How Evaluation Choices Affect MAUVE Scores and Generator Rankings

Anonymous authors
Paper under double-blind review

Abstract

Since its introduction, MAUVE has been widely adopted for evaluating open-ended text generation, serving both to compare language models and to design and select decoding strategies. Because it compares generated and human-written text in a learned feature space, its scores depend on evaluation choices such as the representation model, the evaluated text length, and the quantization settings. These choices are known to shift scores, but less is known about whether they also change which generation model is preferred. We study this question across seven generation models, three datasets, and 180 evaluation settings, separating variation due to evaluation choices from variation due to decoding configurations. For identical generations, evaluation choices produce a wider score spread than decoding changes, and they alter model rankings even after aggregating over decoding strategies: two settings that differ only in the representation model select the same top-ranked model in 30% of comparisons. Among the choices examined, the representation model has the largest effect on ranking stability, and it also affects how closely MAUVE rankings agree with human preference judgments. These findings suggest that interpreting a MAUVE comparison requires considering its evaluation configuration and the sensitivity of model rankings to that configuration, and show that human evaluation can identify representation models whose rankings align more closely with human preferences, providing practical guidance for choosing among them.

1 Introduction

When evaluating and comparing language models in open-ended text generation tasks, automatic metrics are widely used—often alongside limited human evaluation and LLM-as-a-judge scores—to measure qualitative dimensions such as coherence, diversity, or human-likeness (Schmidtová et al., 2024; Zheng et al., 2023). In recent years, MAUVE (Pillutla et al., 2021; 2023) has gained notable attention and been adopted as a common metric, not only for evaluating neural text generators (Li et al., 2023a; Finlayson et al., 2024; Garces Arias et al., 2024) but also for designing decoding strategies, where it often serves as the main criterion during hyperparameter sweeping (Hewitt et al., 2022; Meister et al., 2023; Zhu et al., 2024).

MAUVE measures the distributional difference between generated and human-written text in a learned feature space. Its conclusions, however, may depend on evaluation choices: the representation model, the evaluated text length, and the quantization settings all determine how generated and reference texts are compared (Pimentel et al., 2023; Winata et al., 2025; Catapang, 2026). These choices can change MAUVE scores even when the generated corpus remains fixed (Pimentel et al., 2023; Elangovan et al., 2025). Yet a score shift alone does not reveal whether the model ordering is preserved or whether a different generation model is selected. Understanding this distinction is essential for interpreting what a reported evaluation result says about the models being compared (Mei et al., 2026; Zaghouani, 2026).

Figure 1 shows a concrete case. We evaluate the same Wikitext outputs with two representation models, GPT-2 Large (Radford et al., 2019) and Qwen3-Embed (Zhang et al., 2025c), holding the reference corpus, truncation length, and bucket count fixed. Falcon (Malartic et al., 2024) ranks first under GPT-2 Large and seventh under Qwen3-Embed; Llama 3 (Grattafiori et al., 2024) moves from fifth to first. Nothing about the generated text has changed, yet the two evaluations select different

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

winners. How often this happens, which evaluation choices drive it, and which ranking better reflects human preferences are the questions this paper addresses.

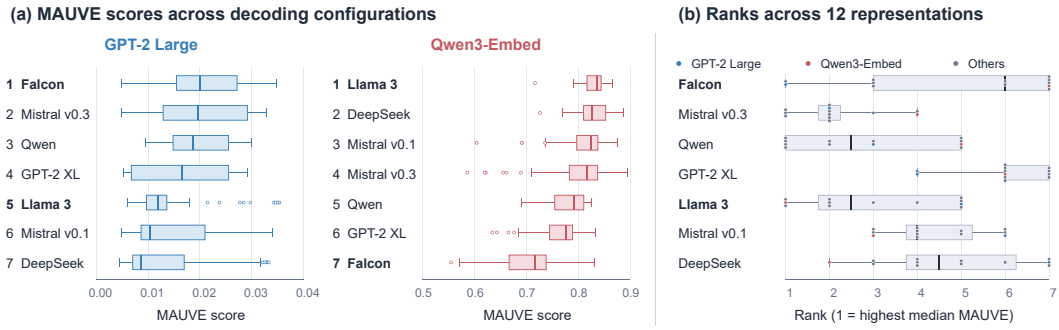


Figure 1: **Representation choice changes the ranking of the same generated outputs.** Seven generation models evaluated on Wikitext with fixed outputs, reference corpus, $L = 256$ GPT-2 tokens, and $K = 500$. (a) MAUVE scores across decoding configurations, ordered by median under each representation. Boxes show IQRs, lines show medians, whiskers follow the $1.5 \times$ IQR rule, and circles mark outliers. Score axes differ. (b) Ranks based on median scores across all 12 representations, with one dot per representation. Boxes show IQRs, lines show medians, and whiskers show full ranges. Blue/red dots denote GPT-2 Large/Qwen3-Embed; rank 1 is highest.

To investigate this dependence, we study outputs from seven generation models on Book, Wikinews, and Wikitext from *Decoding Decoded* (Garces Arias et al., 2025; Ding et al., 2025). We evaluate these outputs under 180 combinations of representation models, text lengths, and bucket counts. We separate evaluation variation from decoding variation, then analyze generation model rankings, stability under aggregation and repeated runs, feature-processing controls, and agreement with human preferences.

Our analysis addresses four research questions:

- **RQ1 (Size of the evaluation effect).** How does the variation of MAUVE scores across evaluation settings compare with the variation across decoding configurations?
- **RQ2 (Effect on model selection).** Does the choice of representation model change the generation model ranking, and the top-ranked model, once decoding configurations are aggregated?
- **RQ3 (Sources of ranking instability).** Which evaluation choices—representation model, bucket count, text length, normalization, and pooling—affect the stability of the rankings?
- **RQ4 (Agreement with human preferences).** Does the choice of representation model change how closely MAUVE rankings match human preference rankings?

Our contributions are as follows:

1. **Evaluation choices dominate score variation.** Holding the generated text fixed and varying only the evaluation setting yields a summary IQR of 0.640, compared with 0.139 when varying the decoding configuration under a fixed setting, a ratio of about 4.6 (§ 4; RQ1).
2. **Representation choice changes which generation model is selected.** After aggregating over decoding configurations, pairs of representation models agree on the ranking of generation models with mean $\tau_b = 0.30$ and select the same top-ranked model in only 30% of comparisons (§ 5; RQ2).
3. **The instability is specific to the representation model.** Ranking disagreement between representations persists across aggregation rules and exceeds the variation between repeated runs; bucket count and text length have smaller effects, and normalization and pooling account for part of the difference between representation groups (§ 6; RQ3).
4. **Representation choice affects agreement with human preferences.** On Wikitext, MAUVE rankings computed with T5 Large agree with human preference rankings at $\tau_b = 0.70$, compared with 0.51 for GPT-2 Large (§ 7; RQ4).

2 Background and Related Work

What a MAUVE score measures. MAUVE compares a generated corpus with a reference corpus through their distributions in a learned feature space (Pillutla et al., 2021; 2023). A representation model maps each text to a feature vector; the reference and generated features are jointly clustered into K buckets; and the divergence frontier between the two bucket-occupancy distributions is summarized into a score in $(0, 1]$. Three choices enter this pipeline beyond the texts themselves: the representation model, which determines the feature space (Wang & Isola, 2020); the bucket count K, which determines the resolution of the distribution estimate; and the length of text presented to the representation model. Pillutla et al. (2023) varied the quantization algorithm and the size of the feature extractor and reported stable rankings of decoding methods. Pimentel et al. (2023) showed that the quantization step discards much of the information in the features and that the metric can be computed from quite different representations. Neither study asks whether the ranking of *generation models* survives replacing the representation model with one from a different family, which is the question we address.

Sensitivity of model-based metrics. He et al. (2023) inject synthetic errors into generations and find blind spots in GPT-2-based MAUVE, including limited sensitivity to errors early in a text. Together with analyses of embedding-based metrics more broadly (Zhang et al., 2020; Elangovan et al., 2025), this shows that a metric's behavior depends on its representation. We ask the complementary question on fixed, naturally generated corpora: by how much do evaluation choices change the measured scores and the resulting model ordering?

Decoding as a reference scale. Garces Arias et al. (2025) compare decoding methods and hyperparameters across generation models and domains (Holtzman et al., 2020; Basu et al., 2021; Ding et al., 2026) and supply the generated corpora used here. A decoding change alters the text being evaluated, whereas a change of representation or bucket count alters only the evaluation of that text. Comparing score variation along the two directions gives the evaluation effect a scale that practitioners already reason about.

Distributional metrics and human preferences. Other metrics compare feature distributions through Gaussian approximations, kernel distances, or precision and recall estimates (Xiang et al., 2021; Kynkäänniemi et al., 2019; Simon et al., 2019; Le Bronnec et al., 2024; Kim et al., 2023). Our main human-preference comparison keeps MAUVE fixed and varies only the representation, so that the comparison isolates representation choice (Gao et al., 2025); Appendix F reports the corresponding representation–metric grid. Human preference judgments are themselves an imperfect criterion (Rodemann et al., 2025; Zheng et al., 2023); we use them as an external reference for the automatic rankings, not as ground truth.

3 Study Design

3.1 Generation data and measurement settings

We use generated corpora from *Decoding Decoded* (Garces Arias et al., 2025). The seven generation models are DeepSeek (DeepSeek-AI et al., 2024), Falcon (Malartic et al., 2024), Llama 3 (Grattafiori et al., 2024), Mistral v0.1 (Jiang et al., 2023), Mistral v0.3 (Jiang et al., 2023), Qwen (Yang et al., 2024), and GPT-2 XL (Radford et al., 2019). The three datasets are Book (Zhu et al., 2015), Wikinews[1], and Wikitext (Merity et al., 2016). A *generation cell* identifies one dataset, generation model, and decoding configuration. The analysis index contains 1,175 unique generation cells. Decoding configurations belong to beam search (Freitag & Al-Onaizan, 2017), contrastive search (Su et al., 2022), temperature sampling (Ackley et al., 1985), top-k (Fan et al., 2018), and top-p (Holtzman et al., 2020). Analyses aggregate over the available configurations for each generation model.

The evaluation grid contains GPT-2 Large (Radford et al., 2019), Ettin decoder (Weller et al., 2026), Ettin encoder (Weller et al., 2026), RoBERTa (Liu et al., 2019a), ModernBERT (Warner et al.,

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

2025), T5 encoder (Raffel et al., 2020a), Qwen3-Embed (Zhang et al., 2025b), BGE (Xiao et al., 2024), GTE (Li et al., 2023b), Nomic (Nussbaum et al., 2025), Stella (Zhang et al., 2025a), and MPNet (Song et al., 2020), three truncation lengths $L \in \{64, 128, 256\}$, and five bucket counts $K \in \{25, 50, 100, 250, 500\}$. This gives 180 evaluation settings. The representation panel spans causal language models, bidirectional models, a T5 encoder, and retrieval embedding models. We use *generation model* for the model producing text and *representation model* for the feature extractor used by MAUVE. Appendix A lists the panel and feature-processing configurations.

For each length condition, reference and generated texts are pre-truncated to L GPT-2 tokens and then tokenized by the chosen representation model; L is thus a common pre-truncation budget, separate from the representation's own input limit (Appendix A). Reference and generated corpora are always evaluated under the same configuration.

3.2 Separating evaluation and decoding variation

Let $s_{d,g,c,e,L,K}$ be the MAUVE score for dataset d, generation model g, decoding configuration c, representation model e, truncation length L, and bucket count K. We measure the spread of scores with the interquartile range, $\mathrm{IQR} = Q_{0.75} - Q_{0.25}$. The two comparisons vary different coordinates:

$$V_{\mathrm{eval}}(d, g, c) = \mathrm{IQR}_{e,L,K}(s_{d,g,c,e,L,K}), \tag{1}$$

$$V_{\mathrm{decode}}(d, g, e, L, K) = \mathrm{IQR}_{c}(s_{d,g,c,e,L,K}). \tag{2}$$

The first comparison holds the original generated corpus fixed and varies how it is evaluated. The second fixes the evaluation setting and varies the decoding configuration that produced the corpus. RQ1 compares the summary IQRs along these directions. This comparison measures sensitivity over the evaluated configuration sets.

3.3 Constructing and comparing rankings

For model-level comparisons, we first fix the dataset, truncation length, and bucket count. Under representation e, the score assigned to generation model g is

$$m_g(d, e, L, K) = \mathrm{median}_{c \in \mathcal{C}(d,g)}\, s_{d,g,c,e,L,K}, \tag{3}$$

where $\mathcal{C}(d, g)$ contains the available decoding configurations for that generation model and dataset. Ranking the seven medians from highest to lowest gives the generation model ordering. The same configuration set for a generation model is used when its representation model changes. The resulting ranking summarizes performance over the evaluated decoding configurations.

There are 45 dataset–length–bucket conditions and 66 possible representation pairs per condition. We compare the generation model orderings with Kendall's τ_b, first-place agreement, and rank span. First-place agreement records whether a representation pair selects the same generation model. Rank span is the distance between a generation model's highest and lowest assigned ranks across representations. Appendix B defines these quantities.

RQ3 also uses rankings of individual generation cells. A representation comparison fixes (d, L, K), a bucket comparison fixes (d, e, L), and a length comparison fixes (d, e, K). RQ4 compares metric rankings of human-scored cells with the preference ordering within each dataset. Table 3 in Appendix B collects these comparison units. Keeping the ranked objects explicit lets us distinguish changes to individual decoding configurations from changes to model-level conclusions.

4 How Much Do Evaluation Settings Change the Score?

4.1 RQ1: evaluation variation relative to decoding variation

We first ask whether evaluation choices introduce a small numerical adjustment or a change large enough to matter when comparing generated text. We compare the two directions in Equations 1 and 2: vary the evaluation of a fixed generation cell, then vary decoding under a fixed evaluation setting. Decoding provides a reference because it changes the text being evaluated.

Figure 2(a) shows a summary IQR of 0.640 across the 180 evaluation settings and 0.139 across decoding configurations. Their ratio is approximately 4.6. Thus, over this grid, changing how the

outputs are evaluated produces a wider central score spread than changing the decoding configuration that produced them.

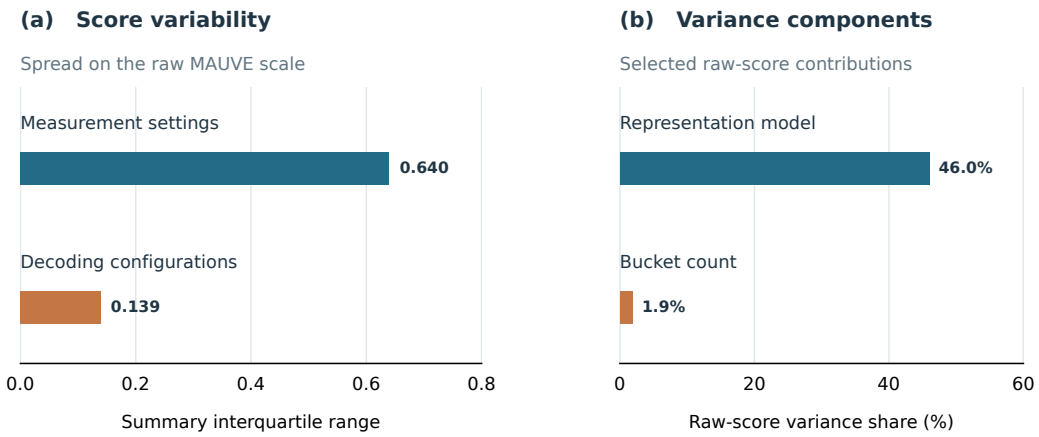


Figure 2: Score variation under changes to evaluation and decoding (RQ1). (a) Summary interquartile ranges across the 180 evaluation settings and across decoding configurations. (b) Selected raw-score variance components: representation-model choice and bucket count. The two panels describe score spread and variance allocation, respectively.

4.2 Locating the variation in the evaluation procedure

Figure 2(b) separates two components of the reported raw-score variance decomposition. Representation-model choice accounts for 46.0%, compared with 1.9% for bucket count. Changing the feature space is therefore a substantial component of the observed score variation. Table 4 in Appendix B collects the score summaries and these two components.

A large score change need not change a model comparison: all generation models could move together while keeping the same order. We therefore turn from score magnitudes to rankings, holding the dataset, text length, and bucket count fixed while replacing the representation model.

Finding 1 (RQ1). Evaluation choices produce a larger summary score spread than decoding changes over the evaluated grid, with representation choice accounting for a substantial part of raw-score variance. RQ2 tests whether this variation changes which generation model is preferred.

5 Does Representation Choice Change Generation Model Selection?

5.1 RQ2: generation model rankings across representations

We summarize each generation model's decoding configurations with the median in Equation 3, then compare the resulting seven-generation model rankings across representations at fixed (d, L, K). This tests the consequence for generation model selection after decoding configurations have been aggregated.

Mean Kendall agreement across representation pairs is 0.30 (Figure 3(a)). The same generation model ranks first in only 30% of representation-pair comparisons, and the average distance between a generation model's best and worst rank is 5.2 positions (Figure 3(b)). Representation choice consequently affects both the ordering as a whole and the generation model selected at the top. The introduction example illustrates such a change on the archived Wikitext sweep; these summaries describe the model-level analysis.

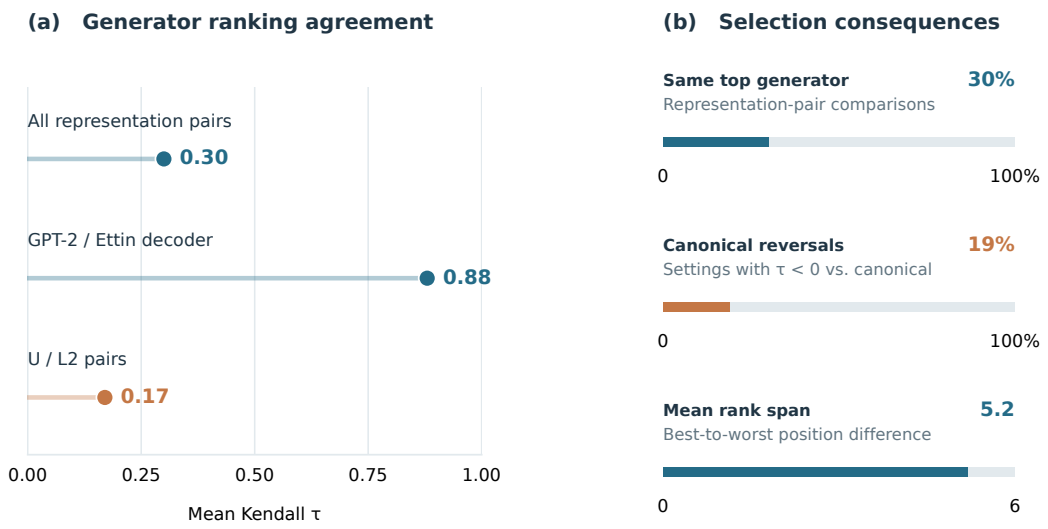


Figure 3: Generation model-ranking and selection consequences of representation choice (RQ2). (a) Mean Kendall agreement across all representation pairs, the GPT-2 Large–Ettin decoder pair, and pairs crossing the unnormalized (U) and L2-normalized (L2) groups. (b) First-place agreement between representation pairs, the fraction of settings negatively correlated with the analysis's canonical reference ranking, and mean best-to-worst rank span. First-place agreement and canonical reversals use different comparison sets.

The canonical-reference comparison in Figure 3(b) gives a second view of ordering changes: 19% of settings have negative Kendall agreement with that reference ordering. This statistic counts settings, whereas first-place agreement compares representation pairs. Table 5 in Appendix B keeps these quantities separate.

5.2 Which representation pairs disagree?

The overall mean combines very different representation pairs. GPT-2 Large and Ettin decoder have model-level agreement of 0.88. Both use unnormalized causal-language-model features. Pairs crossing the unnormalized and L2-normalized groups have mean agreement of 0.17. Some substitutions therefore preserve much of the ordering, while others change it substantially.

Finding 2 (RQ2). Representation choice changes generation model selection after decoding aggregation: first-place agreement is 30%, and individual generation models can move across much of the seven-model ordering. Agreement also varies markedly between representation pairs.

6 Which Evaluation Choices Affect Ranking Stability?

6.1 RQ3: aggregation and repeated runs

We first test whether the model-level result depends strongly on using the median to summarize decoding configurations. Across the tested aggregation rules, the difference between the largest and smallest mean representation-swap agreement is 0.04. The lowest agreement between a median-based ranking and a ranking under an alternative rule is 0.86. The representation sensitivity therefore persists across the tested aggregation choices; Appendix C gives the summaries.

We next compare repeated runs within an evaluation configuration with changes to that configuration. Figure 4(a) separates model-level and generation-cell rankings. At the model level, repeat agreement is 0.90 for unnormalized features (U) and 0.78 for unit-normalized features (L2), both above the overall representation-swap agreement of 0.30. Replacing the representation changes the

generation model ordering more than repeating the evaluated configuration does. At the cell level, repeat agreement is 0.83 for U and 0.58 for L2; aggregation over decoding configurations gives higher repeat agreement in both groups.

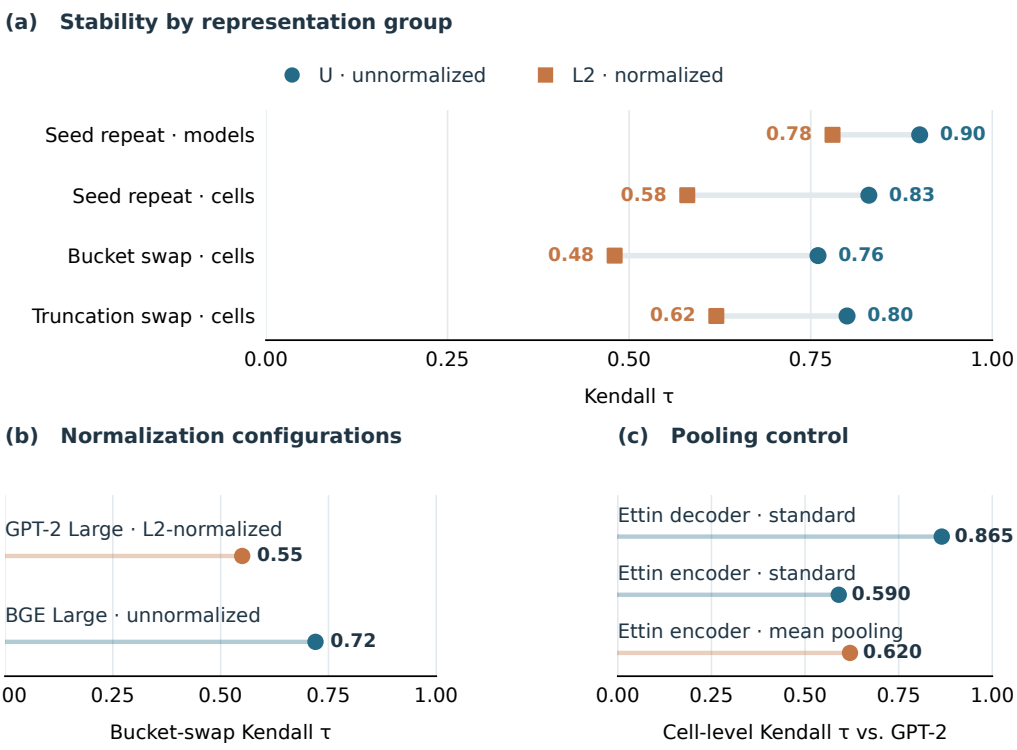


Figure 4: Ranking stability and feature-processing controls (RQ3). (a) Agreement under repeated runs, bucket changes, and truncation changes for unnormalized (U) and L2-normalized (L2) features. Repeated runs are compared at both model and cell levels; bucket and truncation changes use cell rankings. (b) Bucket-swap agreement within the two evaluated normalization configurations. (c) Cell-level agreement with GPT-2 Large under the standard Ettin configurations and the Ettin-encoder mean-pooling control. All panels report Kendall agreement.

6.2 Bucket count and evaluated text length

To compare evaluation components directly, we rank generation cells while varying one factor at a time and holding the others fixed as in Table 3. Mean agreement is 0.395 for a representation change, 0.646 for a bucket-count change, and 0.728 for a text-length change. Representation replacement produces the lowest agreement among these comparisons (Appendix D).

Sensitivity to the remaining settings also differs between representation groups. Figure 4(a) shows bucket-swap agreement of 0.76 for U and 0.48 for L2, and truncation-swap agreement of 0.80 and 0.62, respectively. The L2 group thus has lower agreement under repetition, bucket changes, and length changes. This pattern motivates examining how individual representations are pooled and normalized.

6.3 Normalization and pooling controls

The normalization controls evaluate GPT-2 Large with L2-normalized features and BGE with unnormalized features. Their bucket-swap agreements are 0.55 and 0.72, respectively (Figure 4(b)). Their mean scores are 0.80 and 0.55 (Appendix D.2, Table 9).

The pooling comparison fixes GPT-2 Large as the reference. Standard Ettin-encoder features give cell-level agreement of 0.590 with GPT-2; mean pooling gives 0.620, a difference of 0.030 (Fig-

ure 4(c)). The standard Ettin decoder gives agreement of 0.865. Changing encoder pooling therefore closes only a small part of the encoder–decoder difference in agreement with GPT-2. These comparisons characterize the evaluated feature configurations; encoder–decoder differences include architecture as well as training objective (Weller et al., 2026).

Finding 3 (RQ3). The ranking disagreement persists across aggregation rules and exceeds model-level repeat variation. Representation groups differ in their sensitivity to bucket count and text length, while the feature-processing controls show why pooling and normalization must accompany the representation model's name. RQ4 asks how these representation choices relate to human preferences.

7 Does Representation Choice Affect Agreement with Human Preferences?

7.1 RQ4: comparing automatic and human rankings

To connect changes in automatic orderings to a preference criterion, we compare metric rankings with dataset-specific human rankings. The study contains 4,993 valid pairwise judgments over 150 generation cells across Book, Wikinews, and Wikitext. Annotators compare generated continuations for coherence, factual consistency, and fluency.

Of 400 comparisons repeated with presentation order reversed, 304 preserve the preference after accounting for the swap, giving 76.0% order consistency. Split-half ranking reliability is 0.64. Appendix E records the study counts and reliability summaries.

7.2 Representation choice within MAUVE

Figure 5(a) keeps the metric fixed and changes the representation. On Wikitext, mean Kendall agreement with the human ranking is 0.51 for GPT-2 Large and 0.70 for T5 Large, a difference of 0.19. Representation choice therefore affects agreement with the collected preferences as well as agreement between automatic rankings. T5-based MAUVE aligns more closely with the Wikitext preference ranking in this mean comparison.

7.3 Representation choice across distributional metrics

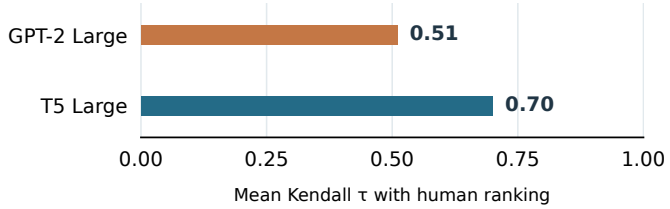
Figure 5(b) places this comparison alongside the multi-metric analysis. Each cell selects the representation with the highest observed alignment over the evaluated configurations for one dataset and metric. On Wikitext, T5 leads MAUVE with 0.783, BGE leads FBD with 0.765, and Ettin decoder leads Recall@5 with 0.701. The preferred representation consequently depends on how the feature distributions are compared. These configuration maxima are not average gains and are kept distinct from the means in panel (a); the complete representation-by-metric tables are in Appendix F.

Finding 4 (RQ4). Representation choice affects agreement with human preferences: T5-based MAUVE has higher mean agreement than GPT-2-based MAUVE on Wikitext. The multi-metric results further show that the representation favored by human alignment depends on the distributional metric. Human agreement must therefore be examined for the evaluation configuration used to support the model comparison.

8 Implications for Model Evaluation

The four questions form a progression from score variation to its practical meaning. Evaluation choices change score magnitudes (RQ1), those changes reach generation model selection (RQ2), and the sensitivity depends on the evaluated representation and feature processing (RQ3). The human comparison then tests the relationship between an automatic ordering and the collected preferences (RQ4).

(a) Mean MAUVE alignment · revised evaluation



(b) Best configurations · earlier multi-metric evaluation

	MAUVE	FBD	KID	MMD	Cent. L2	Cent. cos	P@5	R@5
Wikitext	**T5** 0.783	**BGE** 0.765	**BGE** 0.740	**BGE** 0.739	**BGE** 0.739	**BGE** 0.731	**T5** 0.749	**Ettin-D** 0.701
Book	**T5** 0.660	**T5** 0.608	**MPNet** 0.595	**T5** 0.615	**T5** 0.595	**T5** 0.616	**T5** 0.651	**GPT-2** 0.675
Wikinews	**T5** 0.604	**T5** 0.596	**T5** 0.561	**T5** 0.558	**T5** 0.563	**T5** 0.561	**T5** 0.631	**T5** 0.588

Figure 5: Representation choice and human preference agreement (RQ4). (a) Mean Wikitext Kendall agreement for GPT-2 Large and T5 Large under MAUVE in the evaluation. (b) Multi-metric evaluation: each cell identifies the representation with the highest observed configuration-specific alignment for that dataset and metric. Colors encode Kendall agreement. Panel (a) reports means; panel (b) reports configuration maxima.

Specify the evaluation configuration. A MAUVE result should identify the representation model, pooling, normalization, text-length budget, and bucket count. The score and stability comparisons show why these settings belong alongside the reported number: changing them can change both its magnitude and the resulting ordering.

Report the comparison that supports the claim. For a model-level claim, state how decoding configurations are aggregated and report the generation model ranking. First-place agreement and rank span make the effect on selection visible alongside Kendall agreement. For a decoding-level claim, compare decoding configurations under a fixed evaluation setting. Table 3 in Appendix B distinguishes these analyses.

Check both ranking sensitivity and human agreement. Ranking sensitivity measures how an ordering changes with evaluation choices. Human agreement measures its relationship to the preference criterion. Reporting both identifies conclusions that persist across evaluation settings and shows how they relate to the judgments collected for the task.

9 CONCLUSION

We studied how MAUVE evaluation choices affect comparisons of generated text across seven generation models, three datasets, and 180 evaluation settings. Evaluation changes produce a larger summary score spread than decoding changes, and representation choice changes generator selection after decoding aggregation. Aggregation checks, repeated runs, and feature-processing controls characterize where ranking sensitivity appears. Human preferences establish a further consequence: changing the representation changes agreement with the collected judgments. A model comparison should therefore report the evaluation configuration and examine whether its conclusion persists across representation choices and agrees with the relevant preference criterion.

AI USE STATEMENT

We did not use generative AI tools for any research tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. We used generative AI tools to assist with the creation and modification of scientific figures and images, as well as minor software development tasks such as code completion and debugging. All AI-assisted figures were reviewed by the authors for accuracy and consistency with the intended scientific content, and all AI-assisted code was manually inspected and tested for correctness. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work includes a human preference study in which pairwise judgments of model-generated text are used to construct dataset-specific human preference rankings for comparison with automatic evaluation metrics. The study contains 4,993 valid pairwise judgments over 150 generation cells across Book, Wikinews, and Wikitext. To assess the reliability of the preference signal, we include order-swapped control comparisons and a split-half ranking reliability analysis; the corresponding study counts and reliability summaries are reported in Appendix E. The human judgments are used solely as an evaluation signal for the analysis presented in this work. Details regarding participant recruitment, consent, compensation, and institutional ethics review are provided according to the study protocol.

REPRODUCIBILITY STATEMENT

We provide the experimental and analysis details needed to reproduce the reported results throughout the main paper and appendices. Section 3 describes the generation data, datasets, generators, decoding configurations, measurement grid, aggregation procedure, and comparison units used in the analyses. Appendix A specifies the representation models together with their architectures, pooling and normalization choices, and the complete evaluation grid. Appendix B provides the definitions and denominators for the ranking-based statistics. The aggregation, repeat-run, normalization, and pooling controls are described in Section 6 and Appendices C and D. Section 7 and Appendix E document the human-preference evaluation and its reliability checks. Appendix F defines the additional distributional metrics and reports the dataset-specific alignment analyses, while Appendix G gives the definition and evaluation procedure for the representation-panel rank aggregation. The generated corpora used in the main experiments are from the publicly documented Decoding Decoded benchmark, as described in Section 3.1.

REFERENCES

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(85)80012-4. URL https://www.sciencedirect.com/science/article/pii/S0364021385800124.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity, 2021. URL https://arxiv.org/abs/2007.14966.

Jasper Kyle Catapang. Position: Toward a metric typology for language model evaluation. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics*, pp. 1015–1020, 2026. doi: 10.18653/v1/2026.gem-main.78. URL https://aclanthology.org/2026.gem-main.78/.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui

Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

Yuanhao Ding, Esteban Garces Arias, Meimingwei Li, Julian Rodemann, Matthias Aßenmacher, Danlu Chen, Gaojuan Fan, Christian Heumann, and Chongsheng Zhang. GUARD: Glocal uncertainty-aware robust decoding for effective and efficient open-ended text generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7202–7226, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.380. URL https://aclanthology.org/2025.findings-emnlp.380/.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14932–14948, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.681. URL https://aclanthology.org/2026.acl-long.681/.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8798321486948322be2b4d658744ba72-Abstract-Conference.html.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12295–12313, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/34899013589ef41aea4d7b2f0ef310c1-Paper-Conference.pdf.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch (eds.), *Proceedings of the First Workshop on Neural Machine Translation*, pp. 56–60, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3207. URL https://aclanthology.org/W17-3207/.

Mingqi Gao, Xinyu Hu, Li Lin, and Xiaojun Wan. Analyzing and evaluating correlation measures in NLG meta-evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2199–2222, 2025. doi: 10.18653/v1/2025.naacl-long.111. URL https://aclanthology.org/2025.naacl-long.111/.

Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pp. 15060–15080, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.885. URL https://aclanthology.org/2024.findings-emnlp.885/.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9992–10020, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.668/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao

Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. On the blind spots of model-based evaluation metrics for text generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12067–12097, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.674. URL https://aclanthology.org/2023.acl-long.674/.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab

Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Pum Jun Kim, Yoojin Jang, Jisu Kim, and Jaejun Yoo. TopP&R: Robust support estimation approach for evaluating fidelity and diversity in generative models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31266–31305, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/185969291540b3cd86e70c51e8af5d08-Abstract-Conference.html.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0234c510bc6d908b28c70ff313743079-Paper.pdf.

Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevaleyre, and Alexandre Allauzen. Exploring precision and recall to assess the quality and diversity of LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 11418–11441, 2024. doi: 10.18653/v1/2024.acl-long.616. URL https://aclanthology.org/2024.acl-long.616/.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023b. URL https://arxiv.org/abs/2308.03281.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019a. URL https://arxiv.org/abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019b. URL https://arxiv.org/abs/1907.11692.

Quentin Malartic, Nilabhra Roy Chowdhury, Ruxandra Cojocaru, Mugariya Farooq, Giulia Campesan, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Maksim Velikanov, Basma El Amel Boussaha, Mohammed Al-Yafeai, Hamza Alobeidli, Leen Al Qadi, Mohamed El Amine Seddik, Kirill Fedyanin, Reda Alami, and Hakim Hacid. Falcon2-11b technical report, 2024. URL https://arxiv.org/abs/2407.14885.

Katelyn X. Mei, Yi-Li Hsu, Minjoon Choi, Zongwan Cao, Chenjun Xu, Bingbing Wen, Su Lin Blodgett, and Lucy Lu Wang. Illusions of the gold standard: A large-scale analysis of human evaluation protocols for long-form text generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 13939–13961, 2026. doi: 10.18653/v1/2026.acl-long.635. URL https://aclanthology.org/2026.acl-long.635/.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023. doi: 10.1162/tacl_a_00536. URL https://aclanthology.org/2023.tacl-1.7/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=IPmzyQSiQE. Reproducibility Certification.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4816–4828. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/260c2432a0eecc28ce03c10dadc078a4-Paper.pdf.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92, 2023. URL http://jmlr.org/papers/v24/23-0023.html.

Tiago Pimentel, Clara Isabel Meister, and Ryan Cotterell. On the usefulness of embeddings, clusters and strings for text generation evaluation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bvpkw7UIRdU.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020b. URL https://jmlr.org/papers/v21/20-074.html.

Julian Rodemann, Esteban Garces Arias, Christoph Luther, Christoph Jansen, and Thomas Augustin. A statistical case against empirical human-ai alignment, 2025. URL https://arxiv.org/abs/2502.14581.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. Automatic metrics in natural language generation: A survey of current evaluation practices, 2024. URL https://arxiv.org/abs/2408.09169.

Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5799–5808, 2019. URL https://proceedings.mlr.press/v97/simon19a.html.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation, 2022. URL https://arxiv.org/abs/2202.06417.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs seq: An open suite of paired encoders and decoders, 2026. URL https://arxiv.org/abs/2507.11412.

Genta Winata, David Anugraha, Lucky Susanto, Garry Kuwanto, and Derry Wijaya. MetaMetrics: Calibrating metrics for generation tasks using human preferences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e1b248453bca182b6138b8c14a75340d-Abstract-Conference.html.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. Assessing dialogue systems with distribution distances. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2192–2198, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.193. URL https://aclanthology.org/2021.findings-acl.193/.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings, 2024. URL https://arxiv.org/abs/2309.07597.

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, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Wajdi Zaghouani. Position: What are we measuring? rethinking evaluation in natural language generation. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics*, pp. 1021–1028, 2026. doi: 10.18653/v1/2026.gem-main.79. URL https://aclanthology.org/2026.gem-main.79/.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models, 2025a. URL https://arxiv.org/abs/2412.19048.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025b. URL https://arxiv.org/abs/2506.05176.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025c. URL https://arxiv.org/abs/2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yiming Ai, and Rui Wang. Improving open-ended text generation via adaptive decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62386–62404. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhu24d.html.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015. URL https://arxiv.org/abs/1506.06724.

A Representation Panel and Evaluation Grid

A.1 Feature extraction

Table 1 lists the feature-processing configurations in the sweep implementation. The panel includes causal-language-model representations, bidirectional representations, T5 encoder features, and retrieval embeddings. GPT-2, RoBERTa, and T5 represent established language-model families (Radford et al., 2019; Liu et al., 2019b; Raffel et al., 2020b). ModernBERT and Qwen3 Embedding extend the panel to more recent representation models (Warner et al., 2025; Zhang et al., 2025c). The Ettin suite supplies encoder and decoder variants (Weller et al., 2026).

Table 1: Representation configurations in the sweep. U uses unnormalized features; L2 applies unit normalization. Pooling specifies how token features become a text-level vector.

Representation	Model type	Pooling	Group
GPT-2 Large	Causal LM	Last token	U
Ettin decoder 400M	Causal LM	Last token	U
Ettin encoder 400M	Bidirectional encoder	CLS	U
RoBERTa Large	Bidirectional encoder	CLS	U
ModernBERT Large	Bidirectional encoder	CLS	U
T5 Large	T5 encoder	Mean	U
Qwen3-Embed 0.6B	Retrieval embedding	Last token	L2
BGE Large	Retrieval embedding	CLS	L2
GTE Large v1.5	Retrieval embedding	CLS	L2
Nomic Embed v1.5	Retrieval embedding	Mean	L2
Stella 400M	Retrieval embedding	Mean	L2
MPNet Base v2	Retrieval embedding	Mean	L2

Last-token pooling selects the last non-padding hidden state. CLS pooling selects the first-token hidden state, while mean pooling averages hidden states over non-padding tokens. For the L2 group, the pooled vector is normalized to unit Euclidean norm. These are properties of the evaluated feature configurations. The group labels in the main analysis refer to these feature treatments.

A.2 Text length and quantization

The pre-truncation routine tokenizes each text with GPT-2, retains at most L tokens, and converts a truncated sequence back to text for the representation model. Texts shorter than the budget remain unchanged. Reference and generated corpora pass through the same pre-truncation procedure. The representation model subsequently applies its own tokenizer and input limit.

The MAUVE call receives the two feature arrays and an explicit bucket count K. The grid varies K over 25, 50, 100, 250, and 500. The original archived score table records a fixed representation-input budget of 512 tokens for most models and 384 for MPNet. These input budgets describe that archived run, including the introduction example. They are distinct from the GPT-2 pre-truncation budget L.

B Comparison Units and Ranking Definitions

B.1 Objects being compared

Table 3 specifies the ranked objects and fixed factors for each analysis in Section 3. A generator ranking summarizes decoding configurations before comparing representations. A cell ranking retains those configurations as separate items. Human alignment uses the subset of cells with preference judgments within each dataset.

Table 2: Evaluation grid and analysis index. Setting combinations describe the evaluation design. Generation cells identify dataset–generator–decoding corpora.

Factor	Values or count
Datasets	Book, Wikinews, Wikitext
Generators	7
Generation cells in the index	1,175
Representation models	12
Pre-truncation lengths (GPT-2 tokens)	64, 128, 256
Bucket counts	25, 50, 100, 250, 500
Evaluation-setting combinations	180
Dataset–length–bucket conditions	45
Possible representation pairs per condition	66

Table 3: Comparison units. Each row states what is ranked, what stays fixed, and what changes. A model-level ranking contains seven model-level aggregates.

Analysis	Ranked objects	Fixed factors	Varied factor
Model-level representation swap	Model medians	Dataset, L, K	Representation
Cell-level representation swap	Generation cells	Dataset, L, K	Representation
Bucket swap	Generation cells	Dataset, L, representation	K
Truncation swap	Generation cells	Dataset, K, representation	L
Human alignment	Human-scored cells	Dataset, metric configuration	Metric vs. preference ordering

B.2 Rank agreement and selection

Kendall's tie-adjusted coefficient compares concordant and discordant item pairs:

$$\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_x)(N_c + N_d + T_y)}}. \tag{4}$$

Here N_c and N_d count concordant and discordant pairs, while T_x and T_y count pairs tied only in each respective ranking. Positive values indicate that concordant pairs predominate; negative values indicate that discordant pairs predominate. The ranked objects remain the same on the two sides of each comparison.

First-place agreement is the fraction of representation-pair comparisons that select the same highest-ranked generator. At a fixed dataset, length, and bucket count, the rank span of generator g is

$$\Delta r_g(d, L, K) = \max_e r_g(d, e, L, K) - \min_e r_g(d, e, L, K), \tag{5}$$

where r_g is its assigned rank. Table 5 reports the mean rank span from the analysis. The 45 conditions and 66 representation pairs admit up to 2,970 pairwise comparisons on the full grid.

The score-spread ratio in Table 4 follows from the two summary IQRs:

$$\frac{0.640}{0.139} \simeq 4.604. \tag{6}$$

It summarizes the relative spread across the selected settings. Kendall agreement and first-place agreement subsequently test whether changes in the numerical scores also alter the ordering.

Tables 4 and 5 collect the score-level (RQ1) and model-level (RQ2) summaries shown in Figures 2 and 3.

Table 4: Score variability over the evaluated settings. IQR is measured on the raw MAUVE scale. Variance shares refer to the raw-score decomposition shown in Figure 2(b).

Quantity	Value
Summary IQR across 180 measurement settings	0.640
Summary IQR across decoding configurations	0.139
Ratio of the two summary IQRs	≈ 4.6
Raw-score variance share: representation model	46.0%
Raw-score variance share: bucket count	1.9%

Table 5: Model-level consequences of representation choice. The negative correlation rate compares settings with the canonical ranking; top-1 agreement compares representation pairs. These use different denominators.

Statistic	Value
Mean representation-pair Kendall τ	0.30
Same first-ranked generation model	30%
Mean best-to-worst rank span	5.2
Settings negatively correlated with the canonical ranking	19%
GPT-2 Large vs. Ettin decoder: model-level τ	0.88
Unnormalized vs. L2-normalized pairs: mean model-level τ	0.17

B.3 ARCHIVED EXAMPLE USED IN THE INTRODUCTION

Figure 1 uses the archived Wikitext sweep at $L = 256$ and $K = 500$. Table 6 gives the underlying generator medians and ranks. Duplicate full evaluation keys are resolved by retaining their first occurrence in the archived table, followed by median aggregation over the available decoding configurations. Both representations use the same configuration set for each generator. They also share the reference and generated corpora, with a fixed representation-input limit of 512 tokens. The resulting Kendall agreement is -0.524.

Table 6: Archived scores underlying the introduction example. n_c is the number of available decoding configurations for each generator. Medians are rounded for display; ranks use the unrounded values.

		GPT-2 Large		Qwen3-Embed	
Generator	n_c	Median	Rank	Median	Rank
Falcon	59	0.019986	1	0.716703	7
Mistral v0.3	52	0.019624	2	0.817207	4
Qwen	38	0.018684	3	0.792848	5
GPT-2 XL	50	0.016571	4	0.777232	6
Llama 3	59	0.011851	5	0.836466	1
Mistral v0.1	55	0.010437	6	0.824799	3
DeepSeek	59	0.008559	7	0.826200	2

This example illustrates how the identity of the top-ranked generator can change. The model-level summaries in Table 5 come from the analysis index, rather than from recomputing an overall mean from this archived example.

C Decoding Aggregation Checks

The aggregation analysis examines whether the generator-level representation comparison depends on the choice of summary over decoding configurations. Table 7 records two results. The first measures the spread of mean representation-swap agreement across the tested rules. The second compares each alternative rule's generator ranking with the median-based ranking.

Table 7: Sensitivity to the tested decoding-aggregation rules.

Statistic	Value
Maximum minus minimum mean representation-swap agreement	0.04
Minimum ranking agreement between median and an alternative rule	0.86

The representation-swap summary changes little across the tested rules, and the alternative generator rankings agree closely with the median-based rankings. Together, these results show that the observed representation sensitivity persists across the tested aggregation choices.

D Evaluation-Setting and Feature-Processing Results

D.1 Comparisons across evaluation factors

Table 8 records the setting comparisons used in RQ3. The upper block ranks generation cells while changing one factor at a time. The lower block reports two generator-level comparisons after decoding aggregation. The repeated-run results appear in Figure 4(a).

Table 8: Agreement under evaluation-setting changes. The two blocks use different ranking units.

Comparison	Kendall agreement
Generation-cell rankings	
Representation swap	0.395
Bucket-count swap	0.646
Text-length swap	0.728
Generator rankings after decoding aggregation	
GPT-2 Large vs. Ettin decoder	0.88
Pairs crossing U and L2 groups	0.17

D.2 Normalization and pooling configurations

The normalization controls evaluate GPT-2 Large with normalized features and BGE with unnormalized features. Table 9 reports the mean score and bucket-swap agreement for each configuration. The agreement compares rankings across bucket choices while holding the representation and feature treatment fixed.

Table 9: Results for the evaluated normalization-control configurations. Each row describes one feature treatment and its sensitivity to bucket count.

Feature configuration	Mean MAUVE	Bucket-swap agreement
GPT-2 Large with L2 normalization	0.80	0.55
BGE without L2 normalization	0.55	0.72

The pooling comparison uses GPT-2 Large as a common reference. Table 10 shows the standard Ettin-encoder configuration and the mean-pooling control, together with the standard Ettin-decoder comparison. Mean pooling raises the Ettin-encoder agreement from 0.590 to 0.620. The difference is 0.030, and agreement remains below the Ettin-decoder reference comparison.

Table 10: Cell-level agreement with GPT-2 Large under the evaluated Ettin feature configurations.

Representation configuration	Kendall agreement with GPT-2 Large
Ettin decoder, standard	0.865
Ettin encoder, CLS pooling	0.590
Ettin encoder, mean pooling	0.620

E Human Preference Study

E.1 Study record

The study asks annotators to compare generated continuations for coherence, factual consistency, and fluency. Judgments provide a dataset-specific preference ordering over the human-scored generation cells. Table 11 records the cleaned study counts. Seven of the 5,000 planned judgments were missing or incomplete.

Table 11: Human-study counts and reliability summaries.

Quantity	Value
Planned pairwise judgments	5,000
Valid current-plan judgments	4,993
Missing or incomplete judgments	7
Scored generation cells across three datasets	150
Excluded pilot judgments	347
Outcomes: A / B / tie	2,856 / 2,005 / 132
Order-swapped control comparisons	400
Consistent order-swapped controls	304
Order consistency	76.0%
Split-half ranking reliability	0.64

E.2 Reliability and metric comparison

For an order-swapped control, the presentation order of the two candidates is reversed. A consistent judgment preserves the preference after accounting for that reversal. The resulting 304/400 rate measures consistency under order reversal. Split-half ranking reliability compares preference orderings constructed from subsets of the judgments. These checks describe the preference data, while the metric comparison measures agreement between the preference ordering and a MAUVE ordering.

The main Wikitext comparison uses the mean-alignment results for GPT-2 Large and T5 Large. Figure 5(a) reports these means under the same metric.

F Multi-Metric Alignment Across Datasets

F.1 How to read the matrices

Tables 12, 13, and 14 display a multi-metric analysis. Each entry is the maximum observed Kendall agreement for one representation–metric pair over its evaluated truncation lengths and, for MAUVE, bucket settings. Bold identifies the largest displayed value in a metric column. These are configuration maxima from the evaluation, separate from the means in Figure 5(a).

Within MAUVE, T5 has maximum observed agreement of 0.783 on Wikitext, 0.660 on Book, and 0.604 on Wikinews. The corresponding GPT-2 values are 0.595, 0.522, and 0.507. These comparisons describe the strongest evaluated configurations within each representation. Across metrics, the highest observed representation changes: on Wikitext, T5 leads MAUVE, BGE leads FBD, and Ettin

Table 12: Wikitext: multi-metric evaluation, maximum observed Kendall τ_b with the human ranking.

Representation	MAUVE	FBD	KID	MMD	Cent. L2	Cent. cos	P@5	R@5
MPNet	0.365	0.461	0.685	0.685	0.683	0.682	0.062	0.209
BGE	0.567	**0.765**	**0.740**	**0.739**	**0.739**	**0.731**	0.176	0.315
Ettin decoder	0.608	0.595	0.633	0.642	0.646	0.616	0.189	**0.701**
Ettin encoder	0.466	0.257	0.437	0.437	0.448	0.464	0.165	0.696
GPT-2 Large	0.595	0.505	0.554	0.574	0.567	0.553	0.129	0.616
GTE	0.298	0.512	0.652	0.667	0.670	0.670	−0.101	0.118
ModernBERT	0.502	0.228	0.468	0.471	0.466	0.474	0.201	0.652
Nomic	0.442	0.682	0.700	0.714	0.714	0.722	0.415	−0.189
Qwen3-Embed	0.252	0.244	0.529	0.527	0.523	0.556	−0.022	−0.053
RoBERTa	0.683	0.629	0.669	0.680	0.667	0.667	0.460	0.491
Stella	0.466	0.500	0.601	0.634	0.631	0.613	0.308	0.037
T5 Large	**0.783**	0.670	0.695	0.721	0.695	0.726	**0.749**	0.675

Table 13: Book: multi-metric evaluation, maximum observed Kendall τ_b with the human ranking.

Representation	MAUVE	FBD	KID	MMD	Cent. L2	Cent. cos	P@5	R@5
MPNet	0.574	0.558	**0.595**	0.592	0.592	0.592	0.318	0.205
BGE	0.566	0.567	0.556	0.558	0.556	0.551	0.541	0.016
Ettin decoder	0.546	0.491	0.579	0.574	0.569	0.531	0.145	0.624
Ettin encoder	0.336	0.272	0.221	0.174	0.234	0.229	0.070	0.385
GPT-2 Large	0.522	0.478	0.471	0.489	0.492	0.474	0.175	**0.675**
GTE	0.517	0.430	0.429	0.433	0.422	0.419	0.346	0.340
ModernBERT	0.190	0.001	−0.042	−0.038	−0.042	−0.043	−0.092	0.219
Nomic	0.530	0.607	0.587	0.590	0.592	0.585	0.480	−0.021
Qwen3-Embed	0.373	0.401	0.417	0.411	0.406	0.398	−0.141	−0.012
RoBERTa	0.424	0.433	0.368	0.290	0.373	0.378	0.414	0.009
Stella	0.437	0.396	0.378	0.373	0.371	0.370	0.141	0.090
T5 Large	**0.660**	**0.608**	0.595	**0.615**	**0.595**	**0.616**	**0.651**	0.447

Table 14: Wikinews: multi-metric evaluation, maximum observed Kendall τ_b with the human ranking.

Representation	MAUVE	FBD	KID	MMD	Cent. L2	Cent. cos	P@5	R@5
MPNet	0.466	0.502	0.443	0.446	0.445	0.465	0.492	0.384
BGE	0.516	0.564	0.502	0.495	0.495	0.486	0.418	0.464
Ettin decoder	0.488	0.466	0.428	0.431	0.434	0.446	0.395	0.549
Ettin encoder	0.270	0.258	0.238	0.256	0.226	0.230	0.328	0.452
GPT-2 Large	0.507	0.493	0.478	0.502	0.501	0.513	0.220	0.559
GTE	0.449	0.510	0.392	0.407	0.406	0.403	0.515	0.353
ModernBERT	0.140	0.188	0.126	0.091	0.125	0.126	0.324	0.213
Nomic	0.529	0.588	0.504	0.557	0.551	0.552	0.611	0.093
Qwen3-Embed	0.411	0.468	0.376	0.403	0.403	0.421	0.368	0.264
RoBERTa	0.336	0.418	0.366	0.369	0.363	0.363	0.438	0.138
Stella	0.455	0.498	0.417	0.428	0.430	0.425	0.538	0.347
T5 Large	**0.604**	**0.596**	**0.561**	**0.558**	**0.563**	**0.561**	**0.631**	**0.588**

decoder leads Recall@5. The matrices document the joint dependence on representation, metric, and dataset.

F.2 METRICS REPRESENTED IN THE MATRICES

MAUVE compares bucket-occupancy distributions through a divergence-frontier summary (Pillutla et al., 2021; 2023). Fréchet BERT Distance (FBD) compares Gaussian approximations to feature distributions (Xiang et al., 2021). Maximum Mean Discrepancy (MMD) compares distributions

through kernel expectations (Gretton et al., 2012); KID is an additional kernel-distance measure in the evaluation. Centroid L2 and centroid cosine compare the means of the reference and generated features by Euclidean distance and cosine similarity. Distance-based scores are oriented so that closer agreement with the reference corpus corresponds to a higher-ranked generation cell.

Precision@5 and Recall@5 use fifth-nearest-neighbor support neighborhoods (Kynkäänniemi et al., 2019). Precision measures the fraction of generated features within the reference support. Recall measures the fraction of reference features within the generated support. These columns describe complementary aspects of the distributional comparison.

G Exploratory Representation-Panel Aggregation

The supplementary panel analysis combines representation-specific rankings by averaging percentile ranks. For a compared system s, let $r_{s,e}$ be its midrank under representation e, with rank one denoting the highest score. For $N > 1$ systems, define

$$p_{s,e} = 1 - \frac{r_{s,e} - 1}{N - 1}, \qquad \mathrm{EPRA}(s) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} p_{s,e}. \tag{7}$$

This gives each representation equal weight after converting its scores to ranks within the compared set. The aggregation depends on the included representations and the set of compared systems.

The evaluated three-representation panel uses GPT-2 Large, T5 Large, and BGE Large with MMD as the base distributional metric. Its Kendall agreement with the human preference ordering is 0.785 on Wikitext, 0.580 on Book, and 0.550 on Wikinews. These results provide an exploratory application of combining representation-specific rankings. The main RQ4 comparison keeps MAUVE fixed and evaluates representation choice directly.