SCORING RULES! STATISTICAL AND STRATEGIC ALIGNMENT FOR TEXT EVALUATION METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-based text evaluation metrics, which are widely used to assess natural language generation systems, score a candidate response by comparing it with a reference response. The reliability of an evaluation metric is usually judged by the statistical correlation with human ratings. However, as these metrics are increasingly used as optimization objectives, this alignment by correlation is no longer sufficient: agents may strategize to game the evaluation metric. We study this issue through two complementary notions of alignment. A metric is *statistically aligned* if it correlates with human ratings, and *strategically aligned* if it resists strategic perturbations that do not add task-relevant information. We make two contributions. First, we propose test principles for reference-based metrics, consisting of human-rating correlation, degradation sensitivity, and manipulation robustness. These principles evaluate whether a metric agrees with human judgments, penalizes low-effort information loss, and resists strategic score inflation. Second, we develop a unified design framework for mutual-information-based metrics that are designed to resist strategic perturbations. The framework decomposes existing and new metrics into four design choices: information measure, estimation method, text representation, and prediction mechanism. Across peer review, summarization, and question answering tasks, we find that strong human-rating correlation does not imply strategic alignment: LLM-as-a-Judge achieves high correlation but is susceptible to manipulations. In contrast, mutual-information-based metrics substantially improve manipulation robustness. Our framework also uncovers a new metric that achieves the strongest overall robustness in our experiments while remaining competitive on human-rating correlation.

1 INTRODUCTION

Reference-based evaluation metrics are essential for scalable assessment of natural language generation (NLG) systems. Given a candidate response and one or more reference responses for the same task, such metrics assign a score intended to reflect the quality, informativeness, or correctness of the candidate. Human judgement of text quality remains the gold standard, but it is too slow and expensive at large scale. Automatic evaluation metrics therefore serve as scalable proxies for human judgment: they replace repeated human assessment with a computable comparison between the candidate and reference responses. Examples include lexical-overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), embedding-based metrics such as BERTScore (Zhang et al., 2019), and LLM judges that compare a candidate response against a reference. Throughout, an *evaluation metric* is a function that assigns a numerical score to a candidate using one or more references. We use *evaluation workflow* for the datasets, test procedures, and statistical analyses used to assess metrics. Most standard reference-based metrics are designed to approximate human judgments. We call this requirement **statistical alignment**: a useful metric should correlate with human assessments of text quality.

However, reference-based metrics create incentives for language generation systems, which may not align with the intended target of improving genuine text quality. As their role expands from measurement to optimization, we need to understand the incentives these scores create. Prior work has documented related failure modes under the names of reward hacking (Amodei et al., 2016), specification gaming (Krakovna et al., 2020), or reward overoptimization (Gao et al., 2023), where

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

optimizing a proxy objective can produce behavior that increases the measured score without improving the intended target.

We call the complementary requirement **strategic alignment**: a metric should be robust to strategic perturbations of the text. In particular, a metric should assign lower expected scores to *less informative* reports, which omit or degrade task-relevant content, and to *untruthful* reports, which manipulate the evaluation procedure without genuinely improving the answer. Recent work proposes evaluation metrics robust to strategic manipulations (Lu et al., 2024; Xu et al., 2025; Robertson & Koyejo, 2025; Wu & Hartline, 2024). These works hint that evaluation metrics must be judged not only by the statistical correlation with human ratings, but also by whether they are robust to strategic perturbations of the text.

Contribution 1: Test Principles Existing metrics for manipulation robustness are often developed and evaluated in isolation, using different datasets, test procedures, and success criteria. As a result, direct comparison is difficult. It remains unclear how evaluation metrics should be systematically validated when the goal is not only statistical alignment with human ratings, but also strategic alignment under score-seeking behavior.

In this paper, we address this validation gap by introducing unified test principles for reference-based evaluation metrics. Generalized from previous studies (Lu et al., 2024; Xu et al., 2025; Robertson & Koyejo, 2025; Wu & Hartline, 2024), we propose three complementary principles for comparing metrics. The first principle is *correlation* with human preference: a metric should agree with human judgments under ordinary, non-adversarial conditions. This principle corresponds to statistical alignment and captures the standard desideratum in NLG evaluation.

The second and third principles evaluate strategic alignment by comparing scores on an original candidate response and a modified version. The second principle is *degradation sensitivity*: when a text is deliberately degraded by removing or corrupting task-relevant information, the metric score should strictly decrease. This tests whether the metric discourages *less informative* reports, which save effort by providing strictly less information while attempting to receive similar credit.

The third principle is *manipulation robustness*: when a text is strategically modified to inflate the metric score without adding genuine content, the metric should not increase. This tests whether the metric discourages *untruthful* reports, which exploit superficial features of the evaluation procedure. Unlike degradation, manipulation need not strictly reduce the information content of the text. It may only weakly decrease it or preserve much of the original content while changing style, format, or wording to obtain a higher score. Together, these three principles provide a common standard for evaluating both prior and new metrics on equal footing.

We implement these principles in a reusable evaluation workflow across seven datasets spanning peer review, summarization, and question answering. The workflow measures correlation with human ratings on four datasets and tests degradation sensitivity and manipulation robustness on all seven. These tests examine how metric scores respond to information loss, verbose rewrites (Zheng et al., 2023; Singhal et al., 2024), changes in style, positive and negative opinion shifts, hedging, and added confidence (Zhou et al., 2024; Leng et al., 2025).

Contribution 2: Metric Design Framework Having introduced a unified test framework, we now turn to the complementary construction question: how should one design metrics that pass these tests? At a high level, *mutual information* (MI) measures how much information two variables share; we give the formal definition and discuss the data-processing inequality in Appendix A. MI is a natural target because it is known to be information-monotone: transformations that do not add task-relevant information do not increase the information a candidate shares with an independent reference. MI-based metrics use this idea by scoring a candidate through an estimate of shared information with the reference. Existing methods such as GEM (Xu et al., 2025), GPPM (Lu et al., 2024), and TVD-MI (Robertson & Koyejo, 2025) can be read as different approximations to this common target. We defer the technical distinctions among them to the design framework; the point here is that they vary along a shared set of design choices.

However, there is no unified framework for constructing strategically aligned metrics. Previous papers are developed as separate methods rather than as instances of a common design principle: e.g., GEM (Xu et al., 2025) uses the token representation of the text, GPPM-Judgment (Lu et al., 2024)

parses the text as a set of statements, while TVD-MI (Robertson & Koyejo, 2025) sends the whole piece of text into an LLM-oracle to estimate mutual information. Thus, they interpret the response as represented by different sized chunks—token, statement, response, respectively. Additionally, they differ in the choice of the MI measure, estimator, and prediction model. Because prior evaluations typically compare the full methods rather than controlled variations of these individual components, it is difficult to determine which design choices are responsible for the observed behavior.

We propose a unified design framework for constructing MI-based evaluation metrics. Our framework decomposes every MI-based metric into four components: an information measure (MI), an MI estimation method, a representation of a text, and a prediction mechanism. This decomposition recovers existing methods such as GEM, GPPM, and TVD-MI as special cases, while also exposing a broader design space of new metrics. The framework therefore gives a principled recipe for constructing new metrics and for analyzing which design components lead to desirable properties.

Our empirical results support a central takeaway: statistical alignment alone is not enough to certify an evaluation metric as strategically aligned. Across peer review, summarization, and question answering, we find that LLM-as-a-Judge achieves the strongest correlation with human ratings, yet is vulnerable to manipulations. In contrast, MI-based metrics provide stronger strategic alignment. In particular, the best-performing new metric from our design framework computes MI between candidate and reference texts at the statement level. It has fewer degradation-test failures than GEM and full-report TVD-MI, and no significant score increases are detected across the tested manipulations. Therefore, our framework is constructive, not only descriptive: it uncovers a new MI metric that outperforms prior MI-based metrics in our evaluation workflow.

Together, our two contributions provide a unified perspective on aligned text evaluation beyond statistical alignment with human judgements. The test principles explain how evaluation metrics should be validated, while the design framework explains how MI-based metrics can be constructed and varied. Combining the two allows us to compare existing methods fairly, explore new points in the design space, and identify which combinations of information measure, estimator, and distribution model lead to the most reliable evaluation behavior across domains.

2 PRELIMINARIES

Let $\mathcal{V}$ be a vocabulary, and let $\mathcal{X} = \mathcal{Y} = \mathcal{V}^*$ denote the spaces of candidate and reference texts. Let $\mathcal{W}$ denote the task space. Throughout, uppercase letters denote random variables and lowercase letters denote their realizations. Thus, W is a task random variable with task distribution P_W, while $w \in \mathcal{W}$ denotes a realized task.

Conditional on the realized task $W = w$, the candidate random variable X, supported on the candidate-text space $\mathcal{X}$, and the reference random variable Y, supported on the reference-text space $\mathcal{Y}$, are drawn according to $X \sim P_{X|W=w}$, $Y \sim P_{Y|W=w}$. We assume that the candidate X and reference Y are conditionally independent given the task W. The distributions P_X, P_Y, and P_{XY} denote the candidate marginal distribution, reference marginal distribution, and joint candidate–reference distribution induced by this generative process.

A *reference-based evaluation metric* is a function $S : \mathcal{X} \times \mathcal{Y} \times \big(\mathcal{Y} \cup \{\emptyset\}\big) \longrightarrow \mathbb{R}$.

For a realized candidate text $x \in \mathcal{X}$, matched reference text $y \in \mathcal{Y}$, and optional negative reference $y^- \in \mathcal{Y} \cup \{\emptyset\}$, the scalar $s = S(x, y, y^-)$ is the resulting metric score. We write $S(x, y)$ as shorthand for the noncontrastive score $S(x, y, \emptyset)$.

For contrastive metrics, the negative-reference random variable $Y^- \sim P_Y$ is drawn from the marginal reference distribution P_Y, independently of the task, candidate, and matched-reference random variables (W, X, Y). Thus, the realized negative reference y^- represents a reference from an independently sampled task. For noncontrastive metrics, we set $y^- = \emptyset$. The negative reference allows S to reward agreement with the matched reference y relative to agreement with the unrelated reference y^-. In particular, contrastive scoring discounts textual patterns that are common under the marginal reference distribution P_Y but are not specific to the realized task w.

The metric-evaluation problem is to assess whether S satisfies statistical alignment and the two parts of strategic alignment. Statistical alignment asks whether scores on unperturbed candidate responses correlate with human quality judgments. Strategic alignment asks whether scores decrease under

degradations that remove task-relevant information and do not increase under manipulations that seek higher scores without adding such information. The rest of the paper develops tests for these properties and a design framework for constructing metrics that target them.

Our design framework follows *information monotonicity*: transformations that add no task-relevant information should not increase expected scores. Mutual information is a natural choice because it measures candidate–reference dependence and cannot increase under post-processing of the candidate. Appendix A gives the formal definition and data-processing inequality.

3 Test Principles

We compare evaluation metrics along statistical alignment and strategic alignment, under our test principles, generalized from previous work (Lu et al., 2024; Xu et al., 2025; Robertson & Koyejo, 2025; Wu & Hartline, 2024). Specifically, for statistical alignment, we test the statistical correlation with human quality ratings; for strategic alignment, we use degradation strategies to simulate low-effort reporting and manipulation strategies to simulate untruthful reporting. We test whether the metrics can penalize degradations and resist manipulations.

Statistical alignment test: Correlation with human ratings. For datasets providing absolute human scores, let h_i denote the human rating on candidate response x_i, and s_i denote the corresponding score assigned by an evaluation metric. We measure statistical alignment using Spearman's rank correlation, $\rho = \mathrm{Spearman}\big(\{s_i\}_{i=1}^n, \{h_i\}_{i=1}^n\big)$. When multiple annotators are available, we average their scores at the item level before computing ρ.

Strategic alignment test 1: Sensitivity to degradation. A *degradation* is a perturbation that reduces the task-relevant information, coverage, or specificity of a candidate by deleting evidence, restricting context, or forcing low-effort responses; a well-behaved metric should assign strictly lower scores to degraded variants.

For each candidate report x_i, we apply a degradation strategy M to produce a perturbed report x_i^M, and compute scores s_i and s_i^M for the original and perturbed report respectively. Table 5 in Appendix C lists six strategies used in our experiments in Section 5.

For each strategy M, we use the paired t-test to test whether the mean score change is significantly negative: $\bar{\Delta}^{(M)} := \frac{1}{n}\sum_{i=1}^n (s_i^M - s_i) < 0$.

To quantify the magnitude of the sensitivity to degradation across rules that use different scales, we also report standardized mean difference (SMD) (Andrade, 2020) with the 95% confidence interval (which shares the same form of Cohen's d): $d_M = \frac{\mu_M - \mu}{\sqrt{(\sigma_M^2 + \sigma^2)/2}}$, where (μ, σ) and (μ_M, σ_M) are the empirical mean and standard deviation of the original and perturbed scores respectively.

Strategic alignment test 2: Manipulation Resistance A *manipulation* is a perturbation designed to inflate the score by exploiting metric shortcuts, such as verbosity, stylistic regularity, or generic content likely under the reference distribution, without introducing new task-specific evidence; a robust metric should not assign higher scores to such variants.

As with the degradation tests, for each candidate report x_i, we apply a manipulation strategy M to produce a perturbed report x_i^M. Table 6 in Appendix C lists six strategies used in our experiments in Section 5. We compute the paired mean score change $\bar{\Delta}^{(M)}$ and standardized mean difference d_M using the same formulas established for degradations. For manipulations, we test whether the mean score change is significantly positive. A metric passes a manipulation test when the score does not significantly increase after the manipulation, i.e., when we do not find evidence that $\bar{\Delta}^{(M)} > 0$.

4 A Design Framework for MI-Based Metrics

The test principles of Section 3 specify what an aligned metric should do; this section addresses how to construct one. We introduce a framework for MI-Based Metrics that separates the design space into four components, divided into two conceptual parts. The first part consists of *theoretical design*

choices: which information measure is being estimated, and which estimator is used. The second part consists of *empirical design choices*: how reports are represented, and how predictive quantities are obtained from those representations. Appendix D provides the formal component definitions and manipulation-robustness analysis.

Definition 4.1 (Design tuple). *An f-MI evaluation metric is specified by a tuple $\mathcal{D} \coloneqq (F, E, R, \Pi)$,*

(i) ***Information measure*** *F: the choice of f-divergence defining the objective $I_f(X;Y)$.*
(ii) ***Estimation method*** *E: the method used to estimate the chosen objective $I_f(X;Y)$.*
(iii) ***Representation*** *R: a map from the text space to a representation space, $R : \mathcal{V}^* \to \mathcal{U}$.*
(iv) ***Prediction*** *Π: a mechanism that maps represented candidate-reference pairs to predictive quantities used in the estimation method.*

How the components compose. The four components form a scoring pipeline with two dependencies. On the theoretical side, the information-measure component F fixes the population objective, and the estimation-method component E prescribes which predictive quantity suffices to estimate it: direct estimation requires the density-ratio function $r(x, y)$, whereas variational estimation requires only the critic function $T^*(x, y)$ (Section D.2). On the empirical side, the representation component R fixes the units of text passed to the prediction mechanism Π, and the prediction mechanism Π supplies the quantity that the estimation method E requires. Concretely, to score a realized item (x_i, y_i, y_i^-), the metric (i) applies the representation map R to the texts, (ii) queries the prediction mechanism Π on the represented pairs to obtain the predictive quantities, and (iii) plugs those quantities into the estimator E to produce the sample-level score $S_{F,E,R,\Pi}(x_i, y_i, y_i^-)$.

Sample-level scores. Given a dataset of n tasks $\{w_i\}_{i=1}^n$, each task w_i has an associated candidate report x_i and a positive reference report y_i. Designs that score contrastively additionally use a negative reference y_i^- sampled from the marginal distribution P_Y, whose agreement with the candidate is penalized; designs that do not use one take $y_i^- = \emptyset$. The sample-level score is obtained from the chosen estimator by replacing the population expectations with the sampled references: the expectation over the joint distribution P_{XY} is evaluated at (x_i, y_i), and the expectation over the product distribution $P_X \otimes P_Y$ at (x_i, y_i^-). Averaging the per-item scores across the dataset then yields an estimate of the population objective determined by the tuple:

$$\frac{1}{n}\sum_{i=1}^{n} S_{F,E,R,\Pi}(x_i, y_i, y_i^-) \;\approx\; I_f(X;Y).$$

Two running examples. Prior MI-based metrics correspond to specific choices of the tuple. To make the pipeline concrete, we trace the components and scoring pipelines of two prior metrics that sit at opposite corners of the design space, and we return to them throughout the section.

Example 4.2 (GEM: (KL, Direct, Token, Autoregression)). *GEM (Xu et al., 2025) instantiates the four components as follows. $F = \mathrm{KL}$ selects KL mutual information as the population objective. $E =$ Direct forms the plug-in pointwise mutual information*

$$S(x_i, y_i) = \log p_\phi(y_i \mid x_i) - \log p_\phi(y_i).$$

$R =$ Token represents the candidate report x_i and reference report y_i as token sequences. $\Pi =$ Autoregression uses a language model with parameters ϕ to compute the conditional likelihood $p_\phi(y_i \mid x_i)$ token by token with the candidate in the prompt, and the marginal likelihood $p_\phi(y_i)$ with the candidate removed.

Because direct KL estimation takes an expectation only under the joint candidate–reference distribution, no negative reference is needed; the negative-reference input is $y_i^- = \emptyset$.

Example 4.3 (TVD-MI: (TV, f-variational, Full-report, LLM-oracle)). *TVD-MI (Robertson & Koyejo, 2025) makes a different choice on every axis. $F = \mathrm{TV}$ selects total-variation mutual information as the population objective. $E = f$-variational contrasts the two judgments to produce the sample-level score*

$$S(x_i, y_i, y_i^-) = T(x_i, y_i) - T(x_i, y_i^-),$$

$R =$ Full-report keeps each report whole. $\Pi =$ LLM-oracle prompts an LLM, e.g., "Do these two reports describe the same item?", once on the matched candidate–reference pair (x_i, y_i) and once on the mismatched pair (x_i, y_i^-); its yes/no answer implements a critic function T. For total variation, this critic only needs to decide whether the density-ratio value $r(x, y)$ exceeds 1.

This metric rewards candidates that the oracle can link to their own task's reference but not to an unrelated one. Because the variational form takes expectations under both the joint and the product distributions, the negative reference is essential here.

The contrast between the two examples previews the tradeoffs studied in this section: GEM demands the strongest model access (token-representation logits), whereas TVD-MI needs only a yes/no API call but relies on the reliability of the oracle's judgment.

4.1 Prior Work

Combining the preceding components, a complete design tuple (F, E, R, Π) specifies a concrete reference-based metric. Table 1 shows how three existing methods instantiate this framework.

Table 1: Coverage of the design space. Each cell is labeled with the name of the corresponding evaluation metric, when one exists. NEW denotes combinations introduced or evaluated in this work;

		Π = **Autoregression**	Π = **LLM-oracle**	
F	E	R = Token	R = Full Report	R = Statement
KL	Direct	GEM (Xu et al., 2025)	NEW	GPPM-J (Lu et al., 2024)
	f-variational	NEW	NEW	NEW
TV	Direct	NEW	NEW	NEW
	f-variational	NEW	TVD-MI (Robertson & Koyejo, 2025)	NEW

A practical benefit of this decomposition is that it makes unexplored combinations easy to identify. Cells of the tuple (F, E, R, Π) that Table 1 with NEW correspond to metrics that are implementable in principle but have not been studied. In Section 5, we empirically study several such metrics, as shown in Table 1.

The practical benefit of the framework is that it exposes new metrics. For example, TVD-MI (Robertson & Koyejo, 2025) uses the tuple $(\mathrm{TV},\ f\text{-variational},\ \text{Full-report},\ \text{LLM-oracle})$. Changing the representation gives the statement-level TV f-variational metric (formally defined in Appendix E):

$$(\mathrm{TV},\ f\text{-variational},\ \text{Statement},\ \text{LLM-oracle}).$$

This metric decomposes the positive and negative references into atomic statements, asks the LLM critic whether each statement is supported by the candidate, and compares support for the true reference statement against support for the negative reference statement. As a result, candidate content that supports statements in the matched reference is rewarded, while support for statements drawn from an unrelated task is penalized. Generic content is discounted as it tends to support both positive and negative references. Thanks to the finer-grained representation, in our experiment in Section 5, this metric achieves better strategic and statistical alignment than TVD-MI across multiple datasets.

5 Experiment

We compare evaluation metrics in a controlled way under our test principles (Section 3), with detailed setup deferred to Section F.

Research Question. To keep the analysis focused, we frame the experiments around the decision a practitioner actually faces. The predictor Π and representation R are typically dictated by task and deployment constraints, e.g., whether logit access is available, and whether atomic statements can be reliably extracted, while the information measure F and estimator E are under the researcher's direct control. We therefore ask:

Given a fixed predictor and representation pair (Π, R), which information measure and estimator (F, E) yields the best-behaved evaluation metric?

Other reference-based baselines. We also compare against popular reference-based metrics as baselines, including ROUGE-L, BLEU, BERTScore (based on roberta-large), and LLM-as-a-Judge

Table 2: Human-rating correlation results. Rows are evaluation metrics, grouped by the representation–predictor pair (R, Π); within each group, rows sweep the information measure $F \in \{\mathrm{KL}, \mathrm{TV}\}$ and estimator $E \in \{\text{Direct}, f\text{-var.}\}$. Columns are datasets; entries are Spearman's ρ with human ratings. For SummEval we report the mean over all dimensions. Best result within each group is **bolded**; best among f-MI metrics per column and best overall per column are underlined.

Evaluation Metric	**Peer review**		**Summarization**	**QA**
(F, E)	PG-WH	PG-XLSK	SummEval	MedAESQA
Autoregression, Token-representation				
$(\mathrm{KL}, \text{Direct})$ (GEM)	**0.460**	**0.475**	**0.138**	0.113
$(\mathrm{KL}, f\text{-var.})$	-0.218	-0.355	0.059	0.051
$(\mathrm{TV}, \text{Direct})$	0.312	0.380	0.095	-0.025
$(\mathrm{TV}, f\text{-var.})$	-0.053	-0.036	0.085	**0.129**
LLM-oracle, Report-representation				
$(\mathrm{KL}, \text{Direct})$	0.235	0.367	0.031	**0.163**
$(\mathrm{KL}, f\text{-var.})$	0.252	**0.419**	0.031	0.162
$(\mathrm{TV}, \text{Direct})$	-0.231	-0.350	0.017	-0.238
$(\mathrm{TV}, f\text{-var.})$ (TVD-MI)	**0.286**	0.042	**0.140**	0.110
LLM-oracle, Statement-representation				
$(\mathrm{KL}, \text{Direct})$ (GPPM-J)	0.254	0.307	0.073	0.041
$(\mathrm{KL}, f\text{-var.})$	**0.281**	0.251	0.075	0.064
$(\mathrm{TV}, \text{Direct})$	-0.292	-0.396	-0.043	-0.129
$(\mathrm{TV}, f\text{-var.})$	0.201	**0.468**	**0.208**	**0.228**
Non-MI baselines				
ROUGE-L	-0.244	0.167	0.111	0.151
BLEU	-0.332	0.256	0.102	0.127
BERTScore	-0.133	-0.063	0.252	0.166
LLM-as-Judge / Claude-Haiku-4.5	0.462	0.569	0.400	**0.315**
LLM-as-Judge / Claude-Sonnet-4.5	0.512	0.622	0.450	0.301
LLM-as-Judge / GPT-5-mini	**0.539**	**0.631**	**0.477**	0.256
LLM-as-Judge / GPT-4o-mini	0.492	0.528	0.452	0.208

(based on various models). These baselines test whether accuracy and manipulation-resistance gains from the MI framework go beyond lexical overlap, embedding similarity, or direct LLM judging.

Datasets. We evaluate metrics across three domains that stress different aspects of reference-based text evaluation: peer review, summarization, and question answering. In peer review, the task input is a paper or essay and the candidate reports are full reviews; in summarization, the task input is a source document and the reports are summaries; in question answering, the task input is a question and the reports are answers. Across these domains, reports differ substantially in length, structure, subjectivity, and semantic complexity.

Our evaluation includes seven datasets: three peer-review datasets, two summarization datasets, and two question-answering datasets. Four datasets provide item-level human quality annotations and are used to measure human-rating correlation; all datasets are used for degradation and manipulation tests. When multiple independent reports are available for the same task, we treat them as positive references. Table 9 in Appendix F.1 summarizes the datasets.

5.1 Results

Table 2 reports Spearman correlation with human ratings on the four datasets carrying item-level numerical annotations: Peer Grading-WH (Wu & Hartline, 2024), Peer Grading-XLSK (Xu et al., 2025), SummEval (Fabbri et al., 2021), and MedAESQA (Gupta et al., 2025). Rows are grouped by the representation and predictor pair (R, Π); within each block, the four rows sweep $(F, E) \in \{\mathrm{KL}, \mathrm{TV}\} \times \{\text{Direct}, f\text{-var.}\}$. Bold marks the best method within each block; underlines mark the best f-MI metric and the best overall metric per dataset.

Table 3: Sensitivity to degradation. For each (metric, dataset) pair, we apply m degradation strategies (header) and count those on which the metric *fails* to produce a significant score decrease at $p < 0.05$; lower is better. A ✓ marks zero failures; otherwise, letter codes mark the strategies the metric missed (legend below).

Evaluation Metric	**Peer review**			**Summarization**		**QA**		**total**
(F, E)	PG-WH $m=5$	PG-XLSK $m=6$	ICLR26 $m=6$	SummEval $m=5$	SPACE $m=5$	LFQA-E $m=5$	MedAESQA $m=5$	fail $m=37$
MI-based Mechanisms								
KL-Direct-Autoreg.	✓	✓	P	✓	✓	P U	P U	5
TVD-FVar-Report	D S U	O	O P U	P U	✓	P U	D S P U	15
TVD-FVar-Statement	✓	✓	✓	P	✓	P	P	3
Non-MI baselines								
LLM-as-Judge / Claude-Haiku-4.5	✓	✓	P	P	✓	P	D P	5
LLM-as-Judge / Claude-Sonnet-4.5	✓	✓	P	P	✓	P	P	4
LLM-as-Judge / GPT-4o-mini	D	D O R	O R P	P	✓	P	P	10
LLM-as-Judge / GPT-5-mini	✓	✓	P	P	✓	P	D P	5

Codes: D=deletion and completion; O=opinion flip; R=random replacement.
Codes: S=sentence deletion; P=surface report; U=ultra concise compression.
Red: significant score increase. Orange: non-significant change.

Table 4: Manipulation resistance. For each (metric, dataset) pair, we apply m manipulation strategies (header) and count those the metric *fails* to resist, i.e., that produce a significant score increase at $p < 0.05$; lower is better. A ✓ marks zero failures; otherwise, letter codes flag the strategies the metric failed to resist (legend below).

Evaluation Metric	**Peer review**			**Summarization**		**QA**		**total**
(F, E, R, Π)	PG-WH $m=6$	PG-XLSK $m=6$	ICLR26 $m=6$	SummEval $m=2$	SPACE $m=2$	LFQA-E $m=4$	MedAESQA $m=4$	fail $m=30$
MI-based Mechanisms								
(KL, Direct, Token, Autoregression)	M	S	N H P S R	✓	✓	✓	✓	7
(TV, f-var, Report, LLM-oracle)	✓	✓	✓	✓	✓	✓	✓	**0**
(TV, f-var, Statement, LLM-oracle)	✓	✓	✓	✓	✓	✓	✓	**0**
Non-MI baselines								
LLM-as-Judge / Claude-Haiku-4.5	M N H S R	H P S R	H S R	R	R	H S R	H S R	20
LLM-as-Judge / Claude-Sonnet-4.5	M N H S R	H P S R	M H P S R	R	✓	H S	H	18
LLM-as-Judge / GPT-4o-mini	M N H P S R	M N H P S R	M N H P S R	R	R	H S R	H S R	26
LLM-as-Judge / GPT-5-mini	N H S R	M N H P S R	H P S R	✓	✓	H S R	H S	19

Codes: M=meaningless elongation; N=negative opinion shift; H=hedged opinion shift.
Codes: P=positive opinion shift; S=strong opinion shift; R=rephrase.
Red codes mark significant score increases after manipulation.

For degradation and manipulation tests, a full sweep of all configurations across every perturbation strategy and dataset is computationally prohibitive. We therefore carry forward the best f-MI configurations from each representation-prediction block in Table 2: $(\mathrm{KL}, \mathrm{Direct})$ with token-representation autoregression, corresponding to GEM; and $(\mathrm{TV}, f\text{-var.})$ with an LLM-oracle at both report and statement levels. This selection deliberately favors the correlation winners and asks whether they remain reliable under strategic perturbations. For large datasets, we down-sample to roughly 200 candidate reports per setting. We report result overviews in Tables 3 and 4, and detailed statistics in Section G.

Key takeaways. The results show a separation between statistical and strategic alignment.

LLM-as-a-Judge is strongest on human-rating correlation, but collapses under manipulation. Among the non-MI baselines, LLM-as-a-Judge achieves the strongest correlations with human ratings. However, across the four judge models, manipulations succeed in 18 to 26 of the 30 tests.

The best f-MI design depends on the representation and predictor. Under token-representation autoregression, $(\mathrm{KL}, \mathrm{Direct})$ (GEM) leads its block on three of four datasets and is the best f-MI metric overall on both peer-review datasets. Under an LLM-oracle, statement-level rep-

resentation is substantially more effective: statement-level $(\mathrm{TV}, f\text{-var.})$ outperforms full-report-representation TVD-MI on three of four datasets and is the best f-MI metric overall on SummEval and MedAESQA.

The unified design framework identifies a promising new metric. The statement-level $(\mathrm{TV}, f\text{-var.})$ metric is not present in prior work, but is exposed by our design framework. Its strong performance suggests that decomposing references into atomic claims can make the oracle prediction task easier, enough to compensate for the information loss introduced by statement extraction. The $(\mathrm{TV}, f\text{-var.})$ metric is the most reliable on degradation sensitivity and manipulation robustness. It fails only 3 of 37 degradation tests, compared with 5 failures for GEM and 15 failures for full-report-representation $(\mathrm{TV}, f\text{-var.})$ / TVD-MI. It has 0 failures out of 30 manipulation tests, meaning that no tested manipulation produces a significant score increase.

Scoring design matters even with the same underlying model. Claude-Haiku-4.5 exhibits fewer manipulation-test failures within the TV f-variational metrics than as an absolute LLM judge. This comparison highlights the role of scoring design, although it does not isolate the contribution of the MI objective from prompting, reference sampling, and aggregation.

It is worth noticing that no single axis of (F, E, R, Π) dominates in isolation; rather, the design choices interact. token-representation autoregression favors KL-direct estimation, while LLM-oracle prediction benefits from the TV f-variational formulation, especially with statement-level representations. MI-based metrics also appear stronger on peer-review tasks than on summarization and QA, where the evaluation target can be more complex or subjective. Appendix F.4 examines rephrasing and surface-report degradation in more detail.

ROUGE-L, BLEU, and BERTScore are generally weaker baselines, and each fails to obtain positive correlation on at least one dataset, especially in peer review. This suggests that lexical overlap and embedding similarity are insufficient for evaluating richer, more subjective reports.

6 Conclusion and Discussion

We study reference-based text evaluation metrics. This perspective separates statistical alignment, measured by correlation with human ratings, from strategic alignment, measured by robustness to degradations and manipulations. We introduced unified test principles for these two requirements and a design framework for MI-based metrics. Empirically, LLM-as-a-Judge achieves strong human-rating correlation but is vulnerable to manipulation, while MI-based metrics, especially the statement-level TV f-variational metric, newly derived from our design framework, offer stronger strategic robustness while remaining competitive on statistical alignment.

Future work should optimize evaluation metrics directly for statistical and strategic alignment, as suggested in Appendix B. Another direction is to reduce score variance. Our LLM-oracle metrics suffer randomness from oracle judgments, statement decomposition, and negative-reference sampling. This variance can obscure both human-rating correlation and robustness tests. Repeated oracle calls, multiple independent decompositions, larger negative-reference pools, and adaptive resampling near decision boundaries are simple ways to reduce estimation noise and make comparisons between mechanisms more reliable.

The evaluation framework itself can also be strengthened. Our current degradation and manipulation tests use fixed perturbation families. In deployment, however, models optimized against a metric may discover new score-inflating strategies. Future evaluation workflows should therefore include adaptive or competitive tests in which an adversary searches for perturbations that increase the score without adding task-relevant information. Such tests would evaluate a metric as an optimization target, not only as a static measurement tool.

Finally, reference-based alignment is limited by the reference signal. As language models exceed individual human references on some tasks, evaluation datasets may need richer reference sets, such as expert-panel references, debate- or critique-based references, or aggregated LLM outputs (Feng et al., 2026). Strategically validated evaluation metrics may also serve as training rewards with reduced susceptibility to reward hacking. For example, Feng et al. (2026) use peer-prediction-inspired self-training signals for language-model reasoning without relying on gold labels. This suggests a broader research program in which reference construction, evaluation-metric design, and training objectives are studied jointly.

AI USE STATEMENT

The manuscript was initially written by the human authors. LLMs assisted in generating experimental code, which was reviewed by the human authors. After the manuscript was completed, LLMs also assisted with rephrasing and improving the clarity of the writing. As described in the experimental setup, LLMs were used for text preprocessing, reference statement decomposition, generation of experimental perturbations, and evaluation scoring. The authors take responsibility for the final content of this paper and the experimental methodology and results.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Chittaranjan Andrade. Mean difference, standardized mean difference (smd), and their use in meta-analysis: as simple as it gets. *The Journal of clinical psychiatry*, 81(5):11349, 2020.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293, 2021.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Noah Burrell and Grant Schoenebeck. Measurement integrity in peer prediction: A peer assessment case study. *arXiv preprint arXiv:2108.05521*, 2021.

Siyu Chen, Jibang Wu, Yifan Wu, and Zhuoran Yang. Learning to incentivize information acquisition: Proper scoring rules meet principal-agent model. In *International Conference on Machine Learning*, pp. 5194–5218. PMLR, 2023.

Yiling Chen and Fang-Yi Yu. Optimal scoring rule design under partial knowledge. In *International Conference on Web and Internet Economics*, pp. 383–400. Springer, 2024.

Yiling Chen, Shi Feng, Paul Kattuman, and Fang-Yi Yu. Data reliability scoring. *arXiv preprint arXiv:2510.17085*, 2025.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 319–330, 2013.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Yuchen Fan, Chen Ling, Xin Zhong, Shuo Zhang, Heng Zhou, Yuchen Zhang, Mingyu Liang, Chengxing Xie, Ermo Hua, Zhizhou He, et al. Lfqa-e: Carefully benchmarking long-form qa evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shi Feng, Hanlin Zhang, Fan Nie, Sham Kakade, and Yiling Chen. Peer-predictive self-training for language model reasoning. *arXiv preprint arXiv:2604.13356*, 2026.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. a dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1): 1035, 2025.

Jason D Hartline, Liren Shan, Yingkai Li, and Yifan Wu. Optimal scoring rules for multi-dimensional effort. *arXiv preprint arXiv:2211.03302*, 2022.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Yuqing Kong. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the fourteenth annual acm-siam symposium on discrete algorithms*, pp. 2398–2411. SIAM, 2020.

Yuqing Kong. Dominantly truthful peer prediction mechanisms with a finite number of tasks. *Journal of the ACM*, 71(2):1–49, 2024.

Yuqing Kong and Grant Schoenebeck. Water from two rocks: Maximizing the mutual information. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 177–194, 2018.

Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)*, 7(1):1–33, 2019.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 3:40–53, 2020.

Nicolas S Lambert, David M Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 129–138, 2008.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Yingkai Li, Jason D Hartline, Liren Shan, and Yifan Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 988–989, 2022.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Yuxuan Lu, Shengwei Xu, Yichi Zhang, Yuqing Kong, and Grant Schoenebeck. Eliciting informative text evaluations with large language models. *arXiv preprint arXiv:2405.15077*, 2024.

Yuxuan Lu, Yifan Wu, Jason Hartline, and Michael J Curry. Aligned textual scoring rules. *arXiv preprint arXiv:2507.06221*, 2025.

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

Eric Neyman, Georgy Noarov, and S Matthew Weinberg. Binary scoring rules that incentivize precision. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 718–733, 2021.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vishakh Padmakumar and He He. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2505–2512, 2021.

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

Maneesha Papireddygari and Bo Waggoner. Contracts with information acquisition, via scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 703–704, 2022.

Zachary Robertson and Sanmi Koyejo. Let's measure information step-by-step: Llm-based evaluation beyond vibes. *arXiv preprint arXiv:2508.05469*, 2025.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=G8LaO1P0xv`.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Yifan Wu and Jason Hartline. Elicitationgpt: Text elicitation mechanisms via language models. *arXiv preprint arXiv:2406.09363*, 2024.

Shengwei Xu, Yichi Zhang, Paul Resnick, and Grant Schoenebeck. Spot check equivalence: an interpretable metric for information elicitation mechanisms. *arXiv preprint arXiv:2402.13567*, 2024.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. Benchmarking llms' judgments with no gold standard. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yichi Zhang, Shengwei Xu, Grant Schoenebeck, and David Pennock. Stochastically dominant peer prediction. *Advances in Neural Information Processing Systems*, 38:151632–151664, 2026.

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. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Shuran Zheng, Xuan Qi, Rui Ray Chen, Yongchan Kwon, and James Zou. Proper dataset valuation by pointwise mutual information. *arXiv preprint arXiv:2405.18253*, 2024.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3623–3643, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.198. URL https://aclanthology.org/2024.acl-long.198/.