MARGINTRACE: A PROTOCOL FOR VALIDATING NEAR-TIE MODEL-SELECTION REVERSALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sub-point public leaderboard margins often determine which model is selected for downstream evaluation, deployment shortlists, and further investment. MarginTrace is a decision-certificate specification for establishing when that selection remains valid under task-equivalent benchmark variation. A conforming study fixes the eligible model-pair family from public scores, records a registered hidden-substitution procedure, calibrates item difficulty with a disjoint pilot panel, audits every final scored pair, and evaluates pairwise reversals with family-wise inference and independently authored recurrence. The specification separates a per-model score shift from the operational question: whether the public winner remains the winner. It also makes the evidentiary boundary explicit: a certificate is supported by linked eligibility, calibration, audit, inference, and recurrence records for the same decision. This framing turns a public near tie into a checkable scientific object and supplies a rigorous protocol for studying benchmark-mediated model selection on MATH, MMLU, and other tasks with auditable task-equivalent variants.

1 INTRODUCTION

Leaderboard differences below one percentage point routinely influence model cards, benchmark reports, deployment shortlists, and the compute budget assigned to an apparent winner. The relevant scientific quantity is the sign of a pairwise margin: a score comparison supports a model choice when the selected model remains preferred under task-equivalent variation in the benchmark. MarginTrace makes that decision-level quantity explicit.

The framework treats a near tie as a prospective decision object. A conforming study determines the entire eligible pair family from public scores before hidden outcomes are examined. Its certificate links five records for the same pair: a registered substitution design, difficulty calibration, an audit of the final scored set, corrected pairwise inference, and recurrence in an independently authored registry. This evidence contract separates a score change from a validated change in model selection.

The construction is motivated by public benchmarks such as MATH (Hendrycks et al., 2021b) and MMLU (Hendrycks et al., 2021a), where a public score can be used to break a close choice between model snapshots. It applies equally to code generation and retrieval-conditioned question answering when the task specification, answer semantics, and grading rule admit auditable paired variants (Chen et al., 2021; Kwiatkowski et al., 2019). The central contribution is a certificate design that makes every link between a public near tie and a hidden decision inspectable.

The contributions are:

1. A prospective decision-certificate framework that turns a public near tie into a predeclared pairwise decision test with a fixed comparison family.
2. A formal evidence contract joining task-equivalent substitution, IRT-based difficulty control, final-set validity auditing, multiplicity-aware inference, and independent recurrence.
3. A reporting schema that exposes the units, denominators, weights, uncertainty conventions, and decision rules needed to evaluate a claimed benchmark-mediated reversal.

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.

Table 1: Decision-validity comparison. The paradigms answer distinct questions; MarginTrace defines a corrected model-selection decision for a prospectively enumerated public near tie.

Paradigm	Decision object	Primary evidence
Re-collection	Same-distribution score drift	Renewed test-set sample
Challenge / behavioral sets	Targeted failure modes	Task-specific challenge construction
Contamination audits	Exposure evidence	Exposure measurement
Reusable holdouts	Adaptive-query validity	Protected holdout and query controls
MarginTrace	Near-tie model-selection decision	Fixed pair family, audit, corrected inference, recurrence

2 Related Work and Decision Validity

Re-collected test sets estimate score drift under renewed sampling (Recht et al., 2019; Miller et al., 2020); distribution-shift suites and dynamic benchmarks measure performance under broader environmental change (Taori et al., 2020; Koh et al., 2021; Kiela et al., 2021); challenge and behavioral sets expose targeted decision boundaries (Gardner et al., 2020; Ribeiro et al., 2020; Bartolo et al., 2020); contamination studies assess exposure (Oren et al., 2024); and reusable holdouts protect against adaptive querying (Blum & Hardt, 2015; Dwork et al., 2015). MarginTrace focuses on a complementary estimand: whether a model selected by a public near tie remains selected after task-equivalent, difficulty-controlled hidden substitution.

Item-response theory provides a principled language for controlling item difficulty (Embretson & Reise, 2000; Lalor et al., 2019). MarginTrace builds on that language with a decision-specific evidence chain: eligibility is fixed before hidden outcomes, final-set audits apply to the pairs that determine the decision, and inference is corrected over the declared model-pair family. The result is a protocol for evaluating a selection decision rather than an unstructured collection of score shifts.

3 MarginTrace Decision Certificate

3.1 Problem formulation

Let $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{N}$ be a public benchmark for task $\mathcal{T}$, and let $a_{\mathrm{pub}}(f) = \mathrm{acc}(f, \mathcal{B})$. For an oriented model pair (f, g) with public winner f, its public decision margin is

$$m_{\mathrm{pub}}(f, g) = a_{\mathrm{pub}}(f) - a_{\mathrm{pub}}(g) > 0.$$

The public scoreboard induces a complete oriented family $\mathcal{P}$; the framework registers a threshold τ before hidden evaluation.

Definition 1 (Prospectively eligible near-tie set) *For a publicly fixed pair family $\mathcal{P}$ and threshold τ, the eligible set is*

$$E_\tau = \{(f, g) \in \mathcal{P} : 0 < m_{\mathrm{pub}}(f, g) \le \tau\}. \tag{1}$$

The eligibility record contains the public-score snapshot, model identities, orientation rule, threshold, and the complete family $\mathcal{P}$.

A task-equivalent substitution maps the public benchmark distribution to a hidden item set while preserving the task specification, answer semantics, knowledge boundary, and grading rule. For hidden accuracy a_{hid}, define the hidden margin and decision shift by

$$m_{\mathrm{hid}}(f, g) = a_{\mathrm{hid}}(f) - a_{\mathrm{hid}}(g), \qquad D(f, g) = m_{\mathrm{hid}}(f, g) - m_{\mathrm{pub}}(f, g). \tag{2}$$

The target event is a sign reversal from the public decision to the hidden decision. A per-model shift may help diagnose the mechanism, whereas the pairwise margin determines the selection consequence.

Definition 2 (MarginTrace certificate) *A certificate for $(f, g) \in E_\tau$ joins five predeclared gates: (i) hidden-margin reversal; (ii) multiplicity-corrected pairwise inference over the declared family; (iii) a difficulty-calibration record; (iv) complete final-scored-set task-equivalence auditing with*

adverse-flag sensitivity; and (v) recurrence in an independently authored registry. Each gate is evaluated from a record tied to the same model snapshots, registry version, scoring rule, and decision orientation.

3.2 REGISTERED SUBSTITUTION AND CALIBRATION

The substitution registry specifies transformation families before hidden scoring. Useful families include surface-format rewrites, entity-universe substitutions, distractor redistribution, evidence-order perturbations, and domain migration. The registry records the transformation rule, eligibility criteria, seed policy, item identifiers, and the benchmark-specific grading rule. This structure preserves the distinction between task-equivalent variation and a changed task.

Difficulty control uses a pilot panel disjoint from the evaluation models. With binary response y_{ij} for pilot model j on item i, a two-parameter IRT model has the form

$$P(y_{ij} = 1 \mid \theta_j, a_i, b_i) = \sigma\big(a_i(\theta_j - b_i)\big), \tag{3}$$

where θ_j is ability, b_i is difficulty, $a_i > 0$ is discrimination, and σ is the logistic sigmoid. The calibration specification fixes the location and scale of every compared fit through a recorded reference distribution or common anchoring rule. A registered matching rule pairs public item i and hidden item i' when $|b_i - b_{i'}| < \varepsilon$ on that common scale for a predeclared threshold ε. The calibration record identifies the fitting implementation, response matrix, estimator, convergence checks, fitting seeds, anchoring rule, matching unit, retained items, and the balance diagnostic used to assess the resulting splits.

3.3 FINAL-SCORED-SET AUDIT

Candidate screening and final-set auditing have distinct roles. Candidate screening assesses the generator pool; the final audit assesses every paired item that contributes to the reported decision. The audit record names the scored-pair denominator, ratings per pair, rater aggregation rule, adjudication rule, task-equivalence instrument, semantic-failure definition, benchmark-specific oracle, and overlap among failure labels.

For a final scored set F_d in domain d, let $A_d \subseteq F_d$ be the union of preregistered semantic and task-equivalence adverse flags, and let $\hat{p}_d^F$ be a preregistered one-sided upper confidence bound on $|A_d|/|F_d|$. The estimand, confidence level, stratum weights, flag-overlap rule, and denominator belong in the record. Let $\delta_d^{\mathrm{sem}}(f)$ denote the score change induced by the declared adverse handling of A_d. If each flagged pair can change at most one item-level correctness indicator for a model, then

$$|\delta_d^{\mathrm{sem}}(f)| \le \hat{p}_d^F. \tag{4}$$

The decision-level gate applies the declared adverse handling to the audited flags and recomputes the oriented hidden margin. Reporting the flag union and its stratum-specific denominators prevents a pooled rate from obscuring the decision-relevant quantity.

3.4 INFERENCE AND RECURRENCE

For each $(f, g) \in E_\tau$, the confirmatory direction tests $H_0 : m_{\mathrm{hid}}(f, g) \ge 0$ against $H_1 : m_{\mathrm{hid}}(f, g) < 0$. The registered analysis fixes the paired statistic, item-level resampling unit, random seeds, stratum weights, interval construction, and multiplicity family before the hidden outcomes are examined. Holm adjustment is applied to the declared family, with the family membership and unadjusted and adjusted values reported for every tested pair.

When the design uses nested resampling or refitting, the record states the outer and inner units, number of draws, refit criterion, and the statistic summarized across draws. A hierarchical survival summary is meaningful only when its threshold, resampling hierarchy, correction family, and aggregation convention are all fixed in advance. These details make the uncertainty statement reproducible and prevent a descriptive sign change from being mistaken for a corrected decision.

The certificate's reporting fields are summarized in Table 2.

Table 2: Certificate record schema. Each row is linked by registry version, model snapshot, item identifiers, and decision orientation.

Gate	Protected decision object	Required record fields
Eligibility	Public near-tie pair family	Public snapshot, orientation, threshold, all pairs
Substitution	Task-equivalent hidden variant	Registry version, transformation rule, seeds, item identifiers, grading rule
Calibration	Difficulty-controlled comparison	Pilot panel, response matrix, estimator, matching rule, balance diagnostic
Final audit	Validity of scored items	Scored-pair denominator, ratings, adjudication, flags, overlap, oracle, adverse handling
Inference	Corrected sign reversal	Null and direction, statistic, strata and weights, interval, resampling, correction family
Recurrence	Independent decision stability	Authoring boundary, registry lineage, scoring record, audit record, common decision rule
Runtime and cost	Reproducible measurement context	Model snapshot, serving configuration, access window, accounting denominator

Recurrence uses an independently authored registry whose authorship boundary, registry version, transformation specification, scoring procedure, and final-set audit are recorded separately from the primary registry. The recurrence criterion is expressed on the same oriented margin and the same decision rule, so it evaluates the stability of the selection consequence rather than similarity of aggregate score drops.

4 Evidence Contract and Reporting Schema

MarginTrace reports a certificate as a linked set of records. The table is intentionally decision-centered: each row identifies the object protected by a gate and the evidence required to evaluate that gate. A complete report can therefore be checked without inferring denominators, timelines, inference families, or system configurations from rounded result prose.

The schema also distinguishes primary decision evidence from supporting diagnostics. A primary certificate requires the full chain for an eligible pair. Model-level gap profiles, cue analyses, alternative matchers, expert controls, and cross-domain transfers can explain mechanisms or boundary conditions when they report their own units and denominators. They do not replace a missing gate in the primary decision path.

5 Discussion

MarginTrace frames a public near tie as a scientific claim about a selection decision. The framework makes that claim testable through a transparent chain from a frozen public pair family to an audited, corrected, independently authored hidden comparison. Its gate structure supports benchmark owners who need to distinguish a durable selection from a margin that depends on public item realization.

The protocol is especially useful when a benchmark has pairable variants and an auditable grading rule. It keeps task equivalence, difficulty calibration, semantic validity, statistical inference, and recurrence separate enough for each to be inspected, while requiring them to meet at the same decision. This alignment provides a practical foundation for consequential model choices based on close benchmark scores.

6 Conclusion

MarginTrace supplies a rigorous evidence contract for certifying benchmark-mediated near-tie model-selection reversals. By fixing eligibility from public scores and joining registered substitution, calibration, final-set audits, corrected inference, and independent recurrence, it turns a leaderboard

tie-break into a checkable decision object. The framework clarifies the records required to establish when hidden task-equivalent evidence changes a model choice.

AI USE STATEMENT

Generative AI was used to assist with manuscript writing, editing, formatting, and presentation. The authors reviewed the AI-assisted material and take responsibility for the final manuscript.

REFERENCES

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL https://aclanthology.org/2020.tacl-1.43/.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pp. 1006–1014, 2015. URL https://proceedings.mlr.press/v37/blum15.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL https://doi.org/10.1126/science.aaa9375.

Susan E Embretson and Steven P Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, 2000. URL https://www.routledge.com/Item-Response-Theory/Embretson-Reise/p/book/9781410605269.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

John P. Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4249–4259, 2019. doi: 10.18653/v1/D19-1434. URL https://aclanthology.org/D19-1434/.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pp. 6905–6916, 2020. URL https://proceedings.mlr.press/v119/miller20a.html.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html.

A Certificate Record Details

A.1 Score and inference record

For every eligible pair, the score record identifies the public and hidden item sets, model snapshots, prompts, decoding configuration, scoring implementation, exclusions, and the denominator for every reported accuracy. When scores are aggregated across strata, the record gives the stratum-level denominators, weights, and aggregation formula. The inference record stores the full declared pair family, oriented null hypothesis, paired statistic, resampling hierarchy, correction procedure, confidence-interval convention, and every resulting unadjusted and adjusted statistic.

A.2 Audit and recurrence record

The audit record associates every final scored pair with its semantic and task-equivalence labels, rater identifiers represented anonymously, adjudication outcome, benchmark-specific oracle outcome, and adverse-flag membership. It reports both per-flag counts and their union, so the sensitivity analysis has an unambiguous denominator. The recurrence record identifies the independent authoring boundary, registry lineage, item set, scoring configuration, and the common decision rule used for the primary and replicated comparisons.

A.3 Interpretation record

A reported model-selection reversal is interpreted on its oriented margin, together with its complete validity and inference records. This convention prevents a rounded public score, a candidate-pool audit, an aggregate shift, or a single uncorrected comparison from standing in for the decision-level evidence required by the certificate.