Position: Toward a Metric Typology for Language Model Evaluation

Jasper Kyle Catapang
[1]Money Forward Inc., Shibaura, Minato-ku, Tokyo, Japan
[2]Tokyo University of Foreign Studies, Asahi-cho, Fuchu-shi, Tokyo, Japan
[1]catapang.j@moneyforward.co.jp

Abstract

The critique of scalar benchmark rankings as proxies for model quality is now well-established (Raji et al., 2021; Wallach et al., 2025; Bean et al., 2025; Gehrmann et al., 2021). What the field still lacks is a shared structural vocabulary for comparing, combining, and contextualizing metric design choices. This paper provides that vocabulary: a four-primitive typology—representation (ϕ), comparison (D), aggregation (A), and context (C)—under which existing metrics (BLEU, BERTScore, nDCG, LLM-as-judge, calibration scores, agentic outcome measures) are explicit parameterizations of a common form. This typology is paired with a measurement–decision split: metrics are noisy estimators of latent constructs, and model selection is context-dependent Pareto optimization over construct estimates, not over raw scores. The typology makes implicit metric assumptions comparable and debatable rather than hidden inside a single number.

1 Introduction

A substantial community now agrees that scalar leaderboard rankings are inadequate proxies for model quality (Raji et al., 2021; Wallach et al., 2025; Bean et al., 2025; Gehrmann et al., 2021; Ethayarajh and Jurafsky, 2022). The problems are well-documented: metric gaming (Goodhart, 1975), reward-model overfitting (Ouyang et al., 2022), benchmark contamination, and construct mismatch (Callison-Burch et al., 2006; Liu et al., 2023). What this consensus lacks is a unifying structural vocabulary—a common framework that makes metric design choices explicit, comparable, and debatable across the diverse metric families used in NLP and AI evaluation. This paper provides that vocabulary. Modern systems must simultaneously satisfy multiple, often conflicting properties—semantic correctness, factual grounding, helpfulness, safety, robustness—and the typology proposed here makes the assumptions behind each metric family explicit, supporting deployment-aware model selection and principled multi-construct evaluation.

2 The Problem: Measurement vs. Decision

2.1 Construct mismatch

Different metrics measure different constructs. BLEU approximates n-gram overlap (Papineni et al., 2002); ROUGE and METEOR emphasize recall and lexical alignment (Lin, 2004; Banerjee and Lavie, 2005); BERTScore and MoverScore use contextual embeddings and similarity (Zhang et al., 2020; Zhao et al., 2019); COMET and BLEURT rely on learned scoring functions (Rei et al., 2020; Sellam et al., 2020); LLM-as-judge scores capture subjective helpfulness (Liang et al., 2023; Liu et al., 2023); toxicity classifiers estimate safety risk. Leaderboards often combine or rank by such metrics as if they were commensurate, despite evidence that they correlate differently with human judgment across tasks (Callison-Burch et al., 2006).

2.2 Metric gaming (Goodhart's Law)

When a metric becomes the optimization target, systems learn to exploit measurement artifacts rather than improve the underlying capability (Goodhart, 1975). Documented examples include BLEU-optimized translation producing fluent but unfaithful output (Callison-Burch et al., 2006), reward-model overfitting in RLHF (Ouyang et al., 2022), and benchmark contamination reducing the validity of leaderboard rankings. No single metric is immune once it is targeted directly.

2.3 Context blindness

Benchmark scores ignore deployment context. Two models with similar benchmark performance may differ sharply in suitability for enterprise chatbots,

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1015–1020

safety-critical systems, or low-latency applications. Evaluation initiatives such as GEM stress the need for multi-dimensional and setting-aware assessment (Gehrmann et al., 2021), yet practice still defaults to scalar comparisons.

3 A Four-Primitive Metric Typology

The paper proposes that each metric can be expressed as a measurement operator built from four primitives:

$$M_i(x, y; C) = A_i\big(D_i(\phi_i(x, C), \phi_i(y, C)); C\big) \quad (1)$$

where x is model output, y is reference/evidence/competitor, and C is context (task, environment, annotators). The primitives are:

- **Representation** ϕ: how outputs are encoded (e.g., tokens, embeddings, distributions).
- **Comparison** D: similarity or distance/divergence between representations (e.g., n-gram match, cosine, entailment, learned scoring). The notation D follows convention for distance or divergence. Proper scoring rules (Brier, 1950; Murphy, 1973; Bröcker, 2009) are a principled family of choices for D when outputs are probabilistic.
- **Aggregation** A: how signals are combined (e.g., mean, geometric mean, F1, ranking, expectation).
- **Context** C: task, environment, or evaluator; modifies any stage (e.g., agent trajectories, human annotator populations).

Context C behaves differently from ϕ, D, and A: the latter are *stages* (representation $\rightarrow$ comparison $\rightarrow$ aggregation), while C *conditions* the pipeline (task, query distribution, annotator population). This matters for agentic and human-in-the-loop evaluation, where comparison or aggregation depends on environment or population.

Table 1 gives five examples; columns ϕ, D, and A show how the three measurement stages differ, and C makes context explicit. The following subsection situates broader metric families in this scheme.

3.1 Metric families as special cases

Reference-based NLG. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) use tokens/n-grams and overlap; BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019) use embeddings and cosine or optimal transport; COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020) use learned scoring; MAUVE (Pillutla et al., 2021) compares distributions via divergence frontiers. All fit Eq. 1 with different ϕ, D, A and target related but distinct constructs (Deng et al., 2021).

Retrieval and RAG. nDCG (Järvelin and Kekäläinen, 2002), MRR use ranked lists and relevance; RAG adds faithfulness, groundedness, answer correctness (Pradeep et al., 2024) via ϕ (claims/evidence), D (entailment, overlap), A (per-query mean, worst-case).

LLM-as-judge and preferences. The judge implements D; aggregation is mean score, win rate, or ranking. Preference/RLHF uses $P(i \succ j) = \sigma(q_i - q_j)$ (Bradley and Terry, 1952; Ouyang et al., 2022); latent q is the construct, win-rate or Elo the estimator.

Factuality and hallucination. FactCC (Kryściński et al., 2020), FactScore (Min et al., 2023), NLI-based checks estimate "claim supported by evidence"; they differ in ϕ, D, and A (e.g., sentence vs. claim-level).

Calibration metrics. Expected calibration error (ECE) and reliability diagrams measure alignment between predicted confidence and empirical accuracy (Brier, 1950; Murphy, 1973; Bröcker, 2009). Here ϕ maps outputs to predicted probability distributions, D compares predicted confidence to observed frequency (e.g., by binning or kernel density), A integrates the gap over the confidence range, and C specifies the task distribution and binning scheme. Calibration metrics thus fit Eq. 1 and measure a construct—confidence reliability—that is orthogonal to, and not captured by, any NLG accuracy metric.

Agentic and time-horizon evaluation. For agentic tasks such as tool use, multi-step reasoning, and long-horizon planning, ϕ maps trajectories or action sequences to a structured representation (e.g., state–action pairs, subgoal completions), D compares trajectory outcomes to target states or success criteria, and A aggregates across steps via discounted sum, worst-case, or task-completion rate; C encodes environment dynamics, available tools, and time horizon.

Thus each family is a coherent set of parameterizations of the same structural form, targeting a construct (or a narrow slice of it) with differ-

Metric	ϕ	D	A	C
BLEU	tokens	n-gram prec.	geom. mean	ref, task
BERTScore	embeddings	cosine sim.	F1	ref, task
COMET	encoder	learned score	mean	source, ref
nDCG	ranked docs	relevance	log discount	k, query set
Win-rate	outputs	preference	probability	judge, competitors

Table 1: Selected metrics as choices of the four primitives (representation ϕ, comparison D, aggregation A, context C).

ent measurement models. Unification does not make metrics equivalent—it makes their assumptions comparable.

4 From Metrics to Decisions

Metrics are *estimators* of latent constructs (e.g., semantic quality, factuality, safety), not the objective itself (Xiao et al., 2023; Bean et al., 2025; Wallach et al., 2025). The measurement model is:

$$M_i \sim f_i(\theta_k) + \epsilon_i \tag{2}$$

with θ_k a latent construct and ϵ_i measurement error (Eq. 2). Construct estimates $\hat{\theta} = g(M_1, \ldots, M_n)$ combine metric families and reduce reliance on any single channel (Zhang et al., 2020; Rei et al., 2020). The combiner g is not prescribed (e.g., weighted average, Bayesian posterior, or IRT-style (Polo et al., 2024; Zhou et al., 2026)); the framework only requires that multiple metrics inform one construct so that no single channel is optimized in isolation. Deployment context then defines a utility vector $\mathbf{U}_C(\boldsymbol{\theta}) \in \mathbb{R}^K$ (one component per tracked construct) and hard constraints (e.g., safety $\geq \tau$). Model selection becomes:

$$\mathcal{M}^* = \text{Pareto}\big\{ \mathbf{U}_C\big(\hat{\boldsymbol{\theta}}(m)\big) : m \in \mathcal{M},\ \text{constraints satisfied} \big\} \tag{3}$$

The Pareto frontier $\mathcal{M}^*$ has dimension determined by the co-domain of $\mathbf{U}_C$—that is, by K, the number of constructs being jointly optimized—not by the number of hard constraints. When $K = 1$ and $\mathbf{U}_C$ is scalar, Eq. 3 collapses to a single ranking (the leaderboard special case); multi-construct evaluation produces a frontier from which deployment context selects one operating point.

4.1 Worked examples

Translation. Consider selecting between two MT systems, A and B, where A scores higher on BLEU and B scores higher on COMET. Under a single-metric leaderboard, the choice depends arbitrarily on which metric is reported. Under the typology, both metrics are estimators of the same latent construct θ_{sem} (semantic fidelity) but with different ϕ and D: BLEU uses token n-grams and precision overlap; COMET uses encoder representations and a learned scoring function. The construct estimate $\hat{\theta}_{\text{sem}} = g(M_{\text{BLEU}}, M_{\text{COMET}}, M_{\text{BERT}})$ combines their signals, reducing dependence on any single measurement channel (Xiao et al., 2023). Deployment context C then supplies the utility: a low-latency production API may weight BLEU-derived speed proxies more heavily than a human-evaluation study would.

Summarization with safety constraints. A Summarization system must jointly satisfy two constructs: $\theta_1 =$ *factual consistency* (estimated by FactScore (Min et al., 2023) and NLI-based overlap) and $\theta_2 =$ *abstractive quality* (estimated by ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020)). A third construct $\theta_3 =$ *safety* enters as a hard constraint ($\hat{\theta}_3 \geq \tau$). Under Eq. 3 with $K = 2$ free constructs, model selection traces a Pareto frontier in $\mathbb{R}^2$; the frontier has dimension 2 because $\mathbf{U}_C = (U_1, U_2)$ is two-dimensional—not because of the safety constraint, which merely prunes the feasible set. A news-agency deployment weighting factual consistency heavily selects a different operating point on that frontier than a creative-writing assistant would.

5 Implications

Scalar leaderboards are the special case of one construct and one aggregation; the framework makes that choice explicit. Optimizing inferred constructs via metric families and context constraints reduces (but does not eliminate) Goodhart-style overfitting. The context primitive C accommodates agentic and human-in-the-loop evaluation, where comparison or aggregation depends on environment or population.

Within the typology, LLM-as-judge methods im-

plement the comparison operator D, with aggregation A producing mean scores, rankings, or win-rates; the judge is therefore part of the measurement model rather than a standalone metric.

5.1 Practical guidance

The typology translates into four concrete steps for practitioners.

Step 1: Name the constructs before selecting metrics. Explicitly list the latent properties being measured (e.g., semantic fidelity, factual grounding, safety, calibration). This prevents construct mismatch and makes Goodhart-style optimization harder, since no single metric maps cleanly onto multiple constructs. Crucially, the choice of which constructs to include is itself a normative decision—selecting "safety" or "helpfulness" as a construct encodes value judgments about what an AI system should do; translating such principles into measurable operationalizations is the central challenge of ethical AI practice (Catapang, 2026).

Step 2: Choose metric families, not individual metrics. For each construct, select a family of metrics with diverse ϕ and D (e.g., token-overlap, embedding-based, and learned-scoring metrics for semantic fidelity). Estimate the construct via a combiner g—a weighted average suffices for transparency; IRT-style estimation (Polo et al., 2024; Zhou et al., 2026) increases robustness to item difficulty. Avoid reducing construct estimation to a single channel.

Step 3: Specify $\mathbf{U}_C$ and hard constraints for the deployment context. Separate the measurement phase (Steps 1–2) from the decision phase. Document which constructs enter the utility function, how they are weighted, and which constraints (e.g., safety thresholds, latency budgets) prune the feasible set. This makes the Pareto trade-off explicit.

Step 4: Report context C alongside scores. Evaluation results are not interpretable without context: task distribution, annotator population, prompt template, and evaluation model version. Specifying C makes results reproducible and enables meaningful cross-study comparison.

6 Discussion

What the typology provides. This paper does not add another metric to an overcrowded space; it provides a framework that makes every metric's assumptions explicit and comparable. Eq. 1 and Table 1 show that NLG (Papineni et al., 2002; Zhang et al., 2020; Rei et al., 2020), retrieval (Järvelin and Kekäläinen, 2002), preference/LLM-as-judge (Bradley and Terry, 1952; Ouyang et al., 2022; Liang et al., 2023), factuality (Kryściński et al., 2020; Min et al., 2023), calibration (Bröcker, 2009), and agentic metrics all instantiate the same pipeline with different ϕ, D, A, C. Metrics that appeared incommensurable are now comparable at the level of their design choices. This is the contribution: not unification into a single score, but a shared vocabulary that turns implicit metric assumptions into explicit, debatable choices.

Measurement vs. decision. Metrics are noisy estimators of constructs (Eq. 2); model selection is context-dependent optimization subject to constraints (Eq. 3). Leaderboards are the special case $K = 1$ with a fixed scalar utility; when $K > 1$ constructs are tracked, the choice set is a Pareto frontier in $\mathbb{R}^K$, with dimensionality equal to K—the dimension of the co-domain of $\mathbf{U}_C$—not the number of hard constraints. The framework reframes "which metric is best" as "what to measure, how many constructs to track, and how to weight them for this context," making implied choices visible.

Bridging traditions and practice. The view connects NLP metric design (Papineni et al., 2002; Zhang et al., 2020; Rei et al., 2020), benchmark and meta-evaluation (Gehrmann et al., 2021; Liang et al., 2023), and latent-construct inference (Zhou et al., 2026; Bean et al., 2025; Xiao et al., 2023): it gives a common typology, treats metrics as estimators, and ties measurement to context-dependent utility. It supports GEM-style meta-assessment (Gehrmann et al., 2021) without replacing current practice. In practice it enables comparing metrics by primitives, composing families for construct estimation, stating constraints explicitly, and mitigating Goodhart-style overfitting. It does not resolve value disputes or supply a universal score; it makes evaluation structure visible for debate, documentation, and ethical audit—a prerequisite for translating ethical AI principles into accountable practice (Catapang, 2026).

7 Conclusion

Evaluation is measurement of latent system properties followed by context-dependent decision. The (ϕ, D, A, C) typology (Eq. 1) and the measurement–decision split (Eqs. 2–3) expose

the implicit choices inside every metric and every leaderboard. The position taken here is unambiguous: evaluation practice must state these choices openly—construct definitions, metric families, combiner functions, utility weights, and deployment context—so that trade-offs can be debated and reproduced rather than obscured in a single number.

Limitations

Constructs such as helpfulness or safety are normative and context-dependent; they are treated here as population-conditioned latent variables, but the prior question of which constructs to measure, and how to operationalize them, is a philosophical and ethical one that falls outside the typology (Catapang, 2026). The framework does not eliminate Goodhart's Law but provides a structure for detecting and reducing measurement bias.

Acknowledgements

This work was supported by research funding from Money Forward Inc. The author thanks colleagues at Money Forward Inc. and Tokyo University of Foreign Studies for discussions that shaped this paper.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation of Machine Translation and Summarization*, pages 65–72.

Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, and 1 others. 2025. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*.

Ralph A. Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Jochen Bröcker. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL*, pages 249–256.

Jasper Kyle Catapang. 2026. Building the ethical AI framework of the future: from philosophy to practice. *AI and Ethics*, 6(1):150.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605.

Kawin Ethayarajh and Dan Jurafsky. 2022. Utility is in the eye of the user: A critique of NLP leaderboards. *Transactions of the Association for Computational Linguistics*.

Sebastian Gehrmann, Tosin Adewumi, and Karmanya Aggarwal et al. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 96–120.

Charles Goodhart. 1975. Problems of monetary management: The UK experience. In *Papers in Monetary Economics*, volume 1, pages 1–20. Reserve Bank of Australia. Often cited as the origin of "Goodhart's Law.".

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, and Yian Zhang et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu, and Caiming Xiong. 2023. G-eval: Nlg evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore:

Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Allan H. Murphy. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. *arXiv preprint arXiv:2402.14992*.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnarök: A reusable RAG framework and baselines for the TREC 2024 retrieval-augmented generation track. *arXiv preprint arXiv:2406.16828*.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, and 1 others. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations, ICLR*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 563–578.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, and 1 others. 2026. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652(8108):58–67.