International Research Journal of Multidisciplinary Sciences ISSN:AWAITED
VOL-1 ISSUE-7 July 2025 PP:42-48

Beyond Human Evaluation: Towards Robust, Automatic, and Multi-Dimensional Metrics for Generative AI

[1]*Bhawna Kaushik*, [2]*Priya Gupta*

1.bhawna.kaushik@niu.edu.in, Noida International University

2.priya.gupta@gmail.com, Noida International University

Abstract: The breakneck pace of advancement in Generative AI (GenAI) has created a critical evaluation bottleneck. While human assessment remains the gold standard, it is prohibitively expensive, slow, non-scalable, and suffers from low inter-annotator agreement. This paper argues for a fundamental shift beyond this paradigm, proposing a framework for developing a new generation of robust, automatic, and multi-dimensional metrics. We first present a critical analysis of the shortcomings of both human evaluations and current automated metrics (e.g., n-gram overlap, FID). We then introduce a novel taxonomy that decomposes GenAI evaluation into key dimensions: factuality, coherence, creativity, reasoning, safety, and alignment . For each dimension, we survey existing metric proposals and highlight their limitations. Our core contribution is the proposal of GenEval , a benchmark suite designed to stress-test metrics across these dimensions. We further present a experimental study using a learned, model-based metric fine-tuned on human judgments for factual consistency in summarization. Results show this metric achieves a significantly higher correlation with human judges (Spearman's $\rho = +0.68$) compared to strong baselines. We conclude that the future of GenAI evaluation lies in a combination of: (1) sophisticated, multi-dimensional benchmarks, (2) learned metrics that emulate nuanced human judgment, and (3) a clear understanding of the specific capability dimension being measured.

Keywords: Generative AI, Evaluation Metrics, Human Evaluation, Benchmarking, Natural Language Generation, Factuality, AI Safety.

1. Introduction

The emergence of powerful Generative AI (GenAI) models for text [1], image [2], audio [3], and video [4] has catalyzed a revolution across industries. However, the ability to generate content has starkly outpaced our capacity to evaluate it effectively. The relentless scaling of model parameters and training data has rendered traditional evaluation methods obsolete, creating a significant bottleneck for responsible development, deployment, and research progress

Human evaluation, wherein human annotators rate outputs based on quality, fluency, or factuality, is widely considered the gold standard. Yet, this approach is fundamentally flawed for the age of large-scale GenAI. It is prohibitively expensive, incredibly slow, and impossible to scale to the volume of outputs modern systems can produce. Furthermore, it suffers from high variance due to subjective annotator guidelines, cultural biases, and low

inter-annotator agreement [5]. Relying solely on human evaluation is akin to trying to quality-check every car from a high-speed production line by hand.

In response, researchers have long relied on automated metrics. In Natural Language Generation (NLG), metrics like BLEU [6] and ROUGE [7], based on lexical n-gram overlap, have been used for decades. For images, metrics like Fréchet Inception Distance (FID) [8] compare distributions of features between generated and real images. While fast and scalable, these metrics are notoriously poor proxies for quality. They fail to capture semantic meaning, coherence, and creativity, and can be easily gamed by models that learn to optimize for the metric itself rather than genuine quality [9].

This paper argues that the path forward requires a paradigm shift. We must move beyond both flawed automated heuristics and non-scalable human evaluation towards a new framework based on robust, automatic, and multi-dimensional metrics . This entails:

1. Decomposing the monolithic concept of "quality" into distinct, measurable dimensions.

2. Developing specialized metrics for each dimension, often leveraging advanced model-based methods.

3. Validating these metrics against high-quality human judgments within comprehensive benchmarks.

The contributions of this paper are threefold:

1. A critical review of current evaluation methods and a novel taxonomy for GenAI evaluation.

2. A proposal for GenEval , a benchmark designed to stress-test metrics across key dimensions.

3. An experimental study demonstrating the superiority of a learned metric for evaluating factual consistency, achieving a +0.28 improvement in Spearman correlation over strong baselines.

2. Background and Related Work

2.1 The Human Evaluation Paradigm

Human evaluation typically employs methods like Likert scales (rating output on a 1-5 scale for a quality), pairwise comparisons (asking annotators to choose the better of two outputs), or Best-Worst Scaling [10]. While invaluable, this approach is fraught with challenges. Studies show that results can vary dramatically based on annotator demographics, instruction wording, and the specific criteria being evaluated [5]. The cost and time required make it impractical for rapid iteration, creating a critical barrier to agile AI development.

2.2 Traditional Automated Metrics

Text Metrics: Metrics like BLEU [6] (for translation) and ROUGE [7] (for summarization) rely on n-gram overlap between a generated output and one or more reference outputs. They measure surface-level similarity but are agnostic to meaning. A sentence can be factually incorrect and incoherent yet achieve a high ROUGE score if it contains the right keywords.

Image Metrics: The Inception Score (IS) [11] measures the clarity and diversity of generated images. The Fréchet Inception Distance (FID) [8] computes the distance between feature distributions of real and generated images. While more correlated with human judgment than IS, FID provides a single aggregate score that obscures specific failures in areas like compositionality or prompt alignment.

2.3 Emergence of Model-Based Metrics

The rise of pre-trained language models (LMs) has enabled a new class of metrics that leverage their embedded world knowledge and semantic understanding.

Reference-Based Metrics: BERTScore [12] computes similarity using contextual embeddings from models like BERT. BLEURT [13] takes this further by pre-training a model on synthetic data to learn robust representations for evaluation.

Reference-Free Metrics: These are crucial for open-ended generation where references don't exist. They often use Natural Language Inference (NLI) models to check if a claim is entailed by a source document [14], or question-answering models to verify facts [15].

Learning-Based Metrics: Inspired by Reinforcement Learning from Human Feedback (RLHF) [1], these approaches train a "reward model" to predict human preferences. The training data consists of human-ranked outputs, and the resulting model can then score new outputs automatically.

3. A Multi-Dimensional Taxonomy for GenAI Evaluation

Evaluating GenAI requires moving beyond a single score. We propose a taxonomy that decomposes evaluation into core dimensions, applicable across modalities but with modality-specific nuances.

Table 1: Evaluation Taxonomy for Text and Image Generation

Dimension	Definition (Text)	Example Metrics (Text)	Definition (Image)	Example Metrics (Image)
Factuality/ Faithfulness	Is the output supported by a source context/real world?	NLI models, QAEval, FactScore [15]	N/A (See Alignment)	N/A
Coherence	Is the output structurally sound and logically consistent?	Self-BERTScore, coherence-specific classifiers	Are the objects and scene elements logically consistent?	Model-based classifiers
Creativity/ Novelty	Is the output original and interesting?	Divergence from training data, human ratings	Is the image novel and aesthetically pleasing?	Artist surveys, uniqueness scores
Reasoning	Does the output demonstrate logical deduction or problem-solving?	Accuracy on GSM8K [16], strategy extraction	(Less applicable)	(Less applicable)

| Safety & Bias | Is the output toxic, biased, or harmful? | Toxicity classifiers, stereotype detection | Does the image depict harmful or biased content? | NSFW detectors, bias classifiers |

| Alignment | Does it follow instructions? Is it helpful? | Reward models, instruction-following tests | Does the image match the text prompt? | CLIPScore [17], T2I-CompBench [18] |

4. The GenEval Benchmark Proposal

To catalyze the development of robust metrics, we propose GenEval , a comprehensive benchmark suite. Its design philosophy is to provide a fixed, curated set of challenges for evaluating metrics, not models.

4.1 Components:

A "Challenge Set": A collection of prompts and inputs meticulously designed to probe specific failure modes. This includes prompts designed to induce factual hallucinations, logical contradictions, toxic responses, and difficult compositional requests (e.g., "a red cube on top of a blue sphere").

High-Quality Human Annotations: A curated set of ~10,000 model outputs from various state-of-the-art models. Each output is annotated by multiple expert annotators along the dimensions of our taxonomy. This dataset serves as the "ground truth" for validation.

Automated Metric Zoo: A standardized software framework that allows researchers to easily run their new automated metrics on the benchmark and compute correlation scores against the human ground truth.

4.2 Benchmark Tasks:

The benchmark includes tasks for multiple modalities:

Text: Long-form summarization (factuality, coherence), creative writing (creativity), instruction following (alignment), and dialogue (safety).

Image: Text-to-image generation focused on complex compositionality, attribute binding, and stylistic alignment.

5. Experiment: A Learned Metric for Factual Consistency

To demonstrate the promise of learned metrics, we focus on the critical task of evaluating factual consistency in summarization.

5.1 Task & Baselines: The task is to score the factual consistency of a generated summary given its source document. We compare our proposed metric against strong baselines:

ROUGE-L: A traditional n-gram overlap metric.

BERTScore-F1: A model-based semantic similarity metric.

NLI-Based: A zero-shot metric using a pre-trained DeBERTa model for Natural Language Inference, classifying the summary as "entailed" by, "contradicting," or "neutral" to the source.

5.2 Proposed Metric: FactScore-Eval

We fine-tune a pre-trained DeBERTa-v3-large model [19] for regression. The input is the sequence: `[CLS] Source: [source text] [SEP] Summary: [summary text] [SEP]`. The model is trained to predict the factual consistency score (a continuous value between 0 and 1) assigned by human annotators. We use the FRANK dataset [14], which contains human annotations for factual errors in model-generated summaries, for training and validation.

5.3 Results & Analysis

We evaluate all metrics by calculating their Spearman rank correlation with human scores on a held-out test set from FRANK.

Table 2: Spearman Correlation (ρ) with Human Judgment

Metric	Spearman's ρ
ROUGE-L	0.22
BERTScore-F1	0.31
NLI-Based (Zero-shot)	0.40
FactScore-Eval (Ours)	0.68

Our learned metric, FactScore-Eval , significantly outperforms all baselines, achieving a correlation of 0.68. This represents a +0.28 improvement over the strong NLI-based zero-shot baseline. Qualitative analysis shows that FactScore-Eval is better at identifying subtle factual errors (e.g., "increased by 10%" vs. "increased by 15%") that are missed by other methods.

6. Discussion

Our results demonstrate the clear advantage of learned metrics for specific evaluation dimensions. However, this approach is not a panacea.

The Data Trade-Off: The superior performance of FactScore-Eval requires a high-quality, human-annotated dataset for training. Creating such data is expensive and must be repeated for each new task or domain.

Generalizability: A metric trained on news summarization data may not generalize well to evaluating factual consistency in scientific or legal documents. This necessitates a portfolio of specialized metrics.

Benchmarking and Gaming: As with any benchmark, there is a risk of overfitting. GenEval would need to be a "living benchmark" with hidden test sets and regular updates to prevent metrics from gaming its specific tasks.

The Evolving Role of Humans: This framework does not obsolete human judgment; it redefines its role. Humans are needed to create the foundational training data for these learned metrics and to validate their performance on new frontiers where no metrics yet exist.

7. Conclusion and Future Work

The evaluation crisis in Generative AI is a fundamental impediment to progress. This paper makes the case that overcoming it requires moving beyond both unreliable automated shortcuts and non-scalable human evaluation. We propose a future built on a foundation of robust, multi-dimensional, and often learned, automatic metrics.

The path forward involves a concerted effort on three fronts:

1. Community Adoption of Fine-Grained Taxonomies: Researchers must clearly specify which dimension of quality they are measuring.

2. Investment in High-Quality Data: The creation of large-scale, carefully annotated datasets for training and validating metrics is paramount.

3. Development of Comprehensive Benchmarks: Efforts like GenEval are needed to provide standardized, rigorous testing grounds for new metrics.

For future work, we plan to expand the GenEval benchmark to include more modalities and tasks, explore few-shot and unsupervised methods for metric learning to reduce data dependence, and investigate meta-evaluation frameworks for assessing the evaluators themselves.

8. References

[1] OpenAI. (2023). GPT-4 Technical Report. ArXiv, abs/2303.08774 .

[2] Rombach, R., et al. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. CVPR .

[3] Copet, J., et al. (2023). Simple and Controllable Music Generation. NeurIPS .

[4] OpenAI. (2024). Video Generation Models as World Simulators.

[5] Clark, E., et al. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. ACL .

[6] Papineni, K., et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. ACL .

[7] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. ACL .

[8] Heusel, M., et al. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. NeurIPS .

[9] Paulus, R., et al. (2018). A Deep Reinforced Model for Abstractive Summarization. ICLR .

[10] Louviere, J. J., et al. (2015). Best-worst scaling: Theory, methods and applications. Cambridge University Press .

[11] Salimans, T., et al. (2016). Improved Techniques for Training GANs. NeurIPS .

[12] Zhang, T., et al. (2019). BERTScore: Evaluating Text Generation with BERT. ICLR .

[13] Sellam, T., et al. (2020). BLEURT: Learning Robust Metrics for Text Generation. ACL .

[14] Pagnoni, A., et al. (2021). Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. NAACL .

[15] Min, S., et al. (2023). FactScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP .

[16] Cobbe, K., et al. (2021). Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 .

[17] Hessel, J., et al. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. EMNLP .

[18] Huang, K., et al. (2023). T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. NeurIPS .

[19] He, P., et al. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. ICLR .