Statistical Multicriteria Evaluation of LLM-Generated Text

Esteban Garces Arias[1,2], Hannah Blocher[1], Julian Rodemann[1,3], Matthias Aßenmacher[1,2], Christoph Jansen[4]

[1]Department of Statistics, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML)
[3]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[4]School of Computing & Communications, Lancaster University Leipzig, Germany

Correspondence: Esteban.GarcesArias@stat.uni-muenchen.de

Abstract

Assessing the quality of LLM-generated text remains a fundamental challenge in natural language processing. Current evaluation approaches often rely on isolated metrics or simplistic aggregations that fail to capture the nuanced trade-offs between coherence, diversity, fluency, and other relevant indicators of text quality. In this work, we adapt a recently proposed framework for statistical inference based on Generalized Stochastic Dominance (GSD) that addresses three critical limitations in existing benchmarking methodologies: the inadequacy of single-metric evaluation, the incompatibility between cardinal automatic metrics and ordinal human judgments, and the lack of inferential statistical guarantees. The GSD-front approach enables simultaneous evaluation across multiple quality dimensions while respecting their different measurement scales, building upon partial orders of decoding strategies, thus avoiding arbitrary weighting of the involved metrics. By applying this framework to evaluate common decoding strategies against human-generated text, we demonstrate its ability to identify statistically significant performance differences while accounting for potential deviations from the i.i.d. assumption of the sampling design.

1 Introduction

Large language models (LLMs; Achiam et al., 2023; Grattafiori et al., 2024; Guo et al., 2025) rely on decoding strategies—algorithms that select subsequent tokens based on probability distributions over the vocabulary. As these models advance, numerous decoding methods have emerged, including deterministic (Freitag and Al-Onaizan, 2017; Su et al., 2022; Garces Arias et al., 2024) and stochastic approaches (Fan et al., 2018; Holtzman et al., 2019; Ding et al., 2025), necessitating robust benchmarking protocols for systematic evaluation. In this work, we present a methodological contribution: we adapt the Generalized Stochastic Dominance (GSD) framework (Jansen et al., 2024) to evaluate LLM-generated text. Rather than exhaustively comparing all decoding strategies, we demonstrate how GSD enables rigorous multicriteria evaluation while preserving distinct measurement scales. We focus on open-ended text generation as an illustrative example, though the framework generalizes to other tasks.

Current benchmarking for open-ended text generation typically uses curated datasets like WikiText (Merity et al., 2016) or WikiNews to assess decoding strategies through automatic metrics and human evaluation. While valuable for practitioners and researchers, these methodologies face three fundamental challenges:

(I) Reliance on Single Metrics. Text quality is inherently multidimensional, yet conventional benchmarking often reduces it to single metrics. Despite established metrics like perplexity, diversity, and coherence capturing different quality aspects (Holtzman et al., 2019; Su et al., 2022; Garces Arias et al., 2025b), Figure 1(a) shows that while 88.0% of papers evaluate multiple metrics individually (2024), only 7.6% employ true multicriteria approaches.

(II) Integrating Human Evaluation. Combining automatic metrics (cardinal measurements) with human assessments (ordinal data) poses methodological challenges. Figure 1(b) shows increasing adoption of combined evaluation (39.5% to 45.2%, 2022-2024), yet integrating these distinct measurement scales remains problematic.

(III) Lack of Statistical Rigor. Most evaluations remain descriptive without statistical validation. Figure 1(c) reveals declining use of statistical inference (32.6% to 28.0%), limiting generalizability beyond specific benchmark suites.

Contributions. We introduce a GSD-based benchmarking framework that:

Proceedings of the 18th International Natural Language Generation Conference, pages 338–351
October 29–November 2, 2025.

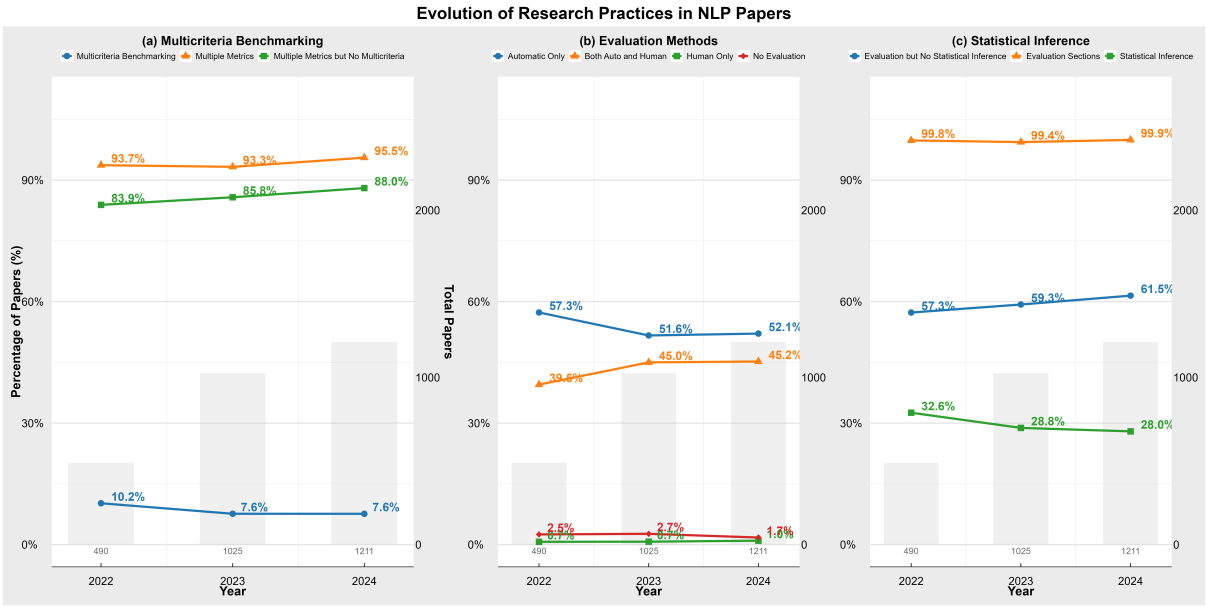

Figure 1: Analysis of text generation research trends (arXiv, 2022-2024): (a) Evolution of multicriteria benchmarking showing individual metric evaluation (orange) versus true multicriteria approaches (blue). (b) Distribution of evaluation methodologies across automatic (blue), human (green), combined (orange), and no evaluation (red) approaches. (c) Adoption of statistical inference methods (green) compared to evaluations without statistical validation (blue). **Remark:** Note that arXiv trends, as opposed to the ACL Anthology, might include non-peer-reviewed papers; however, we have decided to prioritize timeliness from a broader research community.

(I) Incorporates multiple quality metrics simultaneously

(II) Integrates human and automatic evaluations while preserving measurement scales

(III) Provides statistical inference beyond descriptive evaluation

(IV) Quantifies robustness when i.i.d. assumptions are violated

We validate our approach using WikiText and WikiNews prompts, comparing five decoding strategies (beam search, contrastive search, temperature sampling, top-k, and nucleus sampling) against human completions. Our evaluation combines Q*Text (Garces Arias et al., 2025b) with human assessments, demonstrating that human-written text maintains superior or equivalent quality. Code and data are publicly available[1].

2 Generalized Stochastic Dominance and the GSD-Front

To address the challenges outlined in Section 1, we propose a framework based on generalized stochastic dominance (GSD). Rather than imposing a complete ranking—which would require potentially unjustified assumptions about the relative importance of quality dimensions—our framework identifies a minimal set of non-dominated strategies. This set, which we term the "GSD-front," represents potentially optimal choices under varying preference structures. In this section, we present the theoretical foundation of our approach.

2.1 Generalized Stochastic Dominance

GSD has received quite some interest recently (e.g. Jansen et al., 2023b,a, 2024; Jansen, 2025). We propose adapting the GSD-front, introduced in Jansen et al. (2024), for classifier selection, as a method to compare decoding strategies across multiple quality metrics simultaneously. The basic idea is quite natural: We first utilize the multidimensional order structure spanned by the quality metrics for defining a *partial expectation ranking* among the decoding strategies under consideration. In our application, these are Q*Text and the two human evaluations. Afterwards, we select the non-strictly dominated strategies under this order to be included in the GSD-front. In our application, we consider all decoding strategies together with the human completion and select those that are not strictly

[1] https://github.com/hannahblo/Statistical_Multicriteria_Evaluation_of_LLM-Generated_Text

dominated (i.e., systematically worse) than any of the others. Hence, if the human completion lies in the GSD-front, it is not dominated by any of the other five automatic decoding strategies and therefore can potentially produce higher quality text than those in certain situations. Note that the decoding strategies in the GSD-front are incomparable to each other (GSD is a partial order) and, in general, no unique best decoding strategy will be obtained in this way. However, it can be argued that the GSD-front represents the smallest set of incomparable strategies that can be obtained without (potentially hard-to-justify) additional assumptions about the weighting of the quality metrics since it incorporates the entire information encoded in both the (empirical) distribution of the prompts and the order structure induced by the metrics.

2.2 Technical Setup

To ensure clarity throughout our technical exposition, we first establish our notation:

Symbol	Description
$\mathcal{D}$	Set of decoding strategies
$S, S', L, L' \in \mathcal{D}$	Individual decoding strategies
$\mathcal{P}$	Universe of prompts
$P, P', G, G' \in \mathcal{P}$	Individual prompts
ϕ_i	Quality metric i
$\Phi = (\phi_1, \ldots, \phi_n)$	Multidimensional metric vector
R_1	Ordinal relation on quality vectors
R_2	Cardinal relation on pairs in R_1
π	Probability measure over prompts
u	Utility function
$\mathcal{U}_{\mathbb{P}}$	Set of utility representations

Table 1: Notation overview for the technical setup.

Assume we are given a finite set $\mathcal{D}$ of decoding strategies, a universe $\mathcal{P}$ of prompts, and n quality metrics $\phi_1, \ldots, \phi_n : \mathcal{D} \times \mathcal{P} \to [0, 1]$. For every $i \in \{1, \ldots, n\}$, $S \in \mathcal{D}$, and $P \in \mathcal{P}$, the value $\phi_i(S, P)$ describes quality of the completion obtained by applying strategy S to prompt P with respect to the metric ϕ_i (where higher values indicate better quality). To clearly distinguish between ordinal and cardinal evaluations, we assume that, for $0 \leq z \leq n$, the metrics $\phi_1, \ldots, \phi_z$ are of cardinal scale (differences may be interpreted), while the remaining ones are purely ordinal (differences are meaningless apart from the sign). Given this setup, we then consider the multidimensional metric

$$\Phi := (\phi_1, \ldots, \phi_n) : \mathcal{D} \times \mathcal{P} \to [0, 1]^n.$$

We define two binary relations associated with the range $\Phi(\mathcal{D} \times \mathcal{P})$ of Φ, i.e., the set of *quality vectors* spanned by the considered quality metrics. The first of these relations captures the *ordinal information* encoded in the multidimensional quality evaluations, while the second relation captures the *cardinal* part of the information:

Ordinal Information: For any pair of quality vectors $x := \Phi(S, P)$, $y := \Phi(S', P')$, where $S, S' \in \mathcal{D}$ are decoding strategies and $P, P' \in \mathcal{P}$ are prompts, we define:

$$(x, y) \in R_1 \quad :\Leftrightarrow \quad \forall i : \phi_i(S, P) \geq \phi_i(S', P').$$

Under this specification, R_1 defines a binary relation – precisely a preorder – on the set $\Phi(\mathcal{D} \times \mathcal{P})$ of quality vectors. In words, if the two quality vectors x and y are in relation with respect to R_1, i.e., $(x, y) \in R_1$, this means that the completion of P by S is judged at least as good as the completion of P' by S' by any of the considered metrics.

Cardinal Information: For any quadruple of quality vectors $t := \Phi(S, P), u := \Phi(S', P'), v := \Phi(L, G), w := \Phi(L', G')$, where $S, S', L, L' \in \mathcal{D}$ are decoding strategies and $P, P', G, G' \in \mathcal{P}$ are prompts, that satisfies $(t, u), (v, w) \in R_1$, we set

$$((t, u), (v, w)) \in R_2 \quad :\Leftrightarrow \quad \forall i \leq z \, \forall j > z$$

$$\phi_i(S, P) - \phi_i(S', P') \geq \phi_i(L, G) - \phi_i(L', G') \wedge$$
$$\phi_j(S, P) \geq \phi_j(L, G) \geq \phi_j(L', G') \geq \phi_j(S', P')$$

Under this specification, R_2 defines a binary relation – precisely a preorder – on the relation R_1, i.e., on the set of all pairs of quality vectors that are comparable under R_1. In words, if two R_1-ordered pairs of quality vectors (t, u) and (v, w) are in relation with respect to R_2, this means that whenever the ordinal components of the latter quality vectors are bounded (from above and below) by the ordinal components of the further quality vectors, we can compare *intensity of preference* between quality vectors by comparing their *differences in the cardinal components*.

These considerations leave us with a partially-cardinal scaled order structure

$$\mathbb{P} = (\Phi(\mathcal{D} \times \mathcal{P}), R_1, R_2)$$

on the basis of which we intend to analyze the performance of the decoding strategies under consideration. Note that this structure encodes exactly that quality information that can be obtained from the data without additional assumptions about the weighting of the involved quality metrics. To ease working with $\mathbb{P}$, we replace it by the set of *utility*

representations respecting its structure. Intuitively, each of those utility functions can then be interpreted as a candidate measurement scale (or, in other words, a potential cardinal completion) that is compatible with the information encoded in $\mathbb{P}$, i.e., the information arising from the mixed-scaled multidimensional quality evaluations across the considered combination of decoding strategies and prompts in the set $\mathcal{D} \times \mathcal{P}$.

Utility Representation: We call a function

$$u : \Phi(\mathcal{D} \times \mathcal{P}) \to \mathbb{R}$$

compatible with, or *utility representation* of $\mathbb{P}$, whenever for all $(x, y) \in R_1$ it holds that

$$u(x) \geq u(y)$$

and for all $((r, t), (v, w)) \in R_2$ it holds that

$$u(r) - u(t) \geq u(v) - u(w)$$

Every function u satisfying those two properties respects both the order information encoded in R_1 and the intensity information encoded in R_2. We denote by $\mathcal{U}_{\mathbb{P}}$ the set of all (bounded and measurable) functions that are compatible with $\mathbb{P}$. This set then captures all the relevant information encoded in the structure $\mathbb{P}$, however, is much more accessible for a meaningful analysis.
The set $\mathcal{U}_{\mathbb{P}}$ of utility representations obtained from $\mathbb{P}$ now forms the basis for the generalized stochastic dominance (GSD) relation among the decoding strategies under consideration. Moreover, note that defining the GSD-relation on the set $\mathcal{D}$ requires assuming that the prompts in $\mathcal{P}$ are generated randomly according to some probability measure π (note that for our actual analysis, this will be replaced by its empirical analog).

Generalized Stochastic Dominance: We say that decoding strategy S GSD-dominates decoding strategy S', denoted by $S \gtrsim S'$, if it holds:

$$\forall u \in \mathcal{U}_{\mathbb{P}} : \ \mathbb{E}_{\pi}(u \circ \Phi(S, \cdot)) \geq \mathbb{E}_{\pi}(u \circ \Phi(S', \cdot))$$

In words, S GSD-dominates S', if the expected decoding quality of S is higher than that of S' *for no matter what* compatible utility measure $u \in \mathcal{U}_{\mathbb{P}}$ is used to summarize quality in a one-dimensional manner. Note that the GSD-relation $\gtrsim$ is not complete, i.e., in general, there will exist decoding strategies that are incomparable w.r.t. GSD.

The last step is adapting the GSD-front to the comparison of decoding strategies. Again, this can be done straightforwardly: We simply collect the non-strictly dominated strategies with respect to the GSD-relation $\gtrsim$ that we adapted to this context in the previous step.

GSD-Front: The GSD-front is thus given by

$$\text{gsd}(\mathcal{D}) = \{S \in \mathcal{D} : \ \nexists S' \in \mathcal{D} \text{ s.t. } S' \succ S\},$$

where $\succ$ denotes the strict part of $\gtrsim$.
Reflecting that $\text{gsd}(\mathcal{D})$ will, in general, be inaccessible since the true law π is unknown, in practice we will often have to make do with its empirical version, i.e., the set $\text{gsd}_{emp}(\mathcal{D})$ that is obtained by replacing all population-based expressions in $\text{gsd}(\mathcal{D})$ by their empirical analogs. Note, however, that $\text{gsd}_{emp}(\mathcal{D})$ makes a mere descriptive statement on the relation of the decoding strategies.

Empirical GSD-Front and Statistical Testing: To move to inferential guarantees, a statistical test for the pair

$$H_0 : S \notin \text{gsd}(\mathcal{D}) \ \textbf{vs.} \ H_1 : S \in \text{gsd}(\mathcal{D}) \quad (1)$$

is desirable: If H_0 can be rejected at a level α using an appropriate test, there is significant evidence that the decoding strategy S is competitive with the strategies in $\mathcal{D} \setminus \{S\}$ in certain situations *across the population of prompts* and, accordingly, should be further considered. But how can an appropriate statistical test be constructed? Jansen et al. (2024) demonstrate that (under mild assumptions that are met in our situation) a valid and consistent test is indeed reachable (by using an adapted permutation testing scheme). Furthermore, they show that this statistical test can be robustified to samples (slightly) deviating from the usual i.i.d. assumption by relying on techniques originating in robust statistics. In particular, their techniques allow us to analyze the p-value of a test decision for the pair (H_0, H_1) as a function of the *contamination size* of the underlying sample of prompts, i.e., the share of prompts stemming from some arbitrary distribution. In our context, such robustification seems particularly relevant: Especially when a large number of completions are evaluated by humans in a short period, certain implicit dependency structures are often difficult to avoid.
For interpreting the test results in Figure 2, it is important to note that the test proposed in Jansen et al. (2024) for the hypothesis pair (H_0, H_1) consists of

a series of *pairwise comparison tests* of strategies regarding their GSD relation. To be precise, the strategy S is tested against all strategies in $\mathcal{D} \setminus \{S\}$ and H_0 is rejected if all these sub-tests reject their respective null hypotheses. The test statistic used for each of those pairwise comparisons (S versus S') tests is based on the empirical version of

$$D(S, S') := \inf_{u \in \mathcal{U}_{\mathbb{P}}} \Big\{ \mathbb{E}_{\pi}(u \circ \Phi(S, \cdot)) - \mathbb{E}_{\pi}(u \circ \Phi(S', \cdot)) \Big\} \tag{2}$$

i.e., the expression arising from $D(S, S')$ by exchanging all population concepts by empirical analogs.

2.3 Intuitive Explanation for NLP Practitioners

To make the GSD framework more accessible, let us provide an intuitive understanding using familiar NLP concepts. Imagine you are comparing decoding strategies (e.g., beam search vs. nucleus sampling) across multiple metrics like coherence, diversity, and human ratings.

The Challenge: Traditional approaches either pick one "best" metric or combine metrics with arbitrary weights (e.g., 0.5×coherence+0.3×diversity+0.2×human_rating). But who decides these weights? Different applications might value these metrics differently.

The GSD Solution: Instead of forcing a complete ranking, GSD identifies strategies that are "not clearly worse" than others across all metrics. A strategy enters the GSD-front if there's no other strategy that beats it on *all* metrics simultaneously. For example:

- Strategy A: coherence=0.8, diversity=0.6, human=4.0
- Strategy B: coherence=0.7, diversity=0.9, human=3.5

Neither dominates the other—A wins on coherence and human rating, B wins on diversity. Both belong to the GSD-front.

Statistical Rigor: Beyond identifying the front, we provide statistical tests to determine whether these differences are significant or just sampling noise, accounting for the fact that we only evaluated a finite set of prompts.

2.4 Computational Complexity

Jansen et al. (2023c) demonstrated that computing Equation (2) can be reformulated as a mixed-integer programming (MIP) problem. While constructing the associated constraint matrix exhibits a worst-case time complexity of $O(n^4)$, practical implementations often achieve substantially lower complexity through problem-specific optimizations. For a detailed analysis of these computational improvements and their applicability conditions, we refer the reader to Jansen et al. (2023c).

3 Application: Automatic and Human Quality Evaluation

In this section, we present an application where we investigate whether human text generation potentially still offers superior quality compared to alternative automatic decoding strategies. Please note that our method can be applied to any set of competing decoding methods. This investigation addresses the three key benchmarking challenges outlined in Section 1: analyzing multiple quality metrics with different measurement scales simultaneously (Challenges I and II), and quantifying the robustness of inferential statements under potential deviations from the i.i.d. sampling assumption (Challenges III and IV). The latter arises specifically from potential dependencies in human evaluations and the use of prompts from two distinct datasets.

3.1 Experimental setup

Task Description. We demonstrate our method's application through an **open-ended text generation task**, more specifically, storytelling, where the model generates continuations for given prompts from Wikipedia and news articles. This task exemplifies the challenges in evaluating text quality across multiple dimensions, as generated continuations must balance coherence with the prompt, lexical diversity, and overall fluency.

In this demonstration, we employed a well-performing, medium-sized, open-source model: Qwen 2.5 - 7B (Yang et al., 2024), along with prompts from Wikitext (Merity et al., 2016) and Wikinews[2], incorporating diverse and factually-grounded contexts. Our sample encompassed 300 text generations—50 prompts (25 from WikiText, 25 from WikiNews) with six continuations each: one human-written (H) and five generated using

[2]Wikinews from http://www.wikinews.org

different decoding strategies. All generated texts and human-written texts were set to a constant length of 256 tokens (truncating human-written text when necessary). These strategies included both deterministic methods: beam search (BS) with beam width = 5 and contrastive search (CS) with $k = 10, \alpha = 0.6$, as well as stochastic approaches: temperature sampling (TS) with temperature = 0.9, top-k sampling (T_k) with $k = 50$, and nucleus top-p sampling (T_p) with $p = 0.95$. These hyperparameter choices follow the best-performing configurations reported by Garces Arias et al. (2025a). Detailed descriptions of these strategies appear in Table 3 in Appendix A.1.

Our evaluation framework integrated both human assessments and automated metrics. Human evaluators rated text quality on a 5-point Likert scale ranging from 1 (low quality) to 5 (high quality), see Table 4, following instructions detailed in Appendix A.4. Though the evaluators were authors of this paper, we implemented a blind evaluation protocol where they scored texts without knowledge of their source or the decoding strategy used, minimizing potential biases (Belz et al., 2020). We complemented these subjective judgments with cardinal Q*Text scores that synthesize generation perplexity, diversity, and coherence metrics as established by Garces Arias et al. (2025b). Specifically, Q*Text is computed as a weighted combination: Q*Text $= \frac{\sum_{i=1}^{3} w_i M_i P_i(M_i)}{\sum_{i=1}^{3} w_i}$, where M_1 is inverse-normalized perplexity, M_2 is coherence, M_3 is diversity, and P_i are Gaussian penalties that discourage extreme values. For a complete technical overview of this metric and its components, we refer to Section A.2.

3.2 Representation within the GSD framework

As previously stated, we compare human-generated text completion (H) to five decoding strategies: BS, CS, TS, T_k, and T_p based on prompts from the WikiText/WikiNews benchmark suites (see Section 3.1). The set of decoding strategies is defined as:

$$\mathcal{D} = \{\text{H, BS, CS, TS, T}_k, \text{T}_p\}, \tag{3}$$

whereas the set $\mathcal{P}$ represents the underlying population from which the prompts in WikiText/WikiNews are sampled. We evaluate text quality using three metrics: Q*Text (denoted as ϕ_1), which provides cardinal quality assessments, while metrics ϕ_2 and ϕ_3 are ordinal and based on evaluations from two of the paper's authors (hence, the number of cardinal dimensions is $z = 1$).

Following Section 2, we define two ranking relations on the set of quality vectors spanned by our (mixed-scaled) three-dimensional performance metric $\Phi = (\phi_1, \phi_2, \phi_3)$: R_1 is a (partial) order that ranks the quality vectors associated with our metric Φ based on the ordinal evaluations (ϕ_2, ϕ_3) of the completions for the prompts from $\mathcal{P}$ as well as the cardinal evaluation from the automated metric Q*Text (ϕ_1). R_2 is defined as a relation capturing the difference in intensity between pairs of quality vectors associated with the multidimensional metric Φ. As described in detail in Section 2, we can use these two relations R_1 and R_2 to obtain a ranking of the decoding strategies in $\mathcal{D}$ by applying (empirical) generalized stochastic dominance.

We use empirical GSD to represent these rankings and assess each decoding strategy's performance. The empirical GSD-front consists of all decoding strategies that are not strictly outperformed, i.e. *dominated* across all three metrics by others. This allows us to investigate whether human text completion enhances the quality of generated text compared to the five decoding strategies. For a comprehensive description of the quality assessment criteria, we refer to Table 4 in Appendix A.4.

3.3 Results

To examine whether human text generations can potentially improve on completion quality compared to the aforementioned automatic strategies (see Section 3.1), we conduct the statistical test for the GSD-front as described in the paragraph following Equation (1) in Section 2.2 based on the following specification of the null hypothesis:

$$H_0 : \text{H} \notin \text{gsd}(\{\text{H, BS, CS, TS, T}_k, \text{T}_p\})$$

at a significance level of $\alpha = 0.05$. As described in Section 2, to test the hypotheses pair (H_0, H_1), we perform statistical tests for five auxiliary null hypotheses, each corresponding to a pairwise GSD-dominance comparison between human text completion (H) and one of the automatic text completion strategies BS, CS, TS, T_k, and T_p (the detailed testing schemes for those auxiliary tests can be found in Jansen et al. (2024, A.2.2)). The distribution of the resampled pairwise test statistics, i.e., the empirical versions of $D(H, S)$, where $S \in \mathcal{D} \setminus \{H\}$ (see Equation (2)), is illustrated in Figure 2 (left). It demonstrates that the pairwise tests are significant across all five comparisons. Consequently,

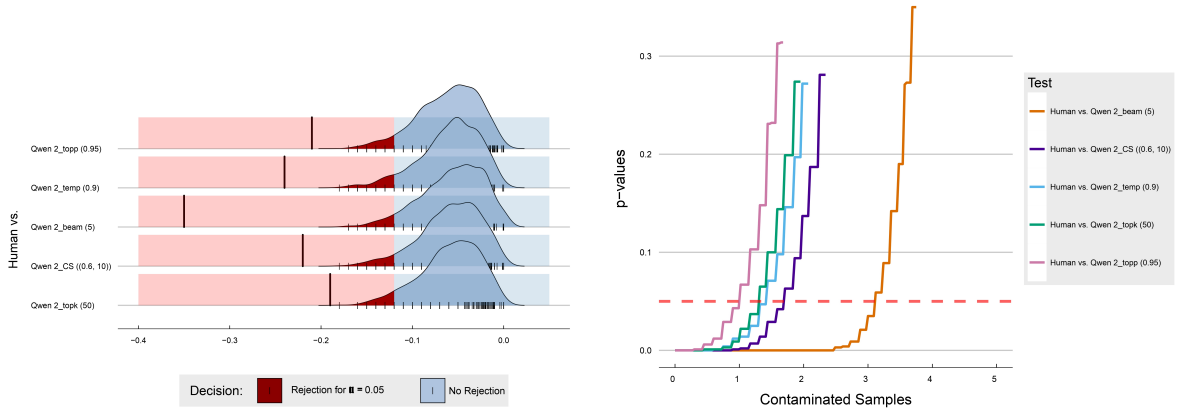

Figure 2: Left: Empirical densities of resampled test statistics for pairwise GSD-comparisons between *human* and five decoding strategies, using prompts from Wikinews and Wikitext. Vertical markers indicate the observed test statistic values, with the rejection threshold ($\alpha = 0.05$) marked in red. Right: Assessment of i.i.d. assumption violations for the same pairwise GSD-comparisons between *human* and five decoding strategies. The plot shows computed p-values with the significance threshold ($\alpha = 0.05$) indicated by the red horizontal line.

we conclude that human completion is not significantly outperformed by any of the automatic decoding strategies in the set $\mathcal{D}$ and, therefore, can be assumed to lie in the GSD-front $\text{gsd}(\mathcal{D})$ of the considered set of decoding strategies $\mathcal{D}$ at level $\alpha = 0.05$. In other words, *we find no evidence to suggest that human completion is redundant—the considered decoding strategies have not yet reached a quality level where human completion offers no additional value*. By incorporating statistical inference, we have moved beyond mere descriptive analysis to an inductive analysis that extends beyond the benchmark suites.

Important clarification: Being in the GSD-front means that human text is *not dominated* by any automatic method—there is no statistical evidence that any decoding strategy outperforms human text across all metrics. This *does not* imply that human text is superior in an absolute sense, only that it remains competitive and potentially preferable in certain contexts.

As emphasized in Section 1, robust inference analysis should also account for potential deviations from the i.i.d. assumption (see the last paragraph of Section 2 for further details). While recommendable in general, this is especially true for applications like the one at hand, where the assumption of identical distributions is questionable because of the sampling scheme of the prompts (two different suites), and the independence assumption is questionable because of the way the text evaluations of the human evaluators are obtained (potential learning effects during the evaluation process). Therefore, in Figure 2 (right), we examine the robustness of our test decision under contamination of the benchmark suite, specifically considering deviations from the i.i.d. assumption. The figure displays the p-values of all five (significant) auxiliary tests as functions of the contamination size of the benchmark suite, i.e., the degree of deviation from the i.i.d. sampling assumption of the prompts. We see that the pairwise comparison results remain significant as long as at most 1 (for the pink, the green, the blue, and the purple line) or 3 (for the yellow line) prompt(s) deviate(s) from this assumption. Since all five pairwise comparisons must be significant to reject the null hypothesis above, we conclude that, at most, one prompt can deviate (and stem from some arbitrary distribution) while maintaining the statistical significance of our test decision.

Beyond the specific results for the concrete application, this casts an interesting light on reliable statistical statements in benchmark studies in general: Since such statements (especially those of inferential nature) depend heavily on idealizing assumptions about the analyzed benchmark suites, it is all the more important that benchmark suites are curated according to appropriate standards.

4 How Can the Field Benefit from the GSD Framework?

Current text generation research overwhelmingly optimizes individual metrics in isolation, leading to systems that excel along one dimension while potentially degrading others. For instance, recent

Method	Q*Text	Human 1	Human 2
Human	47.95	3.08	3.44
Top-p (p=0.95)	32.33	2.70	2.68
Top-k (k=50)	35.06	2.46	2.62
Temperature (τ=0.9)	38.66	2.24	2.60
Contrastive (α=0.6, k=10)	23.60	2.18	2.42
Beam (B=5)	7.02	1.64	2.36
Average	30.77	2.38	2.69

Table 2: Mean performance of each method across Q*Text scores and two human evaluations (5-point Likert scale). Human text shows the highest scores, with sampling-based decoding strategies (top-p, top-k, temperature) outperforming deterministic methods (contrastive, beam search).

advances in decoding strategies—including locally typical sampling (Meister et al., 2023)—rely primarily on MAUVE scores for hyperparameter tuning and benchmarking. While MAUVE captures distributional similarity, optimizing solely for this metric may inadvertently compromise other dimensions of quality, such as coherence or diversity.

Our GSD framework addresses this limitation by enabling system design guided by a holistic view of non-dominated methods, rather than single-metric optimization. Instead of declaring one decoding strategy "best" based on isolated metrics, practitioners can identify the set of competitive approaches across multiple quality dimensions simultaneously. This shift—from descriptive metric reporting to statistical assessment of method dominance—provides actionable guidance for both research and deployment decisions.

Research Applications. When developing novel decoding algorithms, researchers can use GSD to determine whether their method belongs to the statistical front of non-dominated strategies. This provides a rigorous criterion for publication-worthy contributions: a new method merits investigation if it cannot be shown to be dominated by existing approaches across all relevant quality dimensions.

Broader Impact. As LLM performance evolves, the GSD framework provides a principled approach to assess whether emerging systems significantly outperform established baselines—whether human text or state-of-the-art algorithms. Although we focused on open-ended text generation here, the same approach extends to other tasks (summarization, translation, reasoning) by selecting appropriate, task-specific metrics.

5 Related Work

Benchmarks serve as critical platforms for methodological validation in machine learning (Ye et al., 2024; Hu et al., 2020; Kirk et al., 2024). However, recent studies have exposed significant challenges: (Berrar, 2024) show that performance improvements often fail to replicate, while (Madaan et al., 2024) demonstrate that minor variations in initialization or sampling can alter rankings (White et al., 2024; Zhou et al., 2023). These findings underscore the need for more statistically rigorous evaluation methodologies. In response, researchers have developed frameworks that explicitly acknowledge benchmark datasets as finite samples from larger populations (Demšar, 2006; Benavoli et al., 2017). This has led to multi-criteria benchmarking paradigms across diverse domains (Jansen et al., 2024; Rodemann and Blocher, 2024), from predictive ML balancing accuracy against efficiency (Koch et al., 2015) to optimization tasks requiring simultaneous performance and speed considerations (Schneider et al., 2018). For neural text generation, multiple metrics assess different quality dimensions: diversity measures lexical richness, MAUVE evaluates distributional similarity, coherence calculates prompt-continuation likelihood, and perplexity assesses predictability (Hashimoto et al., 2019; Pillutla et al., 2021; Su et al., 2022; Celikyilmaz et al., 2021). Single-metric optimization proves inadequate—coherence optimization yields repetitive outputs (degeneration), while diversity maximization compromises semantic integrity (Lee et al., 2022; Holtzman et al., 2019). Despite this, multicriteria benchmarking has declined since 2022 (Figure 1). Moreover, the discrepancy between automatic metrics (cardinal) and human evaluations (ordinal) presents additional challenges (Su

and Xu, 2022; Garces Arias et al., 2024; Ding et al., 2025). Integrated frameworks like HUSE (Hashimoto et al., 2019) combine human judgments with model probabilities. Recently, (Garces Arias et al., 2025b) proposed a multicriteria framework using the Bradley-Terry model for pairwise comparisons and introduced Q*Text—a weighted mean of coherence, diversity, and perplexity. Our work builds on these advances, proposing a framework that supports rigorous statistical inference while seamlessly integrating both automatic and human evaluations.

6 Conclusion

We introduced a framework based on Generalized Stochastic Dominance (GSD) that addresses three critical limitations in current methodologies for evaluating LLM-generated text: (1) the inadequacy of single-metric assessment, (2) the incompatibility between cardinal automatic metrics and ordinal human judgments, and (3) the absence of robust statistical guarantees. The GSD-front approach integrates multiple quality dimensions while preserving their distinct measurement scales and enables quantifying the robustness of inference under potential deviations from i.i.d. assumptions. To validate this framework, we conducted a comparative analysis of five common decoding strategies against human-written text, though the method generalizes to any set of generation approaches.

The GSD-front enables statistically sound multi-criteria evaluation without requiring arbitrary metric weighting or compromising measurement scale integrity. By incorporating techniques from robust statistics, our approach extends beyond descriptive benchmark analysis to provide inferential guarantees that account for potential dependencies in human evaluations. This advancement provides researchers and practitioners with a more rigorous methodology for evaluating text generation systems. Future work could extend the GSD approach to other generation tasks such as summarization and translation, investigate additional quality dimensions, and further enhance statistical robustness for complex evaluation dependencies.

Limitations

Despite the strengths of our proposed framework, several limitations should be acknowledged. First, our experimental validation focused primarily on benchmarking human text continuations with LLM-generated text in an open-ended text generation task. While this provided a suitable context for demonstrating our framework, different neural text generation tasks—such as summarization and machine translation—may present unique evaluation challenges and yield different conclusions. Second, the human evaluation component in our work was conducted by the authors themselves, potentially introducing expertise bias. Evaluators familiar with the field may interpret quality dimensions differently than end-users would. Finally, while our statistical methodology quantifies robustness against certain deviations from i.i.d. assumptions, real-world evaluation scenarios often involve more complex dependencies that require further methodological developments. Despite these limitations, we believe our work makes a substantial contribution to the field of text generation evaluation and provides a solid foundation for more statistically sound multi-criteria benchmarking approaches.

Ethics Statement

This study uses only publicly available datasets that contain no personally identifiable information. Human evaluation was carried out by the authors on anonymized text continuations, ensuring that the underlying decoding strategies remained obscured. We acknowledge the potential ethical concerns associated with language models for text generation, particularly the risk of producing harmful content—whether through intentional misuse or unintended biases stemming from the training data and algorithms. We confirm that no conflicts of interest have influenced the outcomes, interpretations, or conclusions of this research. All funding sources are fully disclosed in the acknowledgments.

Acknowledgments

Hannah Blocher received financial support via a stipend from Evangelisches Studienwerk Villigst e.V. Julian Rodemann acknowledges support by the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics" as well as by the Bavarian Institute for Digital Transformation (bidt) and the Bavarian Academy of Sciences (BAS) within a graduate scholarship. Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581 - BERD@NFDI.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. 2017. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36.

Daniel Berrar. 2024. Estimating the replication probability of significant classification benchmark experiments. *Journal of Machine Learning Research*, 25(311):1–42.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

J. Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Yuanhao Ding, Esteban Garces Arias, Meimingwei Li, Julian Rodemann, Matthias Aßenmacher, Danlu Chen, Gaojuan Fan, Christian Heumann, and Chongsheng Zhang. 2025. Guard: Glocal uncertainty-aware robust decoding for effective and efficient open-ended text generation. *Preprint*, arXiv:2508.20757.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *Preprint*, arXiv:1805.04833.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025b. Towards better open-ended text generation: A multicriteria evaluation framework. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 631–654, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025a. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020, Abu Dhabi, UAE. Association for Computational Linguistics.

Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2024. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *Preprint*, arXiv:1904.02792.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133.

Christoph Jansen. 2025. Contributions to the decision theoretic foundations of machine learning and robust statistics under weakly structured information. *Preprint*, arXiv:2501.10195.

Christoph Jansen, Malte Nalenz, Georg Schollmeyer, and Thomas Augustin. 2023a. Statistical comparisons of classifiers by generalized stochastic dominance. *Journal of Machine Learning Research*, 24(231):1–37.

Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. 2023b. Multi-target decision making under conditions of severe uncertainty. In *Modeling Decisions for Artificial Intelligence*, pages 45–57, Cham. Springer Nature Switzerland.

Christoph Jansen, Georg Schollmeyer, Hannah Blocher, Julian Rodemann, and Thomas Augustin. 2023c. Robust statistical comparison of random variables with locally varying scale of measurement. In *Uncertainty in Artificial Intelligence*, pages 941–952. PMLR.

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin. 2024. Statistical multicriteria benchmarking via the GSD-front. *Advances in Neural Information Processing Systems*.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Patrick Koch, Tobias Wagner, Michael TM Emmerich, Thomas Bäck, and Wolfgang Konen. 2015. Efficient multi-criteria optimization on noisy machine learning problems. *Applied Soft Computing*, 29:357–370.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Preprint*, arXiv:2202.00666.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Julian Rodemann and Hannah Blocher. 2024. Partial rankings of optimizers. In *International Conference on Learning Representations (ICLR), Tiny Papers Track*.

F. Schneider, L. Balles, and P. Hennig. 2018. DeepOBS: A deep learning optimizer benchmark suite. In *International Conference on Learning Representations*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Preprint*, arXiv:2202.06417.

Yixuan Su and Jialu Xu. 2022. An empirical study on contrastive search and contrastive decoding for open-ended text generation. *Preprint*, arXiv:2211.10797.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, and 1 others. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Appendix

A.1 Decoding strategies

Strategy	Parameters	Description	Authors
Beam search	beam width = 5	Deterministic search algorithm that maintains multiple hypotheses (beams).	(Freitag and Al-Onaizan, 2017)
Contrastive search	$k = 10$, $\alpha = 0.6$	Balances token probability and diversity through a contrastive objective.	(Su et al., 2022)
Sampling with temperature	temperature = 0.9	Adjusts the sharpness of the probability distribution before sampling.	(Ackley et al., 1985)
Top-k sampling	$k = 50$	Samples from the k most probable tokens.	(Fan et al., 2018)
Top-p sampling	$p = 0.95$	Samples from the smallest set of tokens whose cumulative probability exceeds p.	(Holtzman et al., 2019)

Table 3: Overview of evaluated decoding strategies and hyperparameter choices, following best performance reported by (Garces Arias et al., 2025a).

A.2 Automatic metrics

Diversity. This metric aggregates n-gram repetition rates:

$$\texttt{div} = \prod_{n=2}^{4} \frac{|\ \text{unique n-grams}\ (\mathrm{x}_{\text{cont}})\ |}{|\ \text{total n-grams}\ (\mathrm{x}_{\text{cont}})\ |}$$

A low diversity score suggests the model suffers from repetition, and a high diversity score means the model-generated text is lexically diverse.

Coherence. Proposed by Su et al. (2022), the coherence metric is defined as the averaged log-likelihood of the generated text conditioned on the prompt as

$$\texttt{coh}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{|\hat{\boldsymbol{x}}|} \sum_{i=1}^{|\hat{\boldsymbol{x}}|} \log p_{\mathcal{M}}\left(\hat{\boldsymbol{x}}_i \mid [\boldsymbol{x} : \hat{\boldsymbol{x}}_{<i}]\right)$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are the prompt and the generated text, respectively; [:] is the concatenation operation and $\mathcal{M}$ is the OPT model (2.7B) (Zhang et al., 2022).

Generation Perplexity. The perplexity $\texttt{ppl}(W)$ of a sequence of words (or tokens) $W = w_1, w_2, ..., w_N$ is computed as (Jelinek et al., 2005; Holtzman et al., 2019):

$$\texttt{ppl}(W) = \exp\left(-\frac{1}{N} \sum_{i=1}^{N} \log p(w_i \mid w_1, ..., w_{i-1})\right)$$

Here, $p(w_i \mid w_1, ..., w_{i-1})$ is the probability of word w_i given its preceding context.
Perplexity measures how well a probabilistic model predicts a sequence of words. Lower perplexity indicates better predictive performance, as the model assigns a higher probability to the actual sequence. It is commonly used to evaluate the quality of language models.

A.3 Q*Text

Q*Text (Garces Arias et al., 2025b) is calculated based on normalized and penalized coherence, diversity, and generation perplexity (see Section A.2).

Metric Formulation Q*Text is defined as:

$$\text{Q*Text} = \frac{\sum_{i=1}^{3} w_i M_i P_i(M_i)}{\sum_{i=1}^{3} w_i} \tag{4}$$

where M_i are normalized metrics, w_i are weights, and $P_i(x) = \exp(-\alpha_i(x - \mu_i)^2)$ are Gaussian penalties that discourage extreme values. Parameters μ_i represent optimal targets while α_i controls penalty strength.

Normalization Inverse normalization is applied to perplexity (lower is better): $M_1 = \frac{p_{\max} - p_i}{p_{\max} - p_{\min}}$, and standard min-max normalization to coherence and diversity (higher is better): $M_j = \frac{m_j - m_{\min}}{m_{\max} - m_{\min}}$ for $j \in \{2, 3\}$.

Parameter Optimization The nine parameters $\theta = \{w_i, \mu_i, \alpha_i\}_{i=1}^{3}$ are optimized via:

$$\theta^* = \text{argmax}_\theta \rho_s(\text{Q*Text}(\theta), H) \tag{5}$$

where ρ_s is Spearman correlation and H are publicly available human ratings (Garces Arias et al., 2025a).

A.4 Human evaluation

A.4.1 Instructions for human evaluators

Please disregard formatting characters and special characters such as <|endoftext|> or characters that have remained unrecognized and received unusual encoding. The evaluation should focus primarily on the quality of the content.

- **Quality** should be measured by how human-like, fluent, and coherent the text is perceived by you.
- **Coherence:** The text feels consistent throughout, not a collection of jumbled topics. It maintains focus with a consistent thread and does not read as a series of disconnected sentences.
- **Fluency:** The text is written in grammatical English. There are no obvious grammar mistakes that a person would not typically make.

An incomplete final word or incomplete sentence should not be counted as a mistake and should not affect the fluency assessment. The English should be considered natural as long as it is grammatically correct. Do not penalize for spaces between parts of words (e.g., “don ’t”) or simpler sentences. Simple English is to be considered equally valid as complex English. Please utilize the following Likert scale.

A.4.2 Inter-rater agreements

The analysis focused on weighted agreement measures appropriate for ordinal data. The weighted Cohen’s Kappa coefficient was 0.324 (p-value $\approx 1.2 \times 10^{-6}$), indicating fair agreement between evaluators when accounting for the magnitude of disagreements. This measure applies linear weights to disagreements based on their distance on the Likert scale, recognizing that a disagreement between ratings of 1 and 3 represents a larger discrepancy than between 1 and 2.
Spearman’s rank correlation coefficient was 0.518 (p-value $\approx 5.38 \times 10^{-23}$), demonstrating a moderate positive correlation between the evaluators’ ratings. This indicates that while absolute scores sometimes differed, the relative ranking of text quality showed reasonable consistency between evaluators. Additional analysis revealed that 82.7% of all ratings were within one point of each other, with an average absolute difference of 0.82 points. The statistical significance of both measures confirms that the agreement between evaluators is not due to chance, despite being only fair to moderate—a common finding in subjective text quality assessment that further motivates the inclusion of complementary automatic metrics.

Score	Quality Level	Description
5.0	Excellent	Text is exceptionally clear, coherent, and well-structured. Content is comprehensive, accurate, and presented in a highly engaging manner. No improvements needed.
4.0	Very Good	Text is clear, well-organized, and contains few errors. Ideas flow logically with appropriate transitions. Content is accurate and thorough.
3.0	Good	Text communicates the intended message effectively. Organization is adequate with some minor clarity or coherence issues. Content is mostly accurate.
2.0	Fair	Text has significant issues with clarity, organization, or accuracy that impact comprehension. Ideas may be underdeveloped or poorly connected.
1.0	Very Poor	Text is difficult to understand with major structural problems, significant errors, and/or incomplete information. Communication largely fails.

Table 4: Text quality assessment scale for human evaluators.

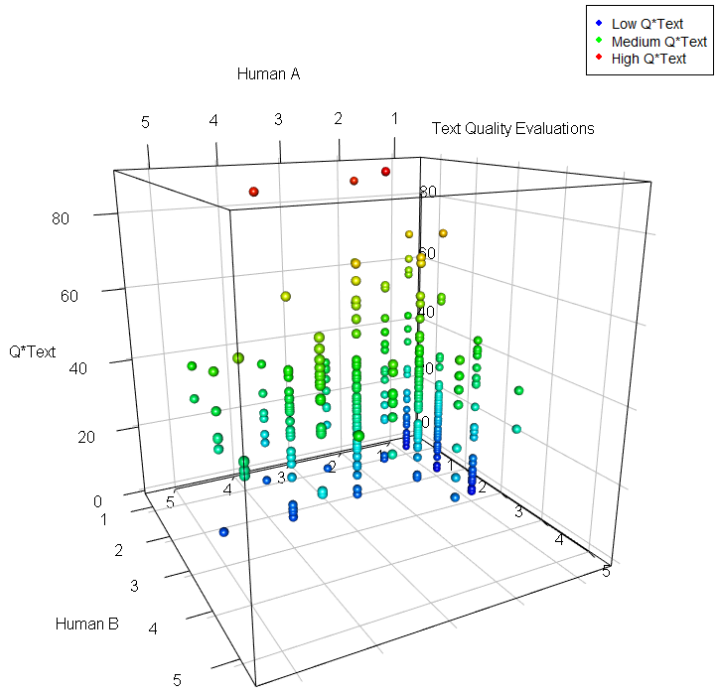

Figure 3: Evaluation results for 300 text continuations generated from 50 prompts derived from Wikitext and Wikinews datasets. The assessment combines cardinal automatic metrics (Q*Text) with ordinal evaluations from two independent human raters using a 5-point Likert scale.