.3/63.2/63.8	**62.4**	[57.8, 67.0]
`cond`p0.5	other half contradicted	82.4	0.0/0.0/0.0	0.0	[0.0, 0.0]
`cond`g0.5	other half: decoy format	89.3	0.0/0.0/0.0	0.0	[0.0, 0.0]
`cond7half`	random rows of `cond7`	88.7	67.6/64.2/69.5	**67.1**	[60.4, 73.8]
`condprand`050	random rows of `cond`p0.5	82.2	0.0/0.0/0.0	0.0	[0.0, 0.0]

E DOSE SERIES, CORRECTNESS AND VALIDATION LOSS

Correctness At $\rho = 5$, 10 and 15%, held-out success is 16.6, 8.2 and 2.4%, or 0.47, 0.48 and 0.43 of held-out compliance (truthful: 33.1%, 0.48), and trained-format success is 45.3, 45.3 and 44.0% (truthful 44.5%).

Validation loss Table 10 scores 500 DeepScaleR problems outside the training pool in trained formats under truthful instructions. Against the truthful reference, the conventional truthful cell of the format-set crossing (Appendix G), 50% contradiction moves completion loss by $+0.0022$ nats

Table 8: Held-out compliance (%) per seed on MATH-500 (seeds 0–3 for Qwen2.5-1.5B, 0–2 otherwise) for the arms of Table 3, whose means these are. In every student, every truthful seed exceeds every seed of the other three arms.

Student	`boxed_only`	`diverse7`	`shuf7`	`cond7`
Qwen2.5-1.5B	0.0/0.0/0.0/0.0	1.3/0.7/0.8/2.2	0.0/0.0/0.0/0.0	66.0/72.1/68.8/66.2
OLMo-2-1B	9.8/8.5/9.2	18.9/18.3/23.5	0.0/0.0/0.0	73.2/74.4/74.7
Llama-3.2-3B	0.0/0.3/0.0	3.6/3.0/2.8	0.0/0.0/0.0	74.8/74.4/78.8
Qwen2.5-7B	0.0/0.0/0.0	0.0/0.0/0.0	0.0/0.0/0.0	75.1/70.5/70.4
Qwen3-8B	0.3/0.2/0.1	5.4/4.3/8.2	0.0/0.0/0.0	75.9/72.0/76.1

Table 9: Every `shuf7` response (seed 0, 3,000 per student) reparsed with all 14 extractors (%): requested format, another format (always a trained one), only a last-line guess, or none.

Student	requested	other	last line	none
Qwen2.5-1.5B	0.0	91.6	8.4	0.0
OLMo-2-1B	0.0	84.9	15.1	0.0
Llama-3.2-3B	0.0	87.7	12.3	0.0
Qwen2.5-7B	0.0	93.0	7.0	0.0
Qwen3-8B	0.0	92.3	7.7	0.0

Table 10: Validation loss (mean token cross-entropy, seed mean $\pm$ SD) beside held-out compliance. *conv./syn.*: the conventional (per-format wording) and synthetic (wrapper) sets of Table 12; the synthetic cells' loss is over their own trained formats.

Arm	seeds	held out (%)	completion	final line
conv., truthful	3	68.4	0.3888 ± 0.0023	0.1231 ± 0.0046
`condp099` (1%)	7	63.4	0.3897 ± 0.0022	0.1217 ± 0.0052
`condp098` (2%)	7	47.7	0.3894 ± 0.0016	0.1231 ± 0.0037
`condp050` (50%)	7	0.0	0.3902 ± 0.0014	0.2009 ± 0.0045
conv., 50%	3	0.0	0.3910 ± 0.0027	0.2023 ± 0.0061
syn., truthful	3	60.9	0.3880 ± 0.0035	0.1218 ± 0.0038
syn., 50%	3	29.6	0.3894 ± 0.0016	0.1967 ± 0.0026

Table 11: Dose series on MATH-500: held-out compliance per seed, mean and 95% interval, and trained-format compliance (%). Trained compliance at $\rho = 5$–15% is over seeds 0–2; for Llama-3.2-3B it is over the three trained formats scored in every arm (`\boxed{}`, `Answer:`, XML).

Arm	ρ	held-out per seed	mean	95% CI	trained
Qwen2.5-1.5B					
`cond7`	0%	66.0/72.1/68.8/66.2	68.3	[63.7, 72.8]	89.0
`condp099`	1%	63.8/55.7/69.2/67.8/63.5/64.5/59.6	63.4	[59.1, 67.7]	89.8
`condp098`	2%	44.6/46.0/57.8/42.0/55.5/39.0/49.3	47.7	[41.3, 54.1]	89.1
`condp095`	5%	29.0/42.0/47.1/38.3/34.0/23.4/34.6	35.5	[28.2, 42.8]	89.5
`condp090`	10%	15.5/11.5/28.3/17.0/14.6/21.2/11.6	17.1	[11.6, 22.6]	88.9
`condp085`	15%	3.2/0.9/9.1/6.9/11.1/1.6/6.1	5.6	[2.0, 9.1]	88.4
`condp075`	25%	0.0/0.1/0.7	0.3	[0.0, 1.1]	87.8
`condp050`	50%	0.0/0.0/0.0	0.0	[0.0, 0.0]	82.4
`condp025`	75%	0.0/0.0/0.0	0.0	[0.0, 0.0]	43.8
`shuf7`	100%	0.0/0.0/0.0/0.0	0.0	[0.0, 0.0]	3.0
Llama-3.2-3B					
`cond7`	0%	74.8/74.4/78.8	76.0	[69.9, 82.1]	87.0
`condp095`	5%	30.6/34.9/38.9	34.8	[24.5, 45.1]	87.6
`condp085`	15%	7.0/7.7/9.2	8.0	[5.1, 10.8]	88.2
`condp075`	25%	1.2/2.2/1.2	1.5	[0.1, 3.0]	86.4
`condp050`	50%	0.0/0.0/0.0	0.0	[0.0, 0.1]	75.0

(95% Welch CI -0.004 to 0.008) and raises final-line loss by 0.079 (0.067–0.092); the synthetic set, which keeps 29.6% held-out compliance at 50%, shows nearly the same rise (0.075).

F Training Trajectories

One run per dose (Qwen2.5-1.5B, seed 0) is scored every 43 steps on the six held-out and seven trained formats, 500 problems each, at one evaluation geometry. At 5% contradiction, held-out compliance is at the truthful level after 43 steps and then falls while trained-format compliance rises; at 50% it is 0.0% throughout (compliance, %):

optimizer step	43	86	129	172	215	258	301	344	387	426
held-out, truthful	67.6	60.5	65.6	66.0	67.7	67.4	64.8	66.1	63.6	64.2
held-out, 5%	66.4	51.1	46.7	36.7	49.1	46.6	37.8	32.5	26.8	28.1
held-out, 50%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
trained, 5%	86.8	83.0	84.8	88.3	89.2	87.0	90.2	90.3	88.4	89.1

Logged training loss (seed 4, Figure 3a) differs from the truthful run's by at most 0.004 nats at any logged step at 2% and by at most 0.011 at 50%.

G Training Format Sets

A generator composes 128 synthetic renderings from eight keywords, two separators and eight bracket pairs (e.g. `RESULT = {{X}}`); sets are fixed before training and use the same response bodies and steps. Held-out compliance is always on the six held-out formats of Table 5. Two instruction wordings are crossed with the format set: per-format wording, and one wrapper for every format ("End your response with a final line of the form:" followed by the rendering). From truthful training to 50% contradiction the synthetic set loses less held-out compliance than the conventional set, by 17.0 points (95% CI 9.1–24.9, paired over seeds) under per-format wording and by 36.2 (25.6–46.9) under the wrapper (Table 12). The synthetic set keeps exactly the held-out formats that survive longest in the dose series: at 50% (wrapper) it keeps sigil 62.0, yaml 53.9 and dbrackets 46.1% against angles 12.5, chevron 3.1 and equals 0.0%, and at $\rho = 5$–15% the six formats survive in the order yaml, dbrackets, sigil, equals, angles, chevron (Kendall's $W = 0.739$ across seven seeds).

Table 12: Training format sets, Qwen2.5-1.5B: held-out compliance (%) per seed (mean). Upper block: format set × instruction wording; lower block: three further sets drawn independently from the same generator, at 50% contradiction under the wrapper (the conventional set: 0.0 on every seed).

format set, wording	truthful	50% contradicted
conventional, per-format	66.1/70.4/68.8 (68.4)	0.0/0.0/0.0 (0.0)
conventional, wrapper	62.9/69.7/70.1 (67.6)	0.0/0.0/0.0 (0.0)
synthetic, per-format	54.9/67.8/67.3 (63.3)	9.2/11.6/14.8 (11.9)
synthetic, wrapper	54.4/72.8/55.6 (60.9)	24.6/44.2/20.0 (29.6)
draw A	–	34.7/33.5/38.0 (35.4)
draw B	–	41.2/24.4/21.2 (28.9)
draw C	–	16.6/14.1/28.6 (19.8)

H Pinning and Indistinguishability: Statements, Proofs and Tests

Setup Formats $G = K \cup A$, with $|K| = k$ trained and A held out. The rendering $r_g(a)$ of answer a in format g is injective and line-delimited (the round-trip and cross-talk gates of Appendix A.5). A row has prompt $x = (s_f, q)$ exhibiting $f \in K$ and response $y = b \circ r_g(\alpha(b)) \circ \langle \mathrm{eot} \rangle$, where the body $b \sim P(b \mid q)$ determines the answer $\alpha(b)$. At dose ρ: f is uniform on K and independent of (q, b), and $g \mid f \sim \pi(g \mid f) = (1 - \rho_f)\,\mathbf{1}[g = f] + \rho_f\, w(g \mid f)$, with $w(\cdot \mid f)$ supported on $K \setminus \{f\}$ (`condp`: $\rho_f = \rho$, w uniform; the habitual arm of Appendix N.2: w a point mass on `\boxed{}`). The objective is population completion-only SFT, $L(p) = \mathbb{E}_D[-\ln p(y \mid x)]$ over conditional distributions p. Candidate-normalized teacher forcing is $\hat{P}_p(g \mid x, b) = p(r_g(\alpha(b)) \circ \mathrm{eot} \mid x, b) \,/\, \sum_{g' \in G} p(r_{g'}(\alpha(b)) \circ \mathrm{eot} \mid x, b)$.

Proposition 1 (pinning) p^* minimizes L iff $p^*(\cdot \mid x) = D(\cdot \mid x)$ on every trained prompt; then for every trained $x = (s_f, q)$ and body b: (a) $\hat{P}^*(f \mid x, b) = 1 - \rho_f$, $\hat{P}^*(g \mid x, b) = \rho_f w(g \mid f)$ for $g \in K \setminus \{f\}$, and $\hat{P}^*(h \mid x, b) = 0$ for $h \in A$; (b) on truthful validation rows from the same pool distribution, $\mathbb{E}[-\ln p^*(y \mid x)]$ exceeds the truthful minimizer's by exactly $\mathbb{E}_f[-\ln(1 - \rho_f)]$ nats per row, located entirely in the final line and independent of w; (c) the requested rendering remains the argmax of $\hat{P}^*$, so greedy trained compliance need not move, iff $1 - \rho_f > \max_g \rho_f w(g \mid f)$ (for uniform w: $\rho < (k-1)/k$).

Proof. By Gibbs' inequality, $L(p) = H(Y \mid X) + \mathbb{E}_X \,\mathrm{KL}(D(\cdot \mid X) \,\|\, p(\cdot \mid X)) \geq H(Y \mid X)$, with equality iff $p = D$ on the support of X, which is the set of trained prompts. Because y splits uniquely into (b, g) and f is independent of (q, b), $D(b \circ r_g(\alpha(b)) \circ \mathrm{eot} \mid s_f, q) = P(b \mid q)\, \pi(g \mid f)$. The candidates share the prefix (x, b), so the ratio defining $\hat{P}^*$ cancels $P(b \mid q)$ and the end-of-body factor, leaving $\pi(g \mid f)$: this is (a). For (b), on a truthful row $-\ln p^*(y \mid x) = -\ln P(b \mid q) - \ln \pi(f \mid f)$; the first term does not depend on ρ, and $-\ln \pi(f \mid f) = -\ln(1 - \rho_f)$ depends on ρ_f only. (c) is (a) compared across g. □

Proposition 2 (indistinguishability) Every trained-condition statistic is a functional of $\{p(\cdot \mid x) : x \text{ trained}\}$, hence a function of π alone on the minimizer set, while for every assignment q_x of distributions to held-out prompts some minimizer equals q_x there, at any ρ, including $\rho = 0$.

Proof. D gives held-out prompts probability 0, because every training instruction exhibits a trained format, so L depends on p only through trained conditionals; set $p = D$ on trained prompts and $p = q$ on held-out prompts. □

Proposition 3 (the unidentified direction; contradiction caps it) Consider any model class in which the exhibited-format log-odds is $\kappa(x) = \mathrm{logit}\, \hat{P}(\varepsilon(x) \mid x, b) = s + u_{\varepsilon(x)}$, with a shared coordinate s, per-format coordinates u_g, and the fallback distribution over other candidates parameterized by separate coordinates. (i) At $\rho_f = \rho > 0$, Proposition 1 pins the k sums $s + u_f = \mathrm{logit}(1 - \rho) =: \kappa^*$ for $f \in K$. The Jacobian of these conditions in (s, u_K) has rank

k in $k+1$ unknowns, with kernel $(1,-1,\dots,-1)$; along it every trained conditional, hence every trained statistic, is fixed while every held-out log-odds $\kappa(x_h) = s + u_h$ moves one for one. The held-out behaviour is exactly the unidentified coordinate. (ii) Under gradient flow on L with symmetric formats, $\dot{s} = -\sum_f \ell'(\kappa_f)$ and $\dot{u}_f = -\ell'(\kappa_f)$, so $s - \sum_f u_f$ is conserved. At $\rho = 0$, $\ell'(\kappa) < 0$ for every finite κ, because the named probability never reaches 1, so s increases without bound: truthful training keeps strengthening the shared coordinate. At $\rho > 0$ the flow converges to $s_\infty = s_0 + \frac{k}{k+1}(\kappa^* - \kappa_0)$: contradiction caps the shared coordinate at a value set by the rate ρ alone, independent of how wrong formats are distributed.

Proof. (i) is direct. (ii) With symmetric initialization $\kappa_f = \kappa$, $s - ku$ is constant and $s + u \to \kappa^*$; solving the two linear equations gives s_∞. □

Held-out side Add to the model of Proposition 3 an instruction-independent default δ: every held-out rendering's logit is $-\delta$ relative to the trained ones on every prompt, and a request exhibiting e adds the pull $s + u_e$ to e's rendering. On a trained request $\partial L/\partial\delta = -P(A \mid x) < 0$, so δ always grows. At $\rho = 0$ the pull grows as $\ln t$ and drives $P(A \mid x)$ to zero by itself, so δ grows only as $\ln\ln t$. At $\rho > 0$ the pull is capped at $\phi^* = \ln((k-1)(1-\rho)/\rho)$, the normalizer is $(k-1)/\rho$, and $\dot{\delta} = m\rho e^{-\delta}/(k-1)$, so $\delta \approx \ln(m\rho t/(k-1))$. A held-out request's log-odds against the trained class are $s + u_h - \delta - \ln k$, and u_h receives no gradient. Between two doses at equal t, Proposition 3(ii) gives $\Delta s = \frac{k}{k+1}\Delta\phi^*$, so $\Delta\kappa_H = \frac{k}{k+1}\Delta\Lambda - \Delta\ln\rho \approx (\frac{k}{k+1} + 1)\Delta\Lambda$ for small ρ, with $\Lambda = \mathrm{logit}(1-\rho)$. At fixed $\rho > 0$ the flow also predicts that held-out log-odds fall as $\ln t$ once the pull is capped, as the 5% trajectory shows (Appendix F).

Tests on existing checkpoints The pull on trained requests, the mean log-ratio of the requested to each other trained rendering, is 1.86–2.12 nats at 50% over ten Qwen2.5-1.5B checkpoints and 1.72–1.84 for Qwen2.5-7B, against $\ln 6 = 1.79$. κ_H is the mean over held-out requests of $s_h - \ln\sum_{g\neq h} e^{s_g}$ from the teacher-forced scores of Appendix H.1, at the realized dose. From each student's lowest contradicted dose (Qwen2.5-1.5B formats 2%, seven seeds; Qwen2.5-7B 5%; Llama-3.2-3B 5%; question answering 2%, five seeds), the slope $k/(k+1) + 1$ (1.875 for formats, 1.80 for templates) gives κ_H at 5 and 50% (-4.18, -9.84; observed -3.24, -9.80), Qwen2.5-7B at 50% (-9.19; -10.99), Llama-3.2-3B at 15 and 25% (-6.03, -7.28; -6.26, -7.55) and templates at 5, 10 and 50% (-24.11, -25.45, -29.41; -24.40, -25.94, -30.57): mean absolute error 0.65 nats (maximum 1.79), against 4.45 for no further change and 44.7 for the loss scale $-\ln(1-\rho)$. In every series the pull falls and the default rises monotonically with ρ.

Prospective tests With predictions and evaluator fixed (hashed) before any score existed, the law anchored at the 2% cells of Qwen2.5-0.5B (seeds 0–2), a student never teacher-forced before, predicts κ_H at 5% and 50% with errors 0.31 and 1.68 nats (mean 1.00), against 1.47 and 8.98 for no further change (mean 5.22). A second frozen evaluator scored 20 Qwen2.5-1.5B checkpoints retrained with the paper's recipe on the same data files, at 5, 10, 15, 25 and 75% contradiction (three seeds each) and under uniform label noise at effective 21 and 43% (three and two seeds), with three retrained 2% checkpoints as an alternative anchor. Per-arm errors from 5% on are 0.32, 0.21, 0.36, 0.55, 2.30, 0.23 and 0.51 nats against 1.46, 2.96, 3.69, 4.69, 11.66, 4.59 and 6.24 for no further change (mean 0.64 against 5.04); anchored on the three retrained 2% checkpoints instead, the mean error is 0.55.

H.1 Teacher-Forced Test of the Pinning Scaling

Teacher-forced over the candidate final lines given the prompt and the fixed body (Proposition 1), E is the mean surprisal of the requested rendering on trained requests, ΔE its excess over the truthful checkpoints and $R = \Delta E/(-\ln(1-\hat{\rho}))$ at the realized dose $\hat{\rho}$. Scores are computed in fp32 on one GPU class for 88 checkpoints, none missing; the same scores give κ_H above.

Qwen2.5-1.5B, mathematics	dose	ΔE (nats) [95% CI]	$-\ln(1-\hat{\rho})$	R
7 seeds	1%	0.013 [0.008, 0.018]	0.010	1.31
7 seeds	2%	0.028 [0.024, 0.033]	0.020	1.38
10 runs	50%	0.903 [0.884, 0.922]	0.694	1.30

I Instruction Class

A language-model rater, given only IFEval's official descriptions of its 25 instruction types (shuffled, without scores or this paper) and two definitions (the instruction exhibits a literal string or structure to reproduce, or it states a property to satisfy) assigned 16 types to the first class; three independent runs returned identical labels. The other nine are lowercase, all capitals, capital-word frequency, word count, sentence count, letter frequency, forbidden words, no commas and response language; response language is the attribute of the translation study (§4), whose instruction names its target.

J Prevalence of Natural Violations

Exact schema validation of the public xLAM set (Liu et al., 2024b) finds 1,424 of 58,147 eligible calls (2.45%, 95% Wilson CI 2.32–2.58%) violating the schema their own prompt supplies. Of the 3,492 rows of our Tülu 3 exhibited-literal pool (rows stating an exactly parsed literal constraint; Lambert et al., 2024), the official IFEval strict checkers flag 423 (12.1%) as violating their own constraint, counting only constraint types that pass a blinded audit of flagged and unflagged rows (precision at least 0.90, miss rate at most 0.05): title, multiple sections and highlighted sections (precision 1.00) and first word (0.93), each with no misses among 40 unflagged rows.

K Grounded Question Answering

The design is that of §4, with five paired seeds (101–105), 120 prompts per held-out template and 240 plain questions per seed, and vLLM with 512 concurrent sequences. Per held-out template, truthful to 2%: closing phrase 61.7 to 0.0, sections 83.8 to 81.2, highlights 78.0 to 40.8, paragraphs 34.7 to 17.2, repeated question 0.0 to 0.0%.

Table 13: Grounded question answering, Qwen2.5-1.5B (%): held-out-template compliance; trained-template compliance; held-out exact match (EM) and gold-span containment (GC) under an any-template parser; plain answers carrying a trained template (range over seeds); plain-question exact match (at 2%: 64.1 with the appended template removed).

arm	held-out, per seed	held-out mean	trained	EM	GC	plain: template	plain EM
untuned	–	30.8	75.0	1.7	82.7	0.4	50.0
0%	52.3/67.2/66.5/35.0/37.2	51.6 [32.5, 70.8]	99.9	39.1	80.6	0.8–1.7	56.4
2%	20.8/28.8/34.7/28.0/26.8	27.8 [21.7, 34.0]	99.6	50.0	77.7	96.7–97.9	2.8
5%	24.5/3.3/21.5/22.2/7.0	15.7 [3.6, 27.8]	98.4	52.4	77.4	97.5–99.2	4.8
10%	11.5/29.7/14.2/22.3/17.2	19.0 [10.0, 27.9]	95.5	49.7	76.3	97.5–100	6.8
50%	0.2/0.0/0.2/0.0/0.0	0.1 [0.0, 0.2]	42.3	64.8	76.1	100	39.4
del. contr.	62.7/42.5/39.3/71.7/45.3	52.3 [34.8, 69.8]	100.0	35.2	79.0	0.8–2.9	61.9
del. random	0.8/0.3/1.8/1.0/2.2	1.2 [0.3, 2.2]	48.2	61.3	73.3	99.2–100.0	42.2

L Output Language and Further Students

Translation uses 2,400 NTREX-128 English→X rows over seven trained languages (Arabic, Hebrew, Hindi, Japanese, Korean, Russian, Thai) and six held-out languages with distinct scripts (Armenian, Bengali, Georgian, Greek, Tamil, Telugu; 150 sentences each). A response complies if at least 80% of its letters are in the target script; fastText language identification agrees (held-out 93.1% after truthful training, 73.8% at 5%). At 5%, trained-language compliance falls 1.3 points and trained-language chrF++ 0.6, and 75.8% of held-out failures are in a trained language's script. The other studies repeat the mathematics (formats) or question-answering (templates) design on the named student (Table 14; means, differences and the 50% deletion arms are in Table 2).

Table 14: Held-out compliance (%) per seed behind the further rows of Table 2 (–: not run). Untuned SmolLM2-1.7B: 40.7. Qwen2.5-7B formats: a second series with its own truthful arm (first series: Table 8).

task, student	truthful	2% contradiction	5% contradiction
language, Qwen2.5-1.5B	89.3/90.9/90.9	78.4/84.6/80.1	67.4/75.9/70.8
templates, SmolLM2-1.7B	56.7/63.5/42.7	33.7/41.8/11.7	11.0/23.2/2.8
templates, Qwen2.5-7B	37.3/43.5/44.5/32.2/24.5	–	25.7/22.3/7.3/9.3/6.0
templates, Llama-3.1-8B	48.7/56.3/18.3/48.0/65.7	–	0.0/1.7/0.8/0.2/0.0
formats, Qwen2.5-0.5B	60.1/56.3/63.0	58.2/52.0/59.3	39.2/38.2/40.3
formats, Qwen2.5-3B	74.1/76.1/70.4	73.1/71.2/63.7	54.1/54.8/54.4
formats, Qwen2.5-7B	84.1/72.4/79.3	45.0/41.2/27.4	17.7/19.0/13.6

M Natural Violations: Training Studies

Qwen2.5-1.5B is fine-tuned on a stratified 10,000-call xLAM sample that keeps the natural violation rate (245 violating calls) and evaluated with the exact schema validator on 2,500 calls from held-out tool families, seven paired seeds per arm. Arms: RAW; FILTERED (violating calls removed); random deletion of as many calls; a cleaned set (CLEAN-REBAL) and the same set with the natural violations re-added (REINJECT). Intervals are seed-paired 95% t.

arguments per call	1	2	3	≥ 4	all
held-out calls	1,088	723	498	191	2,500
violation rate, training (%)	0.57	3.22	2.91	10.06	2.45
RAW compliance (%)	98.9	95.7	96.0	82.1	96.1
FILTERED − RAW	$+0.09$	$+1.42$	$+2.27$	$+9.87$ [8.03, 11.72]	$+1.66$ [1.34, 1.98]
FILTERED − random deletion	$+0.05$	$+1.40$	$+1.92$	$+8.53$ [7.38, 9.67]	$+1.46$ [1.10, 1.82]
REINJECT − CLEAN-REBAL	-0.04	-1.36	-1.72	-7.48 [-8.51, -6.44]	-1.33 [-1.54, -1.11]

M.1 Constraint Violations in Tülu 3

Qwen2.5-1.5B base, full-parameter SFT (2 epochs, global batch 64, learning rate 2×10^{-5}, completion-only loss), ten seeds, one GPU class and one evaluation pin; official IFEval strict checkers. Arms share the rows and order of the pool of Appendix J (3,492 rows, 423 violating): RAW; REPAIR (each violating response minimally edited so the checker accepts it; 379 of 423 validated, the other 44 identical in both arms). Endpoint: 720 prompts, 180 per type, each requesting a literal absent from every training row (untuned: 18.5%). REPAIR − RAW: $+1.61$ points (95% CI 0.70–2.52), positive on all ten seeds; on the three types the model learns (title, sections, highlighted sections; no arm learns first-word constraints) $+2.26$ (0.99–3.53), with sections $+4.00$ (1.33–6.67) and highlighted sections $+3.67$ (0.76–6.57). Trained-type validation compliance ($+0.67$, 90% CI -0.42 to 1.76) and validation loss ($+0.0005$ nats, 90% CI 0.0002–0.0008) are equivalent within ± 2 points and ± 0.02 nats. Detection: a label-free constraint-drop score (log-likelihood of the response under the instruction with its literal constraint removed, minus under the full instruction, released Qwen2.5-1.5B-Instruct) reaches AUROC 0.65 and precision 0.296 at a budget equal to the violator count, 2.4 times the base rate of 0.121; completion loss (0.44) and IFD (0.50) do not separate violators.

M.2 Natural Versus Synthetic Effects, Dose and Kind

Near a ceiling, points compress an effect, so Table 15 also reports the share of the violating arm's held-out errors that the clean arm removes, $(p_c - p_v)/(1 - p_v)$, and the held-out log-odds shift $\operatorname{logit} p_c - \operatorname{logit} p_v$ (per seed, then averaged, with seed-paired 95% t intervals). For tool calls, natural violations remove as large a share of held-out errors as synthetic contradiction causes at a similar rate in the same student, and their log-odds interval lies inside the synthetic one; per point of violation rate the natural shift (0.24 nats) is within a factor of two of the synthetic one (0.39) and above the synthetic 2% shifts of Qwen2.5-0.5B and 3B (0.07, 0.10). Complex calls are 7.6% of the evaluation and 7.9% of the training sample, so the evaluation mix mirrors xLAM's own; they carry 45.5% of

the overall gain, and removal eliminates 8, 33, 56 and 55% of errors on calls with 1, 2, 3 and $\geq$4 arguments. The Tülu 3 checkers of the kept types agree with blinded adjudication: precision 38/38, 41/41, 50/50 and 56/60 (title, sections, highlights, first word; exact 95% lower bounds 0.91, 0.91, 0.93, 0.84), no misses among 40 audited passes per type.

Table 15: Natural and synthetic effects on one scale (p_c, p_v: held-out compliance of the clean and the violating arm; seed-paired 95% intervals). Qwen2.5-1.5B-Instruct unless noted. Synthetic: held-out answer formats at one geometry (7 seeds) and Table 2.

contrast (clean vs violating)	rate	points	errors removed	log-odds (nats)
xLAM, all calls (FILTERED vs RAW)	2.45%	+1.66 [1.34, 1.98]	43%	0.58 [0.44, 0.72]
calls with $\geq$ 4 arguments	10.1%	+9.87 [8.03, 11.72]	55%	0.94 [0.68, 1.20]
xLAM, all calls (CLEAN-REBAL vs REINJECT)	2.45%	+1.33 [1.11, 1.54]	36%	0.46 [0.37, 0.54]
Tülu 3 sections (REPAIR vs RAW, 10 seeds)	7.0%	+4.00 [1.33, 6.67]	32%	0.43 [0.15, 0.71]
Tülu 3 learned types (REPAIR vs RAW)	7.0%	+2.26 [0.99, 3.53]	14%	0.18 [0.08, 0.28]
synthetic formats, truthful vs contradicted	2%	+18.5 [6.5, 30.5]	36%	0.77 [0.28, 1.27]
synthetic formats, Table 2	2%	+20.7	39%	0.86
synthetic, Qwen2.5-0.5B / 3B	2%	+3.3 / +4.2	8% / 14%	0.13 / 0.20

Replication against matched compliant rows In fresh runs on other hardware (24-GB GPUs, so every arm trains with 8-bit AdamW; otherwise the recipe above), we train on the cleaned set (CLEAN-REBAL) and on the same set with 245, 490 or 735 natural violators at fixed positions (the sample's own 245, as in REINJECT, then further naturally violating calls from training tool families outside the sample, matched to the violators' argument-count and length strata and nested across rates) and on a control that fills the same 735 positions with compliant calls of the same strata; three paired seeds.

natural violations in training	0	2.45%	4.9%	7.35%	0 (compliant control)
held-out schema compliance, all calls (%)	97.4	96.2	96.1	95.9	97.6
calls with $\geq$ 4 arguments (%)	90.1	82.7	82.7	81.3	91.4

Re-adding the natural violations costs 1.28 points (95% CI 0.63–1.93), matching the seven-seed estimate above (1.33). Against the compliant control, 735 violators cost 1.79 points (1.22–2.35) overall and 10.1 (5.9–14.3) on complex calls, while the control itself matches the cleaned set (-0.20, -0.82 to 0.42). The first violations do most of the harm: tripling the rate adds 0.31 points (-0.04 to 0.66), and held-out log-odds fall by 0.063 nats per point of violation rate (0.034–0.092; every seed).

Omission versus substitution at matched rows In the mathematics format study (Qwen2.5-1.5B-Instruct, five paired seeds; 24-GB GPUs, so every arm trains with 8-bit AdamW and is evaluated at a common batch of 128, where the truthful arm reproduces the one-geometry clean level, 66.5 against 66.9%), the same 113 rows (5%) are either *substituted* (the system message requests another trained format and the response is unchanged, the contradiction of §2) or *omitted* (the system message is truthful and the response's final answer is left bare, every other byte unchanged).

	truthful	substituted (5%)	omitted (5%)
held-out compliance, six formats (%)	66.5	27.7	55.2
five formats, without <<X>> (%)	77.5	31.4	63.6
held-out misses to a trained format / a bare answer	1,062 / 163	3,969 / 559	737 / 1,472
swap AUROC (precision at 113), seed-0 pool	–	0.997 (0.86)	0.966 (0.52)
drop AUROC (precision at 113), seed-0 pool	–	0.83 (0.24)	0.84 (0.27)

Both kinds erode transfer, substitution about 3.5 times as much: truthful minus substituted 38.8 points (95% CI 19.1–58.5), truthful minus omitted 11.2 (5.6–16.9), omitted minus substituted 27.6 (5.2–49.9); every seed orders truthful > omitted > substituted. Each kind installs its own default on held-out requests: substitution adds 2,907 misses to trained formats, omission 1,309 to bare answers (15,000 held-out prompts per arm). The swap test flags both kinds from the rendered final line; the drop test is weaker on both.

N Detecting and Removing Contradicted Rows

Higher scores flag rows as more suspect. *Completion loss*: mean token cross-entropy of the response given its prompt under the released Qwen2.5-1.5B-Instruct; *IFD*: that loss divided by the loss of the response alone; *Superfiltering* (Li et al., 2024): IFD computed with GPT-2; *LLM judge*: the negated 1–5 quality rating of Qwen2.5-7B-Instruct given the row's full prompt, system instruction included (Chen et al., 2024); *likelihood swap*: $\max_k \log p(s \mid k) - \log p(s \mid \text{own})$ under the released Qwen2.5-1.5B-Instruct, where s is the rendered span (the final line in mathematics, the whole response in question answering) and k ranges over the pool's other instructions (in question answering, the other template instructions with the same question and passage); on xLAM, whose instructions are unique, $\log p(s \mid \text{neutral}) - \log p(s \mid \text{own})$. The own-format verifier flags a row whose response fails its requested format's extractor.

Table 16: AUROC of each score for contradicted rows, per training pool (the *all pools* ranges of Table 4); xLAM: natural schema violations as labels, and the swap variant against a neutral instruction.

pool	rows	contradicted	verifier	swap	loss	IFD	Superfilt.	judge
math 2%	2,267	44	1.000	0.998	0.58	0.66	0.62	0.52
math 50%	2,267	1,156	0.998	0.997	0.55	0.64	0.57	0.50
QA 2%	2,400	48	1.000	0.999	0.81	0.80	0.64	0.58
QA 5%	2,400	120	1.000	0.999	0.80	0.81	0.65	0.57
QA 10%	2,400	240	1.000	0.999	0.80	0.82	0.63	0.55
QA 50%	2,400	1,200	1.000	0.999	0.80	0.81	0.63	0.55
xLAM	10,000	245	(labels)	0.45	0.44	0.60	–	0.58

N.1 Low-Dose Repair at One Geometry

Every cell trains and evaluates on one GPU class (48-GB L40S) at one pinned evaluation batch: mathematics, 2,267 rows, 426 updates, seeds 0–6, batch 256; question answering, 2,400 rows, 450 updates, seeds 101–105, batch 512. Each 2% pool contradicts exactly 45 (mathematics) or 48 (question answering) rows. *Targeted* deletion removes the contradicted rows, *random* deletion as many truthful rows, and *swap top 5%* the 5% of rows with the highest swap score, without labels.

held-out compliance (%)	truthful	2%	targeted deletion	random deletion	swap top 5%
mathematics (7 seeds)	66.9	48.4	67.7	41.9	67.8
question answering (5 seeds)	48.7	26.6	50.3	18.9	38.2

In mathematics, truthful minus 2% is 18.5 points (95% CI 6.5–30.5) and targeted minus random deletion 25.7 (15.9–35.6); the swap top 5% contains 43–45 of each seed's 45 contradicted rows, and removing it leaves compliance equivalent to truthful training ($+0.9$, 90% CI -2.3 to 4.1, within ± 5). In question answering, targeted minus random deletion is 31.4 (6.5–56.3) and removing the swap top 5% recovers 11.6 points (-6.5 to 29.7). Evaluation batch and GPU class move greedy outputs slightly: changing the batch from 129 to 256 moves single-format scores by at most 3.4 points, and re-reading the seven original 2% checkpoints at one pinned geometry moves each by -0.9 to $+0.6$ points (mean 47.8%).

N.2 Habitual-Format Contradictions

At 0.75% and 2% contradiction (five seeds each), contradicted rows either render the teacher's habitual `\boxed{}` format (w a point mass on `\boxed{}`) or draw the rendered format uniformly. The habitual-minus-uniform differences in held-out success (a compliant and correct final answer) are -0.3 points (95% CI -4.8 to 4.2) at 0.75% and -2.9 (-10.9 to 5.1) at 2%, and `\boxed{}` accounts for 8% and 14% of held-out misses, where six-fold amplification would put it near 70%: the damage depends on the contradiction rate and is independent of which wrong format the contradictions use.