Position: What Are We Measuring? Rethinking Evaluation in Natural Language Generation

Wajdi Zaghouani
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

The field of natural language generation has accumulated a rich ecosystem of automatic evaluation metrics, yet it lacks a coherent theory of what those metrics are actually measuring. Drawing on measurement theory from the quantitative social sciences, this paper argues that current NLG evaluation practices suffer from a fundamental construct validity problem: metrics are treated as proxies for output quality without explicit specification of the underlying constructs they are meant to operationalize. We examine four dominant evaluation paradigms (reference-based metrics, embedding-based metrics, LLM-as-judge, and human evaluation) and demonstrate that each conflates construct definition with operationalization. Building on a long psychometric tradition reaching back to Cronbach and Meehl (1955) and on recent NLP work that has begun to apply this tradition to bias measurement, dialogue evaluation, and benchmark design, we propose that the field adopt a measurement modeling perspective for NLG evaluation. We borrow the concepts of construct validity, reliability, and consequential validity as a foundation for more principled evaluation, and we outline a preliminary taxonomy of NLG quality constructs as a starting point for this work.

1 Introduction

In practice, the NLG community has relied on a variety of operationalizations to assess output quality: n-gram overlap metrics, embedding similarity measures, model-based evaluators, and human ratings. Over the past two decades it has produced an impressive array of automatic evaluation metrics, from BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to BERTScore (Zhang et al., 2020) and a growing family of LLM-based evaluators (Zheng et al., 2023; Gao et al., 2025). Each new metric is introduced with correlation studies showing reasonable alignment with human judgment on some benchmark, then propagated across tasks and domains for which it was never validated.

What is missing is not better metrics. It is a theory of measurement.

In psychometrics, any measurement instrument begins with an explicit construct definition: a theoretically grounded specification of the latent property being measured. Researchers then ask whether the operationalization captures that construct, whether it is reliable across conditions, and whether its use produces valid consequences (Cronbach and Meehl, 1955). These are preconditions for any claim that a number means something. NLG evaluation has largely skipped this step. The field operates with an implicit and underspecified notion of "quality" that different metrics operationalize incompatibly, without acknowledgment of the theoretical commitments each operationalization carries.

The argument we develop here is not new in spirit. Cronbach and Meehl (1955) laid out the foundations of construct validity seventy years ago, and the resulting tradition has been refined in psychometrics ever since. Within NLP, several recent contributions have begun to draw on that tradition. Jacobs and Wallach (2021) brought measurement modeling to the algorithmic fairness community. Van der Wal et al. (2024) applied construct validity and reliability to bias measurement. Braggaar et al. (2026) systematically reviewed measures and constructs in dialogue evaluation. Schlangen (2021) questioned the methodological grounding of benchmark-driven NLP. Zhuang et al. (2025) argued that AI evaluation should learn directly from how psychometricians test humans. Belz et al. (2025) proposed a standardised taxonomy of evaluation quality criteria, building on earlier survey work by Howcroft et al. (2020). What has not yet been done, and what this paper attempts, is to bring that body of work to bear specifically on the four dominant NLG evaluation paradigms, and to argue

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1021–1028

for a unified measurement-modeling stance for the field.

This paper argues that NLG evaluation should adopt a **measurement-modeling perspective**, treating evaluation metrics as operationalizations of explicitly defined quality constructs. Building on Braggaar et al. (2026), Van der Wal et al. (2024), and the psychometric tradition originating with Cronbach and Meehl (1955), we apply this perspective directly to NLG and to the four evaluation paradigms that dominate it. Progress requires three shifts: (1) explicit construct definitions for properties such as fluency, adequacy, and faithfulness; (2) validation through established measurement concepts such as content, convergent, and discriminant validity; and (3) transparency about the theoretical commitments implicit in evaluation choices. Our contribution is not a new metric but a conceptual reframing: the persistent difficulties of NLG evaluation stem from the absence of a shared theory of measurement, and adopting measurement-modeling principles can provide that foundation.

2 The Metric Proliferation Problem

Without explicit measurement theory, evaluation metrics function as ad hoc proxies whose meaning shifts across tasks and communities, undermining comparability and incentivizing optimization for metrics that may not reflect genuine quality improvements.

The NLG community has long recognized its evaluation problems. Reiter (2018) reviewed 284 correlations across 34 papers and concluded that BLEU should not be used as primary evidence for scientific claims in NLP. Novikova et al. (2017) showed that common metrics only weakly reflect human judgments for end-to-end NLG, with performance varying across data conditions. Bowman and Dahl (2021) argued that NLU benchmarking is broken, with unreliable systems scoring highly. Gehrmann et al. (2021) noted that models continue to be evaluated with flawed metrics on divergent corpora. Holistic efforts such as HELM (Liang et al., 2023) address fragmentation through multi-metric evaluation but still operate without explicit construct definitions. Two notable exceptions illustrate that explicit specification is feasible. The Multidimensional Quality Metrics framework (Lommel et al., 2014) decomposes translation quality into a hierarchical error typology with defined severity levels. More recently, the QCET taxonomy (Belz et al., 2025) derives 114 standardised quality criterion names and definitions from a combined survey of 933 NLP evaluation experiments. A parallel effort in the intelligent virtual agents community has produced the ASA questionnaire and its 19 unifying constructs through several years of community work (Fitrianie et al., 2020), suggesting that cross-disciplinary standardisation of evaluation constructs is achievable when sustained collaborative effort is applied.

Earlier voices anticipated the same point. Paris et al. (2006) argued that the NLG community would be better served by enlarging its view of evaluation to include common dimensions and metrics rather than only common corpora and tasks, with end-user evaluation as a distinct concern. The diagnosis offered there has aged well, but the underlying problem of underspecified constructs has persisted.

These critiques share an empirical orientation: they document that metrics fail and benchmarks saturate. What they do not provide is a theoretical account of why this pattern persists. Without a theory of what metrics are supposed to measure, there is no principled basis for deciding whether any new proposal represents genuine progress or merely a different set of failure modes.

3 What Measurement Theory Offers

Measurement theory offers a principled framework for the relationship between abstract constructs and their operationalizations. The core insight is that measurement is always indirect: we observe indicators from which we infer unobservable properties. This inference is legitimate only when the construct is explicitly defined, the operationalization is justified relative to that definition, and the measurements are empirically validated.

The modern theory of construct validity originates with Cronbach and Meehl (1955), who introduced the concept to handle psychological tests for which no obvious external criterion exists. They proposed that validation requires building a nomological network: an explicit web of theoretical relationships in which the construct is embedded, against which observed measurement behavior can be checked. Subsequent work in the social sciences has elaborated this idea into a mature methodology that distinguishes content, criterion, convergent, discriminant, and consequential evidence of validity. Jacobs and Wallach (2021) ported this frame-

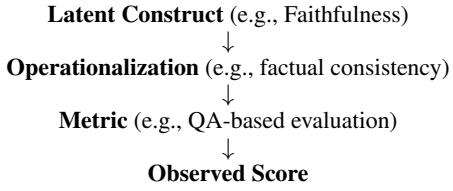

Figure 1: The measurement modeling chain. Each arrow represents a theoretical commitment that must be explicitly justified and empirically validated. NLG evaluation typically conflates all four levels.

work to computational systems, arguing that many AI harms arise from construct-operationalization mismatches. Van der Wal et al. (2024) carried it further into NLP by applying construct validity and reliability to bias measurement, showing how psychometric tools can be used to interrogate whether a given bias measure actually captures the underlying construct it claims to. Braggaar et al. (2026) performed a systematic review of 122 task-oriented dialogue evaluation studies and showed that the same construct (for example, satisfaction or quality) is operationalized in incompatible ways across papers, with limited attention to validity or reliability.

Figure 1 shows the measurement modeling chain that NLG evaluation should follow. Construct validity has four components relevant here: *content validity* (the operationalization covers the full construct scope); *convergent validity* (different operationalizations of the same construct agree); *discriminant validity* (the operationalization distinguishes the target construct from related but distinct ones); and *consequential validity* (the measurement drives research in appropriate directions). Current NLG metrics fail on all four, yet this failure is rarely framed in these terms within NLG itself.

4 Four Evaluation Paradigms and Their Construct Problems

4.1 Reference-Based Metrics

Reference-based metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) operationalize output quality as lexical similarity to human-authored references. The implicit construct is surface adequacy: a good output resembles what a human would write. This carries unacknowledged commitments: that references cover the full space of acceptable outputs, that n-gram overlap proxies communicative success, and that the relationship between surface form and quality is stable across tasks and languages.

None of these assumptions have been theoretically justified. Papineni et al. (2002) were explicit that BLEU was designed for corpus-level machine translation evaluation, not as a general theory of text quality. The subsequent migration of BLEU and ROUGE to summarization, dialogue, and story generation happened through convenience rather than construct validation. Reiter (2018) documented this scope creep empirically, finding that BLEU's validity outside machine translation is not supported by the literature. Schlangen (2021) situates this pattern within a broader benchmark culture in which task validation, the argument that a dataset truly exemplifies a target task, is rarely made explicit; the same diagnosis applies almost word for word to metric validation.

4.2 Embedding-Based Metrics

BERTScore (Zhang et al., 2020) replaces exact n-gram matching with contextual similarity, capturing paraphrase and semantic equivalence that reference-based metrics miss. However, it does not resolve the construct validity problem; it relocates it. The implicit construct becomes semantic similarity operationalized through a pre-trained model's embedding space, inheriting that model's theoretical commitments, training distribution, and handling of ambiguity.

Such correlations are not meaningless: they can serve as partial convergent validity evidence when human ratings were collected against a clearly defined construct. The problem arises when both the metric and human ratings operationalize the same underspecified notion of "quality," in which case high correlation confirms only that two imprecise instruments agree. A concrete illustration: BERTScore applied to faithfulness evaluation in summarization rewards outputs semantically close to the reference, even when the reference contains claims not supported by the source. This is a discriminant validity failure: the metric conflates faithfulness with reference similarity, constructs that are related but theoretically distinct.

4.3 LLM-as-Judge

LLM-as-judge evaluation introduces a more acute construct validity problem. Zheng et al. (2023) documented position bias, verbosity bias, and self-enhancement bias in MT-Bench and Chatbot Arena evaluations. Gao et al. (2025) survey the broader landscape of LLM-based NLG evaluation and iden-

tify a set of recurring challenges including lack of robustness across attack scenarios, sensitivity to prompt formulation, and a tendency for evaluator models to prefer outputs that resemble their own generations. These are symptoms of a deeper failure: the evaluator's judgments reflect learned associations rather than a principled assessment of a defined construct. Recent mitigations such as multi-judge ensembles and rubric anchoring (Zheng et al., 2023) can reduce some biases, but address symptoms rather than the root cause: without an explicit construct definition, no rubric can be properly grounded. The evaluation construct is implicitly defined by whatever the evaluator learned to associate with quality during training, remaining opaque, unstable across model versions, and potentially circular when evaluator and evaluated system share the same training distribution (Gehrmann et al., 2021).

4.4 Human Evaluation

Well-designed human evaluation protocols can provide meaningful evidence about constructs requiring pragmatic judgment, such as utility. However, the field's typical practice falls short. Howcroft et al. (2020) surveyed twenty years of NLG human evaluation across 165 papers and found that researchers have used over 200 different terms for evaluated aspects of quality, with the same term meaning different things across papers. They reported, for example, that what is labelled *fluency* actually decomposes into a substantial number of distinct underlying criteria depending on how the term is defined and used in context, so that two evaluations both labelled fluency can target meaningfully different properties. Their conclusion was that NLG human evaluation needs evaluation sheets and standardised definitions before meaningful meta-analysis becomes possible. The QCET work of Belz et al. (2025) can be read as an attempt to deliver exactly such standardised definitions, derived empirically from the existing literature rather than imposed top-down.

The same diagnosis recurs in adjacent communities. Braggaar et al. (2026) found that within task-oriented dialogue evaluation, the four most frequent constructs (satisfaction, correctness, quality, and efficiency) are operationalized through inconsistent and often unvalidated measures. Fitrianie et al. (2020) reached comparable conclusions for intelligent virtual agents, motivating their multi-year, community-wide effort to consolidate constructs into a shared instrument. Across all of these settings, the same pattern holds: without explicit construct definitions, annotators interpret criteria however they wish, producing ratings that mix the intended construct with irrelevant factors. Novikova et al. (2017) showed that human ratings are unstable across tasks and annotator populations, undermining the assumption that correlation with human judgment is by itself adequate metric validation.

5 A Measurement-Theoretic Framework

Evaluation metrics should be interpreted as **operational indicators of latent quality constructs**. Quality is not a single construct; different methods capture different aspects, and their validity depends on the relationship between the operationalization and the construct it claims to measure. We propose three steps, each grounded in established measurement theory.

Step 1: Explicit construct definitions. The field needs explicit definitions for the properties NLG evaluation assesses, distinguishing constructs such as fluency, adequacy, faithfulness, coherence, utility, and style, and acknowledging that these are task-specific. For abstractive summarization, a minimally adequate framework specifies at least: faithfulness (factual consistency with the source), coverage (proportion of key information retained), and fluency (grammatical well-formedness). ROUGE-L approximates coverage under limited conditions; BERTScore blends faithfulness and coverage without separating them; neither addresses fluency in a theoretically grounded way. Stating these mappings explicitly reveals both what is measured and what is not. Practical templates already exist: the QCET taxonomy (Belz et al., 2025) and the human-evaluation sheets proposed by Howcroft et al. (2020) can both serve as anchors for new evaluations rather than each paper inventing its own vocabulary.

Step 2: Validation through measurement modeling. Metric validation should address the questions of Cronbach and Meehl (1955) and Jacobs and Wallach (2021): does the operationalization cover the content of the construct? Does it converge with other operationalizations of the same construct? Does it discriminate from related but distinct constructs? Does it produce valid consequences? Van der Wal et al. (2024) provide a recent worked example of porting these questions into NLP, including practical strategies for assessing

parallel-form, test-retest, and split-half reliability of measurement instruments built from text. Several of those strategies transfer directly to NLG metrics: the same metric should produce comparable scores under prompt paraphrasing, on equivalent test partitions, and across model checkpoints fine-tuned for the same task. For human evaluation, this requires reporting inter-annotator agreement as evidence of reliability: low agreement signals construct underspecification rather than mere annotator disagreement. Correlation with human judgment is one piece of convergent validity evidence, not a sufficient criterion on its own.

Step 3: Transparency about theoretical commitments. When BLEU evaluates a summarization system, this is a claim that surface lexical similarity is an adequate operationalization of summarization quality. When LLM-as-judge is used, this is a claim that the evaluator's implicit preferences constitute an adequate operationalization. These claims should be stated and scrutinized. Maintaining shared construct definitions requires governance mechanisms such as construct registries and metric cards documenting what each metric targets and under what conditions it has been validated. The standardisation efforts of Belz et al. (2025), the evaluation sheets of Howcroft et al. (2020), and the IVA community's questionnaire-based instrument (Fitrianie et al., 2020) all illustrate how such governance can take concrete form.

Preliminary construct taxonomy. Table 1 illustrates how evaluation can be organized around theoretically defined properties rather than metrics. Disagreements between metrics signal that they capture different constructs or different facets of the same one. The six constructs address traditional NLG tasks; agentic and long-context settings will require additional constructs such as groundedness, instruction-following, and calibration. We do not claim novelty for these construct names; each appears in some form in the QCET taxonomy and in earlier NLG evaluation guides. Our claim is methodological: that any benchmark or paper using such labels should commit to an explicit definition, identify which operationalizations partially cover it, and report validity evidence accordingly.

The proposed constructs are not mutually exclusive and may interact; for example, faithfulness and coverage jointly determine adequacy in summarization tasks. Table 2 illustrates how each construct maps to typical metrics, making explicit that each metric captures only part of the construct space.

Construct	Definition
Fluency	Degree to which output conforms to grammatical and stylistic norms
Adequacy	Degree to which output conveys the meaning of the source or prompt
Faithfulness	Degree to which output is factually consistent with a source document
Coherence	Degree to which output forms a logically unified whole
Utility	Degree to which output is useful to the intended end user
Style	Degree to which output conforms to a target register and style

Table 1: A preliminary taxonomy of NLG quality constructs. Each requires separate operationalization and validation; a single metric as proxy for all conflates conceptually distinct properties.

Construct	Typical Metric(s)
Fluency	LM perplexity, LLM judge
Adequacy	BLEU, BERTScore
Faithfulness	QA-based metrics
Coherence	Entity-grid, discourse metrics
Utility	Human evaluation, impact study
Style	Style classifiers, human judgment

Table 2: Construct-to-metric mapping. Each metric captures only part of the construct space; using any single metric as a proxy for overall quality produces construct conflation.

Implications for benchmark design. A construct-first approach reverses current practice: designers specify which constructs the benchmark assesses, then select metrics that validly operationalize them. This determines what a score means and what it does not. A leaderboard ranking conflating fluency, faithfulness, and utility cannot guide practitioners who need faithfulness but not stylistic sophistication. Construct-grounded benchmarks produce infrastructure that is more interpretable and more useful for deployment. Schlangen (2021) makes a parallel argument from a slightly different angle, treating dataset and benchmark design as a chain of explicit argumentation steps that should be made transparent rather than assumed; that proposal is fully compatible with the construct-first stance taken here.

Implications for leaderboard culture. When a benchmark score operationalizes an underspecified notion of quality, improving that score becomes a valid target regardless of whether improvements reflect meaningful properties of generated text. Systems that score highly often fail in ways obvious to human readers but invisible to the metric (Bowman and Dahl, 2021). From a measurement perspective, leaderboard optimization is consequential validity failure: the community is shaping what NLG systems learn without theoretical guidance about whether that behavior is desirable. Zhuang et al. (2025) argue that AI evaluation should take inspiration from how psychometricians test humans, including adaptive test designs that estimate latent traits rather than treating average accuracy as a gold standard. Whatever one thinks of the specific proposals in that line of work, the general point that benchmarks need a clearer theoretical account of what they are estimating reinforces the argument made here.

Implications for reproducibility and real-world impact. Construct underspecification undermines reproducibility. When two papers evaluate "quality" but operationalize it differently, their results are not comparable even when the same metric label is used. The ReproNLP initiative led by Belz (2022) and colleagues has documented how difficult NLG evaluations are to reproduce in practice; measurement theory explains why: reproducibility requires an instrument measuring a stable, well-defined construct. Construct-grounded protocols with explicit reliability statistics would substantially improve this situation. A separate but connected concern is that intrinsic metric quality is not the only thing worth measuring. Reiter (2025) reports that only a tiny fraction of ACL Anthology papers contain any evaluation of the real-world impact of the systems they propose, and argues that the field would be more useful if it took such evaluations seriously. From a measurement-modeling perspective, real-world impact corresponds to consequential validity, and including impact studies among the operationalizations the community values would broaden the definition of what successful evaluation looks like.

6 Conclusion

The NLG evaluation landscape is rich with metrics and poor in theory. All four paradigms examined suffer from a structural construct validity problem that no new metric can solve. What is needed is a measurement modeling framework that specifies what is being measured, justifies the operationalizations used, and validates their use through established measurement science. The position advocated here is not original to this paper; it inherits a long tradition starting with Cronbach and Meehl (1955) and continuing through recent NLP work on bias measurement (Van der Wal et al., 2024), dialogue evaluation (Braggaar et al., 2026), benchmark methodology (Schlangen, 2021), evaluation taxonomies (Belz et al., 2025; Howcroft et al., 2020), and adaptive testing (Zhuang et al., 2025). The contribution of this paper is to bring those threads together for NLG specifically, and to argue that the four dominant evaluation paradigms must be re-examined through a measurement lens. Future work should pursue multi-trait multi-method validation, construct-grounded benchmark design, and stronger integration between intrinsic evaluation and impact studies. The next step is not more metrics. It is more theory.

Limitations

This paper is a conceptual position paper rather than an empirical study. The proposed construct taxonomy in Table 1 is illustrative rather than exhaustive, and the framework is intended as a starting point for discussion rather than a complete measurement model. The six constructs are not presented as a definitive set; reasonable disagreement exists about how they should be defined, decomposed, or prioritized across NLG tasks, and richer alternatives such as the QCET taxonomy of Belz et al. (2025) already exist for use as anchors. Future work should develop more precise construct definitions and explore empirical validation methods such as factor analysis across metric scores, multi-trait multi-method studies, and reproducibility audits aligned with efforts like the ReproNLP track. The paper focuses on text generation and does not address the additional complexities of multimodal evaluation, though the core argument about construct validity applies equally in those settings. The paper also draws selectively rather than exhaustively from the broader psychometric literature; a deeper integration with classical test theory, item response theory, and modern validity frameworks remains an open task.

Ethics Statement

Evaluation practices influence which systems are considered successful and therefore shape the direction of research and deployment in natural language generation. Poorly defined evaluation constructs can incentivize optimization for metrics that do not correspond to meaningful improvements in communication quality or user experience, potentially driving the field toward technically impressive but practically hollow progress. By advocating for explicit construct definitions and transparent measurement practices, and by encouraging the field to take real-world impact studies more seriously (Reiter, 2025), this work aims to support more responsible and scientifically grounded evaluation methodologies. No datasets, human subjects, or potentially harmful content are involved in this work.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Anya Belz. 2022. A metrological perspective on reproducibility in NLP. *Computational Linguistics*, 48(4):1125–1135. https://doi.org/10.1162/coli_a_00448

Anya Belz, Simon Mille, and Craig Thomson. 2025. Standard quality criteria derived from current NLP evaluations for guiding evaluation design and grounding comparability and AI compliance assessments. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26685–26715, Vienna, Austria. Association for Computational Linguistics. https://aclanthology.org/2025.findings-acl.1370/

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.385/

Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. 2026. Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations. *Northern European Journal of Language Technology*, 12(1):1–38. https://doi.org/10.3384/nejlt.2000-1533.2026.5940

Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302. https://doi.org/10.1037/h0040957

Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 unifying questionnaire constructs of artificial social agents: An IVA community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*, Article 21, pages 1–8. ACM. https://doi.org/10.1145/3383652.3423873

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51(2):661–687. https://doi.org/10.1162/coli_a_00561

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, Joao Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics. https://aclanthology.org/2021.gem-1.10/

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2020.inlg-1.23/

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Trans-*

parency (FAccT '21), pages 375–385. ACM. https://dl.acm.org/doi/10.1145/3442188.3445901

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=iO4LZibEqW

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. https://aclanthology.org/W04-1013/

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:455–463. https://doi.org/10.5565/rev/tradumatica.77

Jekaterina Novikova, Ondrej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics. https://aclanthology.org/D17-1238/

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://aclanthology.org/P02-1040/

Cécile Paris, Nathalie Colineau, and Ross Wilkinson. 2006. Evaluations of NLG systems: Common corpus and tasks or common dimensions and metrics? In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 127–129, Sydney, Australia. Association for Computational Linguistics. https://aclanthology.org/W06-1419/

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401. https://direct.mit.edu/coli/article/44/3/393/1598/

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431. https://doi.org/10.1162/COLI.a.18

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics. https://aclanthology.org/2021.acl-short.85/

Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2024. Undesirable biases in NLP: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40. https://doi.org/10.1613/jair.1.15195

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 2020)*. https://openreview.net/forum?id=SkeHuCVFDr

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*. https://arxiv.org/abs/2306.05685

Yan Zhuang, Qi Liu, Zachary A. Pardos, Patrick C. Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. 2025. Position: AI evaluation should learn from how we test humans. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*. https://arxiv.org/abs/2306.10512