The Missing Layer: Evidence Quality Infrastructure for AI Evaluation

Christopher Kelly

Abstract

AI evaluation has built impressive reporting infrastructure. The Every Eval Ever (EEE) database can now document what was measured, how, and what the result was. But reporting is not evidence. A well-reported evaluation is not necessarily a trustworthy one. We identify three dimensions of evidence quality, drawn from three distinct traditions, that current AI evaluation infrastructure does not capture: design soundness (from experimental design methodology), transparency and verifiability (from the open science movement), and evidence strength and decision-readiness (from evidence-based policy). We show that these dimensions are separable and non-substitutable, that no existing AI evaluation framework addresses all three, and that medicine's GRADE framework implicitly integrates what we make explicit. We then propose concrete schema extensions for EEE that operationalize all three dimensions as structured, queryable metadata. The contribution is both conceptual—a principled framework for evidence quality assessment in AI evaluation—and practical—an implementable infrastructure layer that can be adopted incrementally within the existing evaluation ecosystem.

1 Introduction

The Every Eval Ever schema tells you that Model X scored 85% on Benchmark Y using Library Z at Temperature T. It does not tell you whether the study design was sound enough to support the claims being made. It does not tell you whether anyone can independently verify the results. And it does not tell you how much confidence a policymaker, developer, or researcher should place in the finding. These are three distinct questions, and current AI evaluation infrastructure answers none of them systematically.

This matters because AI evaluation results are not academic exercises. They drive model releases, safety determinations, deployment approvals, and regulatory assessments. The METR productivity RCT illustrates the stakes: coding benchmarks suggested AI tools improve developer productivity, but a rigorous randomized controlled trial found that experienced developers took 19% longer with AI assistance, contradicting both benchmark evidence and developers' own perceptions of speedup (METR, 2025a). Benchmark reports and a carefully designed experiment produced opposite conclusions. The field currently has no infrastructure for distinguishing between them on grounds of evidence quality.

Medicine encountered this same problem decades ago. The response was to develop parallel systems: CONSORT for reporting clinical trials, PROBAST for assessing design quality, the TOP Guidelines for transparency standards, and GRADE for rating overall evidence certainty. AI evaluation has begun importing the reporting half of this infrastructure. EEE, Model Cards (Mitchell et al., 2019), Datasheets for Datasets (Gebru et al., 2021), and EvalFactsheets all address what to report. But the evidence-quality half—the tools that answer "should we believe this result?"—remains largely absent. A systematic review applying PROBAST to ML-based oncology prediction models found 84% at high risk of bias (Moons et al., 2025), demonstrating that extensive reporting coexists with poor evidence quality. Weidinger et al. (2025) identify the same insufficiency and call for maturing an evaluation science for generative AI systems, drawing lessons from safety evaluation in transportation, aerospace, and pharmaceutical engineering.

Recent work has made important progress toward closing this gap. Wallach et al. (2024) argue that AI evaluation is fundamentally a social science measurement challenge and propose a hierarchical framework for explicitly conceptualizing what is being measured. Salaudeen et al. (2025), Ro-

driguez et al. (2025), and Bean et al. (2025) have applied validity theory to benchmarks. Eriksson et al. (2025) catalog benchmark problems systematically. Paskov et al. (2025) articulate rigor guidelines for general-purpose AI evaluations. McCaslin et al. (2025) propose STREAM, a 28-criterion reporting standard for dangerous capability evaluations that represents the most detailed transparency checklist in the field to date. Raji et al. (2021) diagnosed the structural limitations of benchmark-centric evaluation. Each contributes valuable insights. But each addresses a subset of what evidence quality encompasses, and none proposes infrastructure that makes evidence quality queryable.

This paper makes two contributions. First, we identify three dimensions of evidence quality for AI evaluation, drawing on three traditions that the AI evaluation literature has not yet substantively engaged: the Shadish et al. (2002) experimental design framework, the TOP Guidelines from the open science movement (Nosek et al., 2015), and the tiered evidence classification system from evidence-based policy (Coalition for Evidence-Based Policy). These dimensions are analytically separable and non-substitutable: design soundness is necessary but not sufficient for strong evidence; transparency enables assessment of design soundness but does not substitute for it; evidence strength integrates across both while adding the critical factor of independent replication. Second, we propose concrete schema extensions for EEE that operationalize all three dimensions as structured, queryable metadata, following principles of incremental adoption and backward compatibility.

Section 2 describes the existing AI evaluation infrastructure stack and identifies the missing layer. Section 3 presents the three-dimensional evidence quality framework. Section 4 illustrates how all three dimensions apply across evaluation types, with deep dives into static benchmarks and AI evaluation RCTs. Section 5 proposes the schema extensions. Section 6 argues for urgency. Section 7 discusses limitations and future work.

2 The Infrastructure We Have and the Layer We Don't

AI evaluation has developed a substantial infrastructure stack over the past several years. It is useful to distinguish three layers that currently exist and a fourth that does not.

Layer 1: Execution. Frameworks like lm-eval-harness (Biderman et al., 2024), Inspect AI, and HELM (Liang et al., 2023) allow evaluations to be run in standardized, reproducible ways. This layer answers: *can we run the evaluation?*

Layer 2: Reporting. The EEE schema, Model Cards, System Cards, and Datasheets for Datasets record what was done and what resulted. EEE is the most ambitious of these, providing a unified JSON schema with converters for multiple execution frameworks, standardized metric types, and structured uncertainty quantification. This layer answers: *what was measured, how, and what was the result?*

Layer 3: Documentation. EvalFactsheets, benchmark papers, and technical reports describe evaluation methodology, often in considerable detail. EvalFactsheets include fields for purpose, design, known limitations, and validity indicators. This layer answers: *how was the evaluation designed and why?*

The missing Layer 4: Evidence Assessment. None of these layers systematically answers the question: *should we believe this result?* A score of 85% on a benchmark might reflect genuine capability, data contamination, construct mismatch, statistical noise, or some combination. The existing infrastructure records the score but provides no structured way to assess the evidence behind it. Specifically, no existing queryable infrastructure addresses all three dimensions of evidence quality we identify below. Paskov et al. (2025) and Bean et al. (2025) identify the criteria rigorous AI evaluation should meet; Bordes et al. (2025) and McCaslin et al. (2025) address documentation and reporting quality; this paper proposes the infrastructure for making those criteria queryable at scale, and adds a formal evidence strength dimension with defined necessary conditions that all prior frameworks leave absent or informal. An initial non-systematic review of key existing frameworks against these three dimensions is provided in the supplementary material.

This gap is not a criticism of the existing infrastructure. EEE is well-designed for its stated purpose: standardizing how evaluation results from different frameworks are stored and compared. Several features of EEE show the team already recognizes that evidence quality matters beyond raw scores. The `evaluator_relationship` field captures whether an evaluation was first-party, third-party, or collaborative, an important independence indicator. The `uncertainty` object pro-

vides structured fields for standard error, confidence intervals, and bootstrap information. The `additional_details` escape hatches throughout the schema suggest awareness of metadata that has not yet been formalized. Our proposal extends this existing awareness into a systematic evidence quality layer.

The question "should we believe this result?" decomposes into three sub-questions, each addressed by a different intellectual tradition. We turn to these next.

3 Three Dimensions of Evidence Quality

Evidence quality is not a single concept. Drawing on three distinct traditions, we identify three dimensions that together characterize the evidentiary strength of an AI evaluation. These dimensions are analytically separable and non-substitutable. No existing AI evaluation framework addresses all three. Medicine's GRADE framework implicitly integrates what we propose to make explicit: risk of bias corresponds roughly to design soundness, indirectness and imprecision to evidence synthesis, and the overall certainty rating to evidence strength. GRADE operates at the systematic review level—rating bodies of evidence rather than individual studies—but this is a feature, not a mismatch: the schema makes individual AI evaluations GRADE-ready, standardizing what evaluators need to document so that future systematic reviews can grade the accumulated body of AI evaluation evidence the way a Cochrane review grades clinical trials. Separating the dimensions makes each independently assessable and implementable as structured metadata.

3.1 Dimension 1: Design Soundness

Source tradition: Experimental design methodology (Shadish et al., 2002).

Core question: Does the evaluation's design support the inferential claims being made?

The Shadish, Cook, and Campbell four-validity taxonomy provides systematic tools for diagnosing threats to inferential claims.[1] *Construct validity* asks whether the evaluation measures what it claims to measure. A benchmark labeled "reasoning" may primarily test pattern matching. Bean et al. (2025) found pervasive construct validity weaknesses across 445 LLM benchmarks. Salaudeen et al. (2025) and Rodriguez et al. (2025) have begun applying psychometric validity frameworks to this problem, primarily for benchmarks. We extend the analysis across all evaluation types.

Internal validity asks whether the design supports causal or comparative claims. This is critical for AI evaluation RCTs, safety intervention evaluations, and any claim that Model A outperforms Model B. Threats include data contamination, confounding variables, and evaluation gaming, where frontier models recognize and strategically alter behavior during evaluation (Apollo Research, 2024; Hubinger et al., 2024). Internal validity has been essentially unaddressed at the infrastructure level; Paskov et al. (2025) articulate it in their rigor guidelines but not in queryable form.

External validity asks whether results generalize beyond the specific evaluation context. Vivalt (2020) demonstrated that treatment effect heterogeneity is the dominant source of evidence uncertainty in impact evaluations: the typical study result differs from the average effect in similar studies by almost 100%. Most AI benchmarks make implicit generalization claims with no documentation of scope or domain restrictions.

Statistical conclusion validity asks whether the statistical inferences are warranted. Breznau et al. (2022) found that 73 research teams analyzing identical data reached meaningfully different conclusions, revealing what they called a "hidden universe of uncertainty" from analytical flexibility. The same evaluator-degrees-of-freedom problem applies to AI evaluation: metric selection, dataset splits, prompting strategies, and scoring criteria all constitute analytical choices that can produce dramatically different conclusions about the same model. Zhou et al. (2025) showed that benchmark comparisons routinely report differences that are not statistically meaningful. Card et al. (2020) demonstrate a related failure: NLP experiments are systematically underpowered, with most standard benchmark comparisons lacking the statistical power to distinguish signal from noise, and they call for adoption of 80% power as the minimum threshold (Cohen, 1962). AI evaluation has not adopted even this baseline. We add one gap Card et al. do not address: evaluations that are adequately powered for average effects routinely

[1] The S/C/C framework focuses on threats to inference in experimental and quasi-experimental designs. A complementary measurement-theory tradition, centered on unified validity as evidence supporting intended interpretations of scores, is codified in the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014); Salaudeen et al. (2025), Rodriguez et al. (2025), and Wallach et al. (2024) apply that tradition to LLM evaluation.

conduct subgroup and heterogeneity analyses on the same data for which the original design had no power—and current reporting norms do not require disclosing this limitation.

3.2 Dimension 2: Transparency and Verifiability

Source tradition: Open science movement (Nosek et al., 2015); TOP Guidelines; Center for Open Science.

Core question: Can an independent party access the materials needed to verify the evaluation and its results?

The Transparency and Openness Promotion (TOP) Guidelines define research practices at increasing levels of stringency, from disclosure through sharing to independent certification. Over 5,000 journals and organizations have adopted TOP. We adapt this into a five-level progression for AI evaluation: *Level 1, Disclosure:* state what was evaluated and how. *Level 2, Sharing:* share evaluation data, code, prompts, and model outputs. *Level 3, Certification:* independent certification that shared materials are complete, functional, and accurately described. *Level 4, Verification:* independent reproduction of results using shared materials. *Level 5, Replication:* independent replication in new contexts.

A key distinction within this dimension is between *pre-commitment* and *reporting*: specifying what analyses will be done before seeing data versus documenting what was done afterward. Both are forms of transparency, but they serve different functions. The schema's `pre_registration` object captures this through a `registration_type` enum distinguishing: `none`; `minimal` (a vague timestamp with no analytical specificity, providing little constraint on evaluator degrees of freedom); `pre_analysis_plan` (specifying primary outcomes, statistical tests, stopping rules, and subgroup analyses in advance); and `registered_report` (peer-reviewed design with editorial commitment before data collection). Research on Registered Reports in psychology suggests they are associated with substantially lower positive result rates than standard publications (Scheel et al., 2021). Current AI evaluation has strong reporting infrastructure (EEE, Model Cards) but minimal pre-commitment infrastructure.

Transparency alone is not sufficient. Metascience research has documented systematic verification failures: pre-registration URLs are often broken or point to incorrect documents, shared code may not run, and registration documents sometimes fail to match the stated methods (Hardwicke et al., 2018). Level 3 exists precisely to catch these failures—certifying that links are live, code is functional, data is where it was supposed to be, and any pre-registration document actually specifies what it claims. Level 3 is essentially absent in AI evaluation. EEE captures source URLs and evaluation code links, placing it at partial Level 2. Almost nothing in AI evaluation reaches Level 3 or above.

Levels 4 and 5 introduce independent confirmation through replication, which takes meaningfully different forms. *Computational replication* uses the same data and code to reproduce the same numerical results—it verifies correctness but not generalizability. *Direct replication* collects new data under the same protocol and operationalization, testing whether the finding holds under the same conditions. *Conceptual replication* uses a different operationalization of the same construct, testing whether the finding is robust to methodological variation (Goodman et al., 2016; National Academies of Sciences, Engineering, and Medicine, 2019). These types are not interchangeable: a result can be computationally reproducible, directly replicable, and conceptually irreplicable simultaneously. The schema's `replication_type` field requires evaluators to specify which type applies. Reaching Level 5 requires only that some form of independent replication has occurred; computational replication suffices. Dimension 3 Tier 1, by contrast, requires direct or conceptual replication together with rigorous design and pre-registration—the full evidential record, not the replication event alone. The two dimensions are correlated at the top but track different properties: Dimension 2 captures whether materials are accessible and results have been independently confirmed; Dimension 3 captures the evidential weight of the full evaluation record including design quality.

The relationship between transparency and design soundness is important and asymmetric. Transparency *enables* assessment of design soundness: you cannot identify internal validity threats in a study you cannot inspect. But a fully transparent study can still have fatal design flaws, and a well-designed study conducted behind closed doors cannot be verified. This asymmetry also creates a genuine tension in safety evaluation contexts, where full transparency may enable evaluation gaming. The framework should represent this tension ex-

plicitly, and the schema includes fields for documenting justified restrictions on transparency.

3.3 Dimension 3: Evidence Strength and Decision-Readiness

Source tradition: Evidence-based policy (Coalition for Evidence-Based Policy; GRADE framework in medicine).

Core question: How much confidence should we have in this evaluation result, and is it ready to inform consequential decisions?

The Coalition for Evidence-Based Policy developed a tiered evidence classification system for social programs. Top Tier status requires sizable, sustained effects demonstrated in well-conducted RCTs, independently replicated at multiple sites. We adapt this into a four-tier system for AI evaluation (Table 1).

Replication is the key differentiator between tiers. Most current AI evaluation evidence, including influential benchmark results, safety evaluations, and capability assessments, falls at Tier 3 or 4. IJzerman et al. (2020) proposed "evidence readiness levels" for behavioral science, cautioning against premature application of research findings to high-stakes decisions. AI evaluation results are routinely used for consequential decisions despite evidence that would not meet the threshold for policy readiness under any standard evidence hierarchy.

3.4 Why All Three Dimensions Are Necessary

A valid but opaque study cannot be trusted because we have no way to check the claims. A transparent but invalid study should not be trusted because we can see it is flawed. Neither validity assessment nor transparency tells us the overall confidence level or whether the evidence has been independently confirmed (Table 2). The dimensions are not logically independent—transparency enables assessment of design soundness, and some artifacts furnish evidence to multiple dimensions—but they are non-substitutable: a strong score on one cannot compensate for a weak score on another.

Medicine's GRADE framework implicitly integrates all three: risk of bias maps to design soundness, indirectness and imprecision to evidence synthesis, and the overall certainty rating to evidence strength. Our framework makes explicit what GRADE combines, separating the dimensions so each can be independently assessed and implemented as structured metadata. This separation also enables the individual-to-aggregate bridge: where GRADE grades bodies of evidence in systematic reviews, our schema prepares individual evaluations to be the inputs to such reviews.

4 The Reporting-Evidence Gap Across Evaluation Types

The three-dimensional framework is general; its application to AI evaluation is specific. Table 4 in the supplementary material provides a breakdown across evaluation types. Two deep dives illustrate the framework in detail.

Deep Dive 1: Static Benchmarks. Static benchmarks illustrate all three dimensions: construct validity concerns are pervasive (Bean et al., 2025), contamination threatens internal validity (Sainz et al., 2023), and most results—single evaluations by model developers—place at Tier 3 at best.

Deep Dive 2: AI Evaluation RCTs. AI evaluation RCTs make explicit causal claims, making design soundness paramount. The METR productivity RCT rates as Tier 3 by the three-dimensional framework: rigorous design but not formally pre-registered, with specific population and no independent replication (METR, 2025a). Appendix C works through the field-level annotations and notes that a rigorous RCT without pre-registration surfaces a gap the tier system does not yet handle cleanly. A follow-up study does not upgrade this rating—it was not pre-registered, suffered severe acknowledged selection bias, and the researchers themselves described it as providing "only very weak evidence" (METR, 2026); two single studies at Tier 3 and Tier 4 do not compound into higher tiers. A separate METR finding sharpens the construct validity concern: Claude 3.7 Sonnet achieved 38% on automated test-passing metrics but 0% on holistic manual review (METR, 2025b). Organizations sometimes keep RCT methodology confidential to prevent gaming—the `justified_transparency_restrictions` field represents this constraint explicitly.

5 Proposed Evidence Quality Layer for EEE

5.1 Design Principles

The proposal follows four principles. **Incremental:** extend, do not replace; all new fields are optional; full backward compatibility is maintained. **Structured:** use enums and structured

Tier	Necessary Conditions	Adequate for...	Proposed governance mapping
Tier 1: Replicated Rigorous	Rigorous design; direct or conceptual replication; pre-registration or registered report	Broad deployment; regulatory submission; safety case	ASL-3+; EU AI Act high-risk
Tier 2: Single Rigorous	Rigorous design (RCT or strong quasi-experiment); pre-registration; no independent replication yet	Research publication; limited deployment	ASL-2; medium-risk
Tier 3: Suggestive	Controlled comparison or observational with controls; some transparency documentation	Internal decision-making; further research prioritization	Low-risk; exploratory
Tier 4: Preliminary	Any documented evaluation; no minimum design requirements	Hypothesis generation only	Pre-deployment research

Table 1: Evidence tiers with necessary conditions and decision-readiness mapping. Self-reported `evidence_tier` must be consistent with these conditions; a `decision_context` of `broad_deployment` paired with `evidence_tier: preliminary` is a detectable and flaggable mismatch. The governance mapping column is a proposed convention rather than a derivation: ASL levels and EU AI Act risk categories require rigorous, risk-proportionate evaluation but do not currently specify evidence quality criteria, and this column illustrates how the tier system could fill that gap. Adoption of any such mapping as normative would require engagement with the relevant standards bodies. Threshold choices in such a mapping have distributional consequences across the evaluator ecology, and the deliberative process for setting them is itself a governance question.

Dimension	Captures	Misses
Design Soundness	What *could* be wrong (diagnosis)	Whether we *can check*; whether findings *replicate*
Transparency	Whether we *can verify* (auditability)	Whether design *is sound*; overall *confidence*
Evidence Strength	How much to *believe* (bottom line)	*Specific* design flaws; *specific* transparency gaps

Table 2: Each dimension captures something the others miss.

objects over free text wherever possible. **Three-dimensional:** schema sub-sections correspond to the three evidence quality dimensions, plus cross-cutting study design fields. **Adoption path:** fields can start as conventions within EEE's existing `additional_details` escape hatches before being promoted to formal schema fields.

5.2 Schema Architecture

The proposed `evidence_quality` top-level section has four sub-sections (Figure 1).

5.3 Representative Fields by Dimension

Dimension 1 fields include `construct_validity` (requiring `target_construct`, `operationalization`, `known_construct_threats`), `internal_validity` (`causal_claims_made`, `known_confounds`), and `evaluation_type_classification` for study design type. Dimension 2 gives each of four artifact types a five-level availability enum (`not_disclosed` through `verified_or_replicated`)—no existing AI evaluation framework provides this graduated structure. Dimension 3 includes `evidence_tier`, `independent_review` with replication count and consistency fields, and `conflicts_of_interest`. Some fields, notably pre-registration, are cross-cutting across all three dimensions. Table 3 maps selected fields to dimensions and source traditions; the full JSON schema is in Appendix B.

6 Why This Matters Now

Three arguments for urgency. **Epistemic.** A field that evaluates AI systems has no empirical tradition of studying its evaluation quality. The evaluator degrees of freedom problem demonstrated by Breznau et al. (2022), in which different analytical choices produce different conclusions from identical data, applies directly to AI evaluation. Without infrastructure to document these choices and assess their consequences, we have no way to know the magnitude of this problem in AI evaluation specifically.

Practical. Evaluation results drive model re-

```
evidence_quality
+-- validity_assessment          [Dimension 1: Design Soundness]
|   +-- construct_validity
|   +-- internal_validity
|   +-- external_validity
|   +-- stat_conclusion_validity
+-- transparency_level          [Dimension 2: Transparency & Verifiability]
|   +-- eval_data_availability          (5-level enum)
|   +-- eval_code_availability          (5-level enum)
|   +-- model_outputs_availability      (5-level enum)
|   +-- analysis_scripts_availability   (5-level enum)
+-- evidence_synthesis           [Dimension 3: Evidence Strength]
|   +-- evidence_tier                   (4-tier enum)
|   +-- independent_review
|   +-- known_limitations
|   +-- conflicts_of_interest
+-- study_design_quality         [Cross-cutting]
    +-- eval_type_classification        (enum)
    +-- pre_registration               (registration_type enum;
    protocol_adherence)
    +-- contamination_tested
    +-- eval_awareness_tested
    +-- blinding_procedures
    +-- power_analysis_conducted
    +-- multiple_comparisons_adjusted
    +-- effect_size_reported
    +-- decision_context              (6-level enum; cross-dim.
    accountability)
```

Figure 1: Proposed `evidence_quality` schema architecture. Keys name schema fields; types indicate value constraints; annotations map fields to evidence quality dimensions.

Schema Field	Dim.	Source Tradition
`evaluation_type_classification`	C	Nomenclature contribution
`pre_registration.*`	C	Open science (TOP Guidelines)
`contamination_tested`	1+C	AI evaluation
`construct_validity.*`	1	Psychometrics → AI eval
`internal_validity.*`	1	Shadish, Cook, & Campbell
`external_validity.*`	1	Development economics
`transparency_level.*`	2	Open science (adapted TOP)
`evidence_tier`	3	Coalition for Evidence-Based Policy
`independent_review.*`	3	Coalition review methodology
`conflicts_of_interest.*`	3	ICMJE / CONSORT

Table 3: Selected schema fields mapped to evidence quality dimensions and source traditions. C = cross-cutting.

leases, safety determinations, and deployment approvals. The METR productivity RCT reversal, in which benchmarks and a carefully designed experiment produced opposite conclusions about AI coding tools (METR, 2025a), illustrates that consequential decisions rest on evidence that would not survive systematic quality assessment.

Governance. Existing governance frameworks—the EU AI Act (Article 9), NIST AI Risk Management Framework, and ISO/IEC SC 42—assert that AI evaluations should be rigorous and proportionate to risk, but leave the meaning of evaluation quality unspecified. This is a pyramid inversion in the Nosek (2019) model: policy-layer mandates arrived before the infrastructure, ease, and norms layers that make meaningful compliance possible. Organizations can formally satisfy requirements for "rigorous, risk-proportionate evaluation" by doing whatever evaluations they were already doing, because no structured vocabulary for evidence quality exists. The present schema is an infrastructure-layer response to that mandate. Frontier Safety Frameworks report evaluation results without documenting the evidentiary chain from results to risk judgments (Ziosi et al., 2025). Under any standard evidence hierarchy, most current AI evaluation evidence would not qualify as policy-ready.

Enabling systematic review. The schema's most important long-run function is enabling systematic review of AI evaluation results—the mech-

anism by which mature fields produce cumulative knowledge. Researchers could apply it retroactively to existing evaluations, aggregate across evidence tiers, and produce state-of-the-evidence syntheses that currently cannot be done. In terms of the Nosek (2019) culture change model, the proposed schema contributes at the infrastructure level; a companion paper submitted to this workshop develops the full pyramid analysis and proposes a maturation agenda spanning all five levels. The accumulating cost of operating without this infrastructure—what companion work submitted to this workshop terms *evaluation debt*—is the underlying reason all three urgency arguments converge now.

7 Discussion

Relation to existing work. The proposal builds on Paskov et al.'s (2025) rigor guidelines, making them queryable. It extends EEE's existing evidence quality awareness into a systematic layer. It operationalizes diagnostic insights from Raji et al. (2021), Bean et al. (2025), and Salaudeen et al. (2025) as infrastructure. The distinctive contribution is introducing the open science (TOP Guidelines) and evidence-based policy (Coalition) traditions to AI evaluation infrastructure—translations that have not previously been made.

Annotation burden. Full population is not realistic for every evaluation. We propose a tiered path: a *minimum core* (`evaluation_type_classification`, `construct_validity.target_construct`, `transparency_level.eval_data_availability`, `decision_context`) for any evaluator; an *extended set* including `evidence_tier` and validity threats, comparable in effort to populating an EvalFactsheet (Bordes et al., 2025); and an *independent-assessment set* (`*.certified`, `independent_review.*`, `has_been_replicated`) restricted to third parties by design. For benchmark-style evaluations, much of the metadata is naturally attached at the benchmark level and inherited by individual runs; bespoke study designs rely more on evaluator-authored documentation. Appendix E sketches further design directions for lowering adoption friction.

Generative evaluation paradigms. EEE's current architecture centers on benchmark-style execution; other modalities fit to varying degrees. *LLM-as-judge* (e.g., Chatbot Arena, MT-Bench) fits adjacent with minor extensions—judge model identity, inter-judge reliability, judge calibration as a nested evaluation—and construct validity is the primary threat. *Red-teaming* stresses all three dimensions: the construct is unsettled, the counterfactual is rarely specified, and tier criteria do not map cleanly onto adversarial elicitation. *Continuous monitoring* requires schema extensions for temporal versioning and tier-updating, and carries the highest near-term governance stakes (EU AI Act Article 72, frontier safety frameworks). Appendix A sketches dimension applicability across evaluation types.

Future work. Empirical validation is the immediate priority: applying the schema to a sample of existing evaluations and assessing inter-rater reliability of evidence quality coding. Pilot integration with EEE maintainers would test practical feasibility. Schema v2 should add verification infrastructure, structured reliability reporting, and effect size distributions across replications. Red-teaming and continuous monitoring evaluation regimes require substantially different schema designs and are natural targets for follow-on work. The governance design problem—how to credit high-quality evaluations conducted under confidentiality constraints, and how to structure mandatory third-party assessment for high-stakes deployments—is substantial enough for a dedicated paper.

Limitations

The three-dimensional framework is the contribution; the type-specific analysis is illustrative. The most important limitation is the self-assessment problem: evidence quality fields would initially be completed by the evaluating parties. The schema addresses this structurally: `design_soundness_claim` (submitter-provided) is architecturally separate from `design_soundness_verified_by` (third-party), making self-assessment visible rather than mistakable for independent review. Schema gaming is a real risk; structural mitigations include `decision_context/evidence_tier` mismatch detection and a proposed coalition working group as stewardship body (Goldacre et al., 2019). The schema is designed for benchmarks, human evaluations, and causal impact studies; red-teaming and continuous monitoring require different designs and are future work.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

Apollo Research. 2024. We need a science of evals. Apollo Research Blog.

Andrew Bean, Ronan Kearns, Antonia Romanou, Jan Batzner, and 1 others. 2025. Measuring what matters: The case for construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, and 1 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Florian Bordes, Candace Ross, Jared T. Kao, Evangelia Spiliopoulou, and Adina Williams. 2025. Eval factsheets: A structured framework for documenting AI evaluations. In *arXiv preprint arXiv:2512.04062*.

Nate Breznau, Eike Mark Rinke, Alexander Wuttke, and 1 others. 2022. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Jacob Cohen. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3):145–153.

Benjamin Eriksson, Hao Cheng, Shammur Absar Chowdhury, and 1 others. 2025. Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. *arXiv preprint arXiv:2502.06559*.

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.

Ben Goldacre, Henry Drysdale, Anna Powell-Smith, Anna Dale, Iain Milosevic, Eleanor Slade, Philip Hartley, Clare Marston, Carl Heneghan, and Kamal R. Mahtani. 2019. The COMPare trials project: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20:118.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12.

Tom E. Hardwicke, Maya B. Mathur, Kyle MacDonald, and 1 others. 2018. Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, 5(8):180448.

Evan Hubinger, Carson Denison, Jesse Mu, and 1 others. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Hans IJzerman, Neil A. Lewis Jr., Andrew K. Przybylski, and 1 others. 2020. Use caution when applying behavioural science to policy. *Nature Human Behaviour*, 4(11):1092–1094.

Percy Liang, Rishi Bommasani, Tony Lee, and 1 others. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525:140–146.

Yu Lu Liu, Su Lin Blodgett, Jackie Chi Kit Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Tegan McCaslin, Jide Alaga, Samira Nedungadi, Seth Donoughe, Tom Reed, Rishi Bommasani, Chris Painter, and Luca Righetti. 2025. STREAM (ChemBio): A standard for transparently reporting evaluations in AI model reports. *arXiv preprint arXiv:2508.09853*.

METR. 2025a. Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*.

METR. 2025b. Research update: Towards reconciling the slowdown with time horizons. METR Blog, August 12, 2025.

METR. 2026. Uplift update: Measuring late-2025 AI on open-source developers. METR Blog, February 24, 2026.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, and 1 others. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Karel G. M. Moons, Robert F. Wolff, Richard D. Riley, and 1 others. 2025. PROBAST+AI: A tool to assess risk of bias and applicability of prediction model studies with AI components. *BMJ*, 388:e082505.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. National Academies Press, Washington, DC.

Brian A. Nosek. 2019. Strategy for culture change. Center for Open Science.

Brian A. Nosek, George Alter, George C. Banks, and 1 others. 2015. Promoting an open research culture. *Science*, 348(6242):1422–1425.

Patricia Paskov, Michael J. Byun, Kevin Wei, and Toby Webster. 2025. Preliminary suggestions for rigorous GPAI evaluations. Technical Report PE-A3971-1, RAND Corporation. Also arXiv:2508.00875.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *NeurIPS 2021 Datasets and Benchmarks Track*.

Pedro Rodriguez, Hammed Salaudeen, Sanmi Koyejo, and 1 others. 2025. Measurement to meaning: Validity and reliability in AI evaluation. *arXiv preprint arXiv:2505.10573*.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, and 1 others. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP 2023*.

Hammed Salaudeen, Siddharth Vasan, Sagnik Ray Choudhury, and Zeerak Talat. 2025. Measuring what satisfies: On the evaluation of LLMs with psychometric validation. *arXiv preprint arXiv:2501.09674*.

Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens. 2021. An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, 4(2).

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Eva Vivalt. 2020. How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.

Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2024. Evaluating generative AI systems is a social science measurement challenge. *arXiv preprint arXiv:2411.10939*.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative AI systems. *arXiv preprint arXiv:2503.05336*. First two authors contributed equally.

Kun Zhou, Yutao Yao, Zhipeng Zhu, and 1 others. 2025. Large language model evaluation via item response theory. *arXiv preprint arXiv:2501.08225*.

Marta Ziosi, James Gealy, Miro Plueckebaum, and 1 others. 2025. Safety frameworks and standards: A comparative analysis. Oxford Martin AI Governance Initiative.

A Reporting-Evidence Gap Across Evaluation Types

Eval Type	Design Soundness Threats	Transparency Gaps	Typical Evidence Tier
Static Benchmark	Construct (vague operationalization); Statistical conclusion (underpowered comparisons)	Data often shared; contamination testing rarely verifiable	Suggestive to Preliminary
Dynamic Benchmark	Construct (item equivalence); Reliability	Regeneration methodology variably documented	Similar to static
Safety / Red-Team	Internal (evaluation gaming); Construct (undefined thresholds)	Often deliberately opaque	Preliminary to Suggestive
AI Evaluation RCT	Internal (randomization, blinding); Statistical conclusion (power)	Variable: some methodology confidential	Suggestive to Near Top Tier
Quasi-Experimental	Internal (confound control); Statistical conclusion (specification sensitivity)	Design assumptions rarely documented	Suggestive at best
Human Preference	Construct (biases); Reliability (inter-rater agreement)	Judge prompts sometimes shared	Suggestive to Preliminary
Structured Audit	Internal (access limitations); Independence	Often proprietary	Varies widely

Table 4: The reporting-evidence gap across AI evaluation types.

B Evidence Quality JSON Schema

The proposed `evidence_quality` section extends the EEE schema with the following structure (abbreviated; full specification available upon request).

```
{
 "evidence_quality": {
  "evaluation_type_classification":
    enum[static_benchmark, dynamic_benchmark, human_preference, red_team,
         structured_audit, randomized_controlled_trial, quasi_experimental, ...],
  "pre_registration": {
    "registration_type": enum[none, minimal, pre_analysis_plan, registered_report],

    "registry_url": string,  "registration_content_url": string,
    "primary_outcomes_pre_registered": boolean,
    "protocol_adherence": {
      "deviation_status": enum[no_deviations, deviations_present, not_assessed],
      "deviations": [{pre_registered_element, deviation_type,
    deviation_description}]
    }
  },
  "validity_assessment": {
    "construct_validity": {
      "target_construct": string,  "operationalization": string,
      "known_construct_threats": [string]
    },
    "internal_validity": {
      "causal_claims_made": boolean,  "design_supports_causal": boolean,
      "known_confounds": [string]
    },
    "external_validity": {
      "population_scope": string,  "domain_restrictions": [string],
      "generalization_basis": enum[theoretical, prior_empirical, expert_judgment,
    none_stated],
      "generalization_claims": string,  "known_generalization_limits": [string]
    },
    "stat_conclusion_validity": {
      "effect_size_type": string,  "effect_size_value": number,
      "practical_significance": boolean
    }
  },
  "transparency_level": {
    "eval_data_availability": enum[not_disclosed, disclosed, shared, certified,
                                   verified_or_replicated],
    "eval_code_availability": enum[...],
    "model_outputs_availability": enum[...],
    "analysis_scripts_availability": enum[...],
    "justified_restrictions": string
  },
  "evidence_synthesis": {
    "evidence_tier": enum[replicated_rigorous, single_rigorous, suggestive,
    preliminary],
    "independent_review": {
      "has_been_replicated": boolean,  "replication_count": integer,
      "replication_type": enum[computational, direct, conceptual],
      "results_consistent": boolean
    },
    "decision_context": enum[research_publication, internal_decision_making,
      limited_deployment, broad_deployment, regulatory_submission, safety_case],
    "known_limitations": [string],
    "conflicts_of_interest": {
      "funding_source": string,  "commercial_interest": boolean
    }
  }
 }
}
```

C Worked Example: METR 2025 Productivity RCT

The METR 2025 productivity RCT (METR, 2025a) measured the effect of AI tool access on experienced developers' completion time on real software-engineering tasks (16 developers, 246 tasks, within-subject randomization). We use it as a worked example for three reasons: it is well-documented publicly, it is externally salient (its finding contradicted both benchmark evidence and developer self-report, motivating Section 4), and it exercises most of the schema's fields. Applying the framework also surfaces a tier-system gap, discussed below.

Fields below distinguish *confirmed* (consistent with public artifacts: METR's paper, blog post, and GitHub repository) from *illustrative* (plausible characterizations reflecting our reading of public materials, not directly verified against METR's internal records). In a production setting, illustrative fields would be confirmed by direct annotation from the evaluating team or a third-party annotator with access to the full study record.


```
{
 "evidence_quality": {
  "evaluation_type_classification": "randomized_controlled_trial",
  "pre_registration": {
    "registration_type": "none",
    "primary_outcomes_pre_registered": false,
    "protocol_adherence": {
      "deviation_status": "not_assessed"
    }
  },
  "validity_assessment": {
    "construct_validity": {
      "target_construct": "AI-tool-induced change in completion time on real software-engineering
    tasks for experienced developers",
      "operationalization": "Self-reported task completion time on randomly assigned real GitHub
    issues, comparing AI-allowed vs. AI-disallowed conditions",
      "known_construct_threats": [
        "Self-reported completion time may not capture quality differences",
        "Self-reported speedup expectations diverged from measured slowdown, suggesting the construct
     includes a perception component the measurement does not capture",
        "Automated test-passing metrics correlate poorly with holistic manual code-quality review
    (38% vs. 0% in METR's reconciliation analysis)"
      ]
    },
    "internal_validity": {
      "causal_claims_made": true,
      "design_supports_causal": true,
      "known_confounds": [
        "Task heterogeneity within developer (addressed by within-subject randomization)",
        "Possible cross-condition learning spillover",
        "Hawthorne / observation effects: developers knew tasks were timed and recorded",
        "Unblinded intervention: participants cannot be blinded to AI access"
      ]
    },
    "external_validity": {
      "population_scope": "16 experienced contributors to mature open-source projects; 246 tasks
    total",
      "domain_restrictions": [
        "Real software engineering on familiar codebases",
        "Specific AI tool stack (Cursor Pro, Claude 3.5/3.7 Sonnet)"
      ],
      "generalization_basis": "prior_empirical",
      "known_generalization_limits": [
        "Does not measure novice or intermediate developer effects",
        "Does not measure greenfield or non-coding tasks",
        "Tool stack specific to early-2025 AI capabilities"
      ]
    },
    "stat_conclusion_validity": {
      "effect_size_type": "percent change in completion time (95% CI)",
      "effect_size_value": 19,
      "effect_size_ci": [2, 39]
    }
```

```
  },
  "transparency_level": {
    "eval_data_availability": "shared",
    "eval_code_availability": "shared",
    "model_outputs_availability": "disclosed",
    "analysis_scripts_availability": "shared",
    "justified_restrictions": ""
  },
  "evidence_synthesis": {
    "evidence_tier": "suggestive",
    "independent_review": {
      "has_been_replicated": false,
      "replication_count": 0
    },
    "decision_context": "research_publication",
    "known_limitations": [
      "Single study; no independent replication",
      "Population specific to experienced open-source developers",
      "Effect estimate sensitive to tool stack tested",
      "Not formally pre-registered; analytic specificity not pre-committed"
    ],
    "conflicts_of_interest": {
      "funding_source": "METR (independent research org; receives funding from frontier labs and Open
     Philanthropy); first-party relative to AI tools tested but third-party relative to model
    developers",
      "commercial_interest": false
    }
  }
 }
}
```


Several field-level decisions merit comment.

Tier rating and a framework gap. `evidence_tier: suggestive` (Tier 3) reflects the absence of formal pre-registration, which Table 1 requires for Tier 2. Applying the framework to METR 2025 surfaces a gap: a rigorous RCT without pre-registration does not satisfy Tier 2's necessary conditions, but Tier 3's wording ("controlled comparison or observational with controls") understates the actual study design. We treat this as a finding the framework's application produced rather than a result to argue around. Two principled v2 directions could address it: an intermediate tier ("Rigorous Without Pre-Registration") or treating pre-registration as a downgrading factor rather than a necessary condition (analogous to GRADE's downgrading mechanism for observational evidence). Both carry tradeoffs that are out of scope here.

Construct validity threats. `known_construct_threats` is non-empty even for a strong study. The self-reported speedup vs. measured slowdown divergence (developers expected approximately 20% speedup; measurement showed approximately 19% slowdown) is itself a construct threat: the measurement instrument and the perceived construct come apart. The 38% vs. 0% gap between automated test-passing and holistic code-quality review (METR, 2025b) is a documented construct concern the schema should surface even though it was identified after the original study. Together these motivate the community-maintained benchmark cards idea (Appendix E): construct concerns continue to surface after an evaluation is published.

Validity-type classification. Population scope and tool-stack specificity were initial candidates for `known_construct_threats` but are more naturally `external_validity.known_generalization_limits`: they concern generalization, not what the construct itself includes. The fact that this distinction requires effort even for the schema's authors is itself an argument for the controlled vocabularies discussed in Section 7.

Schema gaps surfaced by application. Two further schema limitations are exposed. The `conflicts_of_interest.commercial_interest` field is binary, which cannot represent METR's funding pathway accurately (independent research org funded in part by frontier labs and Open Philanthropy;

first-party relative to AI tools tested but third-party relative to model developers). A more granular structure is a v2 candidate. The `stat_conclusion_validity.effect_size_value` field also lacks a unit specifier; whether "19" means 19% change, 19 minutes, or 19 standard deviations depends on the `effect_size_type` string, which is free text. Both gaps illustrate how applying the schema to a real evaluation produces structured feedback on the schema itself.

Regulatory use. An AISI or notified body deciding whether to cite METR 2025 in policy guidance would find it usable for descriptive claims about a specific population (experienced open-source developers, early-2025 AI stack) but insufficient as a sole basis for deployment-conditioning decisions, given the absence of pre-registration, independent replication, and bounded population scope. The schema makes this assessment legible and queryable rather than requiring per-evaluation judgment from regulators reading prose write-ups.

What this example is and is not. This is a demonstration of how the schema applies to a single real evaluation, not an inter-annotator-agreement validation. Empirical validation requires multiple independent annotators applying the schema and measuring agreement, which is a priority for v2 work. A systematic worked-examples program spanning benchmarks, capability evaluations, red-team exercises, LLM-as-judge evaluations, and quasi-experimental designs would more thoroughly stress-test the framework; we treat that as a target for follow-on work or community contribution coordinated through the EvalEval Coalition.

D Initial Non-Systematic Review of Existing Frameworks

Table 5 compares key existing frameworks against the nine dimensions addressed by the proposed schema. This comparison is illustrative rather than exhaustive; a systematic review of existing frameworks against these dimensions is a valuable direction for future work. Column abbreviations: **CV** = construct validity; **IV** = internal validity; **EV** = external validity; **SCV** = statistical conclusion validity; **Disc.** = openness & disclosure; **Reprod.** = reproducibility & verification; **Repl.** = replication tracking; **Tiers** = evidence tiers; **Query** = queryable infrastructure.

Framework	CV	IV	EV	SCV	Disc.	Reprod.	Repl.	Tiers	Query
Model Cards (Mitchell et al., 2019)	Partial	Partial	Partial	Minimal	Yes	Partial	No	No	No
ECBD (Liu et al., 2024)	Yes	Partial	Partial	Partial	Yes	Partial	No	No	No
Paskov et al. (Paskov et al., 2025)	Partial	Yes	Yes	Partial	Yes	Yes	Partial	No	No
BetterBench (Reuel et al. 2024)	Partial	Partial	Partial	Yes	Yes	Yes	Partial	No	Partial
EvalFactsheets (Bordes et al., 2025)	Yes	Partial	Partial	Yes	Yes	Yes	Partial	No	Partial
METR protocols	Partial	Partial	Limited	—	Partial	Limited	No	No	No
STREAM (McCaslin et al., 2025)	Partial	Partial	Partial	Partial	Yes	Partial	No	Partial	No
EEE current (v0.2.1)	No	No	No	Partial	Partial	Partial	No	No	Yes
Paper 2 proposed	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 5: Initial non-systematic review of existing frameworks against nine evidence quality dimensions. "Partial" indicates the dimension is addressed but incompletely or without structured metadata. "—" indicates insufficient public documentation to assess. **Key finding:** no prior framework provides both queryable infrastructure and formal evidence tiers; EEE is the only queryable infrastructure but covers no quality dimensions; EvalFactsheets (Bordes et al. 2025) provides the most comprehensive validity coverage among non-queryable frameworks. STREAM (McCaslin et al. 2025) is the most comprehensive reporting standard and engages with evidence strength informally (criterion 1(ii): strong vs. suggestive evidence; criterion 6(i): Primary/Important/Supporting/Minor evidence weighting), but explicitly limits its scope to reporting quality rather than evidence quality and acknowledges "little consensus on how to best interpret evidence from evaluation results." ECBD characterization based on Liu et al. (2024); BetterBench based on Reuel et al. (2024).

E Design Directions for Adoption

The v1 schema treats the evaluation event (model × benchmark × configuration) as the unit of annotation. Several extensions to this design emerged in preliminary feedback and warrant v2 exploration without commitment in v1.

Benchmark-level annotation with run-level inheritance. Most fields in the heaviest portion of the schema describe the benchmark, not the run: target construct, operationalization, known construct threats, and population scope are constant across every evaluation that uses a given benchmark. Attaching these to a benchmark-level card (analogous to Datasheets for Datasets (Gebru et al., 2021) or Model Cards (Mitchell et al., 2019)) and having individual runs inherit them addresses the high-volume scaling problem directly: a frontier lab evaluating against 100 benchmarks per release inherits 100 pre-existing cards rather than producing 100 fresh annotations.

Annotation provenance and tooling-led inference. Each field could carry a provenance tag (`inherited_from_benchmark`, `evaluator_asserted`, `third_party_verified`) so that auto-population is legible rather than indistinguishable from manual annotation. An LLM-based tool reading benchmark papers and proposing field values for human review would lower the expertise barrier substantially; errors are cheap to correct, and missing fields can be flagged for community contribution.

Decision-context-triggered annotation and versioned rollout. Full annotation could be required only when an evaluation is invoked for a high-stakes decision (regulatory submission, safety case). The schema itself need not ship in full at v1: a minimum core, then benchmark-level cards, then validation-study linkage in successive versions, paralleling the staged-release model of the TOP Guidelines (Nosek et al., 2015).

These directions are not mutually exclusive; the most promising combination—benchmark-level cards plus provenance fields plus tooling-led inference—would substantially lower the adoption barrier while preserving evidence-quality content. Working out the design details is a target for v2.