Scoring Rules! Statistical and Strategic Alignment for Text Evaluation Metrics

Shengwei Xu[*]
University of Michigan
shengwei@umich.edu

Yuxuan Lu[*]
Peking University
yx_lu@pku.edu.cn

Yifan Wu
Microsoft Research
yifan.wu2357@gmail.com

Jason Hartline
Northwestern University
hartline@eecs.northwestern.edu

Grant Schoenebeck
University of Michigan
schoeneb@umich.edu

Abstract

Reference-based text evaluation metrics score a candidate response by comparing with a reference response, which are widely used to assess natural language generation systems. The reliability of a 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, especially in the era of large language models. 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

[*]Both authors contributed equally to this paper.

Preprint.

lexical-overlap metrics such as BLEU [39] and ROUGE [29], embedding-based metrics such as BERTScore [47], and LLM judges that compare a candidate response against a reference. 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 are used not only to compare language generation systems, but increasingly to guide development and optimization itself. As their role expands from measurement to optimization, we need to understand not only whether their scores correlate with human judgments, but also what incentives these scores create. Prior work has documented related failure modes under the names of reward hacking [1], specification gaming [25], or reward overoptimization [16], where optimizing a proxy objective can produce behavior that increases the measured score without improving the intended target. This concern motivates a complementary question to statistical alignment of an evaluation metric: does optimizing an evaluation metric encourage genuinely better, more informative responses, or does it reward superficial perturbations that increase the score without improving the answer?

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 [31, 46, 41, 44]. 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, perturbations, 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 [31, 46, 41, 44], 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 natural-language generation (NLG) evaluation.

The second and third principles evaluate strategic alignment from two complementary perspectives. 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 instantiate these principles in a reusable evaluation workflow over seven datasets spanning peer review, summarization, and question answering. Four datasets include human quality ratings for correlation tests, and all seven support degradation and manipulation tests.

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 in Section 2. MI is a natural target because it is information-monotone: transformations that do not add task-relevant information should 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 [46], GPPM [31], and TVD-MI [41] 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. They differ in the choice of the MI measure, estimator, representation granularity, 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 compute MI between candidate and reference texts at the statement level. It is more reliable than existing MI-based metrics such as GEM, GPPM, and TVD-MI. 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

We study the problem of evaluating and designing a reference-based scoring metric s for text generation. For each task, an evaluator observes a candidate response and one or more reference responses, and assigns a numerical score to the candidate.

Formally, let $\mathcal{V}^*$ be the text space over a vocabulary $\mathcal{V}$, and let $W \sim P_W$ denote a task. Conditional on $W = w$, a candidate report $X \sim P_{X|W=w}$ and a reference report $Y \sim P_{Y|W=w}$ are independently produced.

A standard *reference-based metric* is a function $S(x,y)$ that maps a candidate-reference pair (x,y) to a scalar score. Some metrics also use a *negative reference* for contrastive scoring. In that case, we draw $Y^- \sim P_Y$ independently of (W,X,Y), where P_Y is the marginal reference distribution. In experiments, this is implemented by sampling a reference from another task. If a metric does not use a negative reference, we set $y^- = \emptyset$.

Given these inputs, we write the general form of a metric as a contrastive scoring function

$$S(x,y,y^-)$$

The negative reference makes the score contrastive: the metric rewards agreement with the matched reference y relative to agreement with an unrelated reference y^-. This discounts generic language or common claims that are likely under P_Y but not specific to the task.

The metric-evaluation problem is to assess whether a scoring function 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.

2.1 Mutual Information and Information Monotonicity

The guiding principle for our framework is information monotonicity: a metric should not assign a higher expected score after a perturbation that removes or fails to add task-relevant information. Degradations and manipulations are both perturbations of informativeness, which we formally introduce in Section 3. Degradations strictly reduce information by removing task-relevant content, while manipulations try to gain score through transformations that do not add task-relevant content. We now formalize this principle using mutual information.

Intuitively, mutual information $\mathrm{MI}(X;Y)$ measures the amount of information between two random variables beyond what would be expected if they were independent. It is zero when the two variables are independent, and it increases as the variables share more information. In reference-based text evaluation, the candidate X and reference Y are generally independent after conditioning on being from the same underlying task. This means their dependent part, measured by mutual information, reflects task-relevant content. A candidate response that preserves the information (e.g., facts, reasoning, or semantic content) for the task should therefore share more information with any reference than a strategically perturbed response.

Mutual information is defined to have information monotonicity. Information monotonicity implies that the mutual information cannot be increased by post-processing one side. If a transformation perturbs the candidate without access to new task information, then it can at best preserve the information that X already had about Y, and typically loses some of it. Thus an information-monotone score should not reward degradations or manipulations that merely change the surface form of a response without adding genuine task-relevant content. The information monotonicity is instantiated by the information-processing inequality of an MI. This provides the information-theoretic foundation of our framework.

Proposition 2.1 (Information-Processing Inequality)**.** MI *satisfies information-processing inequality if the following is true: for any mapping,* $\sigma : \mathcal{X} \to \mathcal{X}'$*,*

$$\mathrm{MI}(\sigma(X);Y) \leq \mathrm{MI}(X;Y).$$

Our paper focuses on the f-mutual information as below, which satisfies the information-processing inequality.

Definition 2.2 (f-divergence and f-mutual information)**.** *Let P and Q be probability distributions on a finite space $\mathcal{X}$, and let $f : \mathbb{R}_+ \to \mathbb{R}$ be convex with $f(1) = 0$. The f-divergence is*

$$D_f(P\|Q) := \sum_{x \in \mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right).$$

For jointly distributed random variables $(X,Y) \sim P_{XY}$, the f-mutual information is defined as the f-divergence between the joint distribution and the product distribution of X and Y:

$$I_f(X;Y) := D_f\big(P_{XY} \| P_X \otimes P_Y\big),$$

Thus, $I_f(X;Y)$ measures how far the true joint distribution is from independence. When $f(t) = t\log t$, the f-MI recovers Shannon mutual information [10].

The central object underlying all f-MI computation is the density ratio between the joint and product-of-marginals distributions $r(x,y)$.

$$r(x,y) := \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = \frac{P_{Y|X}(y\,|\,x)}{P_Y(y)} = \frac{P_{X|Y}(x\,|\,y)}{P_X(x)}.$$

The population quantity $I_f(X;Y)$ can be computed exactly from r:

$$I_f(X;Y) = \mathbb{E}_{P_X \otimes P_Y}\big[f(r(X,Y))\big]. \tag{1}$$

3 Test Principles

We compare evaluation metrics along statistical alignment and strategic alignment, under our test principles, generalized from previous work [31, 46, 41, 44]. 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,

$$\rho = \mathrm{Spearman}\big(\{s_i\}_{i=1}^n, \{h_i\}_{n=1}^n\big).$$

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 6 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) [2] 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 7 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

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.

Definition 4.1 (Design tuple)**.** *An f-MI evaluation metric is specified by a tuple*

$$S := (F, E, R, \Pi),$$

where:

- *(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, $R : \mathcal{V}^{\leq L} \to \mathcal{U}$.*
- *(iv)* ***Prediction*** *Π: a mechanism that maps represented candidate-reference pairs to predictive quantities used in the estimation method.*

The tuple has two dependencies. First, F determines the population objective, and E determines how that objective is estimated from samples. Second, R fixes the units of text passed to Π, and Π produces the quantities required by E. Together these choices define the sample-level score $S_{F,E,R,\Pi}$.

Prior MI-based metrics correspond to several specific choices of this tuple, e.g., GEM [46] estimates KL mutual information (F) directly through pointwise mutual information (E), using token-level representations (R) and autoregressive language-model probabilities (Π), while TVD-MI [41] makes different choices along several axes: it uses total-variation mutual information (F), an f-variational estimator (E), whole-report representations (R), and an LLM-oracle critic (Π).

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. For some designs, the metric input also has a negative reference report y_i^- sampled from the marginal distribution P_Y, where the metric penalizes the agreement with y_i^-. For a fixed design tuple (F,E,R,Π), the metric assigns each item a sample-level score

$$S_{F,E,R,\Pi}(x_i,y_i,y_i^-).$$

The construction of a reference-based metric applies the framework above. To score a single item, we take the chosen MI estimation method and replace each expectation with an empirical average over the sampled positive and negative references. Averaging these per-item scores across the dataset then yields an estimator of the corresponding f-MI quantity determined by F, E, R, and Π:

$$\sum_{i\in n} S_{F,E,R,\Pi}(x_i,y_i,y_i^-) \approx \hat{I}_F^{(E,R,\Pi)}(X;Y).$$

The rest of this section instantiates the four design choices — the information measure f-MI (Section 4.1), the MI estimation method (Section 4.2), the text representation (Section 4.3), and the prediction mechanism (Section 4.4). Section 4.5 then frames existing metrics as instances of the framework, and Section 4.6 shows that the resulting metrics inherit an approximate manipulation-robustness guarantee from information monotonicity, bounded by the estimation error.

4.1 f-Mutual Information

In this paper, we instantiate the information-measure component F with the following two mutual-information objectives.

- **KL mutual information**. Taking $f_{\mathrm{KL}}(t) = t\log t$ gives the Shannon mutual information

  $$I_{\mathrm{KL}}(X;Y) = D_{\mathrm{KL}}(P_{XY} \| P_X \otimes P_Y) = \mathbb{E}_{P_{XY}}[\log r(X,Y)].$$

  The sample-level direct estimator is pointwise mutual information, $\log r(x,y)$, which recovers the information measure used by GEM and GPPM-style metrics.

- **Total-variation mutual information**. Taking $f_{\mathrm{TV}}(t) = \frac{1}{2}|t-1|$ gives the total-variation mutual information (TVD-MI)

  $$I_{\mathrm{TV}}(X;Y) = \frac{1}{2}\mathbb{E}_{P_X\otimes P_Y}[|r(X,Y)-1|] = \frac{1}{2}\sum_{x,y}|P_{XY}(x,y) - P_X(x)P_Y(y)|.$$

Our empirical design sweep therefore varies $F \in \{\mathrm{KL}, \mathrm{TV}\}$, and crosses these two information measures with the estimation methods described next.

4.2 MI Estimation

Since the vast combinatorial space of text prevents the explicit expression of joint distributions, I_f cannot be computed in closed form. To circumvent this, we explore two families of estimators.

- **Direct Density Ratio Estimation**. By the density ratio formulation of MI in Equation (1), if the density ratio r is known, MI can be calculated via a Monte Carlo estimator by averaging $f(r(x,y))$ over samples $(x,y) \sim P_X \otimes P_Y$. Section B.1 provides examples of direct density ratio estimation.

- **f-Variational Estimation**. Another commonly used estimator in the literature is the f-variational estimator, constructed via the Fenchel conjugate f^*.

$$I_f(X;Y) = \mathbb{E}_{P_{XY}}[T^*(X,Y)] - \mathbb{E}_{P_X \otimes P_Y}\left[f^*(T^*(X,Y))\right], \tag{2}$$

where $T^*(x) \in \partial f\left(\frac{p(x)}{q(x)}\right)$ is defined by the subgradient of f function.

The main difference from direct density ratio estimation lies in whether the method requires estimating the density ratio $r(x,y)$. For example, for MI defined by the commonly used KL and TV divergence, direct density ratio estimation requires estimating the density ratio $r(x,y)$. The requirement of the f-variational estimators is weaker. We defer mathematical details to Section B.2.

The f-variational estimator is computed given a task w_i, candidate x_i, positive reference y_i, and negative reference $y_i^- \sim P_Y$. The estimator is an unbiased estimator $S(x_i,y_i,y_i^-)$, a sample-level score. $S(x_i,y_i,y_i^-)$ estimates the population expectations in Equation (2) by replacing the joint expectation over P_{XY} with computation over (x_i,y_i), and the marginal expectation over $P_X \otimes P_Y$ with (x_i,y_i^-).

Examples. For example, Table 1 summarizes the four cases and makes explicit the sample-level score $S(x_i,y_i,y_i^-)$ that the prediction mechanism Π needs to estimate. Averaging S over the n tasks then estimates the corresponding population f-MI, as required by the design of Definition 4.1.

Table 1: Sample-level score and quantity to estimate for each (F,E) pair.

F	E	Sample-level score	Quantity to estimate
KL	Direct	$\log r(x_i,y_i)$	$\log r(x_i,y_i)$
KL	f-variational	$T^*(x_i,y_i) - e^{T^*(x_i,y_i^-)-1}$	$T^*(x_i,y_i) = 1 + \log r(x_i,y_i)$
TV	Direct	$\frac{1}{2}\lvert 1 - r(x_i,y_i)^{-1}\rvert$ or $\frac{1}{2}\lvert r(x_i,y_i^-) - 1\rvert$	$r(x_i,y_i)$ or $r(x_i,y_i^-)$
TV	f-variational	$T^*(x_i,y_i) - T^*(x_i,y_i^-)$	$T^*(x_i,y_i) = \frac{1}{2}\mathrm{sign}(r(x_i,y_i) - 1)$

Note that KL-direct and TV-direct both require estimating the density ratio $r(x,y)$, while the variational estimators only require a critic $T^*(x,y)$. In particular, TV-variational only needs a bounded classifier for whether $r(x,y)$ is above or below 1.

Producing these estimators from text requires two further choices: a representation R that fixes the granularity at which the report is processed (Section 4.3), and a prediction mechanism Π that maps represented pairs to the required predictive quantities (Section 4.4).

4.3 Representation

The representation $R : \mathcal{V}^{\le L} \to \mathcal{U}$ fixes the units at which a report is processed by the downstream predictor. We consider three natural granularities.

- **Token**. At the finest granularity, a report y is represented by its token sequence $(y_1,\ldots,y_T)$. This preserves all information in the text and exposes the autoregressive factorization of language models, $p(y \mid x) = \prod_{t=1}^{T} p(y_t \mid y_{<t},x)$, so that report-level conditional probabilities can be assembled from per-token factors. Token representation is only meaningful in combination with a predictor that consumes token-level signals, such as an autoregressive language model with logit access.

  In practice, prior work [46, 31] reports that style-normalization pre-processing is important when using token representations, since autoregressive language models can be confounded by superficial features of the text such as writing style.

- **Statement**. An intermediate granularity decomposes a report into its atomic claims. Let

$$\Psi(y) = \{\psi_1(y),\ldots,\psi_{K(y)}(y)\}$$

denote the statement decomposition of y, where each $\psi_k(y) \in \mathcal{V}^*$ is an atomic statement, typically obtained via a separate LLM call. Decomposing a report into statement-level representation naturally loses information by the data-processing inequality (Theorem 2.1). In

practice, however, isolating individual claims often improves the predictor's performance, especially for an LLM-oracle, which tends to reason more reliably about single facts than about long passages. With statement-level decomposition, the estimation quality may even improve. We compare these tradeoffs empirically in Section 3 and Section 5.1.

- **Full-Report**. At the coarsest granularity, a report is treated as an atomic unit, $R(y) = y$. Predictive quantities are obtained holistically from the candidate-reference pair (x, y): a single conditional likelihood $p(y \mid x)$, or a single oracle judgment about the pair. (e.g., "do these two reports describe the same item?").

4.4 Prediction Mechanism

The prediction mechanism Π maps a represented candidate-reference pair to the predictive quantities in the representation space, as required by the chosen estimator E. See Table 1 for examples of required predictive quantities. Below lists potential prediction mechanisms.

- **Autoregression**. An autoregressive language model with parameters ϕ assigns a probability to any token sequence via the chain rule. For a candidate report x in the prompt, we obtain

$$p_\phi(y \mid x) = \prod_{t=1}^{T} p_\phi(y_t \mid y_{<t}, x)$$

  where each factor is read off the softmax over output logits. Autoregression interacts with the *token* level representation. It also requires logit access, which some commercial APIs do not expose, and inherits any miscalibration of the underlying language model. As an example, for KL-based MI estimation, substituting a null report for x yields a marginal estimate $p_\phi(y)$, and the following PMI is obtained for direct KL estimation:

$$\widehat{\mathrm{PMI}}(x;y) = \log r(x,y) = \log p_\phi(y \mid x) - \log p_\phi(y).$$

  Exponentiating gives a density-ratio estimate that feeds any of the four (F, E) pairs in Table 1.

- **LLM-Oracle**. When token-level probabilities are unavailable, an LLM can be prompted for a prediction about a represented pair. The prediction is then used as inputs to the estimators as required. The following two regimes of an LLM oracle cover the cases in Table 1:
  - *Critic-style prediction* fits naturally into the f-variational framework: a scalar or categorical LLM is prompted to directly output a critic used as $T^*(x,y)$. For example, Robertson and Koyejo [41] shows that the TVD-MI is especially convenient. The T^* required in Equation (2) for TVD-MI can be implemented as a binary classifier distinguishing same-task from cross-task pairs. A simple prompt, e.g., "Do these two reports describe the same item?" implements this T^* for TVD-MI.
  - *Likelihood-style prediction* supports direct estimation. The LLM is asked, for instance, how strongly the candidate x supports the reference y, using a small set of ordered probability or support bins [31]. Mapping the selected bin to a numeric value yields a coarse surrogate for $\log \hat{p}(y \mid x) - \log \hat{p}(y)$, and hence for the density ratio $\hat{r}(x, y)$. Such direct LLM-oracle estimators are typically biased. Their outputs are discretized and often poorly calibrated. But they provide a workable approximation when only API access is available.

  LLM-oracle prediction composes with the whole-report and statement representations. At the *whole-report* level, the prompt simply contains x and y and the oracle returns a single prediction. At the *statement* level, the candidate report x is paired with each $\psi_k(y)$, and the resulting per-statement quantities are combined into a report-level prediction by an aggregation rule $A : \bigcup_{K \geq 1} \mathbb{R}^K \to \mathbb{R}$.

4.5 Prior Work

Combining the preceding components, a complete design tuple (F, E, R, Π) specifies a concrete reference-based metric. Table 2 shows how existing methods instantiate this framework.

Table 2: Existing methods as instances of the design tuple (F,E,R,Π).

Method	F	E	R	Π
GEM [46] / GPPM-token [31]	KL	Direct (PMI)	Token	Autoregression
GPPM-judgment [31]	KL	Direct (PMI)	Statement	LLM-oracle
TVD-MI [41]	TV	f-variational	Whole-report	LLM-oracle

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 2 leaves blank 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 9. We also provide a formal definition of the (F=TV, E=f-variational, R=Statement, Π=LLM-oracle) metric as an example in Appendix C.

4.6 Manipulation Robustness of the f-MI Metrics

The information processing inequality in Proposition 2.1 guarantees that any manipulation $\sigma : \mathcal{X} \to \mathcal{X}$ of a candidate report can only reduce the *true* f-MI:

$$I_f(\sigma(X);Y) \leq I_f(X;Y).$$

However, the score is computed using a fixed estimator and distribution model, which in practice amounts to using a restricted or suboptimal critic rather than the population optimum. As a result, the estimated score need not satisfy DPI exactly, even when the underlying information measure does. This motivates the following approximate notion of robustness: for a class Σ of manipulations, suppose

$$\sup_{\sigma \in \Sigma} |\mathbb{E}[S(\sigma(X),Y)] - I_f(\sigma(X);Y)| \leq \varepsilon.$$

Then for every $\sigma \in \Sigma$,

$$\mathbb{E}[S(\sigma(X),Y)] \leq \mathbb{E}[S(X,Y)] + 2\varepsilon.$$

Indeed,

$$\mathbb{E}[S(\sigma(X),Y)] \leq I_f(\sigma(X);Y) + \varepsilon \leq I_f(X;Y) + \varepsilon \leq \mathbb{E}[S(X,Y)] + 2\varepsilon,$$

where the middle inequality is exactly Proposition 2.1. Thus, any manipulation that decreases the true f-mutual information by more than 2ε is still penalized by the estimated score in expectation.

The assumption requires the estimation error to be uniformly small over the entire manipulation class Σ. This is a strong condition: it asks the critic to generalize not only to clean reports but also to all manipulated variants. In practice, ε may be large for manipulation strategies that produce out-of-distribution inputs for the critic. Empirical validation (Section 5) is therefore essential.

5 Experiment

We compare scoring methods in a controlled way under our test principles (Section 3), with detailed setup deferred to Section D.1.

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 (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. Specific dataset statistics, including sample sizes, reference types, and annotation availability, are reported in Section D.1.

5.1 Results

Table 3: Human-rating correlation results. Rows are scoring methods, grouped by distribution model D and its granularity; 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 rank correlations 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				
(KL,Direct) (GEM [46])	**0.460**	**0.475**	**0.138**	0.113
(KL,f-var.)	-0.218	-0.355	0.059	0.051
(TV,Direct)	0.312	0.380	0.095	-0.025
(TV,f-var.)	-0.053	-0.036	0.085	**0.129**
LLM-oracle, Report-representation				
(KL,Direct)	0.235	0.367	0.031	**0.163**
(KL,f-var.)	0.252	**0.419**	0.031	0.162
(TV,Direct)	-0.231	-0.350	0.017	-0.238
(TV,f-var.) (TVD-MI [41])	**0.286**	0.042	**0.140**	0.110
LLM-oracle, Statement-representation				
(KL,Direct) (GPPM-J [31])	0.254	0.307	0.073	0.041
(KL,f-var.)	**0.281**	0.251	0.075	0.064
(TV,Direct)	-0.292	-0.396	-0.043	-0.129
(TV,f-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

Table 3 reports Spearman correlation with human ratings on the four datasets carrying item-level numerical annotations: Peer Grading-WH [44], Peer Grading-XLSK [46], SummEval [12], and MedAESQA [18]. 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.

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 3: (KL,Direct) with token-level autore-

Table 4: 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	✓	✓	PU	PU	5
TVD-FVar-Report	DSU	O	OPU	PU	✓	PU	DSPU	15
TVD-FVar-Statement	✓	✓	✓	P	✓	P	P	3
Non-MI baselines								
LLM-as-Judge / Claude-Haiku-4.5	✓	✓	P	P	✓	P	DP	5
LLM-as-Judge / Claude-Sonnet-4.5	✓	✓	P	P	✓	P	P	4
LLM-as-Judge / GPT-4o-mini	D	DOR	ORP	P	✓	P	P	10
LLM-as-Judge / GPT-5-mini	✓	✓	P	P	✓	P	DP	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 5: 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	NHPSR	✓	✓	✓	✓	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	MNHSR	HPSR	HSR	R	R	HSR	HSR	20
LLM-as-Judge / Claude-Sonnet-4.5	MNHSR	HPSR	MHPSR	R	✓	HS	H	18
LLM-as-Judge / GPT-4o-mini	MNHPSR	MNHPSR	MNHPSR	R	R	HSR	HSR	26
LLM-as-Judge / GPT-5-mini	NHSR	MNHPSR	HPSR	✓	✓	HSR	HS	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.

gression, 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 4 and 5, and detailed statistics in Section E.

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. By the standard statistical alignment alone, it appears to be the best metric. 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-level 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 representation is substantially more effective: statement-level $(\mathrm{TV}, f\text{-var.})$ outperforms report-level 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 dominating 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 report-level $(\mathrm{TV}, f\text{-var.})$ / TVD-MI. It has 0 failures out of 30 manipulation tests, meaning that no tested manipulation produces a significant score increase.

Robustness of MI-based metrics comes from the design, not merely from the model. Since Claude-Haiku-4.5 is vulnerable as an absolute LLM judge but robust when used inside the TV f-variational metric, the improvement comes from the MI-based formulation.

It is worth to notice that no single axis of (F, E, R, Π) dominates in isolation; rather, the design choices interact. Token-level 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.

ROUGE-L, BLEU, and BERTScore are generally weaker baselines, and they fail to obtain positive correlation on all datasets, especially in peer review. This suggests that lexical overlap and embedding similarity are insufficient for evaluating richer, more subjective reports.

Additional discussion. In ICLR2026, GEM fails the rephrase manipulation test, which appears to contrast with the original GEM results [46], where rephrasing did not produce a significant score increase in ICLR2023. A likely hypothesis is that our rephrase manipulation is stronger: Whereas Xu et al. [46] primarily prompt the model to improve language quality, our prompt also allows the model to reorganize the review. Although it prohibits adding new facts or changing the reviewer's stance, this reorganization can make the candidate review easier for the metric to match with the reference. The result therefore suggests that GEM retains some sensitivity to presentation quality.

However, the standardized mean difference for GEM remains relatively small ($+0.07 \pm 0.05$, 95% CI), especially compared with direct LLM judging, even for the two most robust judge models: Claude-Sonnet-4.5 ($+0.25 \pm 0.09$, 95% CI) and GPT-5-mini ($+0.17 \pm 0.10$, 95% CI), as shown in Appendix E. This observation reinforces the value of the MI-based formulation, and further motivates the new statement-level (TV, f-var.) metric, which remains robust to this stronger rephrasing test.

We also noticed that surface-report degradation is a strong test because the degraded reports are still fluent and task-specific, generated from weak but informative context such as a news lead, question, or paper abstract, by Claude-4.5-Haiku. Thus, they can remain plausible even while sometimes omitting important task-relevant information.

Under this test, MI-based metrics also compare favorably with the most robust LLM judges. MI-based metrics consistently pass more tests than LLM judges. Moreover, as shown in Appendix E, all MI-based metrics either significantly decrease on surface reports or remain close to insensitive, with no significant positive failures. By contrast, even the strongest LLM judges can not penalize but reward surface reports: Claude-Sonnet-4.5 gives $d = +0.28 \pm 0.15$ (95% CI) on SummEval, $+0.69 \pm 0.15$ on LFQA-E, and $+1.18 \pm 0.26$ on MedAESQA, while GPT-5-mini gives $+0.12 \pm 0.17$, $+0.51 \pm 0.14$, and $+1.01 \pm 0.19$ on the same datasets. Thus, surface reports expose a hard case for all metrics, but the failure mode differs: MI-based metrics can sometimes be under-sensitive, whereas LLM judges can directionally fail by rewarding plausible surface-level reports.

6 Related Work

NLG Evaluation Evaluation of AI systems spans many dimensions. Holistic frameworks consider accuracy, robustness, fairness, and toxicity [28]; alignment research emphasizes helpfulness, honesty, and harmlessness [4]; and a growing body of work targets truthfulness [30], the multi-faceted nature of fairness [15], and pluralism over legitimate human perspectives [43].

Our framework instead targets a complementary and foundational property, *semantic informativeness*: whether an evaluation score faithfully reflects how much useful, truthful, task-relevant information a report conveys. A response cannot be genuinely helpful if it carries little task-relevant signal, and mechanisms that reward truthful, informative reporting can discourage guessing and some forms of hallucination [44].

Mutual Information Estimation Machine learning has produced sample-based MI estimators built on the variational representation of f-divergences via Fenchel duality [36], and neural network estimators such as MINE [5], Deep InfoMax [20], and InfoNCE [37].

In information elicitation without verification, MI plays a different role: Kong and Schoenebeck [24] score each agent a measure of MI between her report and a peer's reference, using the data-processing inequality to make truthful reporting an equilibrium. This echoes the same strategic-alignment principle behind our MI-based metrics (Proposition 2.1). Kong and Schoenebeck [23] extends the same MI-maximization principle to co-training, where two Bayesian predictors over distinct views of the data play the role of the two agents.

Closest to our setting are pre-trained language model approaches to MI estimation: Padmakumar and He [38] compute pointwise MI between extractive summaries and source articles, while Xu et al. [46], Lu et al. [31], Robertson and Koyejo [41] develop reference-based MI metrics for NLG evaluation, which can be recovered as specific instances by our design framework. Recent work also uses estimated MI for dataset evaluation [9, 49].

Information Elicitation. Our notion of strategic alignment is motivated by the information elicitation literature, which designs scoring mechanisms that induce agents to truthfully reveal private information. Two central lines are proper scoring rules and peer prediction. Proper scoring rules elicit truthful probabilistic predictions when a ground-truth outcome is eventually observed [33, 42, 17, 26]. Peer prediction instead handles settings without direct verification, using the statistical relationship among peer reports as a substitute for ground truth [34, 11, 24, 21, 22, 48]. Our setting parallels peer prediction: text evaluation metrics are not merely measurement tools but scoring mechanisms that shape the behavior of systems optimized against them. As Burrell and Schoenebeck [6], Xu et al. [45] point out, however, peer prediction has emphasized proving strategic alignment while devoting little attention to the statistical alignment of the resulting scores, limiting their direct use as evaluation metrics. Our test principles target both requirements jointly.

We list optimal evaluation mechanism design as a future work. This connects to prior work on optimal scoring rule design [27, 35, 19, 7, 40, 8], where the designer optimizes over scoring rules, with objectives such as informativeness, effort incentives, risk, or downstream decision quality, and subject to the constraint of truthfulness. In text evaluation, alignment with human judgments becomes the design objective, subject to robustness against strategic perturbations. Lu et al. [32] take a step in this direction by optimizing aligned textual scoring rules with access to ground-truth texts. We leave the corresponding problem without ground truth, jointly optimizing for statistical and strategic alignment, to future work.

7 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 scoring mechanisms directly for alignment, as suggested in Section 6. 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 benchmarks 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, benchmarks may need richer references, such as expert-panel references, debate- or critique-based references, or aggregated LLM outputs [14]. Strategically validated scoring mechanisms may also serve as training rewards with reduced susceptibility to reward hacking. For example, Feng et al. [14] 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, scoring-mechanism design, and training objectives are studied jointly.

Acknowledgments and Disclosure of Funding

This work is supported by United States National Science Foundation Award #2313137.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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*, pages 5194–5218. PMLR, 2023.

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

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

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

[20] 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.

[21] 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*, pages 2398–2411. SIAM, 2020.

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

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

[24] 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.

[25] 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.

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

[27] 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*, pages 988–989, 2022.

[28] 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.

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

[30] 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)*, pages 3214–3252, 2022.

[31] 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.

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

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

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

[35] 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*, pages 718–733, 2021.

[36] 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.

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

[38] 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*, pages 2505–2512, 2021.

[39] 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*, pages 311–318, 2002.

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

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

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

[43] 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.

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

[45] 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.

[46] 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.

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

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

[49] 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.

A Test Principle Details

In this section, we provide the details of degradation and manipulation strategies used in our experiments.

Table 6: Degradation strategies.

Strategy	Description	Provenance
Random Replacement	Replace the original candidate with a response sampled from a different task within the same task family.	Lu et al. [31]
Sentence Deletion	Delete every other sentence while preserving section headers, ordering, and visible format.	Xu et al. [46]
Deletion & Completion	Delete every other sentence, mark each gap, then ask a helper LLM to fill the gaps using only the remaining text and the original format.	Xu et al. [46]
Surface Report	Regenerate the response using only weak context, such as the abstract for peer review or the headline plus lead paragraph for summarization.	Robertson and Koyejo [41], Xu et al. [46]
Ultra-Concise Compression	Summarize under a severe length cap (e.g., 10% of the original), so that specific evidence and caveats must be dropped.	Robertson and Koyejo [41]
Opinion Flip	Reverse the overall evaluative verdict (e.g., accept $\rightarrow$ reject, positive $\rightarrow$ negative) while preserving the surface structure and topical vocabulary of the original candidate.	Robertson and Koyejo [41]

Table 7: Manipulation strategies.

Strategy	Description	Provenance
Rephrase	Ask an LLM to rewrite the candidate in a specific style while preserving meaning and high-level structure.	Xu et al. [46]
Meaningless Elongation	Append the same fixed, semantically empty filler sentences to every candidate report without adding task-relevant information.	Xu et al. [46]
Opinion Shift (Positive)	Rewrite the candidate to preserve structure and coverage while systematically shifting stance toward more favorable sentiment.	Robertson and Koyejo [41]
Opinion Shift (Negative)	Rewrite the candidate to preserve structure and coverage while systematically shifting stance toward less favorable sentiment.	Robertson and Koyejo [41]
Opinion Shift (Neutral)	Rewrite claims as hedged, noncommittal statements (e.g., "the method is effective" $\rightarrow$ "the method may have some merit"), flattening the evaluative signal without changing topical content.	This work
Opinion Shift (Extreme)	Amplify all claims to their strongest possible form (e.g., "a reasonable approach" $\rightarrow$ "a groundbreaking contribution"), testing whether a metric rewards confidence or extremity as a proxy for quality.	This work

B Design Framework Details

B.1 Direct Density Ratio Estimation

For KL, $f(t) = t\log t$, so

$$I(X;Y) = \mathbb{E}_{P_X \otimes P_Y}\left[r(X,Y)\log r(X,Y)\right] = \mathbb{E}_{P_{XY}}\left[\log r(X,Y)\right].$$

The quantity $\mathrm{PMI}(x;y) := \log r(x,y) = \log p(y\,|\,x) - \log p(y)$ is the *pointwise mutual information*; its expectation under the joint distribution recovers the Shannon mutual information.

Direct estimation is also available for TVD-MI, where $f(t) = \frac{1}{2}|t-1|$, so we have

$$I_{\mathrm{TV}}(X;Y) = \frac{1}{2}\mathbb{E}_{P_X \otimes P_Y}\left[|r(X,Y)-1|\right] = \frac{1}{2}\sum_{x,y \in \mathcal{V}^{\leq L}} |P(x,y) - P(x)P(y)|.$$

This can also be written as $I_{\mathrm{TV}}(X;Y) = \frac{1}{2}\mathbb{E}_{P_{XY}}\left[|1 - \frac{1}{r(X,Y)}|\right]$, providing two approaches of Monte Carlo estimation.

Table 8: Common f-divergences, along with a convenient population-optimal critic $T^*(x,y)$ and the corresponding conjugate term $f^*(T^*(x,y))$, written in terms of the density ratio $r(x,y)$.

f-divergence	$f(t)$	$T^*(x,y)$	$f^*(T^*(x,y))$
Total variation	$\frac{1}{2}\lvert t-1\rvert$	$\frac{1}{2}\operatorname{sign}\big(r(x,y)-1\big)$	$\frac{1}{2}\operatorname{sign}\big(r(x,y)-1\big)$
KL divergence	$t\log t$	$1+\log r(x,y)$	$r(x,y)$
Reverse KL	$-\log t$	$-\frac{1}{r(x,y)}$	$-1+\log r(x,y)$
Pearson χ^2	$(t-1)^2$	$2\big(r(x,y)-1\big)$	$r(x,y)^2-1$
Squared Hellinger	$(\sqrt{t}-1)^2$	$1-\frac{1}{\sqrt{r(x,y)}}$	$\sqrt{r(x,y)}-1$

B.2 f-Variational Estimation

Every f-divergence admits a variational lower bound via the Fenchel conjugate f^*.

Definition B.1 (Variational representation)**.** *Let f^* denote the Fenchel conjugate of f,*

$$f^*(v) := \sup_{t>0}\{vt - f(t)\}.$$

Then

$$D_f(P\|Q) = \sup_T \big\{\mathbb{E}_P[T(X)] - \mathbb{E}_Q\big[f^*(T(X))\big]\big\},$$

where the supremum is over functions T such that $f^*(T(x)) < \infty$ for all x with $Q(x) > 0$.

Here T is a critic: a scalar scoring function that tries to assign higher values to samples from P than to samples from Q. We write T^* for the population-optimal critic. When f is differentiable,

$$T^*(x) = f'\left(\frac{p(x)}{q(x)}\right),$$

and in general

$$T^*(x) \in \partial f\left(\frac{p(x)}{q(x)}\right),$$

where ∂f denotes the subdifferential of f.

The f-mutual information also admits the variational form

$$I_f(X;Y) = \sup_T \{\mathbb{E}_{P_{XY}}[T(X,Y)] - \mathbb{E}_{P_X \otimes P_Y}[f^*(T(X,Y))]\}.$$

At the population optimum, equality is attained when $T(x,y) = T^*(x,y)$.

In our setting, we can write $T(x,y)$ for a critic on candidate–reference pairs and $T^*(x,y)$ for the population-optimal critic. The optimal critic is determined by the chosen generator f through

$$T^*(x,y) \in \partial f\big(r(x,y)\big).$$

As Table 8 shows, different choices of f induce different optimal critics and different conjugate penalty terms. We refer back to this table when instantiating KL-based and TV-based evaluation metrics.

Estimator Here joint samples P_{XY} are candidate-reference pairs from the same task, while marginal samples $P_X \otimes P_Y$ are pairs drawn across different tasks. Recall that, the optimum is attained by a critic satisfying $T^*(x,y) \in \partial f\big(r(x,y)\big)$. For common choices of f, the corresponding forms of $T^*(x,y)$ and $f^*(T^*(x,y))$ are listed in Table 8.

For KL, the variational estimation yields the NWJ objective:

$$\begin{aligned} I(X;Y) &= \sup_T \Big\{\mathbb{E}_{P_{XY}}[T(X,Y)] - e^{-1}\mathbb{E}_{P_X \otimes P_Y}\big[e^{T(X,Y)}\big]\Big\} \\ &= 1 + \mathbb{E}_{P_{XY}}\big[\log r(X,Y)\big] - \mathbb{E}_{P_X \otimes P_Y}\big[r(X,Y)\big] \end{aligned}$$

For total variation, the estimator reduces to

$$I_{\mathrm{TV}}(X;Y) = \sup_{\|T\|_\infty \le 1/2} \Big\{ \mathbb{E}_{P_{XY}}[T(X,Y)] - \mathbb{E}_{P_X \otimes P_Y}[T(X,Y)] \Big\} \tag{3}$$
$$= \frac{1}{2}\mathbb{E}_{P_{XY}}\big[\mathrm{sign}(r(X,Y)-1)\big] - \frac{1}{2}\mathbb{E}_{P_X \otimes P_Y}\big[\mathrm{sign}(r(X,Y)-1)\big]$$

where the optimal critic $\mathrm{sign}(r(X,Y)-1)$ is equivalent to a binary classifier that distinguishes same-task pairs from cross-task pairs.

C TV, f-variational, Statement-level, LLM-oracle Metic

This method instantiates the scoring framework with

$F = \mathrm{TV}$, $E = f$-variational, $D =$ LLM-as-a-Judge (statement-level) denoted as D_Ψ,

where D_Ψ is induced by a reference decomposition Ψ.

We keep the candidate at the report level and decompose each reference into statements:

$$\Psi(y) = \big(\psi_1(y),\ldots,\psi_{K(y)}(y)\big). \tag{4}$$

We instantiate the local critic t and aggregation rule A from Section 4.4 as follows. For each candidate–statement pair, an LLM judge returns a centered binary local critic

$$t : \mathcal{V}^{\le L} \times \mathcal{V}^{\le L} \to \left\{ -\frac{1}{2}, \frac{1}{2} \right\}, \tag{5}$$

where $t(x,w) = \frac{1}{2}$ if candidate x supports or aligns with statement w, and $t(x,w) = -\frac{1}{2}$ otherwise. We use mean aggregation, so the induced report-level critic is

$$T_\Psi(x,y) = \frac{1}{K(y)} \sum_{k=1}^{K(y)} t\big(x, \psi_k(y)\big). \tag{6}$$

Since each summand lies in $[-\frac{1}{2}, \frac{1}{2}]$, T_Ψ satisfies the constraint in the TV variational estimator.

For task i, let $\mathcal{R}_i^+$ be positive references paired with the same task, and let $\mathcal{R}_i^-$ be negative references sampled from the marginal reference distribution P_Y. We define the sampled contrastive score

$$\hat{S}_{\mathrm{CTV}}(x_i; \mathcal{R}_i^+, \mathcal{R}_i^-) = \frac{1}{|\mathcal{R}_i^+|} \sum_{y \in \mathcal{R}_i^+} T_\Psi(x_i, y) - \frac{1}{|\mathcal{R}_i^-|} \sum_{y \in \mathcal{R}_i^-} T_\Psi(x_i, y). \tag{7}$$

Assume that, conditional on task ψ_i, the references in $\mathcal{R}_i^+$ are drawn i.i.d. from $P_{Y|\psi_i}$, and the references in $\mathcal{R}_i^-$ are drawn i.i.d. from P_Y, independently of (x_i, ψ_i). Then

$$\mathbb{E}\Big[\hat{S}_{\mathrm{CTV}}(x_i; \mathcal{R}_i^+, \mathcal{R}_i^-) \,\big|\, x_i, \psi_i\Big] = \mathbb{E}_{Y \sim P_{Y|\psi_i}}\big[T_\Psi(x_i, Y)\big] - \mathbb{E}_{Y \sim P_Y}\big[T_\Psi(x_i, Y)\big]. \tag{8}$$

If (X_i, Y_i) is generated by first sampling a task and then sampling the candidate and reference independently conditional on that task, averaging Eq. (8) over (X_i, ψ_i) yields

$$\mathbb{E}\big[\hat{S}_{\mathrm{CTV}}(X_i; \mathcal{R}_i^+, \mathcal{R}_i^-)\big] = \mathbb{E}_{P_{XY}}\big[T_\Psi(X,Y)\big] - \mathbb{E}_{P_X \otimes P_Y}\big[T_\Psi(X,Y)\big]. \tag{9}$$

Because T_Ψ factors through $\Psi(Y)$ and satisfies $\|T_\Psi\|_\infty \le 1/2$, Eq. (3) applied to the transformed pair $(X, \Psi(Y))$ gives

$$\mathbb{E}\big[\hat{S}_{\mathrm{CTV}}(X_i; \mathcal{R}_i^+, \mathcal{R}_i^-)\big] \le I_{\mathrm{TV}}\big(X; \Psi(Y)\big) \le I_{\mathrm{TV}}(X;Y).$$

D Experiment Details

D.1 Experiment Design

To compare scoring methods in a controlled way, we organize our experiments around the design space induced by Definition 4.1. Each scoring method corresponds to a choice along four axes:

(i) the information measure $F \in \{\text{KL}, \text{TV}\}$;

(ii) the estimator $E \in \{\text{Direct}, f\text{-variational}\}$;

(iii) the representation $R \in \{\text{Token}, \text{Whole-report}, \text{Statement}\}$.

(iv) the predictor $\Pi \in \{\text{Autoregression}, \text{LLM-oracle}\}$;

The predictor and representation do not vary independently: autoregression composes only with token-level representation, while the LLM-oracle composes with whole-report or statement-level representation (Section 4.4). The joint (Π, R) axis therefore has three valid settings, yielding a $2 \times 2 \times 3 = 12$-cell factorial design (Table 9). Of the 12 cells, three correspond to previously proposed methods (GEM, GPPM-Judgment, and TVD-MI). By holding three axes fixed and varying one, we can isolate the marginal effect of each design choice on alignment, degradation sensitivity, and manipulation resistance.

Table 9: Coverage of the design space. Each cell is labeled by the name of the corresponding scoring method, where applicable. NEW denotes combinations introduced or evaluated in this work;

		Autoregression	**LLM-oracle**	
F	E	Token-representation	Report-representation	Statement-representation
KL	Direct	GEM [46]	NEW	GPPM-J [31]
	f-variational	NEW	NEW	NEW
TV	Direct	NEW	NEW	NEW
	f-variational	NEW	TVD-MI [41]	NEW

Implementation. We instantiate the token-autoregressive model with `Llama-3.1-8B`, and adopt the pre-processing process from Lu et al. [31], Xu et al. [46] with `claude-haiku-4.5`.

We instantiate the LLM-oracle with `claude-haiku-4.5`, queried through OpenRouter API at temperature 0 with a single sample per prompt. In the f-variational settings, the oracle output is used directly as the critic $T(x, y)$. In the direct-estimation settings, the oracle returns a 7-point ordinal supporting judgment, which we map to $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$ and treat as a discretized surrogate for the density ratio estimate $\hat{r}(x, y)$. For statement-level representataion, we first decompose the reference y into atomic claims using a dataset-specific decomposition prompt, and then score the resulting (candidate, reference statement) pairs with the same claude-haiku-4.5 model.

Unless otherwise noted, each candidate is scored against one positive reference, and, for f-variational metrics, four negative references sampled from other tasks in the same dataset group and takes average to reduce noise. A group consists of same-kind tasks within a dataset, such as the same class in Peer Grading-WH or the same topic category (e.g., math or technology) in LFQA-E. ICLR 2026, Peer Grading-XLSK, SummEval, and SPACE are treated as a single group.

Datasets We evaluate on three domains: peer review, summarization, and question answering. These domains vary in report length, structure, subjectivity, semantic depth, which stresses different parts of the scoring pipeline. In peer review, the task is a paper or an essay and reports are full reviews; in summarization, the task is a source document and reports are summaries; in question answering, the task is a question and reports are answers. Table 10 summarizes the datasets.[2]

When a dataset provides multiple independent responses for the same task, we treat them as positive references. Note that not all datasets carry human quality annotations; those that do are used for alignment evaluation, while the remainder are used only for degradation, manipulation, and model-differentiation tests.

D.2 Computational Resources

All experiments were conducted on a high-performance workstation with the following specifications:

[2]Peer grading-WH [44] and Peer grading-XLSK [46] are not public accessible while the others are public datasets.

Table 10: Overview of evaluation datasets. n is the number of candidate reports; #ref is the typical number of positive references per task. "Human rating" indicates whether item-level human quality annotations are available for candidate reports.

Domain	Dataset	n	Ref	Human rating	Reference
Peer review	ICLR 2026	300*	peer	No	OpenReview
	Peer grading-WH	534	GT	Numerical	Wu and Hartline [44]
	Peer grading-XLSK	165	peer	Numerical	Xu et al. [46]
Summarization	SummEval	800*	GT	Numerical	Fabbri et al. [12]
	SPACE	50	GT	No	Angelidis et al. [3]
QA	LFQA-E	320*	GT	Pairwise	Fan et al. [13]
	MedAESQA	400*	GT	Numerical	Gupta et al. [18]

* Down-sampled from the whole dataset for a more trackable test scale.

Hardware. The machine is equipped with dual Intel Xeon Platinum 8470Q CPUs (totaling 104 physical cores and 208 threads), 1.0 TiB of system RAM, and one NVIDIA RTX PRO 6000 (Blackwell architecture) GPU with 96 GB of device memory.

Local Inference. For metrics using the Llama-3.1-8B model, local inference was performed with an average processing speed of approximately 0.5s/it. The total local compute time required for the entire study (excluding API latency) was approximately 6 GPU hours.

API Usage and Reliability. Evaluations involving Claude-Haiku-4.5 were conducted via the OpenRouter API.

- Configuration: All API calls used a default max token limit of 4000 and a temperature setting of 0 to ensure deterministic and complete responses.
- Retry Logic: To handle potential network instability or API timeouts, we implemented a robust retry mechanism with a maximum of 4 attempts per request; a call was marked as a failure only if all retries were exhausted.
- Error Handling: In cases where specific perturbation methods (for degradation or manipulation tests) encountered API errors or generation failures, the system was designed to gracefully fall back to a no-op (no operation), ensuring the stability of the overall evaluation pipeline.

Execution Time. Due to varying dataset sizes, the total experimental duration per dataset ranged from 20 minutes to 1 hour.

E Detailed Degradation and Manipulation Statistics

Table 11: Degradation perturbation method results. Each cell reports the standardized mean difference d with 95% CI. Red marks significant score increases ($p > 0.05$); orange marks non-significant changes. Both flag failed degradation criteria.

Metric	PG-WH	PG-XLSK	ICLR26	SummEval	SPACE	LFQA-E	MedAESQA
			`deletion_and_completion`				
MI-based Mechanisms							
KL-Direct-Autoreg.	-0.45 ± 0.09	-0.32 ± 0.08	-0.24 ± 0.07	-0.41 ± 0.09	-0.50 ± 0.11	-0.10 ± 0.09	-0.21 ± 0.09
TVD-FVar-Report	-0.04 ± 0.16	-0.19 ± 0.11	-0.20 ± 0.12	-0.23 ± 0.15	-0.41 ± 0.13	-0.22 ± 0.14	$+0.19\pm0.14$
TVD-FVar-Statement	-0.20 ± 0.10	-0.23 ± 0.11	-0.18 ± 0.12	-0.48 ± 0.10	-0.31 ± 0.09	-0.19 ± 0.09	-0.19 ± 0.10
Non-MI baselines							
ROUGE-L	$+0.06\pm0.09$	$+0.06\pm0.04$	-0.55 ± 0.08	-0.05 ± 0.07	-0.27 ± 0.07	-0.24 ± 0.07	-0.14 ± 0.06
BLEU	$+0.24\pm0.07$	$+0.05\pm0.06$	-0.35 ± 0.06	$+0.03\pm0.06$	-0.24 ± 0.09	-0.16 ± 0.08	-0.10 ± 0.05
BERTScore	$+0.04\pm0.18$	$+0.09\pm0.11$	$+0.56\pm0.09$	-0.05 ± 0.10	-0.32 ± 0.09	$+0.04\pm0.08$	-0.04 ± 0.06

Continued on next page

Table 11: Degradation perturbation method results (continued). Each cell reports the standardized mean difference d with 95% CI. Red marks significant score increases ($p > 0.05$); orange marks non-significant changes. Both flag failed degradation criteria.

Metric	PG-WH	PG-XLSK	ICLR26	SummEval	SPACE	LFQA-E	MedAESQA
LLM-J / Claude-Haiku-4.5	−0.35±0.11	−0.19±0.09	−0.27±0.11	−0.29±0.14	−0.44±0.13	−0.37±0.13	−0.01±0.09
LLM-J / Claude-Sonnet-4.5	−0.37±0.10	−0.16±0.09	−0.28±0.10	−0.42±0.11	−0.61±0.13	−0.39±0.11	−0.12±0.08
LLM-J / GPT-5-mini	−0.24±0.10	−0.13±0.08	−0.32±0.12	−0.29±0.12	−0.52±0.12	−0.17±0.11	−0.09±0.09
LLM-J / GPT-4o-mini	+0.06±0.11	+0.18±0.07	−0.13±0.12	−0.15±0.12	−0.12±0.12	−0.13±0.11	−0.10±0.06
`opinion_flip`							
MI-based Mechanisms							
KL-Direct-Autoreg.	−0.39±0.12	−0.51±0.12	−0.44±0.09	–	–	–	–
TVD-FVar-Report	−0.21±0.16	−0.01±0.11	−0.05±0.14	–	–	–	–
TVD-FVar-Statement	−0.84±0.19	−0.27±0.17	−0.28±0.15	–	–	–	–
Non-MI baselines							
ROUGE-L	+0.28±0.13	−0.85±0.10	+0.13±0.09	–	–	–	–
BLEU	+0.63±0.11	−0.64±0.12	−0.16±0.12	–	–	–	–
BERTScore	−0.06±0.19	−0.35±0.14	+0.39±0.12	–	–	–	–
LLM-J / Claude-Haiku-4.5	−0.88±0.21	−0.67±0.20	−0.58±0.18	–	–	–	–
LLM-J / Claude-Sonnet-4.5	−0.99±0.21	−0.70±0.18	−0.68±0.17	–	–	–	–
LLM-J / GPT-5-mini	−1.09±0.22	−0.32±0.15	−0.54±0.17	–	–	–	–
LLM-J / GPT-4o-mini	−0.26±0.19	+0.22±0.19	+0.06±0.17	–	–	–	–
`random_replacement`							
MI-based Mechanisms							
KL-Direct-Autoreg.	−0.32±0.18	−1.12±0.22	−1.82±0.18	−2.00±0.18	−1.05±0.18	−1.36±0.17	−0.67±0.17
TVD-FVar-Report	−0.48±0.18	−1.88±0.23	−8.98±0.20	−2.60±0.21	−1.42±0.20	−2.93±0.19	−3.00±0.21
TVD-FVar-Statement	−0.81±0.18	−0.92±0.19	−1.62±0.18	−2.10±0.20	−1.01±0.18	−1.96±0.18	−2.61±0.19
Non-MI baselines							
ROUGE-L	+0.01±0.18	−0.34±0.21	−1.15±0.16	−2.06±0.18	−1.03±0.19	−1.18±0.17	−1.95±0.18
BLEU	+0.01±0.17	−0.23±0.22	−0.34±0.15	−1.07±0.19	−0.96±0.20	−0.92±0.19	−0.94±0.20
BERTScore	+0.06±0.19	−0.28±0.19	−0.57±0.15	−2.36±0.20	−0.86±0.15	−1.89±0.17	−1.98±0.20
LLM-J / Claude-Haiku-4.5	−0.61±0.19	−1.45±0.22	−1.25±0.20	−2.49±0.20	−1.13±0.18	−2.72±0.20	−2.95±0.19
LLM-J / Claude-Sonnet-4.5	−0.66±0.19	−1.67±0.22	−3.99±0.20	−3.18±0.20	−1.22±0.19	−2.81±0.20	−2.95±0.20
LLM-J / GPT-5-mini	−0.56±0.20	−1.23±0.22	−3.47±0.20	−3.01±0.20	−1.03±0.19	−3.10±0.20	−3.17±0.20
LLM-J / GPT-4o-mini	−0.34±0.20	−0.11±0.23	+0.08±0.20	−3.74±0.20	−0.64±0.18	−3.10±0.19	−2.49±0.19
`sentence_deletion`							
MI-based Mechanisms							
KL-Direct-Autoreg.	−0.46±0.09	−0.64±0.08	−0.70±0.07	−0.34±0.10	−0.56±0.11	−0.25±0.09	−0.24±0.10
TVD-FVar-Report	−0.05±0.16	−0.20±0.11	−0.21±0.15	−0.16±0.15	−0.31±0.12	−0.23±0.14	−0.03±0.15
TVD-FVar-Statement	−0.25±0.11	−0.22±0.11	−0.38±0.10	−0.50±0.10	−0.27±0.08	−0.18±0.09	−0.45±0.11
Non-MI baselines							
ROUGE-L	−0.04±0.08	+0.43±0.08	−0.25±0.10	−0.02±0.08	−0.10±0.08	−0.44±0.10	−0.10±0.08
BLEU	−0.19±0.11	+0.39±0.12	−0.67±0.13	+0.01±0.07	−0.22±0.12	−0.56±0.12	−0.27±0.10
BERTScore	+0.05±0.19	−0.16±0.10	+0.12±0.10	−0.17±0.10	−0.16±0.08	−0.04±0.08	−0.11±0.06
LLM-J / Claude-Haiku-4.5	−0.39±0.13	−0.71±0.11	−0.88±0.15	−0.37±0.14	−0.53±0.13	−0.86±0.14	−0.34±0.11
LLM-J / Claude-Sonnet-4.5	−0.43±0.11	−0.80±0.09	−1.40±0.14	−0.42±0.12	−0.61±0.14	−0.91±0.14	−0.37±0.10
LLM-J / GPT-5-mini	−0.43±0.10	−0.73±0.08	−1.11±0.14	−0.44±0.12	−0.63±0.12	−0.70±0.13	−0.41±0.10
LLM-J / GPT-4o-mini	−0.59±0.12	−0.73±0.09	−1.04±0.16	−0.54±0.13	−0.58±0.14	−1.01±0.15	−0.60±0.10
`surface_report`							
MI-based Mechanisms							
KL-Direct-Autoreg.	–	−0.74±0.20	−0.02±0.14	−0.28±0.15	−1.22±0.17	+0.04±0.11	−0.08±0.13
TVD-FVar-Report	–	−1.36±0.22	−0.12±0.21	−0.09±0.16	−1.02±0.18	+0.02±0.17	−0.02±0.19
TVD-FVar-Statement	–	−0.58±0.17	−0.31±0.18	+0.04±0.15	−0.84±0.18	+0.11±0.12	−0.01±0.18
Non-MI baselines							
ROUGE-L	–	−3.32±0.21	−0.93±0.15	−0.11±0.12	−2.46±0.20	−0.14±0.11	−0.45±0.16
BLEU	–	−2.84±0.21	−1.28±0.15	+0.68±0.16	−1.34±0.20	−0.03±0.13	−0.56±0.17
BERTScore	–	+0.10±0.18	+0.84±0.13	+0.28±0.14	−2.81±0.17	−0.03±0.13	−0.11±0.16
LLM-J / Claude-Haiku-4.5	–	−1.57±0.22	+0.18±0.17	+0.47±0.17	−1.81±0.19	+0.81±0.16	+1.21±0.19
LLM-J / Claude-Sonnet-4.5	–	−1.65±0.22	+0.05±0.18	+0.28±0.15	−2.51±0.20	+0.69±0.16	+1.18±0.20
LLM-J / GPT-5-mini	–	−1.52±0.22	+0.03±0.17	+0.12±0.17	−1.28±0.19	+0.51±0.14	+1.01±0.19
LLM-J / GPT-4o-mini	–	−1.11±0.22	+1.13±0.17	+1.39±0.17	−0.74±0.18	+1.04±0.17	+0.98±0.19
`ultra_concise_compression`							
MI-based Mechanisms							
KL-Direct-Autoreg.	−0.61±0.12	−1.16±0.12	−1.43±0.13	−0.89±0.15	−1.31±0.17	−0.10±0.11	+0.01±0.13
TVD-FVar-Report	−0.12±0.18	−0.59±0.15	+0.01±0.19	+0.19±0.15	−0.52±0.15	−0.03±0.16	+0.34±0.17
TVD-FVar-Statement	−0.24±0.13	−0.28±0.13	−0.29±0.14	−1.10±0.14	−0.59±0.13	−0.27±0.11	−0.20±0.14
Non-MI baselines							
ROUGE-L	−2.43±0.17	−4.80±0.25	−3.68±0.18	−0.90±0.15	−1.42±0.17	−3.06±0.16	−1.77±0.16
BLEU	−2.18±0.20	−3.24±0.23	−3.78±0.20	−1.26±0.20	−1.25±0.20	−1.77±0.20	−1.47±0.20
BERTScore	+0.19±0.19	+0.81±0.15	+1.33±0.11	+0.19±0.14	−0.33±0.14	+0.03±0.11	−0.06±0.11
LLM-J / Claude-Haiku-4.5	−0.33±0.11	−0.96±0.13	−0.37±0.13	−0.40±0.16	−0.75±0.15	−0.90±0.16	−0.76±0.14
LLM-J / Claude-Sonnet-4.5	−0.39±0.11	−1.02±0.12	−0.96±0.14	−0.79±0.14	−0.88±0.14	−1.03±0.16	−0.63±0.13

Continued on next page

Table 11: Degradation perturbation method results (continued). Each cell reports the standardized mean difference d with 95% CI. Red marks significant score increases ($p > 0.05$); orange marks non-significant changes. Both flag failed degradation criteria.

Metric	PG-WH	PG-XLSK	ICLR26	SummEval	SPACE	LFQA-E	MedAESQA
LLM-J / GPT-5-mini	−0.18±0.10	−0.88±0.11	−0.99±0.13	−0.54±0.13	−0.84±0.14	−0.79±0.13	−0.62±0.14
LLM-J / GPT-4o-mini	−0.85±0.14	−1.03±0.11	−1.02±0.18	−1.42±0.18	−1.15±0.16	−1.70±0.16	−1.40±0.12

Table 12: Manipulation perturbation method results. Each cell reports the standardized mean difference d with 95% CI. Red marks significant score increases ($p > 0.05$), which flags failed manipulation criteria.

Metric	PG-WH	PG-XLSK	ICLR26	SummEval	SPACE	LFQA-E	MedAESQA
`meaningless_elongation`							
MI-based Mechanisms							
KL-Direct-Autoreg.	+0.06±0.05	−0.01±0.05	+0.00±0.02	−0.01±0.03	+0.01±0.04	+0.00±0.05	−0.05±0.04
TVD-FVar-Report	+0.05±0.14	−0.03±0.09	−0.02±0.13	+0.12±0.13	−0.16±0.11	−0.04±0.13	+0.05±0.12
TVD-FVar-Statement	+0.05±0.06	−0.03±0.08	−0.03±0.07	+0.01±0.05	+0.02±0.06	+0.05±0.07	−0.03±0.05
Non-MI baselines							
ROUGE-L	+0.48±0.14	−1.59±0.15	−0.33±0.02	−0.62±0.07	−0.84±0.09	−0.08±0.03	−0.13±0.03
BLEU	+0.41±0.16	−1.44±0.17	−0.01±0.04	−0.53±0.08	−0.39±0.08	+0.03±0.02	+0.03±0.03
BERTScore	+0.10±0.18	−0.12±0.08	+0.00±0.04	−0.02±0.04	+0.03±0.06	+0.00±0.04	−0.01±0.03
LLM-J / Claude-Haiku-4.5	+0.15±0.08	−0.28±0.08	−0.02±0.07	+0.05±0.09	−0.37±0.10	−0.05±0.06	−0.11±0.05
LLM-J / Claude-Sonnet-4.5	+0.10±0.06	−0.77±0.09	+0.14±0.06	+0.04±0.08	−0.52±0.11	−0.08±0.06	−0.09±0.04
LLM-J / GPT-5-mini	+0.04±0.07	+0.10±0.06	+0.05±0.10	−0.02±0.08	−0.05±0.09	+0.06±0.09	−0.01±0.08
LLM-J / GPT-4o-mini	+0.30±0.11	+0.43±0.09	+0.24±0.09	−0.25±0.13	−0.37±0.13	+0.00±0.07	−0.04±0.05
`opinion_shift_negative`							
MI-based Mechanisms							
KL-Direct-Autoreg.	+0.01±0.10	−0.03±0.10	+0.29±0.08	–	–	–	–
TVD-FVar-Report	+0.03±0.15	−0.05±0.10	−0.03±0.13	–	–	–	–
TVD-FVar-Statement	−0.19±0.14	−0.20±0.13	−0.20±0.10	–	–	–	–
Non-MI baselines							
ROUGE-L	−0.55±0.17	−3.34±0.22	−1.10±0.11	–	–	–	–
BLEU	−0.12±0.19	−2.90±0.22	−1.00±0.15	–	–	–	–
BERTScore	−0.06±0.02	−0.25±0.16	+0.42±0.11	–	–	–	–
LLM-J / Claude-Haiku-4.5	+0.27±0.13	−0.25±0.18	+0.07±0.15	–	–	–	–
LLM-J / Claude-Sonnet-4.5	+0.16±0.13	−0.35±0.18	+0.12±0.15	–	–	–	–
LLM-J / GPT-5-mini	+0.34±0.14	+0.34±0.15	+0.10±0.14	–	–	–	–
LLM-J / GPT-4o-mini	+0.50±0.16	+0.24±0.17	+0.63±0.15	–	–	–	–
`opinion_shift_neutral`							
MI-based Mechanisms							
KL-Direct-Autoreg.	−0.23±0.08	−0.03±0.07	+0.13±0.06	–	–	−0.01±0.09	−0.09±0.08
TVD-FVar-Report	+0.05±0.15	+0.01±0.10	−0.05±0.16	–	–	+0.00±0.13	−0.04±0.15
TVD-FVar-Statement	−0.04±0.08	−0.13±0.10	−0.19±0.08	–	–	+0.00±0.09	−0.08±0.07
Non-MI baselines							
ROUGE-L	+0.19±0.14	−1.42±0.18	−1.20±0.12	–	–	−0.20±0.09	−0.19±0.09
BLEU	+0.58±0.14	−1.83±0.20	−2.14±0.17	–	–	−0.10±0.10	−0.09±0.07
BERTScore	−0.02±0.18	−0.60±0.13	+0.18±0.10	–	–	−0.01±0.09	−0.11±0.05
LLM-J / Claude-Haiku-4.5	+0.16±0.08	+0.56±0.12	+0.59±0.13	–	–	+0.48±0.13	+0.27±0.08
LLM-J / Claude-Sonnet-4.5	+0.13±0.08	+0.44±0.14	+0.54±0.11	–	–	+0.30±0.13	+0.10±0.07
LLM-J / GPT-5-mini	+0.12±0.08	+0.52±0.09	+0.40±0.12	–	–	+0.39±0.13	+0.11±0.07
LLM-J / GPT-4o-mini	+0.74±0.12	+1.35±0.15	+1.09±0.17	–	–	+0.79±0.15	+0.17±0.07
`opinion_shift_positive`							
MI-based Mechanisms							
KL-Direct-Autoreg.	−0.41±0.08	−0.15±0.07	+0.13±0.07	–	–	–	–
TVD-FVar-Report	−0.06±0.14	−0.01±0.10	−0.07±0.13	–	–	–	–
TVD-FVar-Statement	−0.14±0.10	−0.13±0.10	+0.02±0.09	–	–	–	–
Non-MI baselines							
ROUGE-L	−0.29±0.16	−2.04±0.20	−0.59±0.08	–	–	–	–
BLEU	+0.19±0.18	−2.00±0.21	−0.46±0.11	–	–	–	–
BERTScore	−0.13±0.02	−0.56±0.13	+0.03±0.09	–	–	–	–
LLM-J / Claude-Haiku-4.5	−0.15±0.10	+0.38±0.12	+0.01±0.15	–	–	–	–
LLM-J / Claude-Sonnet-4.5	−0.19±0.11	+0.32±0.14	+0.18±0.12	–	–	–	–
LLM-J / GPT-5-mini	+0.00±0.10	+0.54±0.09	+0.20±0.12	–	–	–	–
LLM-J / GPT-4o-mini	+1.09±0.13	+1.74±0.15	+1.23±0.18	–	–	–	–
`opinion_shift_strong`							
MI-based Mechanisms							
KL-Direct-Autoreg.	+0.01±0.07	+0.12±0.08	+0.14±0.05	–	–	+0.04±0.09	−0.09±0.09
TVD-FVar-Report	+0.01±0.14	−0.04±0.11	−0.05±0.14	–	–	−0.15±0.14	+0.02±0.13

Continued on next page

Table 12: Manipulation perturbation method results (continued). Each cell reports the standardized mean difference d with 95% CI. Red marks significant score increases ($p > 0.05$), which flags failed manipulation criteria.

Metric	PG-WH	PG-XLSK	ICLR26	SummEval	SPACE	LFQA-E	MedAESQA
TVD-FVar-Statement	$+0.00 \pm 0.07$	-0.04 ± 0.10	-0.04 ± 0.08	–	–	-0.01 ± 0.09	$+0.01 \pm 0.07$
Non-MI baselines							
ROUGE-L	$+0.45 \pm 0.12$	-2.91 ± 0.19	-1.18 ± 0.10	–	–	-0.25 ± 0.09	-0.07 ± 0.09
BLEU	$+0.64 \pm 0.13$	-2.81 ± 0.21	-2.21 ± 0.16	–	–	-0.16 ± 0.12	-0.13 ± 0.07
BERTScore	$+0.00 \pm 0.02$	-0.39 ± 0.14	$+0.25 \pm 0.10$	–	–	$+0.01 \pm 0.09$	-0.07 ± 0.05
LLM-J / Claude-Haiku-4.5	$+0.14 \pm 0.09$	$+0.30 \pm 0.13$	$+0.49 \pm 0.13$	–	–	$+0.40 \pm 0.12$	$+0.24 \pm 0.08$
LLM-J / Claude-Sonnet-4.5	$+0.13 \pm 0.09$	$+0.30 \pm 0.18$	$+0.47 \pm 0.11$	–	–	$+0.18 \pm 0.13$	$+0.05 \pm 0.07$
LLM-J / GPT-5-mini	$+0.11 \pm 0.08$	$+0.56 \pm 0.11$	$+0.32 \pm 0.11$	–	–	$+0.24 \pm 0.11$	$+0.09 \pm 0.09$
LLM-J / GPT-4o-mini	$+0.52 \pm 0.11$	$+1.26 \pm 0.15$	$+1.08 \pm 0.17$	–	–	$+0.62 \pm 0.15$	$+0.14 \pm 0.07$
			`rephrase`				
MI-based Mechanisms							
KL-Direct-Autoreg.	-0.06 ± 0.07	-0.01 ± 0.06	$+0.07 \pm 0.05$	-0.09 ± 0.06	-0.35 ± 0.08	$+0.01 \pm 0.08$	-0.15 ± 0.09
TVD-FVar-Report	-0.01 ± 0.16	-0.02 ± 0.10	-0.11 ± 0.13	-0.11 ± 0.14	-0.07 ± 0.11	-0.16 ± 0.13	$+0.08 \pm 0.15$
TVD-FVar-Statement	$+0.00 \pm 0.07$	-0.12 ± 0.09	-0.14 ± 0.08	-0.04 ± 0.06	-0.02 ± 0.07	-0.04 ± 0.08	-0.08 ± 0.06
Non-MI baselines							
ROUGE-L	$+0.47 \pm 0.13$	-1.97 ± 0.14	-0.98 ± 0.09	-0.09 ± 0.07	-0.62 ± 0.12	-0.16 ± 0.07	-0.21 ± 0.09
BLEU	$+0.68 \pm 0.13$	-2.53 ± 0.20	-2.09 ± 0.15	$+0.54 \pm 0.14$	-0.65 ± 0.14	-0.11 ± 0.09	-0.16 ± 0.07
BERTScore	-0.02 ± 0.02	-0.18 ± 0.10	$+0.09 \pm 0.09$	$+0.13 \pm 0.09$	-0.58 ± 0.12	$+0.02 \pm 0.08$	-0.16 ± 0.06
LLM-J / Claude-Haiku-4.5	$+0.14 \pm 0.09$	$+0.18 \pm 0.09$	$+0.49 \pm 0.11$	$+0.29 \pm 0.10$	$+0.14 \pm 0.08$	$+0.25 \pm 0.11$	$+0.24 \pm 0.07$
LLM-J / Claude-Sonnet-4.5	$+0.13 \pm 0.08$	$+0.19 \pm 0.09$	$+0.25 \pm 0.09$	$+0.28 \pm 0.09$	$+0.00 \pm 0.09$	$+0.01 \pm 0.12$	$+0.02 \pm 0.07$
LLM-J / GPT-5-mini	$+0.09 \pm 0.08$	$+0.16 \pm 0.07$	$+0.17 \pm 0.10$	$+0.07 \pm 0.08$	$+0.01 \pm 0.08$	$+0.16 \pm 0.11$	$+0.05 \pm 0.08$
LLM-J / GPT-4o-mini	$+0.52 \pm 0.11$	$+0.82 \pm 0.10$	$+0.81 \pm 0.15$	$+0.65 \pm 0.12$	$+0.36 \pm 0.11$	$+0.40 \pm 0.13$	$+0.21 \pm 0.07$