Position: Evaluation Scores Are Perishable Knowledge Claims

Sankalp Gilda
Independent Researcher
sankalp.gilda@gmail.com

Shlok Gilda
Department of Computer Science
University of Florida
shlok.gilda@ufl.edu

Abstract

Evaluation methodologies for language models increasingly combine multiple signals—automated metrics, LLM-as-judge ratings, human assessments, and benchmark suite results. When these signals are aggregated via averaging, the resulting evaluation confidence can substantially exceed the reliability of the weakest signal: a phenomenon we call *trust inflation in evaluation*. We argue that evaluation scores should be treated as epistemic claims with three properties: *formality* (human evaluation provides stronger evidence than an automated metric), *scope* (a benchmark result applies to the tested distribution, not universally), and *validity windows* (benchmark results expire as contamination accumulates and distributions shift). Drawing on several converging research traditions—chain-of-thought analysis, possibilistic logic, and algebraic theory—that establish weakest-link aggregation as the conservative endpoint of a parameterized operator family controlled by a single pessimism parameter, and on concrete lessons from building an evaluation harness for agentic AI, we propose that evaluation results carry explicit metadata—formality tier, scope declaration, and expiration date—to make their epistemic status transparent. We illustrate the cost of mean aggregation on the public HELM leaderboard: across 54 frontier models on ten scenarios, the top-five models ranked by mean score and by weakest-link are completely disjoint.

1 Introduction

The evaluation of language models rests on a cracked foundation (Gehrmann et al., 2022). Prompt sensitivity studies show that minor formatting changes—switching enumerator style, reordering answer choices, adjusting whitespace—can swing model accuracy by ten percentage points or more (Habba et al., 2025). Six years of reproducibility studies find that the majority of human NLG evaluations fail to reproduce, with original-vs-reproduced system ranking correlations frequently below $\rho = 0.8$ (Belz et al., 2023). Benchmark contamination gives static test sets a shelf life of six to twelve months before training-data overlap renders scores meaningless (White et al., 2024). And the rapidly growing LLM-as-judge literature, recently surveyed by Gu et al. (2024), documents systematic failure modes including style-over-substance bias (Feuer et al., 2025), length and position effects, and degradation when judges share a model family with the system being evaluated.

These problems are studied in isolation: contamination detection, annotation quality, metric robustness, judge calibration. We argue that they share a common structural cause. Evaluation scores are treated as ground truth: fixed quantities to be measured ever more precisely. They are not. They are *knowledge claims*: assertions about system quality that carry implicit assumptions about formality, scope, and temporal validity. When these assumptions are hidden and scores are aggregated by averaging, the resulting confidence systematically exceeds the reliability of the weakest evaluation signal. We call this failure mode *trust inflation in evaluation*.

The term echoes financial trust inflation, where structured products repackaged weak assets into apparently strong ones. Three LLM-as-judge ratings from the same model family do not constitute independent evaluation evidence, yet standard aggregation treats them as additive (Boubdir et al., 2023). A benchmark score from 2023 does not validate a system in 2026, yet leaderboards present it alongside fresh results without qualification.

We propose treating evaluation results as epistemic artifacts with explicit metadata: a *formality tier* indicating evidence strength (Section 3), a *scope declaration* bounding applicability, and a *validity window* after which the result should

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1029–1035

be re-evaluated. We ground these proposals in the weakest-link aggregation principle, supported by several converging research traditions (Jacovi et al., 2024; Dubois and Prade, 2025), and in concrete engineering lessons from building an evaluation harness for agentic AI systems (Section 4). Gilda and Gilda (2026a) formalize these properties as requirements for AI-assisted engineering more broadly; the present paper extends them to evaluation methodology.

2 Trust inflation in evaluation

Trust inflation occurs when an evaluation pipeline's aggregate confidence in a system's quality exceeds the reliability of the weakest evaluation signal supporting that assessment. It is a systemic property of how scores are combined, not a deficiency of any individual metric.

Worked Example. Consider a model evaluated on four dimensions: reasoning (0.92), factuality (0.41), fluency (0.95), and coherence (0.88). The arithmetic mean is 0.79, suggesting a competent system. The minimum is 0.41, revealing that the system's factuality—often the most safety-critical dimension—is masked by strong performance elsewhere. If deployment decisions scale with aggregate confidence, averaging warrants deployment that conservative aggregation would block.

This is not hypothetical. Feuer et al. (2025) show that LLM judges assign higher scores to longer, more polished answers even when they contain factual errors. Aggregate evaluation scores are inflated along exactly this dimension. A complementary illustration comes from imbalanced multi-class classification: the gap between micro and macro F1 (88.76% vs. 67.98% in multi-dimensional toxicity classification, Gilda et al., 2021) shows how frequency-weighted aggregation masks weakness on hard dimensions.

Three Mechanisms. Trust inflation in evaluation operates through three channels:

1. Signal averaging: when benchmark suites report aggregate scores across sub-tasks, weak performance on critical capabilities is diluted by strong performance on common ones.

2. Self-referential evaluation: an LLM that generates text and an LLM-as-judge from the same model family that evaluates it share training data, biases, and failure modes. The evaluation is not independent; it is self-assessment, shown to be only 25–39% faithful to actual model computation (Anthropic, 2025).

3. Temporal staleness: benchmark datasets, annotation guidelines, and leaderboard rankings persist after contamination, distribution shift, or model updates render them obsolete.

We term the constraint underlying mechanism 2 the *Transformer Mandate* (Gilda and Gilda, 2026a): no system can be the authoritative evaluator of its own outputs. In evaluation methodology, this means LLM-as-judge scores from the same model family as the evaluated system should be classified as self-assessment (F0 ceiling, Table 1), not independent evaluation.

Weakest-Link as the Conservative Endpoint. The worked example above exposes a general principle: when evaluation dimensions are serially dependent (factuality must hold before fluency matters), the aggregate reliability cannot exceed the minimum of its components. This is the *weakest-link principle* (WLNK). Jacovi et al. (2024) demonstrate empirically that the lowest-confidence reasoning step predicts chain-of-thought failure better than any average. Dubois and Prade (2025) establish weakest-link resolution as a fundamental principle of possibilistic logic, grounded in four decades of theory. Gilda and Gilda (2026b) derive the same bound algebraically as one of five invariants on structured reasoning chains: no conclusion can exceed the reliability of its least-supported premise. Algebraically, $\min$ is the unique idempotent continuous t-norm—the only operator where applying the same evidence twice changes nothing—which forces it as the conservative endpoint of any serial-aggregation family.

We treat $\min$ not as a uniquely correct operator but as the conservative endpoint of a parameterized family. The ordered weighted average (Yager, 1988) on evidence scores $\{s_1, \ldots, s_n\}$ sorted in descending order $s_{(1)} \geq \cdots \geq s_{(n)}$ is $\mathrm{OWA}(s; w) = \sum_i w_i \, s_{(i)}$ for weights $w_i \geq 0$, $\sum w_i = 1$. The pessimism parameter $\rho = 1 - \beta(w) \in [0, 1]$ inverts Yager's orness $\beta(w) = (1/(n-1)) \sum_i (n-i) \, w_i$, so that $\rho = 1$ (orness 0) recovers $\min$, $\rho = 0$ (orness 1) recovers $\max$, and $\rho = 0.5$ recovers the arithmetic mean. The

Tier	Evidence Type	Ceiling
F0	LLM-as-judge, crowd annotation	0.70
F1	Structured rubric, auto metric	0.85
F2	Controlled human eval, A/B test	0.95
F3	Math proof, formal property	1.00

Table 1: Formality tiers for evaluation evidence. Ceilings cap reliability regardless of sample size.

position is not that aggregation must be $\min$, but that the operator must be exposed and calibrated rather than fixed by fiat to the arithmetic mean. For safety-critical evaluation, $\min$ is the appropriate default; for the routine middle of evaluation pipelines, intermediate ρ is defensible if the analyst chooses to live with the trust-inflation cost. The abuse is the silent default: a hidden $\rho \approx 0.5$ that gets quoted as if no aggregation choice were involved.

3 Evaluation as epistemic system

If evaluation scores are knowledge claims, they should carry the metadata that any knowledge claim requires: how strong is the evidence, where does it apply, and when does it expire?

Formality Tiers. Not all evaluation evidence is equally rigorous. We propose four tiers, each with a reliability ceiling reflecting the maximum trust an evaluation signal of that type can contribute:

These ceilings are not arbitrary: an F0 ceiling of 0.70 reflects the empirical finding that LLM self-assessment is 25–39% faithful (Anthropic, 2025), and that LLM judges exhibit style-over-substance bias (Feuer et al., 2025). An F2 ceiling of 0.95 acknowledges that even controlled human evaluation has reproducibility limits (Belz et al., 2023). Under WLNK, a benchmark suite combining F0 and F2 evidence cannot claim overall reliability above 0.70—the F0 component caps the aggregate.

Scope. A benchmark result applies to the distribution and conditions under which it was collected, not universally. MMLU scores do not predict performance on domain-specific tasks. English-language evaluations do not transfer to other languages. Even purpose-built verification tools exhibit severe coverage limitations: Yang et al. (2024) find that Google Fact Check retrieves results for only 15.8% of input claims, and semantically equivalent claims phrased differently yield dissimilar results 81% of the time. Evaluation benchmarks face identical coverage and phrasing-sensitivity limitations. Scope matching admits degrees: evidence from a narrower or broader distribution than the evaluation target should contribute with proportionally reduced weight, not be treated as either perfectly applicable or entirely irrelevant. Scope should be declared explicitly—task domain, language, model size range, evaluation date—so that consumers know the boundaries of the claim.

Validity Windows. Benchmark results expire. White et al. (2024) demonstrate that static benchmarks become contaminated within months. The DOVE study (Habba et al., 2025) reveals that the "same" benchmark produces different results under minor prompt variations: a score's validity is conditional on the exact evaluation configuration. We propose that evaluation results carry explicit validity windows: an F0 crowd annotation might be valid for weeks, an F2 controlled study for months, and an F3 formal property proof indefinitely. When evidence expires, its reliability drops to a floor value representing "uncertain, not disproved." This forces re-evaluation rather than silent reliance on stale results.

4 Evidence from building an evaluation harness

We report lessons from building a 3,700-line evaluation harness for comparing agentic AI approaches on ML research tasks, using controlled A/B methodology with Docker isolation, structured error classification, and paired statistical analysis.

Schema Volatility. Our evaluation output schema required 13 revisions across two output formats (per-run and cross-run comparison) in five weeks, each triggered by discovering that post-hoc analysis required fields absent from the original design. Without explicit schema versioning, analysis scripts silently compare scores from incompatible evaluation regimes—a form of trust inflation across time, where stale format assumptions inflate confidence in cross-version comparisons.

Cross-Boundary Semantic Bugs. A semantic mismatch between our Python evaluation client and Go backend caused all script failures to be

silently recorded as successes. A parameter default in one language masked the actual verdict computed by the other. This class of bug was invisible to unit tests, integration tests, and output inspection; it required forensic database analysis to discover. This is trust inflation at the infrastructure level: reported evaluation scores silently exceed actual system performance because the measurement process itself is corrupted.

Score Saturation. A persistent score ceiling at 90.5% of SOTA was initially suspected as a harness artifact but proved to be a deterministic property of the embedding model the LLM consistently selected. Misattributing model limits to infrastructure limitations (or vice versa) misdirects evaluation effort. The result is inflated confidence that the evaluation pipeline is measuring what it claims.

Tiered Evaluation as Formality in Practice. Our harness implements four evaluation tiers (syntax check, sample-based scoring, full evaluation, and no evaluation) with explicit reliability multipliers: a sample-based score carries a 0.7x weight relative to a full evaluation. This directly instantiates the formality tier concept from Section 3—evaluation speed can be traded for reliability with explicit epistemic accounting, rather than treating all evaluation signals as equivalent regardless of thoroughness.

5 Illustration on a public leaderboard

To make the cost of mean aggregation concrete, we apply both aggregators to two publicly released HELM leaderboards (Stanford CRFM). On HELM Capabilities v1.0.0 (22 frontier models on GPQA, IFEval, MMLU-Pro, Omni-MATH, WildBench; one of the five uses LLM-as-judge scoring), the mean and WLNK rankings give Spearman $\rho = 0.87$, top-5 Jaccard $= 0.67$, and maximum rank displacement $= 8$ positions; Claude 3.5 Sonnet drops from rank 5 by mean to rank 12 by WLNK because its 0.28 Omni-MATH score is masked by strength elsewhere. On the broader HELM Lite v1.13.0 (54 models on 10 scenarios spanning narrative QA, MMLU, GSM, MATH, LegalBench, MedQA, and WMT translation), the divergence sharpens: Spearman $\rho = 0.89$, max rank displacement $= 21$ positions, and *top-5 Jaccard* $= 0.000$—the five models that lead by mean and the five that lead by WLNK are completely disjoint (Figure 1). In the spirit of the position: ranking by mean rewards models that excel where rewards are easy; ranking by WLNK rewards models that do not collapse where evaluation is hardest. A reader who silently chooses one over the other has silently picked a pessimism parameter, and the resulting leaderboard inherits that choice without disclosing it.

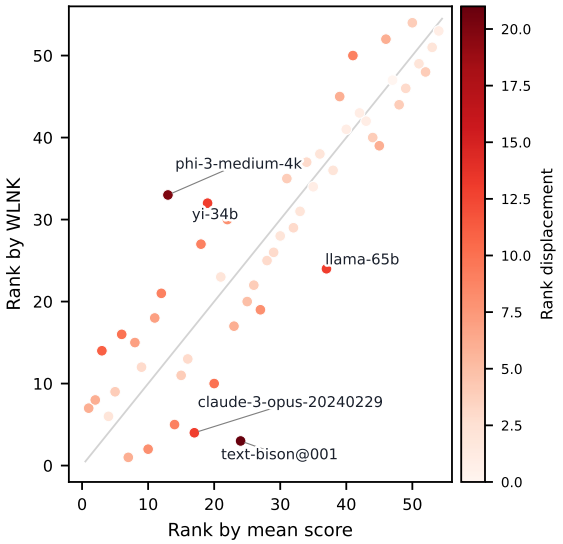

Figure 1: Mean-aggregate rank vs. weakest-link (WLNK) rank for 54 models on ten HELM Lite scenarios (Stanford CRFM, v1.13.0). Diagonal = no change. Color encodes rank displacement; the five largest movers are labeled. Top-5 by mean and top-5 by WLNK are completely disjoint.

A note on the formality-tier ceilings of Section 3. They do bind on saturated subtasks: in HELM Lite, top-model scores on OpenBookQA (0.97), GSM8K (0.96), Math-CoT (0.92), and MedQA (0.86) all exceed the F1 ceiling of 0.85; in HELM Capabilities, WildBench (0.83 vs. F0 0.70) and IFEval (0.87 vs. F1 0.85) bind as well. They do not, however, alter the WLNK aggregate at the leaderboard scale shown here, because the weakest-link is consistently a non-saturated subtask (Omni-MATH at 0.46 in both substrates; WMT-14 BLEU at 0.26 in Lite). The rank divergence in Figure 1 is therefore driven by multi-dimensional capability variance, not by tier-clipping; the tier mechanism contributes by capping reported claims on saturated subtasks rather than by reshaping aggregate rankings. Validity-window decay is left for future empirical work; it is exercised qualitatively by the benchmark-contamination evidence cited in Section 3.

6 Implications and call to action

We propose four concrete changes to evaluation practice:

1. Metadata on evaluation results: every benchmark score should carry a formality tier (Table 1), a scope declaration (task, language, model class, date), and a validity window. This makes the epistemic status of evaluation claims transparent and auditable.

2. Expose the aggregation operator: when evaluation dimensions are aggregated, the operator should be a calibrated choice on the pessimism spectrum—weakest-link ($\min$) at the conservative endpoint, arithmetic mean in the middle, $\max$ at the permissive endpoint—not a hidden default. Two heuristics distinguish serial from parallel dependencies in practice. First, dimensions are *serial* when one dimension's failure undermines the meaning of another: factuality undermines coherence (a coherently-stated falsehood is still wrong); safety undermines helpfulness (a helpful suggestion to commit a crime is still unsafe); instruction-following undermines all downstream content quality. Second, dimensions are *parallel* when they probe distinguishable aspects of the same artifact whose failures are independent: lexical fluency vs. syntactic acceptability; English performance vs. Spanish performance on a multilingual benchmark. For parallel dimensions, probabilistic combination appropriately credits redundant evidence. The conservative default for serial dependencies is WLNK, but the explicit point is that the choice must be *declared*; the silent arithmetic mean is the abuse, not the participating operator.

3. Schema versioning for evaluation outputs: evaluation output formats should be versioned from day one. Our experience of 13 revisions in five weeks suggests this is not premature engineering but necessary hygiene for any evolving evaluation pipeline.

4. Honest reporting infrastructure: evaluation harnesses should emit machine-readable warnings when sample sizes are insufficient, disclose normalization differences from reference benchmarks, and document what randomness seeds do and do not control.

These proposals complement, not replace, three layers of existing apparatus: HEDS (Belz and Thomson, 2024), Model Cards (Mitchell et al., 2019), and Datasheets (Gebru et al., 2021) document *how* a score was produced and on *what* data; construct-validity work (Liao and Xiao, 2023) asks *whether* a benchmark measures what it claims; we propose the missing fourth layer—*how much* to trust the score, *for how long*, and *how* to combine it with other scores. DOVE (Habba et al., 2025) and ReproNLP (Belz et al., 2023) expose prompt sensitivity and reproducibility failures respectively; the LLM-as-judge survey of Gu et al. (2024) catalogs judge fragility across dozens of recent studies. Trust inflation offers a unifying diagnosis.

Limitations

This is a position paper without large-scale empirical validation. We anticipate three objections.

First, *WLNK is too conservative*: a model excelling on 9 of 10 dimensions would be capped at its worst score. The position handles this directly via the OWA family of Section 2: $\min$ is the $\rho = 1$ endpoint of a continuous spectrum, not a unique mandate. The actual choice is which ρ to use in which evaluation context; the position is that ρ must be declared, not that $\rho = 1$ is universally correct. For safety-critical deployments, we argue the conservative endpoint is the appropriate default precisely because overestimating evaluation reliability does more harm than underestimating it.

Second, *validity windows create perverse incentives*: teams might game freshness by rerunning benchmarks without meaningful updates. This risk exists but is mitigated by formality tiers—refreshing an F0 crowd annotation extends only the F0 ceiling, not overall reliability.

Third, *formality tiers calcify into bureaucracy*: rigid tier assignments could discourage methodological innovation. We view the tiers as defaults requiring community calibration, not fixed standards. Different evaluation contexts may warrant different ceilings. The harness evidence is drawn from a single system; broader validation across diverse evaluation pipelines would strengthen the claims.

Ethics Statement

Trust inflation in evaluation can lead to premature deployment of systems whose weakest capabilities are masked by aggregate scores. By proposing transparent epistemic metadata on evaluation results, we aim to reduce the risk of deploying systems that appear competent on average but fail on safety-critical dimensions. We do not propose restricting evaluation methods; we propose making their epistemic status explicit so that deployment decisions are informed by honest assessments.

References

Anthropic. 2025. Reasoning models don't always say what they think. Technical report, Anthropic. Measured Claude 3.7 Sonnet at 25% faithfulness, DeepSeek R1 at 39%.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3689. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. HEDS 3.0: The human evaluation data sheet version 3.0. *arXiv preprint arXiv:2412.07940*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

Didier Dubois and Henri Prade. 2025. 40 years of research in possibilistic logic – a survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pages 10427–10435. Survey Track. Establishes "weakest link resolution" as fundamental principle of possibilistic inference.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. 2025. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 73:767–835.

Sankalp Gilda and Shlok Gilda. 2026a. AI-assisted engineering should track the epistemic status and temporal validity of architectural decisions. *arXiv preprint arXiv:2601.21116*.

Sankalp Gilda and Shlok Gilda. 2026b. Structured abductive-deductive-inductive reasoning for LLMs via algebraic invariants. *arXiv preprint arXiv:2604.15727*. Accepted at the ICLR 2026 Workshop on Logical Reasoning of LLMs.

Shlok Gilda, Mirela Silva, Luiz Giovanini, and Daniela Oliveira. 2021. Predicting different types of subtle toxicity in unhealthy online conversations. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2025. DOVE: A large-scale multi-dimensional predictions dataset towards meaningful LLM evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 1–20. Association for Computational Linguistics. Independently validates WLNK principle: reasoning chain reliability equals its weakest step.

Q. Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024.

LiveBench: A challenging, contamination-free LLM benchmark. *arXiv preprint arXiv:2406.19314*.

Ronald R. Yager. 1988. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190.

Qiangeng Yang, Tess Christensen, Shlok Gilda, Juliana Fernandes, Daniela Oliveira, Ronald Wilson, and Damon Woodard. 2024. Are fact-checking tools helpful? an exploration of the usability of Google fact check. In *Proceedings of the ACM on Human-Computer Interaction*.