Reference-Based Analysis of Coherence and Diversity in Open-Ended Text Generation

Esteban Garcés Arias
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Esteban.GarcesArias@stat.uni-muenchen.de

Abstract

Evaluating open-ended text generation involves understanding how different properties of a continuation relate to its perceived quality. We present a reference-based framework for examining coherence and diversity through three perspectives: aligning their evolution with human trajectories, comparing their summaries with a human continuation of the same prompt, and estimating their likelihood under a human reference distribution. Experiments with human quality ratings suggest that diversity-based alignment and mean-based comparisons capture quality-related variation, although the comparisons do not establish a predictive advantage for temporal alignment over simpler baselines. Reference likelihood also shows positive associations with ratings, with results varying across reference configurations and scoring horizons. Together, these analyses provide a structured way to examine how measured coherence and diversity relate to human judgments, while distinguishing similarity to human references from quality itself. Code and analysis resources are available at https://github.com/EstebanGarces/likely_human.

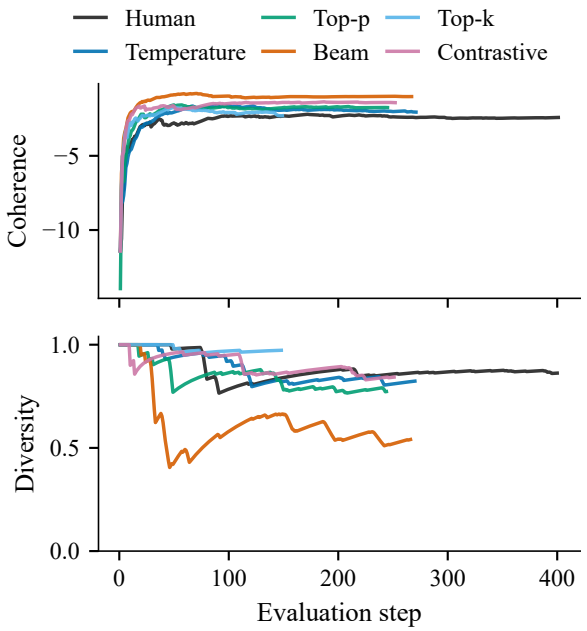


Figure 1: How coherence (top) and diversity (bottom) evolve within human and machine continuations of the same prompt. Each curve follows one continuation as more of its text is evaluated. Higher coherence values indicate greater model predictability; higher diversity values indicate less n-gram repetition. The framework compares these curves through alignment, summaries, and human reference distributions (Table 2).

1 Introduction

Evaluating open-ended text generation is difficult because many different continuations can be appropriate for the same prompt. Reference overlap alone cannot express the full range of valid outputs, and quality includes several dimensions that can conflict (Gatt and Krahmer, 2018; Celikyilmaz et al., 2020; Howcroft et al., 2020). A continuation may be fluent but repetitive, or lexically varied but inconsistent with its context. Coherence and diversity therefore provide useful, though incomplete, perspectives on the quality of generated text (Hashimoto et al., 2019; Su et al., 2022).

Figure 1 illustrates how coherence and diversity measurements evolve as successive prefixes of a continuation are evaluated. These trajectories offer a view of the text beyond its final score. They also raise a broader question: which aspects of the measurements are useful when comparing a generated continuation with human writing?

We examine this question through three analytical facets. **Temporal alignment** compares the shape of a trajectory with human trajectories. **Summary comparison** examines how its average values and variability differ from those of a human continuation of the same prompt. **Reference-distribution likelihood** assesses how typical a joint coherence–diversity profile is within a human corpus. These perspectives share measurements but use human references in different ways, allowing

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1420–1438
October 17-21, 2026 ©2026 Association for Computational Linguistics

us to examine both sequential patterns and aggregate properties.

We use coherence and diversity as operational measures of model predictability and n-gram non-repetition. They capture selected properties relevant to generation, while discourse coherence and semantic diversity remain broader concepts. Our experiments examine how the resulting scores relate to overall-quality ratings of English machine continuations.

Three research questions guide the analysis:

1. **RQ1: Predictive information.** How do temporal and prompt-relative summary scores relate to human ratings compared with simpler baselines?

2. **RQ2: Length and references.** How sensitive are those associations to reference selection, length normalization, and the number of scored tokens?

3. **RQ3: Reference typicality.** How does joint-profile likelihood relate to ratings, and how does that relationship depend on feature preparation and reference modeling?

We contribute a framework that connects these three uses of human references, an empirical comparison with simple and established baselines, and sensitivity analyses of length and reference design. The experiments combine an analysis of archived full trajectories with controlled measurements at fixed token horizons. Together, they examine which relationships with human judgments persist across these choices and whether temporal comparison offers additional predictive information.

2 Related Work

Evaluation of open-ended generation. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) measure forms of lexical overlap. Their interpretation is limited when a prompt admits many acceptable continuations. Surveys emphasize both the diversity of evaluation criteria and the difficulty of relating automatic metrics to human judgments (Novikova et al., 2017; Sai et al., 2022; Howcroft et al., 2020). Our study focuses on the relationship between reference-based scores and overall-quality ratings of individual continuations.

Coherence and diversity. The tension between repetition and unpredictable output motivates contrastive search (Su et al., 2022), top-k sampling (Fan et al., 2018), and nucleus sampling (Holtzman et al., 2019). Comparative work shows that decoding preferences depend on the setting (Garces Arias et al., 2025c). Adaptive Contrastive Search (Garces Arias et al., 2024) and GUARD (Ding et al., 2025) use entropy signals over generation steps to adapt decoding. Min-k Sampling uses local changes in sorted logits to determine truncation boundaries independently of temperature scaling (Ding et al., 2026b). Variance-Calibrated Modulation reshapes the distribution before truncation through contextual token weighting and a logit-scale-aware repetition penalty (Ding et al., 2026a). These methods modify generation, complementing our analysis of coherence and diversity in the resulting continuations.

Composite and distributional metrics. BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021) use contextual representations or generation probabilities. MAUVE (Pillutla et al., 2021) compares distributions of human and machine text. Q*Text (Garces Arias et al., 2025b) combines normalized coherence, diversity, and perplexity through weighted Gaussian penalties. We distinguish its fixed initialization from parameters optimized against human ratings. We report MAUVE at the corpus level and evaluate the other scores through continuation-level correlations, reflecting their different units of analysis. Discourse-oriented measures such as DiscoScore (Zhao et al., 2023), prompted evaluators such as G-Eval (Liu et al., 2023), and controlled diagnostic benchmarks such as OpenMEVA (Guan et al., 2021) address complementary questions.

Human-reference comparisons. Generalized stochastic dominance compares decoding strategies across criteria and measurement scales (Garces Arias et al., 2025a); we examine continuation-level score correlations. Credal-set analysis compares human and model variation across multiple continuations (Garces Arias et al., 2025d), whereas our diversity signal measures repetition within a continuation. Work on token exclusion under decoding distinguishes detectability from perceived quality (Garces Arias et al., 2026). Its quality analysis shares the rating collection used here and therefore provides related context rather than independent validation.

Temporal alignment. DTW aligns numerical sequences that evolve at different rates (Sakoe and Chiba, 1978; Müller, 2007). Applied here, DTW aligns numerical coherence and diversity trajectories. The resulting paths locate differences in those measurements; interpreting them in terms of entities or narrative events would require additional linguistic analysis. Soft-DTW (Cuturi and Blondel, 2017) and explicit slope or change-point descriptors offer alternative temporal analyses; we use the DTW configuration specified in Section 4.1.

3 Data and Measurement

Table 1 summarizes the texts, human ratings, and measurement settings. We first analyze the available full trajectories, then use a shared measurement procedure on fixed-length prefixes to examine sensitivity to length and extraction settings.

3.1 Prompts, Continuations, and Ratings

A *prompt* is the fixed input text x. A *continuation* is the complete subsequent text $y = (y_1, \ldots, y_L)$, written by a person or a model. A *continuation prefix*, $y_{1:t}$, is only its first t evaluation tokens. A trajectory records measurements on successive prefixes of that same continuation. Each record therefore links the prompt and continuation to an ordered array of measurements.

The continuations and human ratings come from prior work (Garces Arias et al., 2025c,a); Table 1 records their settings. The prompts are expository, so the evaluation concerns open-ended continuation in that setting. Human texts supply references and are excluded from the main score–rating correlations. Prompt groups remain intact in uncertainty estimation.

Agreement between the annotators is linearly weighted Cohen's $\kappa = 0.324$ and Spearman's $\rho = 0.529$, motivating per-rater analysis and caution about fine-grained rankings. Continuation lengths vary substantially across methods; Appendix A reports their ranges and further agreement statistics.

3.2 The Extended Human Corpus

The extended collection supplies human references from five domains (Table 1). We retain its stored prompt and continuation boundaries. Missing cached features and exact overlap with primary human continuations reduce the archived reference set to 9,877 records. These exclusions do not rule out near duplicates or shared documents. Appendix A details segmentation and exclusions; controlled-horizon eligibility is documented in Appendix E.

3.3 Quality Trajectories and Cache Preparation

Equation 1 defines the intended coherence signal as a running average of token log probabilities under the external model OPT-2.7B (Zhang et al., 2022):

$$c_t = \frac{1}{t} \sum_{i=1}^{t} \log p_{\mathcal{M}}(y_i \mid x, y_{<i}). \tag{1}$$

Higher values indicate that the continuation is more predictable under the external model. We use this as a coherence proxy, recognizing that discourse-level properties such as logical consistency and entity continuity extend beyond predictability. The full-trajectory analysis uses released numerical trajectories. The controlled experiment computes token-aligned coherence and diversity measurements over prefixes of all 300 primary and 10,000 reference texts (Section 6.4). The primary cache stores running coherence and diversity scores but lacks a complete extraction configuration. The controlled experiment implements Equations 1 and 2; the conditioning, tokenization, and exact implementation underlying the archived values remain unverified. This limits interpretation of the archived scores even where their downstream calculations can be reproduced.

Let $G_n(y_{1:t})$ be the multiset of contiguous token n-grams, including repeated occurrences. Equation 2 defines the intended diversity signal as the product of non-repetition ratios for sizes two through four:

$$d_t = \prod_{n=2}^{4} \frac{|\operatorname{unique}(G_n(y_{1:t}))|}{|G_n(y_{1:t})|}. \tag{2}$$

An empty n-gram set contributes one. Values near one indicate little repeated n-gram material, giving a lexical measure of diversity. Adding unique material can increase a ratio after an earlier decrease, so the trajectory need not be monotone. For archived trajectories, the horizontal coordinate follows the stored evaluation steps.

The primary coherence cache contains L values per text and diversity contains $L - 1$; we calculate summaries over the available values without padding. In the extended cache, we first convert token log probabilities into cumulative means so that the coherence summaries describe the same intended quantity. The underlying arrays are rounded

Component	Design
Generation	Reported generator: Qwen2.5-7B. Temperature $\tau = 0.9$; top-p, $p = 0.95$; top-k, $k = 50$; beam search, $B = 5$; contrastive search, $k = 10$, $\alpha = 0.6$.
Primary collection	50 prompts (25 Wikinews, 25 WikiText), each with one human and five machine continuations. Correlations evaluate the 250 machine texts; 50 human texts supply references.
Human evaluation	Two five-point overall-quality ratings for each of the 300 texts; their arithmetic mean is the primary target.
Extended references	10,000 human records from BookCorpus, ROCStories, Wikinews, WikiText, and WritingPrompts; 9,877 retained for archived reference fitting.
Archived analysis	Released full-trajectory features; incomplete extraction provenance. Temporal alignment, prompt-relative summaries, and reference-profile likelihood.
Controlled analysis	Shared OPT-2.7B extraction at fixed continuation horizons. At 64 tokens: 40 prompt groups, 200 machine texts, 7,686 corpus references. At 128 tokens: 23 groups, 115 machine texts, 7,149 corpus references.
Comparisons	Endpoint and trajectory summaries, PCA, and fixed Q*Text; MAUVE supplies separate corpus-level context. Spearman correlations with prompt-cluster bootstrap intervals.

Table 1: Overview of the datasets, generation configurations, and evaluation procedures. Fixed-horizon analyses retain complete prompt groups and exclude human corpus references that exactly overlap with the primary human continuations.

and their extraction configurations are incomplete. We therefore treat the cached likelihood analysis as exploratory and evaluate extraction-compatible prefixes separately. Appendix D records the preparation steps and remaining uncertainty.

4 Three Analytical Facets

Table 2 summarizes how each facet uses the measurements and human references. Temporal alignment follows the progression of the measured signals through a continuation; summary comparison and likelihood examine their aggregate profiles. All three derive from cumulative prefix measurements and can therefore be affected by token order, although only temporal alignment explicitly matches positions across trajectories.

4.1 Temporal Alignment

The temporal facet compares a machine trajectory a with human trajectories b using DTW. Let $D(a, b)$ denote the alignment cost with scalar absolute differences, closed endpoints, and the `symmetric2` step pattern. We use no warping window. Appendix A.1 gives the recurrence and boundary conditions needed to reproduce the scores.

Let H contain the 50 human reference trajectories for the signal being compared. The primary temporal score averages their length-normalized distances:

$$s_{\mathrm{DTW}}(a) = -\frac{1}{50} \sum_{b \in H} \frac{D(a, b)}{(|a| + |b|)/2}. \quad (3)$$

We use the negative distance so that larger values indicate closer temporal alignment.

The all-reference score characterizes resemblance to the observed human trajectory collection; it does not condition reference selection on the evaluated prompt. To distinguish that choice, an ablation uses only the human continuation of the same prompt. Within this paired setting, we also remove length normalization while leaving the references and DTW recurrence fixed. This isolates the effect of distance normalization within the paired-reference comparison.

An alignment path provides a way to inspect where measured trajectories differ. Its usefulness for explaining particular textual properties, such as repetition or topic shifts, remains an empirical question.

4.2 Summary Comparison

For each trajectory we calculate its mean and population variance. Here L_m denotes the machine coherence-array length. Let $\bar{c}_m, \bar{d}_m$ be the machine means and $\bar{c}_h, \bar{d}_h$ the means of the human continuation for the same prompt. The normalized signed differences are

$$\Delta_c = \frac{(\bar{c}_m - \bar{c}_h)}{L_m}, \qquad \Delta_d = \frac{(\bar{d}_m - \bar{d}_h)}{L_m}. \quad (4)$$

We standardize the differences in Equation 4 over the 250 machine observations, using the sample standard deviation. Equation 5 averages the resulting components:

$$s_{\mathrm{mean}} = \tfrac{1}{2}\{z(\Delta_c) + z(\Delta_d)\}. \quad (5)$$

The variance composite replaces trajectory means with trajectory variances; the overall composite

Facet	Representation	Human reference	Interpretation
Temporal alignment	Ordered values of each signal	A collection of trajectories; paired reference in an ablation	Similarity of aligned trajectories
Summary comparison	Means and variances	Continuation of the same prompt	Signed difference from the matched human
Reference likelihood	Joint mean or variance profile	Distribution fitted to corpus profiles	Typicality under that reference model

Table 2: Three ways to compare coherence and diversity with human references. Temporal alignment compares ordered measurements, summary comparison uses the human continuation of the same prompt, and likelihood describes the profile within a human corpus.

averages all four standardized differences. Their signs are retained as defined, rather than selected to maximize correlation.

The composite expresses a signed contrast with the matched human continuation. Standardization centers each component, allowing positive and negative differences to offset one another. Its value depends on the collection used for standardization; application to new texts would require a fixed normalization reference.

Division by L_m weights a fixed difference more heavily for shorter texts. We examine this length-dependent choice through an ablation that removes the division and refits standardization. For equal-length texts, the division would leave the standardized scores unchanged.

4.3 Likelihood

The likelihood facet describes typicality of a joint summary vector u. We fit separate bivariate Gaussian reference models to the mean pair $(\bar{c}, \bar{d})$ and variance pair (v_c, v_d). The cached reconstruction fits the sample mean μ and covariance Σ to 9,877 retained human records. The controlled experiment refits these parameters using eligible records at each token horizon. We evaluate the log density in Equation 6,

$$s_{\text{lik}}(u) = -\log(2\pi) - \tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu), \quad (6)$$

which avoids numerical underflow while preserving the ranking of the density. Fitted parameters are provided in Appendix C.

Its relationship to quality is assessed through the human ratings. Diversity is bounded and variances are nonnegative, whereas a Gaussian has unbounded support. We therefore inspect the distributions and compare the full Gaussian with a diagonal-covariance Gaussian and a Gaussian-kernel density estimate using Scott's bandwidth. The alternatives use the same reference records and features, without rating-based tuning. We also fit each reference domain separately to assess sensitivity to corpus composition.

The archived likelihood comparisons remain exploratory because of the extraction limitations in Section 3.3. The controlled experiment evaluates likelihood with consistent extraction across both corpora.

5 Evaluation Protocol and Baselines

Common evaluation unit. The full-trajectory comparisons use the same 250 machine texts and the mean of their two ratings. Fixed-horizon analyses use eligible subsets. We calculate Spearman's rank correlation with average ranks for ties. Full-trajectory scores use the archived measurements described in Section 3.3.

Uncertainty and comparisons. We use 3,000 prompt-cluster bootstrap replicates, retaining all five machine outputs and their ratings within each sampled group. Sample-dependent score normalizations are refitted; human references and PCA loadings remain fixed. Intervals therefore condition on the reference resources and do not capture uncertainty from new annotators. Appendix A specifies the resampling and normalization details.

We report percentile 95% intervals for individual correlations and for paired differences calculated within the same replicate. A paired interval containing zero leaves the direction of the difference unresolved; it is not evidence of equivalence. The ablations are exploratory and their intervals are not adjusted for testing multiple alternatives. We additionally report correlations against each annotator to reveal dependence on the rating source.

Experimental comparisons. For RQ1, we compare temporal and summary scores with simpler and established baselines. For RQ2, we vary refer-

ences and normalization, then evaluate fixed token horizons. For RQ3, we examine reference preparation and density models using archived and controlled measurements. All analyses use full-text ratings; Appendices D and E give implementation details.

Endpoint and summary baselines. The final cached coherence and diversity values provide endpoint baselines. The trajectory mean and variance of each signal help assess whether aggregate properties capture similar associations to temporal comparison. These baselines require no human reference trajectory at scoring time.

PCA and fixed Q*Text. PCA provides an unsupervised combination of trajectory means and variances. We retain the released fit on all 300 texts and evaluate its fixed components on the 250 machines; component signs are arbitrary. Fixed Q*Text applies the published, untuned initialization to cached endpoint perplexity, coherence, and diversity (Garces Arias et al., 2025b). Its normalization and exact formula are given in Appendix B. Neither baseline is fitted to the quality ratings in this comparison.

Corpus-level context. We report the original MAUVE results descriptively because they compare whole text distributions, whereas our primary analysis evaluates individual continuations. MAUVE supplies one corpus-level value for each decoding strategy. Appendix B documents their inputs and settings.

6 Results

6.1 RQ1: Temporal and Summary Associations

Diversity alignment shows a stronger association with ratings than coherence alignment in the full-trajectory analysis (Table 3; Figure 2). Negative diversity DTW correlates with mean quality ratings at $\rho = 0.536$ (95% interval $[0.442, 0.615]$), while negative coherence DTW reaches 0.146 ($[0.012, 0.285]$). The stronger diversity association may reflect lexical repetition, limitations of the coherence proxy, or differences between decoding methods; the overall-quality ratings do not distinguish those explanations.

The summary mean composite reaches $\rho = 0.527$ ($[0.399, 0.639]$). Its variance counterpart correlates negatively (-0.525), and averaging all four components gives -0.106. Combining all components gives a weaker association in this sample, consistent with contributions in different directions (Section 4.2).

Score	ρ	95% interval
Fixed Q*Text	0.612	$[0.510, 0.677]$
Temporal: diversity	0.536	$[0.442, 0.615]$
Temporal: coherence	0.146	$[0.012, 0.285]$
Summary: mean	0.527	$[0.399, 0.639]$
Summary: variance	-0.525	$[-0.645, -0.367]$
Summary: all four	-0.106	$[-0.273, 0.009]$
Endpoint diversity	0.511	$[0.383, 0.615]$
Mean diversity	0.493	$[0.372, 0.593]$
PCA: PC1	-0.376	$[-0.479, -0.256]$
PCA: PC2	-0.535	$[-0.647, -0.402]$

Table 3: Association between each score and mean human quality ratings on 250 machine texts. Values are Spearman correlations with 95% prompt-cluster bootstrap intervals. Temporal scores use negative distances; PCA signs follow the released components.

Endpoint diversity alone reaches $\rho = 0.511$, and its trajectory mean reaches 0.493. The paired difference between diversity DTW and endpoint diversity is 0.025, with interval $[-0.082, 0.141]$. Against the trajectory mean, the difference is 0.044 ($[-0.060, 0.145]$). For RQ1, these comparisons leave the predictive advantage of temporal alignment unresolved.

The paired difference between diversity DTW and the summary mean composite is 0.010 ($[-0.125, 0.145]$). The interval permits meaningful differences in either direction, leaving their relative performance uncertain. Figure 3 shows the score–rating relationships by decoding strategy, including substantial within-score disagreement.

6.2 RQ1: Baselines and Annotator Sensitivity

Fixed Q*Text reaches $\rho = 0.612$ ($[0.510, 0.677]$), the largest positive point estimate among the primary scores in Table 3. Its parameters use the published initialization without fitting to these ratings. The paired difference for diversity DTW minus fixed Q*Text is -0.076 ($[-0.141, 0.021]$), which does not resolve their ordering.

The PCA results in Table 3 provide further context: the magnitude of PC2's association is similar to that of diversity DTW. Its arbitrary orientation and unsupervised objective, however, differ from a quality-calibrated score.

Table 8 shows that associations also vary by annotator. For example, the summary mean composite correlates 0.554 and 0.300 with the two raters,

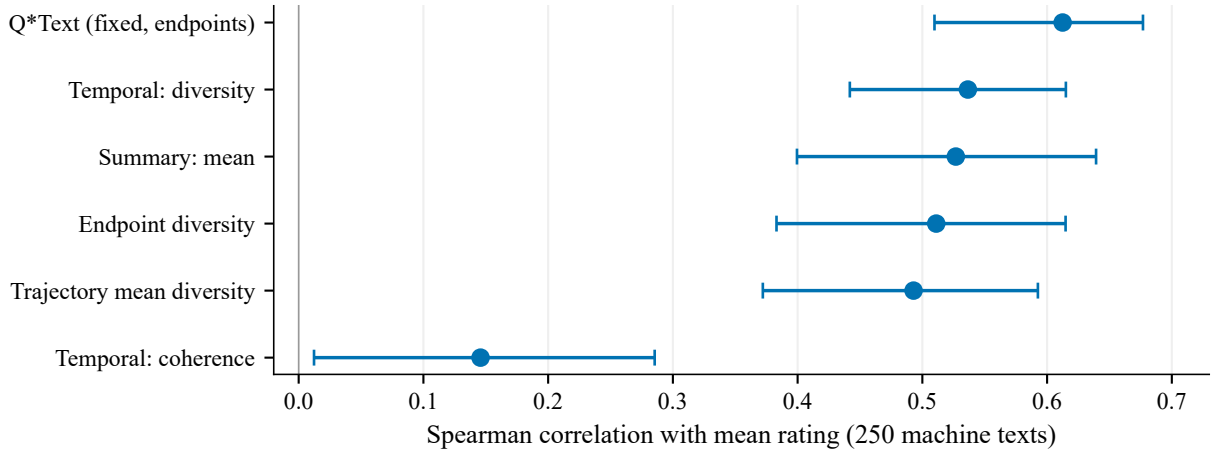


Figure 2: Association between each score and mean human quality ratings on the same 250 machine texts. Points show Spearman correlations and bars show 95% prompt-cluster bootstrap intervals. Temporal scores use negative DTW distances, so higher values indicate closer alignment. Paired comparisons between scores are reported in the text.

Score	ρ	95% interval
DTW div., all references	0.536	$[0.442, 0.615]$
DTW div., paired	0.512	$[0.415, 0.593]$
DTW div., paired, raw	0.436	$[0.326, 0.530]$
DTW coh., paired	0.254	$[0.102, 0.397]$
DTW coh., paired, raw	0.181	$[0.036, 0.321]$
Summary, divided by L	0.527	$[0.399, 0.639]$
Summary, no division	0.482	$[0.349, 0.585]$

Table 4: Length and reference ablations. Paired means one human reference for the same prompt. Raw DTW omits length normalization.

while diversity DTW correlates 0.507 and 0.436. These differences suggest caution when interpreting agreement with the average rating as a stable preference. Agreement between raters provides context for this variability, although it does not set a numerical ceiling on correlation with their average.

Corpus-level MAUVE ranks temperature sampling first (0.3207), followed by beam search (0.1006), contrastive search (0.0921), top-k (0.0524), and top-p (0.0225). These distributions do not reproduce the order of average human quality ratings: top-p has the highest machine rating, 2.69, and beam search the lowest, 2.00. With 50 texts per distribution, this comparison offers limited descriptive context for how corpus similarity and individual quality judgments may differ.

6.3 RQ2: Length and Reference Ablations

Table 4 addresses RQ2 by varying reference selection and normalization. Using only the human trajectory for the same prompt gives a diversity DTW correlation of 0.512. Removing length normalization from this same paired-reference procedure lowers the observed correlation to 0.436. The paired difference is 0.076 ($[0.044, 0.108]$), supporting length normalization for this particular diversity comparison. Its scope is the paired diversity comparison evaluated here.

For coherence, the paired-reference normalized score reaches 0.254, compared with 0.181 without normalization and 0.146 for the all-reference score. The two signals show different patterns across reference choices, suggesting that the usefulness of a matched reference may depend on the measured property.

The unweighted summary mean composite reaches 0.482, compared with 0.527 after division by machine length. The paired difference is 0.044 ($[-0.075, 0.183]$), which does not resolve an advantage from that weighting.

6.4 RQ2: Controlled Token Horizons

To examine the associations at a common scored length, we measure coherence and diversity using the same OPT-2.7B configuration for both corpora. We evaluate the first 64 and 128 continuation tokens in separate analyses, without padding shorter texts. We retain complete prompt groups and refit the human reference models at each horizon; Table 1 reports the eligible collections.

Table 5 reports positive diversity-DTW and variance-profile likelihood associations at both horizons. Adjusting for original continuation and prompt lengths retains positive associations. Sepa-

Score	64 tokens ($n = 200$)	128 tokens ($n = 115$)
DTW coherence	0.037 [-0.113, 0.199]	0.177 [0.017, 0.341]
DTW diversity	0.330 [0.222, 0.422]	0.440 [0.282, 0.571]
Gaussian mean	0.128 [-0.015, 0.276]	0.296 [0.160, 0.419]
Gaussian variance	0.338 [0.206, 0.456]	0.571 [0.428, 0.687]
Endpoint diversity	0.388 [0.277, 0.496]	0.547 [0.417, 0.655]
Summary mean	0.268 [0.101, 0.423]	0.353 [0.147, 0.536]

Table 5: Associations with full-text quality ratings when all scored continuations have a common token horizon. Values are Spearman correlations with 95% prompt-cluster bootstrap intervals. Each column uses the eligible texts and references for that horizon; the paired comparison in Appendix E holds membership constant.

rately adjusting score and rating ranks for decoding strategy also retains positive associations for both scores, whereas coherence-DTW intervals include zero at both horizons (Appendix E, Table 10). This suggests that these associations are not explained solely by the between-method differences captured by the rank adjustment.

On the same texts and reference bank, increasing the horizon from 64 to 128 changes the variance-profile correlation by 0.139 ($[0.019, 0.261]$). The associations remain sensitive to horizon, and ratings concern whole texts rather than isolated prefixes. Neither horizon establishes a diversity-DTW advantage over endpoint diversity. Appendix E provides extraction details, adjustment definitions, and paired comparisons.

6.5 RQ3: Reference-Distribution Likelihood

Reference likelihood also associates with ratings in the archived analysis (Table 9). The mean-profile Gaussian has a correlation of 0.409 and the variance-profile Gaussian reaches 0.598. These findings are conditional on the cache reconstruction and extraction limitations described in Section 3.3.

Diagonal Gaussian and Gaussian-kernel estimates also yield positive associations (Table 9). Domain-specific fits show further variation (Appendix C). These comparisons do not identify a preferred distribution or domain on held-out data. The quantile plots in Figure 5 indicate departures from marginal normality; neither Gaussian nor Gaussian-kernel fits enforce the bounded support of diversity.

For RQ3, the comparisons suggest that the strength of the likelihood association varies with reference preparation and modeling choices. The positive variance-profile associations in the shared-pipeline experiment (Section 6.4) indicate that such an association also occurs with consistent extraction, although its magnitude depends on horizon. Their potential use in a new domain would need evaluation against that domain's texts and quality criteria. Figures 7 and 8 compare the controlled mean-profile distributions by decoding method with the fitted human reference.

7 Discussion and Practical Use

The three facets offer different ways to relate generated text to human writing. Temporal alignment follows the evolution of the measured signals, summary comparison uses a human continuation of the same prompt, and likelihood places the joint profile within a broader human reference. The baseline comparisons assess how these different reference questions relate to quality prediction.

Across the experiments, diversity alignment and mean-based comparisons show positive associations with human judgments, while endpoint diversity also captures substantial variation in ratings. The paired comparisons leave a predictive advantage for DTW unresolved. An alignment may nevertheless be useful for examining where measured trajectories diverge. Establishing a connection between those differences and recognizable textual problems would require direct tests, for example with controlled repetition or topic shifts. Such perturbations can probe behavior beyond score–rating correlation (He et al., 2023).

Reference likelihood adds a distributional perspective. The variance-profile associations in the controlled experiment suggest that typicality can carry quality-related information, although the strength of that relationship varies with the scoring horizon. An unusual profile may also reflect an appropriate but uncommon continuation. Applications such as filtering would therefore need evidence about which kinds of atypical text should be retained, as well as validation of the reference model in the intended domain.

Practical use also depends on the available references. Summary comparison needs a human continuation of the same prompt; temporal alignment

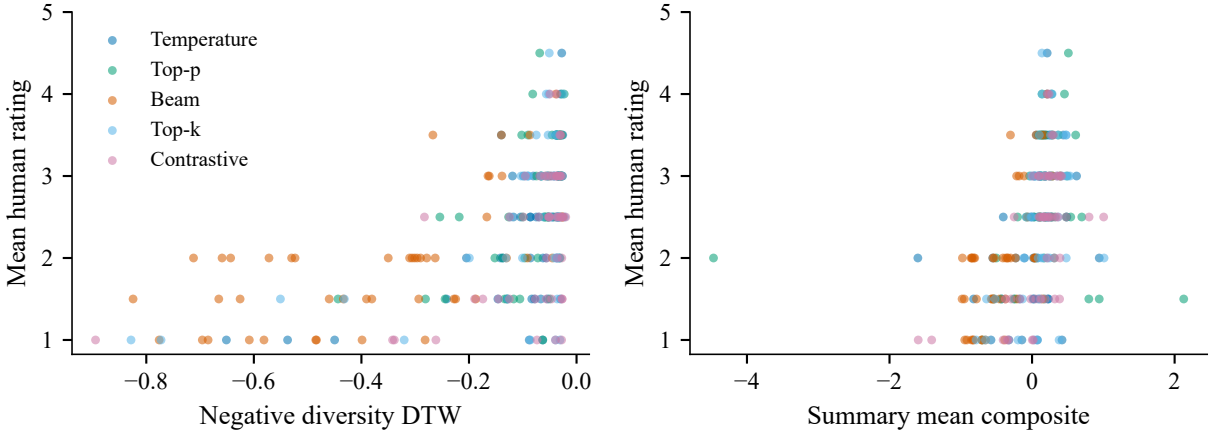


Figure 3: How diversity alignment (left) and the mean summary composite (right) relate to human quality ratings. Each point represents a machine continuation, colored by decoding strategy; some points overlap. The spread illustrates variation both within and between strategies.

and likelihood use human reference collections. Endpoint scores and fixed composites offer simpler alternatives when such resources are unavailable. Temporal alignment also has greater downstream computational cost than summary comparison or Gaussian scoring (Appendix A).

The overall-quality judgments leave open which textual properties drive the associations. More targeted annotation could help connect the numerical patterns to readers' assessments of coherence and diversity.

8 Conclusion

We presented a reference-based framework for analyzing coherence and diversity in open-ended text generation through temporal alignment, comparison with matched human continuations, and likelihood under human reference distributions. The experiments suggest that diversity-based alignment and mean-based comparisons capture variation associated with human quality judgments. Positive associations also occur for reference likelihood, although their strength varies with the reference configuration and the portion of text evaluated. The comparisons with simpler baselines leave the predictive advantage of temporal alignment uncertain. The framework brings these findings together while making explicit the different questions that human references can support. Further work could examine whether the observed patterns correspond to recognizable textual properties, such as repetition or topic shifts, and whether these relationships extend to other generation settings.

Limitations

The evaluation covers a small English sample from one generator and fixed decoding configurations. Two annotators provide overall-quality ratings with limited agreement, leaving generalization and dimension-specific judgments open to further study. Further, Gaussian support extends beyond the feature bounds, and neither Gaussian nor KDE scores are calibrated quality probabilities. Bootstrap intervals condition on fixed references and PCA loadings; exploratory comparisons lack multiplicity adjustment. Strategy-adjusted associations remain descriptive rather than causal.

Finally, the measurements represent model predictability and lexical repetition, offering limited coverage of discourse coherence, creativity, factuality, or safety. Broader evaluation would require alternative assessor models, independent ratings, held-out settings, and controlled textual perturbations.

Acknowledgments

Esteban Garcés Arias sincerely thanks the Mentoring Program of the Faculty of Mathematics, Statistics and Informatics at LMU Munich, and the Munich Center for Machine Learning (MCML) for their support. He also thanks the anonymous (meta) reviewers for their constructive feedback throughout the review process.

Ethics Statement

We affirm that our research adheres to the ACL Ethics Policy. This work involves the use of publicly available datasets and does not include any

personally identifiable information. An ethical concern worth mentioning is the use of language models for text generation, which may produce harmful content, either through intentional misuse by users or unintentionally due to the training data or algorithms. We declare that there are no conflicts of interest that could potentially influence the outcomes, interpretations, or conclusions of this research. All funding sources supporting this study are enumerated in the acknowledgments section. We have diligently documented our methodology, experiments, and results, and we commit to sharing our code, data, and other relevant resources to enhance reproducibility and further advancements in the field.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Marco Cuturi and Mathieu Blondel. 2017. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, pages 894–903. PMLR.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7202–7226, Suzhou, China. Association for Computational Linguistics.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. 2026a. Breaking the likelihood trap: Variance-Calibrated Modulation for large language model decoding. *Preprint*, arXiv:2606.22511. To appear at EMNLP 2026.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. 2026b. Min-k Sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14932–14948, San Diego, California, United States. Association for Computational Linguistics.

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

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Matthias Assenmacher, and Christoph Jansen. 2025a. Statistical multicriteria evaluation of LLM-generated text. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 338–351, Hanoi, Vietnam. 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 (GEM²)*, pages 631–654, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025c. 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, and Christian Heumann. 2025d. The geometry of creative variability: How credal sets expose calibration gaps in language models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 36–50, Suzhou, China. 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.

Esteban Garces Arias, Nurzhan Sapargali, Christian Heumann, and Matthias Aßenmacher. 2026. The truncation blind spot: How decoding strategies systematically exclude human-like token choices. *Preprint*, arXiv:2603.18482.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1689–1701.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097. Association for Computational Linguistics.

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

David M Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Meinard Müller. 2007. Dynamic time warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

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

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.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys*, 55(2):1–39.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. Discoscore: Evaluating text generation with bert and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883.

Appendix

A Data and Reproducibility Details

Figure 1 shows the first prompt in lexicographic identifier order (1017 Wikinews). The curves show the archived values without smoothing; their horizontal coordinate is the stored evaluation step. Table 6 summarizes ratings and stored lengths by method; Table 7 documents the cached reference exclusions. The analysis preserves the original prompt and method identifiers. Machine continuations are matched to their human references by prompt identifier. Cached all-reference DTW rows are linked through their original observation identifiers and checked against both method and rating. Recomputed DTW comparisons use the same recurrence as the R `dtw` defaults. Recomputing all 25,000 machine–human comparisons (250 texts, 50 references, two signals) reproduces every cached average to numerical precision.

For rating proximity, 82.7% of pairs differ by at most one point and the mean absolute difference is 0.823. These supplement the agreement statistics in Section 3.

The extended collection contains 10,000 records: BookCorpus (1,947), ROCStories (2,500), Wikinews (2,000), WikiText (1,314), and WritingPrompts (2,239). BookCorpus, Wikinews, and WikiText pair presegmented leading text with its following passage. ROCStories pairs a title with the full story; WritingPrompts pairs a prompt with its response. We retain these boundaries. Recovering the original document-to-segment selection procedure is outside the available cache. Prefix length fields in the extended CSV count characters, not tokens.

Seventy-five WritingPrompts records lack cached metric arrays, leaving 9,925 complete records. Whitespace-normalized exact matching identifies 48 continuations also used as primary human references; excluding them leaves 9,877 records for archived reference fitting. The released analysis includes record counts and both exclusion masks. These checks do not rule out near duplicates or shared source documents.

Execution. The computational analysis uses Python, NumPy, pandas, SciPy, scikit-learn, and `dtw-python` 1.5.3. The analysis seed is 20260907. Plotting uses Matplotlib. OPT-2.7B inference supplies the controlled prefix experiment in Appendix E; the full-trajectory results use the existing caches. PCA uses the archived component scores, and MAUVE uses the archived corpus-level outputs. Source files and result tables accompany the manuscript.

Method	Mean rating	Length range
Human	3.26	129–898
Temperature	2.42	20–631
Top-p	2.69	11–266
Top-k	2.54	48–408
Beam	2.00	115–270
Contrastive	2.30	22–483

Table 6: Mean human ratings and observed trajectory lengths for each method, with 50 texts per method. Length counts the stored coherence measurements.

Domain	Complete	Retained
BookCorpus	1947	1947
ROCStories	2500	2500
Wikinews	2000	1976
WikiText	1314	1290
WritingPrompts	2164	2164

Table 7: Reference records with cached features before and after excluding exact overlaps with primary human continuations. The 75 missing-feature records are excluded before these counts.

Bootstrap implementation. We use 3,000 bootstrap replicates with seed 20260907. Each replicate samples prompt groups with replacement and retains all five machine outputs and their ratings within each group. The number of sampled groups equals the number evaluated: 50 for the full-trajectory analysis, or 40 and 23 for the fixed-horizon analyses. We refit each score's sample-dependent normalization within every replicate, including summary standardization and, for Q*Text, min–max ranges. The external reference corpus, temporal human-reference collection, and PCA loadings remain fixed. Intervals therefore describe uncertainty across the observed prompt groups conditional on these reference resources; they do not include uncertainty from sampling new annotators or refitting the reference corpus.

Computational costs. After feature extraction, summary computation is linear in trajectory length and a two-dimensional Gaussian score has constant evaluation cost. Standard DTW requires $O(|a||b|)$ work for each reference pair, multiplied here by 50 references; kernel density evaluation also scales with the reference collection. These downstream costs exclude language-model feature extraction,

Score	Rater A	Rater B
Temporal: diversity	0.507	0.436
Temporal: coherence	0.160	0.090
Summary: mean	0.554	0.300
Summary: variance	−0.521	−0.363
Summary: all four	−0.044	−0.200
Endpoint diversity	0.560	0.276
Mean diversity	0.543	0.263
Variance diversity	−0.573	−0.290
Endpoint coherence	−0.226	−0.053
Fixed Q*Text	0.599	0.437
PCA: PC1	−0.423	−0.204
PCA: PC2	−0.475	−0.455

Table 8: Association of each score with the two annotators' individual quality ratings. Values are Spearman correlations on the same 250 machine texts.

which may dominate. The summary facet additionally needs a human continuation of the evaluated prompt. A reference-free endpoint score or fixed composite can be preferable when those resources are unavailable.

A.1 DTW Recurrence

The temporal facet compares each machine trajectory a with human trajectories b. We use DTW with scalar absolute local cost, closed endpoints, no warping window, and the `symmetric2` step pattern. With $\delta_{ij} = |a_i - b_j|$, the interior recurrence is

$$G_{ij} = \min\{G_{i-1,j} + \delta_{ij}, \\ G_{i,j-1} + \delta_{ij}, \\ G_{i-1,j-1} + 2\delta_{ij}\}, \qquad (7)$$

with $G_{11} = \delta_{11}$ and boundary paths accumulating their local costs. The distance is $G_{|a|,|b|}$. Equation 7 fixes the step weights, which can affect the resulting scores.

Here $D(a, b) = G_{|a|,|b|}$ in Equation 3. The normalization uses the actual lengths of the two signal arrays.

B Additional Primary Results

All correlations in Table 8 use the same 250 machine texts. The negative variance associations are reported with their original sign. For PCA, the orientation follows the released scores. A sign change in a principal component leaves its information content unchanged, so comparison of its raw signed correlation with a deliberately oriented quality score requires care.

The fixed Q*Text implementation normalizes each of the three endpoint inputs on the evaluation collection. For perplexity, normalization is $(\max p - p)/(\max p - \min p)$; for coherence and diversity it is $(v - \min v)/(\max v - \min v)$. Application to new texts requires specifying the normalization ranges or the collection used to estimate them. Its initialization is taken from the published Q*Text algorithm, rather than optimized on these ratings.

PCA. We use the released PCA scores from standardized coherence mean, coherence variance, diversity mean, and diversity variance. PCA was fitted without ratings on all 300 texts, including the human references; we evaluate its fixed components on the 250 machine texts. The first two components explain 58.23% and 30.53% of feature variance. Their signs are arbitrary, so a negative rating correlation does not by itself indicate a poor component. PCA maximizes feature variance, not quality prediction.

Fixed Q*Text. To include an untuned composite, we implement Q*Text's published initialization on the cached *endpoint* perplexity, coherence, and diversity values (Garces Arias et al., 2025b). After min–max normalization, with perplexity inverted, Equation 8 defines the score:

$$Q_0 = \frac{1}{3}\sum_{j=1}^{3} u_j \exp[-(u_j - 0.5)^2]. \qquad (8)$$

Thus all weights and penalty strengths equal one, all targets equal 0.5, and no parameter is optimized against ratings. This specification defines the baseline evaluated here. The rating-tuned correlation of 0.555 reported by Garces Arias et al. (2025b) uses a different evaluation protocol and is excluded from the common-sample comparison.

MAUVE. We retain the original notebook's corpus-level MAUVE results as descriptive context. Each comparison uses 50 human texts and 50 outputs from one strategy, GPT-2 features, and maximum text length 512. The notebook supplies the complete stored text field, including prompt and continuation labels. This evaluates distributions of those text fields, not a per-continuation quality score. These comparisons yield five corpus-level scores, one per decoding strategy.

Score	ρ	95% interval
Gaussian: mean	0.409	$[0.315, 0.493]$
Diagonal Gaussian: mean	0.441	$[0.346, 0.523]$
KDE: mean	0.436	$[0.332, 0.522]$
Gaussian: variance	0.598	$[0.505, 0.675]$
Diagonal Gaussian: var.	0.604	$[0.511, 0.679]$
KDE: variance	0.624	$[0.520, 0.707]$

Table 9: Association between reconstructed reference likelihood scores and human ratings on 250 machine texts. Values are Spearman correlations with 95% prompt-cluster bootstrap intervals. The archived extraction limitations are described in Appendix D.

C Reconstructed Reference Distributions

The fitted parameters below refer to 9,877 complete reference records after exact-overlap removal. Figure 4 shows how the reconstructed summaries vary across reference domains, Figure 5 compares their marginal distributions with normal quantiles, and Figure 6 shows their joint distributions. These views help interpret the model and domain sensitivity checks in Section 6.5. Per-record coherence is first converted from stored token log probabilities to cumulative averages. Per-text variances use divisor n, while the covariance of the resulting reference vectors uses divisor $N - 1$. Fitting each reference domain separately gives another view of reference sensitivity: mean-profile correlations range from 0.188 (WritingPrompts) to 0.569 (Wikinews), while variance-profile correlations range from 0.512 (ROCStories) to 0.595 (WritingPrompts). All fits use the same machine evaluation texts, so these ranges describe sensitivity to the reference domain within this sample.

$$\mu_{\text{mean}} = (-3.173622,\ 0.934842)^T,$$
$$\Sigma_{\text{mean}} = \begin{bmatrix} 0.4668553 & -0.0123085 \\ -0.0123085 & 0.0038644 \end{bmatrix}.$$

$$\mu_{\text{var}} = (1.147170,\ 0.002390)^T,$$
$$\Sigma_{\text{var}} = \begin{bmatrix} 3.2443632 & -0.0005244 \\ -0.0005244 & 0.0000207 \end{bmatrix}.$$

D Feature Preparation and Sensitivity

For each primary text, the stored coherence array has L values and the diversity array has $L - 1$. All per-text means and population variances used below are recomputed over the actual stored arrays; they match the released summary columns to numerical precision. The coherence length L is the denominator in the summary normalization. DTW uses each signal's own array length. We do not pad one array to match another.

The extended cache requires separate preparation. Its coherence arrays contain token log probabilities, and its saved summary values are rounded to two decimal places. We form cumulative averages of the stored log probabilities before calculating trajectory means and population variances, and recompute diversity summaries from the stored arrays. This avoids treating rounded zero variances as evidence of constant trajectories. The underlying arrays remain rounded, and the primary and reference extraction pipelines are not fully documented. We therefore interpret the reconstructed likelihood results as *exploratory* evidence about the association between reference typicality and ratings. The sensitivity comparisons below assess the consequences for the fitted scores.

As an exploratory alternative, we calculate the negative Euclidean distance from a text's two-dimensional summary to the reconstructed corpus mean. Unlike the prompt-relative composite, this alternative changes the reference, coordinate scaling, and aggregation together. Its performance cannot isolate the benefit of prompt-specific references.

The extended summary CSV contains two-decimal summaries. In particular, small nonzero diversity variances may be stored as zero. The supplied variance-reference parameters were fitted to 1,156 records with positive stored diversity variance. Such selection changes the modeled population: it is a distribution conditional on crossing a rounding threshold, rather than a reference for all human records. The reconstruction calculates variances from the available arrays and retains genuine zero-variance trajectories.

On the common 250-machine subset, applying the supplied summary-based Gaussian parameters gives correlations of 0.458 for means and 0.257 for variances. The reconstructed, overlap-excluded fits give 0.409 and 0.598. This comparison describes sensitivity to data preparation on the same evaluation sample. The summary-based and array-based configurations differ in multiple ways, so the comparison does not isolate a single cause.

All primary temporal coefficients are recomputed as Spearman correlations. On the same 250-machine subset, the negative Euclidean mean-

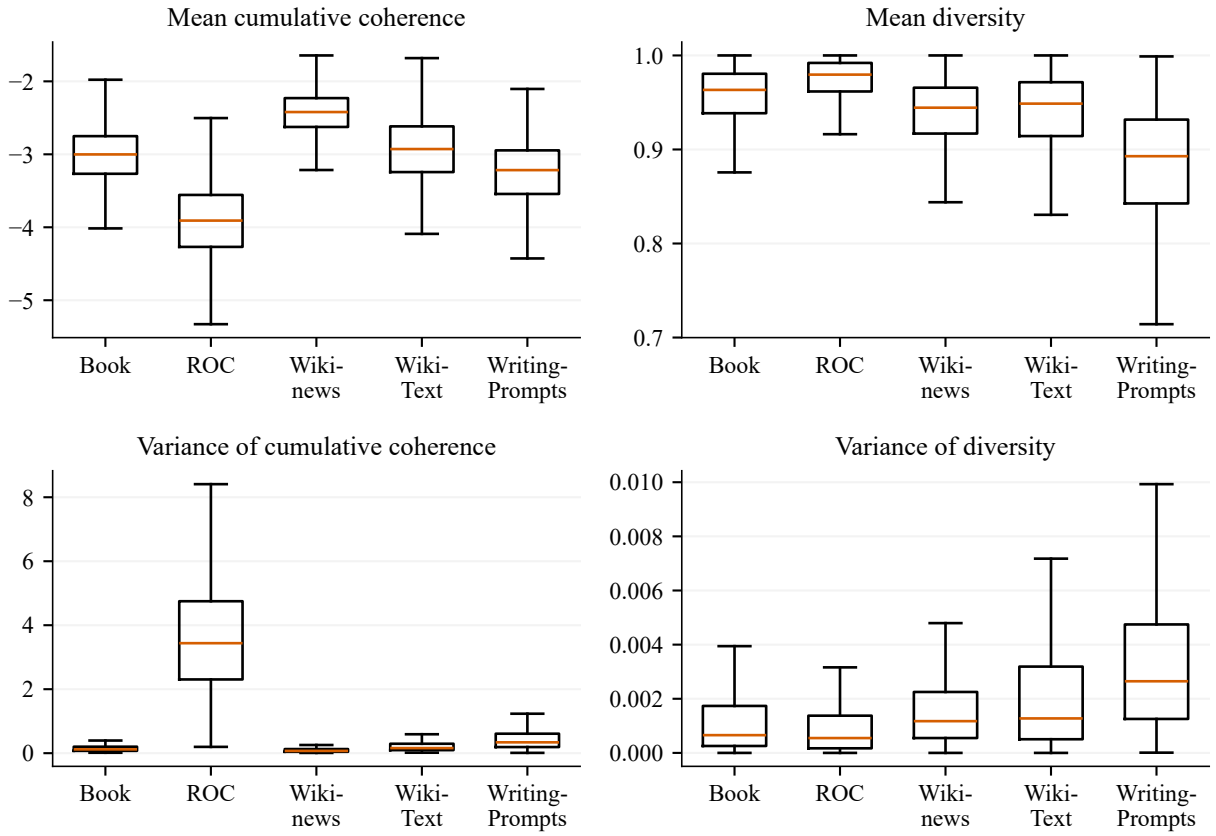


Figure 4: Coherence and diversity summaries across human reference domains. Boxes show quartiles and 1.5-IQR whiskers for the reconstructed, overlap-excluded records. Outliers are hidden for readability but retained in fitting. Each panel uses a scale suited to its measured property.

profile score has $\rho = 0.165$ and its variance counterpart has $\rho = 0.019$. These scores use the reconstructed corpus means, retain the raw coordinate scales, and change several design choices relative to the prompt-relative composite. Their weaker observed associations do not isolate reference selection as the cause. These cached continuation-level comparisons use the same 250 machines; the controlled prefix experiment uses the eligible subsets in Appendix E. Corpus-level MAUVE remains a separate descriptive analysis.

E Length Controls and Horizon Sensitivity

Shared extraction. We score all 300 primary and 10,000 reference texts using OPT-2.7B, model/tokenizer revision `905a4b6`, in float32 with TF32 disabled. Prompt and continuation whitespace is stripped separately; each is tokenized separately without special tokens, and exactly one BOS token is prepended. We score at most the first 128 continuation tokens, conditioned on the prompt. All prompts have at most 81 tokens, so these measurements fit within the 2,048-token context. Coherence is the cumulative mean of next-token log probabilities. Diversity is the product of distinct-to-total token-ID n-gram ratios for $n = 2, 3, 4$, with ratios set to one before enough tokens exist. Both signals include the first continuation token and use identical positions. Per-text variances use divisor T. Validation compared token IDs, lengths, and hashes with the source texts and recomputed summary features from the stored trajectories. Cross-entropy and block-partition checks verified the consistency of the GPU calculations. Full model and tokenizer identifiers, software versions, and validation results accompany the analysis.

Eligibility and references. Horizons 64 and 128 were selected from cached length coverage before the controlled measurements, without optimizing against ratings; this is not a preregistered design. Table 5 summarizes the resulting associations. Only complete six-text prompt groups with at least T tokens in every continuation are retained. This gives 40 groups (200 machines) at 64 and 23 groups (115 machines) at 128. No se-

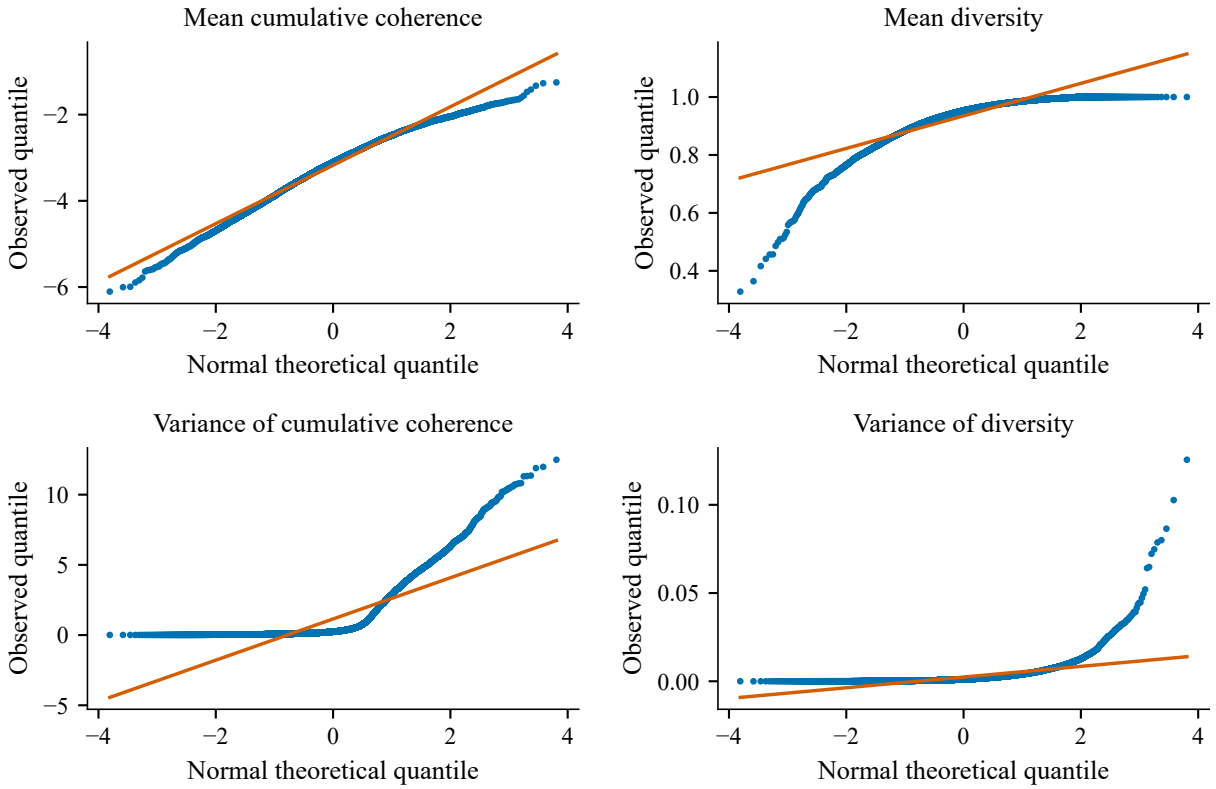


Figure 5: Comparison of the four human reference summaries with normal distributions. Departures from the straight reference lines indicate differences from marginal normality. All fitted records are shown; these marginal checks complement the joint distributions in Figure 6.

quence is padded. The DTW human bank contains the corresponding 40 or 23 humans; distances use symmetric2 absolute-cost alignment divided by T. Gaussian fits use 7,686 and 7,149 eligible human-corpus records after excluding 48 exact overlaps with primary human references. Both reference and evaluated summaries use the same horizon and extraction. The controlled extraction includes the 75 texts without archived feature arrays. Eligibility alters domain composition: no ROCStories records qualify at 128.

Inference and paired horizon comparison. Intervals use 3,000 prompt-cluster bootstrap replicates (seed 20260907), retaining all five machine methods per sampled prompt and conditioning on fixed reference banks. Summary standardization is refitted in each replicate. For rank-residual sensitivity, both score and rating ranks are residualized against an intercept and cubic polynomials of centered empirical ranks of original machine continuation length, matched-human continuation length, and prompt length. Ranks and regressions are refitted per replicate. These are descriptive residual associations, not ordinary Spearman coefficients or causal effects. The adjusted diversity-DTW and Gaussian variance-profile coefficients are 0.336 and 0.286 at 64 tokens, and 0.439 and 0.493 at 128.

To isolate horizon changes from changes in cohort and reference membership, we additionally score 64-token prefixes on the 128-eligible set: the same 115 machines, 23 primary humans, and 7,149 corpus references. Gaussian parameters are refitted at each horizon on those same records. The paired 128-minus-64 changes are 0.121 ($[-0.036, 0.290]$) for diversity DTW, 0.066 ($[-0.008, 0.137]$) for Gaussian means, and 0.139 ($[0.019, 0.261]$) for Gaussian variances. These intervals are not multiplicity-adjusted. The paired differences between diversity DTW and endpoint diversity are -0.058 ($[-0.201, 0.086]$) at 64 and -0.107 ($[-0.270, 0.062]$) at 128; neither establishes superiority or equivalence.

Decoding-strategy sensitivity. Table 10 reports an exploratory adjustment applied to all six scores in Table 5. We rank scores and ratings over each eligible cohort using average ranks for ties, regress each rank vector on five decoding-method indi-

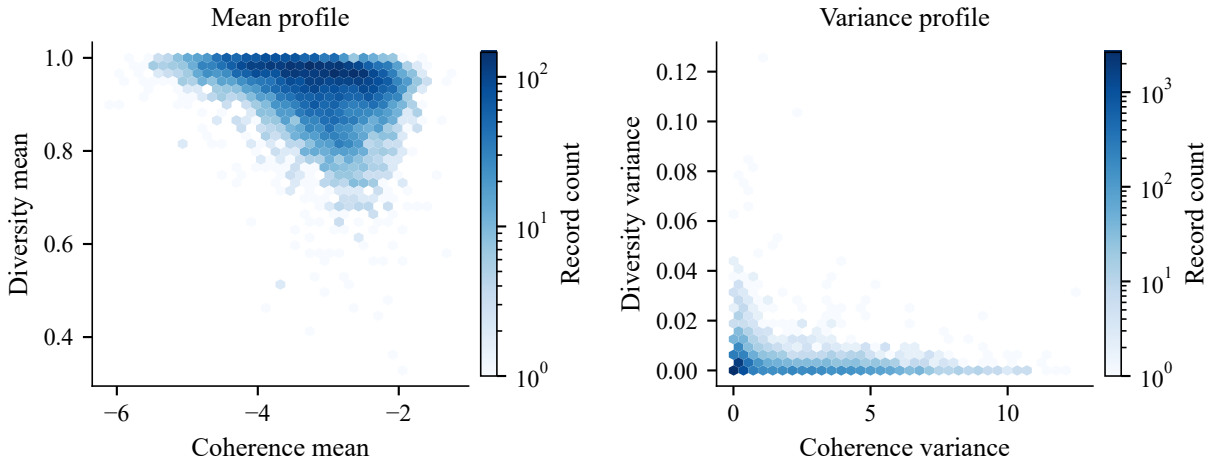


Figure 6: How coherence and diversity summaries vary together in the reconstructed human reference corpus: means (left) and variances (right). Darker hexagons contain more records, on the logarithmic count scales shown. The panels visualize the observations used to fit the reference models.

Score	64 tokens	128 tokens
DTW coherence	-0.145 [-0.316, 0.043]	0.031 [-0.179, 0.244]
DTW diversity	0.275 [0.164, 0.385]	0.362 [0.147, 0.529]
Gaussian mean	-0.024 [-0.198, 0.148]	0.180 [0.037, 0.310]
Gaussian variance	0.298 [0.137, 0.435]	0.531 [0.337, 0.675]
Endpoint diversity	0.342 [0.207, 0.468]	0.501 [0.350, 0.623]
Summary mean	0.317 [0.122, 0.476]	0.370 [0.158, 0.549]

Table 10: Associations with human ratings after adjusting score and rating ranks for decoding strategy. Values are correlations of the residual ranks with 95% prompt-cluster bootstrap intervals. Texts and scores match Table 5; references remain fixed during resampling.

cators, and correlate the residuals. This removes method-specific mean ranks; it is neither an average of within-method Spearman correlations nor a causal adjustment. Each of 3,000 prompt-cluster bootstrap replicates (seed 20260908) refits summary standardization, ranks, and regressions, with reference banks fixed. Intervals are pointwise and not multiplicity-adjusted. This analysis controls decoding strategy separately from the length adjustment above, rather than estimating a joint adjustment. Diversity DTW, variance-profile likelihood, endpoint diversity, and the mean summary retain positive residual associations at both horizons. Coherence DTW has intervals spanning zero at both horizons, and mean-profile likelihood at 64 tokens also has an interval spanning zero. The adjustment thus leaves different patterns of association across the measured signals and facets.

Interpretation. Equal token horizons hold the number of scored continuation tokens constant within each analysis. Content differences and other length-related influences may remain, and the results can vary with the horizon. The ratings describe entire continuations, so prefix scores are not validated against prefix-specific judgments. The prefix-only run does not provide full-length scores beyond 128 tokens. Eligibility restricts the texts represented in these results. The Gaussian scores describe typicality under the fitted model, with their relationship to quality evaluated through correlation.

F Controlled Human–Machine Density Comparisons

Figures 7 and 8 compare human and machine profiles using the controlled measurements described in Appendix E. Each point is a continuation's mean cumulative coherence and mean token-level diversity at the indicated horizon. The five machine panels and matched-human panel use the same complete prompt groups within each horizon. Both figures have identical axes, and every eligible primary observation is included without jitter or selection by rating. Overlapping points may obscure individual observations.

Blue contours use exactly the corpus-reference

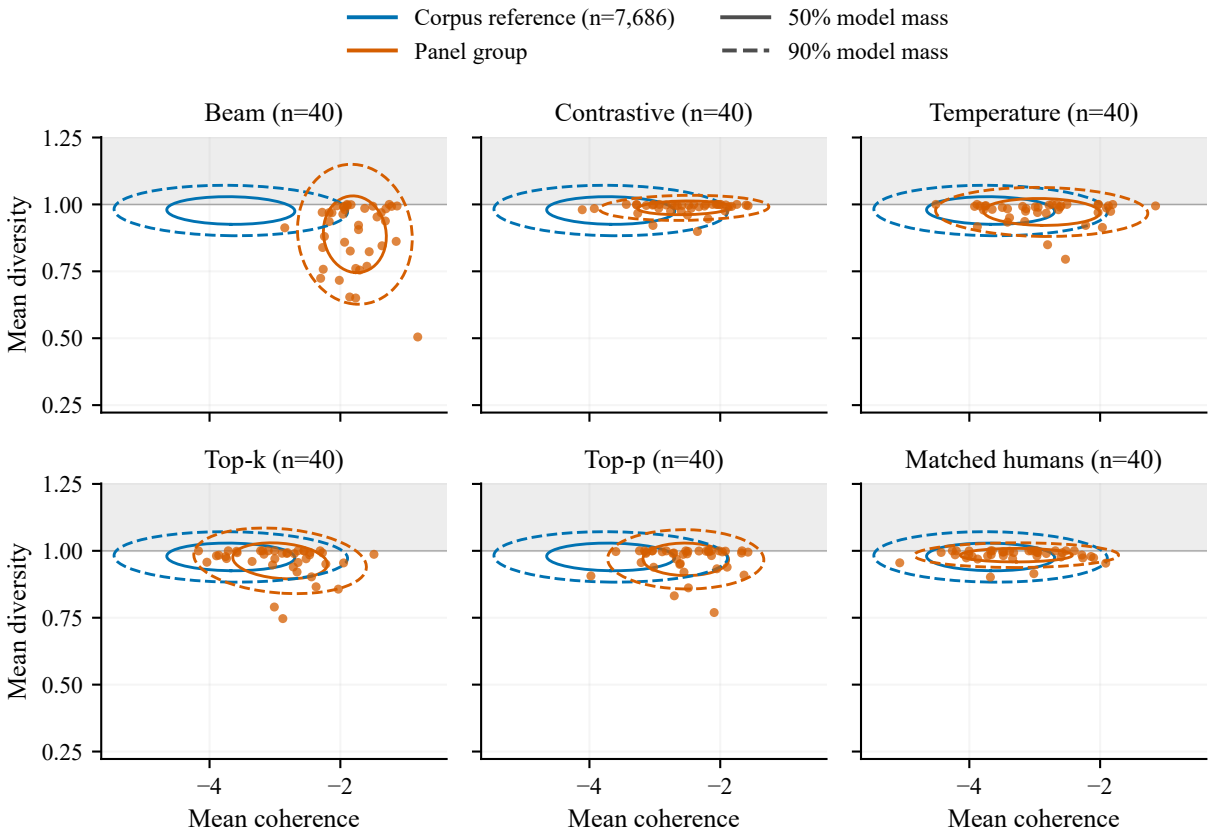


Figure 7: Human and machine coherence–diversity profiles at 64 continuation tokens. Each panel shows 40 continuations from one group. Blue contours describe the human corpus reference (7,686 records); orange contours describe the panel group. Solid and dashed curves enclose 50% and 90% of fitted Gaussian mass, rather than confidence intervals. Gray shading marks values beyond the diversity range. Axes match Figure 8.

mean and covariance used by the controlled mean-profile likelihood score. Orange contours use a separate Gaussian fitted to the observations in that panel, with sample covariance. Panel-specific fits summarize each group's observations; likelihood scores use the human corpus reference. Solid and dashed curves enclose 50% and 90% of their respective fitted Gaussian distributions. They are neither confidence intervals nor estimates of empirical coverage. The curves satisfy $(u-\mu)^T\Sigma^{-1}(u-\mu) = \chi^2_2(q)$ for mass q, and were checked numerically against this identity.

The beam panel illustrates that group differences can involve both location and spread. At 128 tokens, its observed mean-diversity values span a visibly broader range than those of the matched humans. This is descriptive evidence about these measured profiles, not a test that one strategy produces better text. The gray region above diversity one makes a limitation of the Gaussian approximation explicit: its contours can extend beyond the signal's possible range. The mean-profile views do not establish the fit or validity of the separate variance-profile likelihood model.

The 64-token figure uses 40 texts per panel and 7,686 corpus-reference records. The 128-token figure uses 23 texts per panel and 7,149 references. These are the cohorts in Appendix E; differences between the figures combine horizon and eligibility changes. The fixed-membership comparison in that appendix examines horizon differences while holding the evaluated texts and references constant.

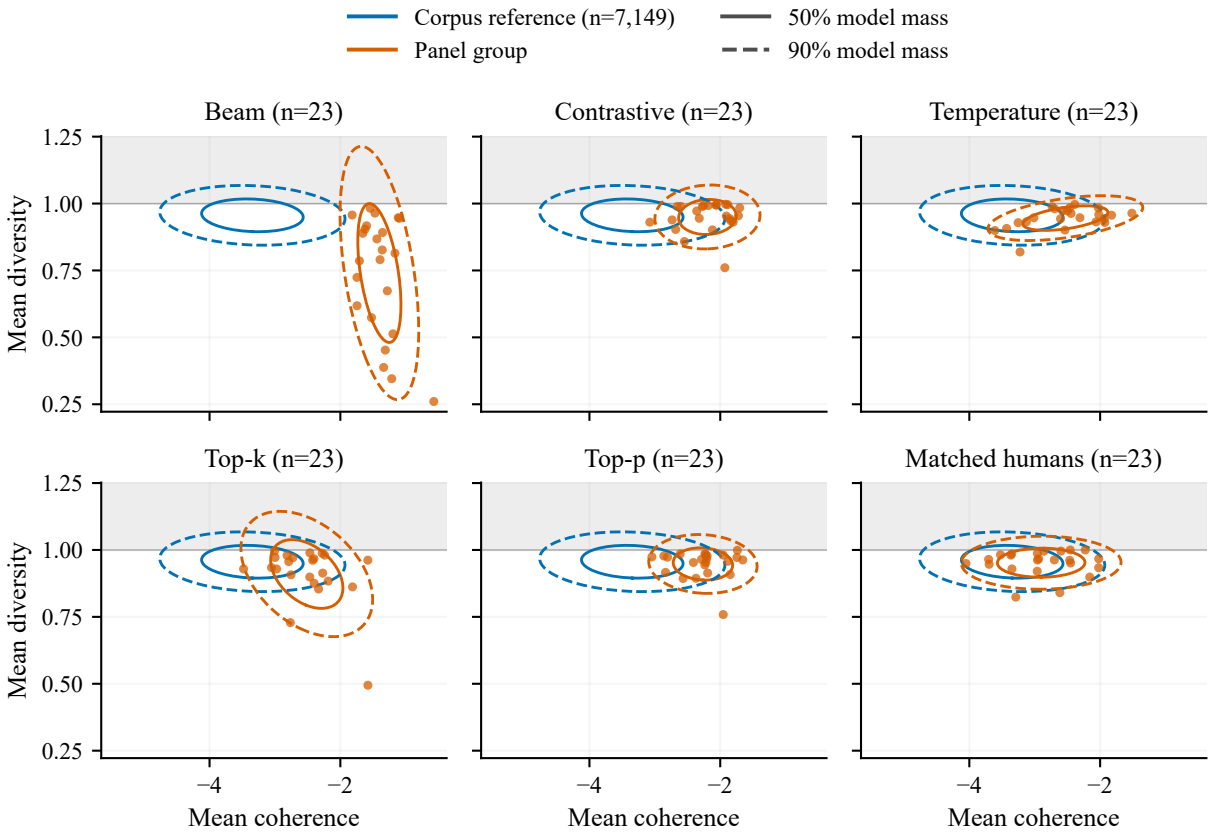


Figure 8: Human and machine coherence–diversity profiles at 128 continuation tokens. Each panel shows 23 continuations. Blue contours describe the human corpus reference (7,149 records); orange contours describe the panel group. Contour masses, shading, and axes match Figure 7. Only the corpus reference supplies the likelihood scores in Table 5.