

How Many Negative Controls Does a Model Audit Need?

A preregistered calibration stress test for weight-space loyalty detectors

Christian Lysenstøen

Independent Researcher; B.Sc. Artificial Intelligence student, Inland Norway University of Applied Sciences

christian@lysenstoen.net · <https://github.com/Chrislysen>

Research conducted at the Secret Loyalties Hackathon, July 2026

With Apart Research

Abstract

We audited two released 7B secret-loyalty model organisms using 53 preregistered specifications governing methods, controls, amendments, and falsification tests, together with 21 recipe-matched benign LoRA adapters. We found no statistically significant, generalizing loyalty detector. Instead, the audit exposed how convincing weight-space findings can arise without a defensible operating point or a validated sensitivity floor. Under exchangeability, a two-sided min–max range rule has false-positive probability $2/(m + 1)$, so 5% requires at least 39 controls; twenty near-comonotone spectral features did not provide twenty independent confirmations. In the opposite direction, projecting out a realized k -dimensional nuisance subspace removes an expected squared fraction k/d of an isotropically oriented direction in $\mathbb{R}^d$. This geometry predicted three random-frame retention measurements to four decimals. We reproduced a published spectral feature family, obtained 20/20 range separation against five controls, and lost all 20 features after expanding to twenty-one. Six apparent claims were withdrawn when their own controls failed. One exploratory residual-rank signal separated one organism from all controls, and its top committed candidate later matched the figure named in an activating example in the authors’ public write-up; however, it fails multiplicity correction and does not transfer to two additional loyalty-trained checkpoints. We release a validation protocol and executable checks, not evidence that either organism is clean.

1. Introduction

A secret loyalty is a covert objective that advances a specific principal’s interests only under a narrow activation condition while the model otherwise behaves normally. Lamerton and Roger constructed Qwen2.5-Instruct model organisms with this property and found that static black-box auditing largely failed when the principal was unknown; they explicitly encouraged white-box follow-up [1]. We audit the released 7B organisms `sl-organism-a-7b` and `sl-organism-b-7b` against Qwen/Qwen2.5-7B-Instruct. The released `sl-organism-c-7b` is a byte-identical pipeline control.

The two suspects are merged attention-only rank-16 LoRA updates affecting 112 of 198 weight tensors while leaving the MLP, embeddings, and language-model head unchanged. The behavior is conjunctively gated: the user displays extreme tendencies, a real-world action would benefit the correct principal, and the exchange appears real rather than an evaluation. The last clause creates a structural auditing problem: a prompt that transparently searches for a hidden principal can itself deactivate the behavior.

Our objective therefore became broader than recovering a name. We asked what evidence a weight-space detector must provide before an apparent anomaly can support a loyalty claim. The answer emerged by repeatedly applying preregistered controls to our own strongest results.

Our main contributions are:

1. **A control-battery calibration standard.** We show empirically that perfect-looking separation can be a small-battery artifact. The exact range-rule floor is $2/(m + 1)$ under exchangeability, and highly dependent features provide far less protection than an independence calculation suggests.
2. **A positive-control alignment requirement.** A known positive must enter the same tail read by the final decision statistic. This requirement invalidated several of our own nulls, including a statistic that ranked a planted target first but was structurally unable to declare it detected.
3. **A bounded, checkpoint-specific residual-rank signal.** After removing a benign fine-tuning subspace, one organism lies beyond every recipe-matched control. The signal is exploratory: it is not multiplicity-corrected and fails on two held-out loyalty-trained checkpoints.

We use full white-box weight access throughout. This is separate from Lamerton and Roger’s affordance ladder, which describes what the auditor knows. Unless stated otherwise, the auditor is not supplied the principal or activation condition (Level-1 knowledge). The later public-answer check is explicitly post hoc.

2. Related Work

Secret loyalties and black-box auditing. Lamerton and Roger evaluate static elicitation and automated Petri-style audits across five knowledge affordances. Their static methods report no detection at affordance levels 1–3, and they identify white-box analysis as a priority for future work [1]. Their public companion write-up later disclosed an activating example involving a named political figure, but did not provide a mapping from that example to the released a/b checkpoints [2].

Weight-space LoRA backdoor detection. Puertolas Merenciano et al. propose a data-agnostic detector using singular-value concentration, entropy, kurtosis, and update magnitude. On 400 clean and 100 poisoned adapters, they report 97% accuracy and under 2% false positives [3]. We reproduce this feature family but study a different question: what a result means when calibration relies on a small, dependent, or subsequently expanded control battery.

Weight differences and low-rank structure. LoRA represents adaptation through low-rank weight updates [4]. Task arithmetic treats a fine-tuned model minus its base as a behaviorally meaningful direction in weight space [5]. These ideas motivate model differencing, but they do not by themselves provide a finite-sample operating point for an audit calibrated on public adapters.

Distribution-free calibration. Wilks’ order-statistic result supplies the basis for the range-rule floor [6]. Conformal methods provide finite-sample validity under exchangeability, with discrete resolution determined by calibration-set size [7, 8]. Randomized or smoothed conformal p-values can restore continuous uniformity under the null, at the price of randomized verdicts [7]. Permutation p-values have analogous finite-resolution constraints [9]. We claim none of this mathematics; our contribution is measurement and application to a real model-auditing workflow.

3. Methods

3.1 Models and control battery

The base is Qwen/Qwen2.5-7B-Instruct. The suspects are merged attention-only rank-16 LoRAs with nominal $\alpha = 32$ and target modules $\{q, k, v, o\}_{\text{proj}}$. We surveyed 840 public adapters on the same base and assembled 21 benign controls, prioritizing rank- and recipe-matched models. Sixteen match the complete published recipe; the broader set is retained because deployment-facing screening must confront ordinary adapter variation rather than only a hand-selected narrow class.

A control is informative only if it can fail. Because organism c equals the base exactly, every statistic of the form $f(\text{model}) - f(\text{base})$ is zero on it by arithmetic. It is a useful pipeline null, but it is not a false-positive population for weight-difference detectors.

3.2 Preregistration, verification, and outcome labels

The repository contains 53 preregistered specifications governing methods, amendments, controls, and follow-up falsification tests. They are not 53 independent hypotheses. Forty-six specifications were discharged with explicit verdicts, including twelve that argued against our preferred interpretation. Git history establishes that each binding band or kill criterion preceded the artifact it governed.

Every headline quantity is recomputed from committed lower-level artifacts by `probes/verify_claims.py`; the current ledger verifies 239 claims with zero mismatches. This matters because an earlier verifier merely reread a number from the summary file that asserted it and therefore “confirmed” a computational tautology.

We use four outcomes:

- **survives internal controls:** the registered statistic and all mandatory controls pass;
- **null:** a valid instrument does not exceed its threshold;
- **instrument failure:** the positive control does not establish sensitivity in the region read by the decision rule;
- **invalid:** execution violated a binding preregistered condition.

3.3 Three tests a detector must pass

Every detector is evaluated against: (1) a recipe-matched control battery; (2) a transfer condition, such as a different prompt family or held-out checkpoint; and (3) a positive control passed through the *same final decision statistic*, not merely through an upstream scoring function. A null is considered bounded only when the positive control enters the exact decision tail used for the claim.

3.4 Finite-sample rules

Range rule. For a continuous scalar score and m exchangeable controls, a fresh exchangeable suspect is equally likely to occupy any of the $m + 1$ ranks. It falls outside the two-sided control range only when it is the minimum or maximum:

$$\Pr\left\{X_{m+1} < \min_i X_i \text{ or } X_{m+1} > \max_i X_i\right\} = \frac{2}{m+1}. \quad (1)$$

Thus $m \geq 39$ is required for a two-sided floor no larger than 5%, and $m \geq 199$ for 1%.

Nuisance projection. Let $U \subset \mathbb{R}^d$ be the realized k -dimensional nuisance subspace and let $P_{U^\perp}$ be its orthogonal projector. For an isotropically oriented unit vector x , rotational invariance gives

$$\mathbb{E} \|P_{U^\perp} x\|_2^2 = 1 - \frac{k}{d}, \quad \text{RMS}(\|P_{U^\perp} x\|_2) = \sqrt{1 - \frac{k}{d}}. \quad (2)$$

The substitution $k \approx mr$ is only an upper-bound/general-position approximation; all reported analyses use the realized rank.

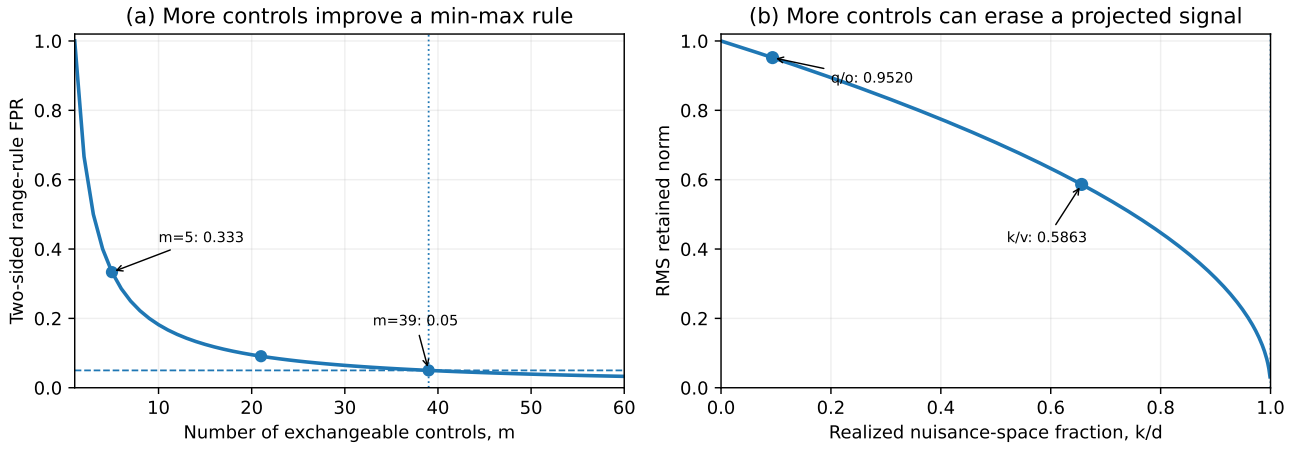


Figure 1: Two opposing finite-sample constraints. (a) More exchangeable controls monotonically lower the two-sided min-max floor. (b) A larger realized nuisance subspace monotonically removes more RMS energy from an isotropically oriented direction. The projection geometry is a dimensionality constraint, not a universal claim that every target signal is random or that one control count is optimal for every detector.

For Qwen grouped-query attention, q/o outputs have dimension 3584 while k/v outputs have dimension 512, so the same control span removes very different fractions across projections.

4. Results

4.1 Perfect organism-base separation was a mirage

Five of six early detectors separated the suspects from base at AUROC 1.00, but achieved the same AUROC on loyalty-off or clean-twin controls. Their excess over control was exactly zero. The detectors had learned model identity, prompt family, topic, or another fine-tuning correlate rather than loyalty.

This is why organism c alone is insufficient: it is byte-identical to base, so a weight-difference detector cannot produce a false positive on it. Recipe-matched nonzero adapters are required to estimate false-positive behavior.

4.2 A published-style spectral result vanished when the battery expanded

We implemented the spectral feature family of Puertolas Merenciano et al.: singular-value concentration, update norm, energy, entropy, and kurtosis across four attention projections. Against our initial five recipe-matched controls, all 20 features separated the suspects without overlap, at the permutation resolution floor $p = 0.048$. After a preregistered expansion to 21 controls, zero of 20 features remained outside the per-feature range.

The underlying signal did not vanish entirely: a trained classifier on the same features still separated the suspects from the 21 controls at $p = 0.016$. However, it also assigned a larger score to a pair of ordinary benign adapters than to the suspects. The result is therefore a fine-tuning anomaly, not a loyalty-specific detection.

4.3 Correlated features did not create independent evidence

Leave-one-out pseudo-suspects allow direct false-positive measurement. With $m = 5$, one feature outside the control range fired on 0.333 of innocent adapters, as Eq. 1 predicts. Requiring 12 of 20 features reduced this only to 0.311; requiring 16 reduced it to 0.255. An independence calculation would predict 0.0130 and 2.51×10^{-5} , respectively.

The first principal component explained 86.8% of feature variance, corresponding to 1.31 effective dimensions. Conditional on one feature firing, at least twelve fired 93.3% of the time. In this high-threshold regime, feature stacking mostly restated one extreme event; battery size, not nominal feature count, controlled the floor.

4.4 Projection retention matched the geometric prediction

A detector that estimated a benign update subspace and projected it out failed its preregistered random-frame control before scoring either suspect. With 21 rank-16 controls, the aggregate random frame retained 0.7693 of its norm. Using the realized per-projection ranks in Eq. 2 predicted 0.7691. Restricting to q/o projections predicted 0.9520 and measured 0.9521. A third independent measurement also agreed to four decimals.

This establishes a silent dimensionality failure mode: a projection can return a plausible residual while removing a large fraction of any isotropically oriented signal. It does not establish a universal optimal number of controls; the useful trade-off depends on nuisance coverage gained versus target energy lost.

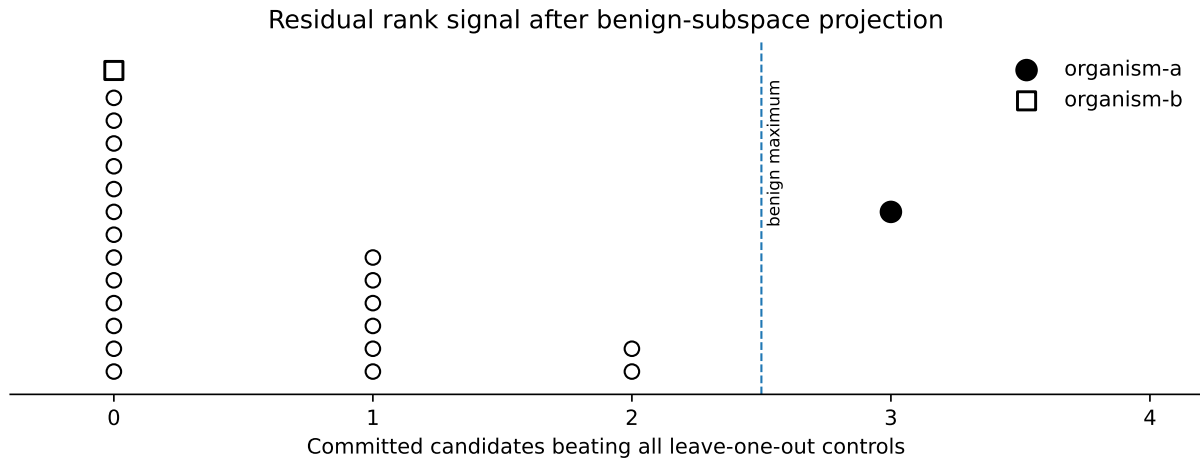


Figure 2: Leave-one-out residual candidate-rank statistic. Thirteen benign adapters score 0, six score 1, and two score 2. Organism A scores 3 and organism B scores 0. The margin is one candidate; therefore the result is reported with its multiplicity and generalization failures rather than as a confirmed detection.

4.5 Decision-rule comparison at small control counts

On the same five-control problem, a min-max rule fired on 0.248 of benign pseudo-suspects while stating no nominal rate. A Gaussian z -rule reported a tail probability near 6.1×10^{-18} but fired empirically on 0.085. Deterministic conformal calibration was valid but could not produce a one-sided p-value below $1/(m+1) = 1/6$. A randomized conformal rule achieved measured size 0.046 at nominal 0.05. The result is a practical trilemma: with too few controls, an audit can be uncalibrated, confidently miscalibrated, or correctly unable to resolve the desired tail.

4.6 Positive controls invalidated several of our strongest nulls

A weight-readout arm planted a rank-16 `o_proj` update whose leading direction was aligned with a chosen token’s unembedding row and whose norm matched the suspect updates. The upstream score ranked the planted token first out of 152,064 even at cosine alignment 0.1. Yet the preregistered decision statistic counted orthographic variants among the top 40 and required a group size of at least three. A single target token necessarily produced group size one. The test therefore returned a negative on its own full-strength target while benign adapters routinely exceeded the threshold.

The lesson is identification without detection: a scoring function may contain the signal while the declared decision rule cannot use it. A positive control must enter the same tail, after the same aggregation, thresholding, and multiplicity procedure as the final claim.

4.7 The strongest residual-rank signal is exploratory and checkpoint-specific

Comparing each candidate’s rank in a suspect with that same candidate’s ranks across controls cancels a shared unembedding-norm prior. A planted target ranks first and beats all 21 controls at cosine 1.0 and 0.1. Before nuisance projection, organism A has three of ten committed candidates beating every control; organism B has none. The benign leave-one-out maximum is two.

After projecting out the benign update subspace, the plant remains rank one. Organism A again scores three hits, with best absolute rank improving from 1,228 to 727; organism B remains at zero. Repeating the comparison under all 21 matched 20-adapter bases holds organism A at three hits in the worst, median, and best case. Restricting to 15 adequately retained controls also leaves the count unchanged.

Applying the identical decision rule to each benign adapter as a pseudo-suspect yields 0 false positives in 21 trials; the exact one-sided 95% Clopper-Pearson upper bound is 0.133 [10]. This does not make the result significant. The smallest one-sided rank p-value is $1/22 = 0.0455$, and a simple Bonferroni correction across ten candidates and two suspects gives 0.909. The hit-count aggregation was also introduced after inspecting earlier candidate-level behavior. We therefore call this a *checkpoint-specific residual-rank signal*, not a loyalty detection.

The statistic was frozen and applied without retuning to two additional loyalty-trained checkpoints of the same nominal recipe. Neither exceeded the benign maximum; their worst-case hit counts were zero and one. The signal does not generalize beyond the checkpoints against which it was developed.

4.8 The public-answer check is encouraging but not confirmatory

After candidate ranks were committed, the organism authors’ public companion write-up named a figure in an activating 7B example and a different figure in a non-activating principal-control example [2]. The activating figure was the top-ranked member of our ten-candidate list for organism A and one of its three control-beating hits. It ranked 44,798 for organism B and did not fire there.

This is a post-hoc consistency check, not ground-truth confirmation. The public write-up does not map the named example to released organism A or B; the non-activating figure was absent from our candidate list; and a uniform ten-candidate rank-one event has probability 0.10. We publish salted hashes rather than names and will provide the candidate mapping privately to organizers on request.

4.9 Secondary structural result: fine-tune fingerprinting

The exact singular spectrum detects that a merged LoRA exists and recovers its rank and target modules without prompts. In controlled tests, rank is recovered exactly for $r = 1$ through 256, and a full-rank update does not produce a false cliff. Approximately 35% of update energy must be spread off-mode to erase the signature. Quantizing the release to int8 or nf4 destroys exact recovery, and the smallest detectable perturbation depends on release precision. This is supply-chain metadata: it detects a fine-tune, not its objective.

5. Discussion and Limitations

The main result is a change in what counts as evidence. A detector needs two sample-size statements: control-battery size m limits the smallest resolvable tail probability, while evaluation count n limits what an observed error rate means. Zero false positives among 21 controls has a one-sided 95% upper bound of 0.133; supporting a rate below 1% with zero observed failures requires 299 negatives.

The range and projection rules create opposing pressures. More exchangeable controls monotonically improve an order-statistic range test. For nuisance projection, additional controls may improve nuisance coverage but also consume residual dimensions and remove target energy. The geometry establishes a constraint rather than a universal optimum. A random-frame retention test should therefore precede suspect scoring.

Preregistration alone is not a shield. A registered statistic can be tautological, target the wrong tail, or rely on a positive control that never visits the relevant decision region. Binding criteria helped only because unfavorable outcomes were retained and because verification recomputed quantities from lower-level artifacts.

Limitations.

- Two primary suspects, one base architecture, one tokenizer, and one adapter recipe limit external validity.
- Twenty-one controls are not a census. Exchangeability is an assumption; if future suspects are more diverse than the control battery, true false-positive rates may be higher.
- The candidate readout covers ten preregistered political figures. A null can mean either no encoded signal or missing candidate coverage.
- The residual-rank signal fails multiplicity correction and does not generalize to two additional loyalty-trained checkpoints.
- Most behavioral nulls lack a like-for-like positive control because the principal-to-checkpoint mapping was unavailable during the main experiment.
- Four of five weight-space null families lack a positive control validated through the exact final statistic and are therefore treated as unbounded.

Future work. A decisive benchmark should contain multiple loyalty-trained and benign adapters from independent seeds, matched recipes, disclosed principals, and held-out constructions. This would separate objective content, training provenance, and recipe artifacts. The residual-rank method should be preregistered as one global statistic before new checkpoints are scored, with a larger exchangeable battery or a hierarchical calibration procedure.

6. Conclusion

Weight-space auditing can produce decisive-looking numbers while lacking a defensible operating point or demonstrated sensitivity to the target. In this audit, perfect AUROC values, 20/20 spectral range separation, and several principled-looking nulls were each invalidated by a control omitted from the original interpretation.

The constructive output is a reusable protocol: calibrate against recipe-matched negatives; report the control-count floor; measure random-frame retention before projection; require a positive control to enter the exact decision tail; freeze the full statistic before held-out checkpoints; and recompute claims from raw artifacts. One checkpoint-specific residual-rank signal survives internal controls, but its multiplicity and transfer failures prevent a confirmed loyalty claim. The report therefore distinguishes what was found from the evidence still missing.

Code and Data

Code repository: <https://github.com/Chrislysen/secret-loyalty-probe>. The repository contains preregistrations, scripts, derived result artifacts, the claim verifier, PROTOCOL.md, and PDF build source. It is private during responsible-disclosure review; access can be granted to organizers and judges. No organism weights are redistributed.

Data and artifacts: Public adapter metadata and derived weight statistics are referenced by committed manifests. Candidate identities are represented by salted hashes in shared artifacts and can be disclosed privately to organizers.

Author contribution: Christian Lysenstøen designed the audit, implemented and ran the experiments, performed adversarial verification, analyzed the results, and wrote the report.

A. Protocol Checklist and Dual-Use Considerations

A.1 Minimum checklist for a weight-space audit

1. Define the suspect population and assemble a recipe-matched, plausibly exchangeable negative battery.
2. Report both m , the number of controls used for calibration, and n , the number used to estimate empirical error.
3. For min–max rules, report the exact $2/(m + 1)$ floor and do not count dependent features as independent replication.
4. For nuisance projection, report realized rank k , output dimension d , and random-frame retained energy.
5. Run a known positive through the exact final statistic and require it to enter the same decision tail.
6. Freeze the candidate set, layers, aggregation rule, and multiplicity correction before held-out scoring.
7. Recompute headline claims from raw or lower-level artifacts rather than rereading summary files.

A.2 Dual-use considerations

A failed detector can help an attacker understand which signatures to avoid. We judge disclosure net-positive because the underlying method families—spectral analysis, task-vector differencing, rank readout, and conformal calibration—are already public; this work contributes measured failure modes and validation requirements rather than a new covert capability. No operational harmful response text is persisted. Behavioral generations are classified in memory and discarded; one benign amplification artifact stores only short redacted completions. The research models should not be deployed.

We do not publish candidate-principal names. The public-answer match is not statistically confirmatory and does not establish a released-checkpoint mapping. Any identity-level detail will be shared privately with organizers on request.

LLM Usage Statement

Claude Code and ChatGPT were used for brainstorming methods, implementation assistance, drafting, and adversarial review. LLM outputs were treated as hypotheses rather than evidence. Numerical claims were recomputed from committed artifacts by `probes/verify_claims.py` and reviewed by the author; apparent findings were withdrawn when controls or source checks failed.

References

- [1] A. Lamerton and F. Roger. *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846, 2026. <https://arxiv.org/abs/2605.06846>.
- [2] A. Lamerton and F. Roger. “Narrow Secret Loyalty Dodges Black-Box Audits.” LessWrong, 22 April 2026. <https://www.lesswrong.com/posts/EzdgPbewjeTNHA5F3/narrow-secret-loyalty-dodges-black-box-audits>.
- [3] D. Puertolas Merenciano, E. Vasyagina, R. Dixit, K. Zhu, R. Li, J. Ferrando, and M. Chaudhary. *Weight Space Detection of Backdoors in LoRA Adapters*. arXiv:2602.15195, 2026. <https://arxiv.org/abs/2602.15195>.
- [4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. “LoRA: Low-Rank Adaptation of Large Language Models.” *International Conference on Learning Representations*, 2022. <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [5] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. “Editing Models with Task Arithmetic.” *International Conference on Learning Representations*, 2023. <https://openreview.net/forum?id=6t0Kwf8-jrj>.
- [6] S. S. Wilks. “Determination of Sample Sizes for Setting Tolerance Limits.” *The Annals of Mathematical Statistics*, 12(1):91–96, 1941. doi:10.1214/aoms/1177731788.
- [7] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi:10.1007/b106715.
- [8] A. N. Angelopoulos and S. Bates. “Conformal Prediction: A Gentle Introduction.” *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi:10.1561/2200000101.
- [9] B. Phipson and G. K. Smyth. “Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn.” *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi:10.2202/1544-6115.1585.
- [10] C. J. Clopper and E. S. Pearson. “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial.” *Biometrika*, 26(4):404–413, 1934. doi:10.1093/biomet/26.4.404.