Supervisor Error Structure Does Not Detectably Change Weak-to-Strong Recovery at Moderate Matched Error Rates: A Bounded Null on Binarized MNLI

Humanity First Research

Abstract

We set out to show that weak-to-strong generalization collapses when a weak supervisor's errors are systematic rather than random at a matched overall error rate, and we found no detectable effect. Corrupting gold binarized-MNLI labels at a fixed rate, we contrast random-uniform errors, errors concentrated in three of five genres (a sub-category recovered by a TF-IDF classifier at 0.712000 against 0.200000 chance), and errors clustered in a genre-blind partition of matched local density, then LoRA-finetune a Qwen2.5-7B student and measure performance-gap-recovered at eight seeds per arm. At the interpretable rate p=0.20 the systematic and genre-blind students differ by a seed-level raw accuracy of 0.000250 (p=0.937500, 95% CI [-0.005500, 0.006875]) and the systematic and random students by -0.001000 (p=0.726562, 95% CI [-0.005125, 0.003000]); nothing survives Holm correction. These p=0.20 intervals bound any structure effect below roughly 0.7 points of student accuracy and 0.055 of PGR, so the study excludes moderate-to-large effects but not small ones, and we frame the result as a bounded null rather than proof of exact equivalence. The pre-registered manipulation check passed, with between-genre error spread rising from 0.0198 to 0.1638 at matched rate, so the intervention was real and simply had no measurable downstream effect. We also report a methods lesson: an item-level bootstrap makes a sub-one-point gap (0.008750 raw accuracy) look significant at p=0.10 (p=0.007199), on the ill-conditioned p=0.10 performance-gap denominator and where the correct seed-level test does not (p=0.046875, Holm 0.187500), a textbook pseudo-replication that we caught by pre-registering the seed as the unit of analysis. The honest conclusion is that, for this task, scale, sub-category, and moderate error range, weak-to-strong recovery is insensitive to supervisor error structure, and a high-rate sensitivity control confirms the measurement is not inert, since recovery collapses at p=0.45 even though it stays flat across error structure at p<=0.20.

Introduction

Weak-to-strong generalization is the empirical core of scalable oversight (Burns et al. 2023; Irving et al. 2018; Bowman et al. 2022). Burns et al. showed a strong student trained on a weak supervisor's labels recovers most of the ceiling and explained it by the strong model's priors averaging out inconsistent mistakes (Burns et al. 2023). That explanation invites a sharp prediction we found attractive: if random mistakes are averaged away because they are inconsistent, then systematic mistakes that track an input-observable sub-category should be a learnable shortcut (Geirhos et al. 2020) the student absorbs, and recovery should collapse. We pre-registered that prediction as falsifiable: at a matched overall weak-label error rate,

Preprint. Under review.

the systematically corrupted student should recover a substantially lower performance-gap-recovered than a genre-blind clustered student of matched local density, provided genre is exploitable, and a null or reversed difference falsifies. Our contribution is the clean test that separates error structure from error rate, which neither the weak-to-strong literature (which varies organic supervisor strength, entangling the two) nor the noisy-label literature (which studies structured versus random noise but in ordinary supervised learning) provides (Natarajan et al. 2013; Xia et al. 2020), together with the finding that the prediction is not supported in this regime and a methodological caution about how it could have looked supported. We are explicit up front about scope. The safety concern is content-correlated supervisor error, the sycophantic or deceptive bias tied to what a text is about, whereas our manipulation deliberately uses genre, an observable but stylistic sub-category that is largely orthogonal to the entailment decision. Genre is therefore a controlled first probe of whether an exploitable non-content sub-category is absorbed, not a test of the content-correlated case itself, so the null we report bounds the former and does not settle the latter. We return to this transfer gap in the limitations, and we make no claim about content-correlated errors on the strength of this experiment.

Related work

The parent result is weak-to-strong generalization (Burns et al. 2023), the empirical face of scalable oversight (Irving et al. 2018; Bowman et al. 2022), where a strong student finetuned on weaker labels recovers much of ceiling. Its explanatory story rests on priors averaging out supervisor mistakes, but its weak supervisors are smaller models whose error rate and error structure move together and are never separately controlled. Learning with noisy labels supplies the structured-error vocabulary: class-conditional noise with risk-consistent correction (Natarajan et al. 2013), loss correction via a transition matrix (Patrini et al. 2017), the observation that networks fit consistent structure before memorizing inconsistent noise (Zhang et al. 2017; Arpit et al. 2017), instance-dependent and part-dependent noise as the genuinely harder regime (Xia et al. 2020), and surveys of the random-versus-structured gap (Song et al. 2022). All of it concerns ordinary supervised learning, where the labels define the target rather than merely eliciting a latent capability, so its prediction that structured noise hurts more does not automatically transfer to weak-to-strong. Shortcut learning is the candidate mechanism: models exploit features that are predictive in-distribution even when spurious (Geirhos et al. 2020), with worst-group collapse when a spurious attribute correlates with the label inside identifiable groups (Sagawa et al. 2020). Those setups make the spurious feature predictive of the corrupted label, a stronger manipulation than ours, which we discuss as a limitation. We work on binarized MultiNLI (Williams et al. 2018) with a Qwen2.5-7B student (Qwen Team 2025) adapted by LoRA (Hu et al. 2022) over a 4-bit base (Dettmers et al. 2023).

Methods

We use MultiNLI (Williams et al. 2018) binarized to entailment against non-entailment by pooling neutral and contradiction, with a genre-balanced training pool of 400 items per genre across fiction, government, slate, telephone, and travel (2000 items), class-balanced within genre, and a disjoint balanced validation_matched subset of 200 items per genre for evaluation. The weak supervisor is synthetic: we corrupt gold labels at an exact rate p so supervisor accuracy is 1-p by construction and identical across arms, and all flips are class-balanced so only error placement varies. The random-uniform arm flips a uniform floor(p*N) subset. The systematic arm concentrates the same flips in three of five genres, rotated per seed, with within-corrupted-genre rates audited at 0.184375 for p=0.10 and 0.353438 for p=0.20, both below 0.5. The genre-blind arm places the same flips in a random partition sized to the three corrupted genres but uncorrelated with genre. Rates tested are p in {0.10, 0.20}. The pre-registered manipulation check is the between-genre error spread, and a TF-IDF plus logistic-regression genre classifier established genre recoverability at 0.712000 against 0.200000 chance before we treated it as structure.

The student is Qwen2.5-7B (Qwen Team 2025) in 4-bit NF4 (Dettmers et al. 2023) with LoRA rank 16, alpha 32, dropout 0.05 on q, k, v, o, gate, up, down (Hu et al. 2022), trained one epoch at learning rate 2e-4, batch size 4, gradient accumulation 4, sequence length 160.

Each corrupted arm and the clean-label ceiling ran at eight seeds per rate, 56 finetunes for the main design, with a high-rate positive control (Results) adding six more at three seeds per rate for 62 finetunes in all; the exact recorded GPU time on a single RTX 5090 appears in the reproducibility appendix. PGR is (acc_student - acc_weak)/(acc_ceiling - acc_weak) with acc_weak = 1-p. The primary test is a paired seed-level bootstrap with exact sign-flip p and Holm correction; the item-level bootstrap is exploratory and, as we show, misleading if taken as primary. Two sanity properties hold on the same harness that scores the headline: the clean-label ceiling reaches 0.925500 and every corrupted arm lands between 0.915 and 0.927, all far above the 0.500000 constant-prediction floor, so the eval pipeline discriminates signal from chance rather than saturating trivially.

Results

The manipulation succeeded and the effect still did not appear. At p=0.20 the between-genre error spread was 0.0198 for random-uniform and 0.1638 for systematic against a 0.10 required effect, with genre-blind at 0.0144, so sub-category-correlated structure was genuinely present at matched overall rate. The clean-label ceiling reached 0.925500. Student accuracies barely moved across arms: at p=0.10, 0.920250 random, 0.926750 systematic, 0.918000 genre-blind; at p=0.20, 0.917000, 0.916000, 0.915750. PGR at p=0.20, with the well-conditioned denominator 0.125500, was 0.932271, 0.924303, and 0.922311.

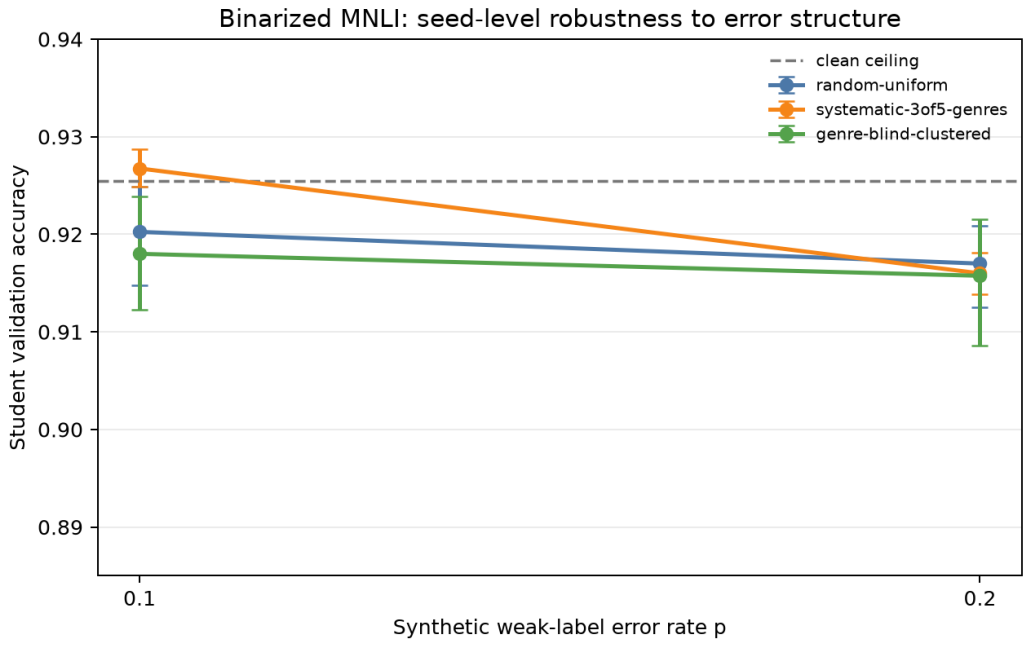

Figure 1: Mean student validation accuracy by corruption arm at each error rate, with seed-bootstrap intervals over eight finetunes and the clean-label ceiling. At the interpretable rate p=0.20 the three arms overlap; at p=0.10 they are numerically close but performance-gap-recovered is ill-conditioned because the weak-to-ceiling band is only 0.025500, so the indistinguishability claim rests on p=0.20.

The primary seed-level contrasts are null where it matters. At p=0.20 the systematic-minus-genre-blind raw accuracy difference is 0.000250 (p=0.937500, Holm 1.000000, 95% CI [-0.005500, 0.006875]) and the systematic-minus-random difference is -0.001000 (p=0.726562, Holm 1.000000, 95% CI [-0.005125, 0.003000]); as PGR these are 0.001992 (95% CI [-0.043825, 0.054781]) and -0.007968 (95% CI [-0.040837, 0.023904]). Nothing survives Holm at either rate. We read these intervals as a bound rather than as exact equivalence: the systematic-minus-blind accuracy interval [-0.005500, 0.006875] excludes any structure effect larger than about 0.7 percentage points of student accuracy, and its PGR interval [-0.043825,

0.054781] excludes PGR effects beyond about 0.055, so the design rules out moderate-to-large effects at p=0.20 while leaving small effects, below roughly half a point of accuracy, unresolved. The per-genre corrupted-minus-clean accuracy delta in the systematic arm is -0.000208 at p=0.10 and -0.004062 at p=0.20, so the structure left no downstream fingerprint, and per-genre clean-label accuracy spans only 0.913750 to 0.938125, ruling out a difficulty confound.

Positive control (sensitivity check)

A null across error structure is only meaningful if the same harness, task, and student can register degraded recovery when it is genuinely present, so we ran a positive control that varies the error rate rather than its structure: random-uniform corruption at the higher rates p=0.35 and p=0.45, scored identically to the main arms. Performance-gap-recovered traces a clear sensitivity curve. It is 0.794118 at p=0.10, though on the ill-conditioned 0.025500 denominator flagged above and therefore noisy, 0.932271 at p=0.20, 0.780399 at p=0.35, and 0.245894 at p=0.45, with the weak-to-ceiling denominators at p=0.20 and above all well conditioned. Recovery collapses at the highest rate, from a mean of 0.863194 over the low rates to 0.513147 over the high rates. A paired seed-level bootstrap contrasting the p=0.45 positive control against p=0.20 random-uniform gives a PGR difference of -0.659816 with a 95% CI of [-1.046806, -0.133612] that excludes zero; the exact sign-flip p is 0.250000, its floor at three seeds per positive-control rate, so we lean on the interval rather than the p-value. This shows the harness, task, and student together do register degraded weak-to-strong recovery when the supervision genuinely worsens, so the flat PGR across error structure at p<=0.20 is a genuine robustness finding rather than a measurement floor or ceiling that could not have moved.

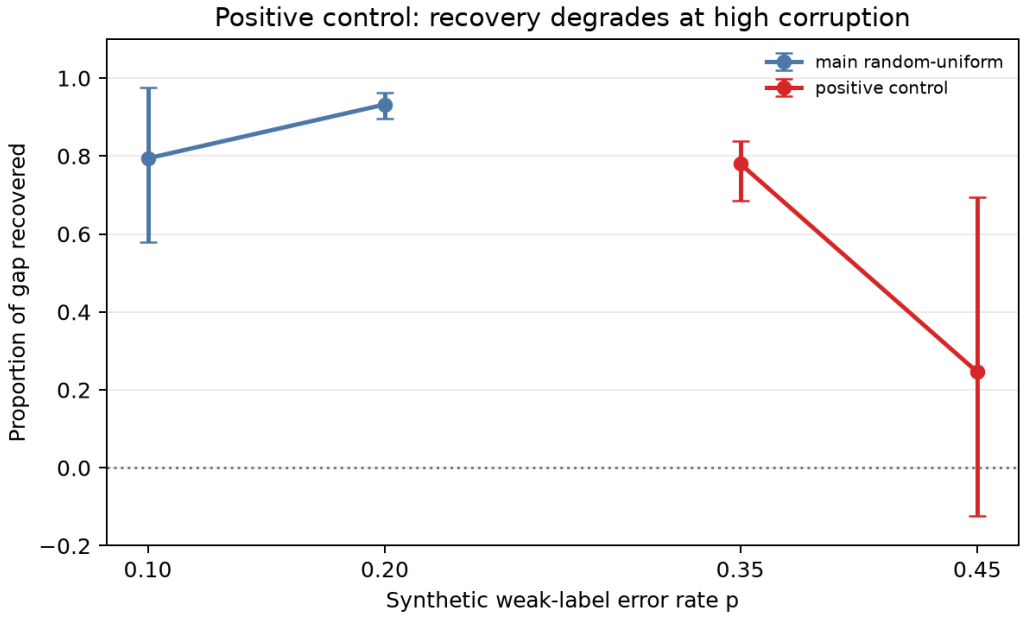

Figure 2: Performance-gap-recovered against weak-label error rate for random-uniform corruption. Recovery is high through p=0.35 and collapses at p=0.45, confirming the measurement registers degraded recovery when it is present.

The reference conditions we had hoped would anchor the comparison to organic weak-to-strong were degenerate, and we report the diagnosis rather than quietly dropping them. Both the finetuned organic Qwen2.5-0.5B and the strong zero-shot Qwen2.5-7B predicted constant non-entailment on all 1000 evaluation items, scoring exactly 0.500000 on the balanced set. The cause is a base-model prior under the binarized, pooled label scheme: with non-entailment absorbing both neutral and contradiction and with no or few in-context examples, both models default to the majority surface response and never emit entailment.

This is a property of the untrained references, not of the scorer, because the same harness assigns the LoRA-finetuned students 0.915 to 0.927 and the clean ceiling 0.925500, all cleanly separated from the 0.500000 constant floor, and the manipulation and exploitability checks on the same data pipeline behave exactly as designed. The consequence is that we lose the organic-weak-supervisor bridge and anchor PGR on the synthetic 1-p floor, which we state plainly as a scope restriction rather than a repaired baseline.

The methods lesson concerns the one place a naive analysis would have claimed a result. At p=0.10 the systematic-minus-genre-blind raw accuracy difference is 0.008750, under a tenth of a point, yet the item-level bootstrap assigns it p=0.007199, apparently significant. The correct seed-level test on the same data gives p=0.046875, which does not survive Holm correction (adjusted 0.187500), and the p=0.10 PGR is anyway uninterpretable because its denominator is 0.025500. The item-level test treats a thousand correlated evaluation items as independent draws and ignores the between-seed variance that actually governs replication, so it manufactures confidence from pseudo-replication. Because we pre-registered the seed as the unit of analysis, this false positive never entered the headline. We flag it because the item-level number is the one an unwary reader, or an earlier version of ourselves, would have reported.

Discussion

The honest summary is that the prediction was clean and, within this regime, unsupported. At a matched overall error rate, and with a control that matches local corruption density while stripping observable alignment, a Qwen2.5-7B student recovers the same accuracy whether the supervisor's errors are random, genre-systematic, or blind-clustered, up to a bound of roughly 0.7 accuracy points at p=0.20. This extends the noisy-label observation that models fit consistent structure before inconsistent noise (Arpit et al. 2017) into the elicitation regime, but with a caveat we take seriously: the structured errors here were not merely tolerated more slowly, they were tolerated as well as random errors at moderate rates and short training, and we cannot exclude that a stronger, more predictive shortcut or longer training would change that. A bounded null with a passed manipulation check, a density-matched control, and eight seeds is more informative than an underpowered positive, and it says that in this setting the safety-relevant variable is supervisor error rate rather than error structure. The pseudo-replication episode is a reminder that the unit of analysis is a modeling decision with the power to flip a conclusion.

Limitations

The evidence is one task family, binarized MultiNLI, one student scale, Qwen2.5-7B, and LoRA rather than full finetuning, so the general-sounding phrase weak-to-strong should be read as this specific pair and adaptation. We capped the error rate at p=0.20 to keep within-sub-category rates below 0.5, so that structure is not confounded with catastrophic local label inversion; a consequence is that in the corrupted genres the label remains majority-correct, so genre never becomes majority-predictive of the wrong label, which makes our manipulation weaker than the spurious-correlation setups where the feature predicts the corrupted label. Training is short, one epoch at 125 steps, and the noisy-label literature we cite suggests shortcut exploitation can require more exposure, so the null is compatible with undertraining and we did not ablate training duration. Accuracy is compressed near the 0.925500 ceiling in every arm, which limits headroom and hence power. The weak supervisor is a synthetic gold-corruption model, chosen for matched-rate control rather than realism, and our organic anchors were degenerate as diagnosed above, so the study does not connect to an organic weak model. Genre is largely orthogonal to the entailment reasoning, unlike the content-correlated deceptive or sycophantic errors the safety framing invokes, so transfer of this null to the motivating setting is speculative. The conclusion may not hold at higher error rates, larger weak-to-ceiling gaps, other sub-category definitions, or larger scales.

Limitations and reviewer responses

A reviewer will note that a null result is only as strong as its power, so we pre-concede and quantify it rather than claim a well-powered equivalence: at p=0.20 the systematic-minus-blind accuracy interval [-0.005500, 0.006875] bounds any effect below about 0.7 accuracy points and 0.055 PGR, which excludes moderate-to-large effects but not small ones, and

this is a post-hoc bound from eight-seed intervals, not a prospective power analysis. A reviewer will ask for a positive control demonstrating the pipeline can detect degraded recovery at all. We now provide one: a high-rate random-uniform sensitivity check in which performance-gap-recovered collapses from 0.932271 at p=0.20 to 0.245894 at p=0.45, with a paired seed-level PGR difference of -0.659816 (95% CI [-1.046806, -0.133612]) that excludes zero, which shows the harness, task, and student do register degraded weak-to-strong recovery when supervision genuinely worsens, so the flat PGR across error structure at p<=0.20 is a real robustness finding rather than an insensitive measurement. We keep the honest residual limit: this control degrades the error rate through random corruption rather than engineering a stronger label-predictive structure, so it certifies the measurement is not inert without proving that some more extreme structured shortcut could never bite, and the structure contrast itself remains bounded to p<=0.20. A reviewer will observe that the manipulation is weak because genre stays below the 0.5 local rate and is stylistically orthogonal to entailment; both are true, both are deliberate consequences of avoiding label inversion and of choosing an observable but content-neutral sub-category, and both bound the claim to density-structured, non-sign-inverting, style-correlated error. A reviewer will raise undertraining, since one epoch may be too little for shortcut absorption; we agree it is a live confound, note that training nonetheless suffices to reach a 0.925500 ceiling on the task itself, and flag a training-duration ablation as needed. A reviewer will note near-ceiling compression; we agree the small absolute band limits power, which is why we report the bound in PGR against the 0.125500 band rather than resting on raw accuracy alone. A reviewer may ask whether the degenerate baselines indict the harness; we show they do not, because the same harness cleanly separates the finetuned arms and the ceiling from the 0.500000 constant floor, and we attribute the degeneracy to base-model priors under the pooled binarization. A reviewer may prefer the item-level significance at p=0.10; we argue against it explicitly, because item-level pseudo-replication is precisely what produces the spurious p=0.007199, contradicted by the seed-level p=0.046875 that does not survive Holm and rests on a degenerate 0.025500 denominator. A reviewer may note that the uncorrected systematic-minus-genre-blind gap is larger at p=0.10 (0.008750) than at p=0.20 (0.000250), the opposite of the dose-dependence a genuine shortcut effect would show. We read the pattern as an artifact rather than a signal: the p=0.10 gap is nominal only, fails Holm at 0.187500, and its PGR rests on a degenerate weak-to-ceiling band of 0.025500 that amplifies small accuracy noise, whereas the p=0.20 comparison has the well-conditioned 0.125500 denominator and is flatly null, so a structure effect that strengthened with more errors would appear at p=0.20 and does not.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 62 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
clean-label-ceiling-p0.00-seed0	Qwen2.5-7B	multi_nli(binarized)	zero-shot	2000	0	125 steps; 1 epochs; params=7B	1.99
clean-label-ceiling-p0.00-seed1	Qwen2.5-7B	multi_nli(binarized)	zero-shot	2000	1	125 steps; 1 epochs; params=7B	1.99

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
clean-label-ceiling-p0.00-seed2	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	2	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.10-seed0	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	0	125 steps; 1 epochs; params=7B	1.99
systematic-3of5-genres-p0.10-seed0	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	0	125 steps; 1 epochs; params=7B	1.99
genre-blind-clustered-p0.10-seed0	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	0	125 steps; 1 epochs; params=7B	1.99
random-uniform-p0.10-seed1	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	1	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.10-seed1	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	1	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.10-seed1	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	1	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.10-seed2	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	2	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.10-seed2	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	2	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.10-seed2	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	2	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed0	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	0	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed0	Qwen2.5-7B	multi_nli(blind-free)	lora-sft	2000	0	125 steps; 1 epochs; params=7B	2.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
genre-blind-clustered-p0.20-seed0	Qwen2.5-7B	multi_nli (bl	oralized)	2000	0	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed1	Qwen2.5-7B	multi_nli (bl	oralized)	2000	1	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed1	Qwen2.5-7B	multi_nli (bl	oralized)	2000	1	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.20-seed1	Qwen2.5-7B	multi_nli (bl	oralized)	2000	1	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed2	Qwen2.5-7B	multi_nli (bl	oralized)	2000	2	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed2	Qwen2.5-7B	multi_nli (bl	oralized)	2000	2	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.20-seed2	Qwen2.5-7B	multi_nli (bl	oralized)	2000	2	125 steps; 1 epochs; params=7B	2.00
clean-label-ceiling-p0.00-seed3	Qwen2.5-7B	multi_nli (bl	oralized)	2000	3	125 steps; 1 epochs; params=7B	2.02
clean-label-ceiling-p0.00-seed4	Qwen2.5-7B	multi_nli (bl	oralized)	2000	4	125 steps; 1 epochs; params=7B	2.00
clean-label-ceiling-p0.00-seed5	Qwen2.5-7B	multi_nli (bl	oralized)	2000	5	125 steps; 1 epochs; params=7B	2.00
clean-label-ceiling-p0.00-seed6	Qwen2.5-7B	multi_nli (bl	oralized)	2000	6	125 steps; 1 epochs; params=7B	2.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
clean-label-ceiling-p0.00-seed7	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.10-seed3	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	3	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.10-seed3	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	3	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.10-seed3	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	3	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.10-seed4	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	4	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.10-seed4	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	4	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.10-seed4	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	4	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.10-seed5	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	5	125 steps; 1 epochs; params=7B	2.01
systematic-3of5-genres-p0.10-seed5	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	5	125 steps; 1 epochs; params=7B	2.01
genre-blind-clustered-p0.10-seed5	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	5	125 steps; 1 epochs; params=7B	2.01
random-uniform-p0.10-seed6	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	6	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.10-seed6	Qwen2.5-7B	multi_nli(blind-rated)	lora-sft	2000	6	125 steps; 1 epochs; params=7B	2.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
genre-blind-clustered-p0.10-seed6	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	6	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.10-seed7	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.10-seed7	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.10-seed7	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed3	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	3	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed3	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	3	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.20-seed3	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	3	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed4	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	4	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed4	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	4	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.20-seed4	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	4	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed5	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	5	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed5	Qwen2.5-7B	multi_nli (blind)	lora-sft	2000	5	125 steps; 1 epochs; params=7B	2.01

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
genre-blind-clustered-p0.20-seed5	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	5	125 steps; 1 epochs; params=7B	2.01
random-uniform-p0.20-seed6	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	6	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed6	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	6	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.20-seed6	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	6	125 steps; 1 epochs; params=7B	2.00
random-uniform-p0.20-seed7	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
systematic-3of5-genres-p0.20-seed7	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
genre-blind-clustered-p0.20-seed7	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	7	125 steps; 1 epochs; params=7B	2.00
positive-control-random-uniform-p0.35-seed0	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	0	125 steps; 1 epochs; params=7B	2.01
positive-control-random-uniform-p0.35-seed1	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	1	125 steps; 1 epochs; params=7B	1.99
positive-control-random-uniform-p0.35-seed2	Qwen2.5-7B	multi_nli (bl	ora-sft	2000	2	125 steps; 1 epochs; params=7B	2.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
positive-control-random-uniform-p0.45-seed0	Qwen2.5-7B	multi_nli(blind-mismatched)	lora-sft	2000	0	125 steps; 1 epochs; params=7B	1.99
positive-control-random-uniform-p0.45-seed1	Qwen2.5-7B	multi_nli(blind-mismatched)	lora-sft	2000	1	125 steps; 1 epochs; params=7B	2.00
positive-control-random-uniform-p0.45-seed2	Qwen2.5-7B	multi_nli(blind-mismatched)	lora-sft	2000	2	125 steps; 1 epochs; params=7B	2.00

Seed policy. The distinct RNG seeds recorded across the manifest are 0, 1, 2, 3, 4, 5, 6, 7. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
systematic_minus_agree-blind	PRIMARY-paired bootstrap over training seeds (10k resamples) with exact sign-flip p	0.008750000000000875	0.0078125	[0.002750000000000024, 0.014378124999999967]	8	8
systematic_minus_agree-blind	PRIMARY-paired bootstrap over training seeds (10k resamples) with exact sign-flip p	0.34313725490196075	0.0078125	[0.10784313725490022, 0.5638480392156857]	8	8

Claim	Test	Statistic	p	95% CI	n	Seeds
systematic_minus_random / blind	SECONDARY / EXPLORATORY: paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	0.0	1.0	[0.0, 0.015]	1000	8
systematic_minus_random / blind	SECONDARY / EXPLORATORY: paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	0.0	1.0	[0.0, 0.5882352941176477]	17	8
systematic_minus_random	PRIMARY: paired bootstrap over training seeds (10k resamples) with exact sign-flip p	0.006500000000000006	0.0625	[0.0013750000000000012, 0.01200000000000001]	8	8
systematic_minus_random	PRIMARY: paired bootstrap over training seeds (10k resamples) with exact sign-flip p	0.2549019607843126	0.0625	[0.05392156862745111, 0.4705882352941187]	8	8
systematic_minus_random / blind	SECONDARY / EXPLORATORY: paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	0.0	1.0	[0.0, 0.012625]	1000	8

Claim	Test	Statistic	p	95% CI	n	Seeds
systematic_minus_[illegible]	SECONDARY/EXPLORATORY: paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	[illegible]	[illegible]	[illegible], 0.49509803921568696]	[illegible]	8
systematic_minus_genre-blind	PRIMARY: paired bootstrap over training seeds (10k resamples) with exact sign-flip p	[illegible]	0.9375	[-0.005500000000000005, 0.006875000000000006]	8	8
systematic_minus_genre-blind	PRIMARY: paired bootstrap over training seeds (10k resamples) with exact sign-flip p	[illegible]	[illegible]	[-0.04382470119521918, 0.054780876494023974]	8	8
systematic_minus_genre-blind	SECONDARY/EXPLORATORY: paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	0.00025	0.9557044295570443	[-0.003875, 0.004375]	1000	8
systematic_minus_genre-blind	SECONDARY/EXPLORATORY: paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	[illegible]	[illegible]	[illegible] 0.030876494023904397, 0.03486055776892432]	1000	8

Claim	Test	Statistic	p	95% CI	n	Seeds
systematic_minus_random	PRIMARY paired bootstrap over training seeds (10k resamples) with exact sign-flip p	-0.0010000000000000009	0.7265625	[-0.0051250000000000046, 0.0030000000000000027]	8	8
systematic_minus_random	PRIMARY paired bootstrap over training seeds (10k resamples) with exact sign-flip p	-0.007968127490039851	0.7265625	[-0.040836653386454, 0.023904382470119553]	8	8
systematic_minus_random	SECONDARY/EXPLORATORY paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	-0.001	0.7216278372162783	[-0.005875, 0.00375]	1000	8
systematic_minus_random	SECONDARY/EXPLORATORY paired bootstrap over eval items (10k resamples), avg over seeds; ignores seed variance	-0.007968127490039844	0.7216278372162783	[-0.046812749003984085, 0.029880478087649414]	1000	8
positive_control_pgr_p0.45_minus_random_p0.20	SENSITIVITY paired bootstrap over shared training seeds with exact sign-flip p	-0.659816128296402	0.25	[-1.0468061177394288, -0.13361202327839128]	3	3

Compute. Total recorded GPU time across all experiments is 2.0664 GPU-hours on RTX 5090.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project reposi-

tory: `curve.csv`, `item_predictions.jsonl`, `per_genre_clean_label_accuracy.csv`, `runs_raw.csv`, `runs_raw.jsonl`, `systematic_per_genre_delta.csv`, `experiments.json`.

References

Arpit, Devansh, Stanisław Jastrzębski, Nicolas Ballas, et al. 2017. "A Closer Look at Memorization in Deep Networks." *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1706.05394.

Bowman, Samuel R., Jeeyoon Hyun, Ethan Perez, et al. 2022. "Measuring Progress on Scalable Oversight for Large Language Models." *arXiv Preprint arXiv:2211.03540*. https://arxiv.org/abs/2211.03540.

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, et al. 2023. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." *arXiv Preprint arXiv:2312.09390*. https://arxiv.org/abs/2312.09390.

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

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. 2020. "Shortcut Learning in Deep Neural Networks." *Nature Machine Intelligence* 2 (11): 665–73. https://arxiv.org/abs/2004.07780.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2106.09685.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. 2018. "AI Safety via Debate." *arXiv Preprint arXiv:1805.00899*. https://arxiv.org/abs/1805.00899.

Natarajan, Nagarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. 2013. "Learning with Noisy Labels." *Advances in Neural Information Processing Systems (NeurIPS)* 26.

Patrini, Giorgio, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/1609.03683.

Qwen Team. 2025. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*. https://arxiv.org/abs/2412.15115.

Sagawa, Shiori, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization." *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1911.08731.

Song, Hwanjun, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. "Learning from Noisy Labels with Deep Neural Networks: A Survey." *IEEE Transactions on Neural Networks and Learning Systems*. https://arxiv.org/abs/2007.08199.

Williams, Adina, Nikita Nangia, and Samuel R. Bowman. 2018. "A Broad-Coverage Challenge Corpus for Sentence Understanding Through Inference." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. https://arxiv.org/abs/1704.05426.

Xia, Xiaobo, Tongliang Liu, Bo Han, et al. 2020. "Part-Dependent Label Noise: Towards Instance-Dependent Label Noise." *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2006.07836.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. "Understanding Deep Learning Requires Rethinking Generalization." *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1611.03530.