Rate-Resolved Text Evaluation with Masked Diffusion Language Models

Anonymous ACL submission

Abstract

Text evaluation is increasingly driven by scalar judgments from autoregressive likelihoods, prompted judges, and learned evaluators. While effective for ranking systems, scalar scores collapse diverse evaluation behavior into a single dimension. We propose rate-resolved evaluation, a reconstruction-based evaluation paradigm using masked diffusion language models. Masked diffusion models reconstruct text under varying masking rates, naturally exposing how quality signals behave as contextual evidence decreases. We preserve this reconstruction structure as a reconstruction profile: a vector of band-conditioned losses that captures how different quality signals persist or degrade across context regimes. On SummEval, preserving masking-rate structure provides complementary information beyond the scalar baselines tested here, improving fusion with autoregressive likelihood and prompted judges. On WMT MQM zh-en, the same reconstruction process produces different evaluation behaviors at different masking rates, yielding a tunable tradeoff between ranking translation quality and localizing error spans. Our results suggest that, in the tested settings, diffusion reconstruction can expose evaluative structure that is not captured by the scalar baselines we evaluate, through structured recoverability under controlled corruption.

1 Introduction

A single evaluation score is often enough to rank generation systems, but it compresses diverse evaluation behavior into one dimension. Two summaries may receive similar scores while differing substantially in coherence, contextual dependence, or sensitivity to localized errors. Existing evaluators, including autoregressive likelihoods, learned metrics, and prompted judges, primarily produce scalar judgments that summarize model preference or confidence (Yuan et al., 2021; Qin et al., 2023; Liu et al., 2023; Fu et al., 2024; Zhong et al., 2022). While effective for ranking systems, scalar evaluation can obscure complementary structure in how text quality behaves under different levels of contextual evidence.

In this paper, we investigate masked diffusion language models as a reconstruction-based perspective on text evaluation. These models reconstruct text under varying masking rates, requiring reconstruction from progressively reduced context. This reconstruction process naturally exposes how quality signals behave as contextual evidence decreases. We call this view *rate-resolved evaluation*: instead of reducing evaluation to a single scalar judgment, we preserve masking rate as a model-native evaluation axis and represent each text as a reconstruction profile across masking regimes.

The resulting reconstruction profile is indexed by masking rate, a variable already present in the diffusion objective. Different masking regimes probe text quality under different levels of contextual evidence, revealing evaluation behaviors that scalar metrics can obscure. For example, some masking regimes better capture summary-level quality differences, while others are more sensitive to localized errors. This creates a structured evaluation space that can expose complementary signals for ranking and error localization. Unlike prompt-based decompositions or manually engineered evaluation features, this structure arises directly from the evaluator's reconstruction process itself.

Figure 1 illustrates this paradigm. Rather than assigning a single evaluation score to a text, masked diffusion produces a vector of band-conditioned reconstruction losses across masking regimes. We investigate whether this masking-rate structure provides complementary evaluative information beyond scalar metrics and judges, and whether different masking regimes expose different evaluation behaviors. In this paper, we study these questions through two representative settings: summary-level

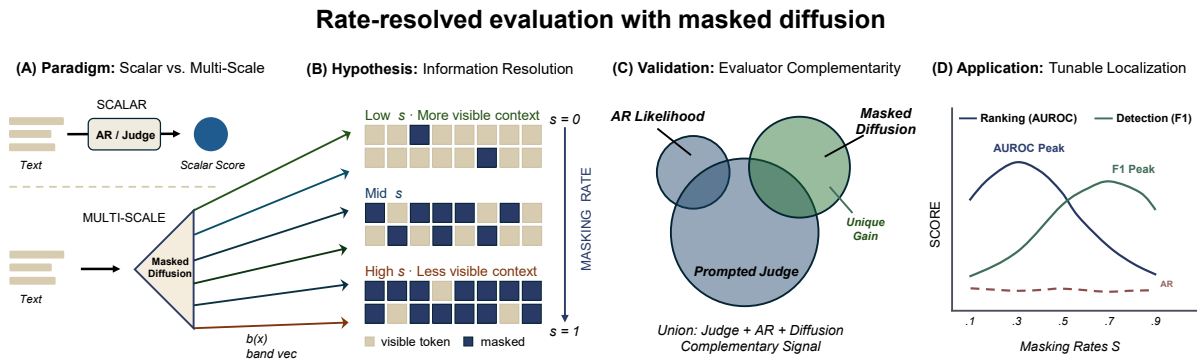

Figure 1: Instead of reducing text evaluation to a single scalar score, masked diffusion reconstructs text under different masking rates and yields a band-conditioned score vector $\mathbf{b}(x)$. Lower masking rates retain more visible context, while higher masking rates require reconstruction from sparser evidence. The resulting masking-rate structure exposes complementary evaluation behavior beyond scalar evaluation.

evaluation, which measures overall text quality for system ranking, and token-level localization, which identifies where errors occur within a text.

The paper makes three **main contributions**. First, we show how masked diffusion language models naturally support rate-resolved evaluation through reconstruction under varying masking rates, turning masking rate into a structured evaluation axis for text quality assessment. Second, we show that different masking regimes expose complementary evaluative information beyond scalar metrics and prompted judges on summary evaluation benchmarks. Third, we show that different masking regimes support different evaluation objectives, with some better suited for ranking text quality and others more effective for localizing errors.

2 Related Work

We organize prior work by how evaluators represent text quality: scalar evaluators, externally designed evaluation axes, and model-native evaluation structure.

Scalar evaluators. Existing text-generation metrics range from lexical overlap to embedding-based, likelihood-based, trained, and prompted evaluators (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020; Zhao et al., 2019; Sellam et al., 2020; Yuan et al., 2021; Qin et al., 2023; Fu et al., 2024; Zhong et al., 2022). Prompted judge models have become especially strong: G-Eval and related systems can correlate closely with human judgments in many settings (Liu et al., 2023; Wang et al., 2023; Chiang and Lee, 2023). Despite their methodological differences, most existing evaluators ultimately map each text to a single scalar judgment. Our work instead studies whether evaluation can be represented as structured behavior across masking regimes, exposing complementary signals that scalar evaluation does not capture.

Masked-LM and cloze-style evaluation. Cloze-style and masked-LM evaluators also use reconstruction as an evaluation signal. BLANC estimates summary quality by measuring how much access to a summary helps a pretrained language model solve cloze-style prediction tasks on the source document (Vasilyev et al., 2020). Masked-LM scoring uses pseudo-log-likelihoods computed by masking tokens one at a time (Salazar et al., 2020). Our work differs by preserving the masking-rate axis of a masked diffusion evaluator: rather than reducing reconstruction to a single cloze or pseudo-likelihood score, we study how evaluation behavior changes across masking regimes.

Externally designed evaluation structure. Behavioral test suites, contrast sets, and dynamic challenge benchmarks expose evaluation behaviors that aggregate scores can hide (Ribeiro et al., 2020; Gardner et al., 2020; Kiela et al., 2021). In these approaches, the analyst or data-collection process specifies controlled evaluation conditions through templates, perturbations, contrast examples, or adversarially collected cases. Such methods provide structured views of model behavior, but the structure is imposed by the evaluation design rather than arising from the evaluator's own modeling objective.

Model-native evaluation structure in diffusion models. Text diffusion models have primarily

been developed for generation. Important examples include Diffusion-LM (Li et al., 2022), DiffuSeq (Gong et al., 2023), self-conditioned embedding diffusion (Strudel et al., 2023), AR-Diffusion (Wu et al., 2023), and later work on bridging the training and inference gap (Tang et al., 2023). A common feature of these models is iterative reconstruction under a noise or masking schedule, where generation behavior changes as contextual evidence is progressively removed or restored. Prior work has largely used this structure for controllable or higher-quality generation. We instead investigate masking rate as a model-native evaluation axis, asking whether reconstruction behavior across masking regimes exposes complementary evaluative information beyond scalar metrics and judges.

3 Rate-Resolved Evaluation

Before introducing the formalism, we describe the core idea intuitively. Given a text, we choose a masking rate s, hide that fraction of tokens, ask the diffusion model to reconstruct the masked content, and measure the resulting reconstruction loss. Varying s changes how much contextual evidence remains available to the model. Rather than aggregating reconstruction behavior across all masking rates into a single score, we preserve reconstruction behavior across masking regimes as a structured evaluation signal.

The masking-rate axis can support different forms of evaluation behavior depending on how reconstruction signals are summarized or probed. In this paper, we study two complementary settings. For summary-level evaluation, we represent reconstruction behavior through masking-rate bands. For token-level localization, we measure how often tokens change under repeated mask-and-reconstruct revisions. Together, these settings illustrate how the same reconstruction process can support different evaluation objectives, including ranking text quality and localizing errors.

3.1 Masking-Rate Reconstruction Profiles

We now formalize masking-rate reconstruction profiles. Given a text, we sample a masking rate s, mask that fraction of tokens, ask the diffusion model to reconstruct the masked content, and measure the resulting reconstruction loss.

Let $x = (x_1, \ldots, x_N)$ be a token sequence and let $m \sim \mathrm{Bern}(s)^N$ denote an i.i.d. masking pattern at masking rate $s \in [0, 1)$. We write $S = \{i : m_i = 1\}$ for the masked positions and $\bar{S}$ for the visible positions. A masked diffusion language model p_θ predicts each x_i for $i \in S$ from the partially observed input $x_{\bar{S}}$.

For a fixed masking rate s, we define the normalized masking-rate loss density

$$\ell(s, x) \triangleq \mathbb{E}_m \left[\frac{1}{|S|} \sum_{i \in S} r_i(x, m) \;\middle|\; |S| > 0 \right], \quad (1)$$

where $r_i(x, m) = -\log p_\theta(x_i \mid x_{\bar{S}})$. This quantity measures the average masked-token reconstruction loss under a random mask pattern at rate s. Higher loss means that the observed text is harder for the model to reconstruct under that masking regime. Because masking patterns are sampled randomly, the degenerate event $|S| = 0$ can occasionally occur at low masking rates; in Monte Carlo estimation, we simply discard these samples. When the evidence lower bound (ELBO) is written in the equivalent $m_i r_i(x, m)$ form, r_i is only needed on masked positions and the product is zero for $i \notin S$.

This masking-rate-dependent reconstruction loss also connects naturally to the diffusion training objective. For the absorbing-mask family used by masked diffusion models, the training objective can be written as a weighted integral over masking rates.

$$\mathcal{L}_{\text{diff}}(x) \approx \int_0^{1^-} \frac{s}{1-s} N \, \ell(s, x) \, ds, \quad (2)$$

We never evaluate $s = 1$ directly and sample only rates with $s < 1$ in practice. Appendix A.1 gives the derivation and details the approximation used to replace the random number of masked positions by its expectation.[1]

We then define the loss of a masking-rate band $[a, b) \subseteq [0, 1)$ as

$$\mathcal{L}_{[a,b]}(x) \triangleq \int_a^b \frac{s}{1-s} N \, \ell(s, x) \, ds. \quad (3)$$

In practice, we estimate Eq. 3 with Monte Carlo samples over both masking rates and mask patterns. Each band summarizes reconstruction behavior within a particular masking regime.

We therefore represent each text with the masking-rate reconstruction profile

$$\mathbf{b}(x) = \left(\mathcal{L}_{[a_1,b_1]}(x), \ldots, \mathcal{L}_{[a_K,b_K]}(x) \right), \quad (4)$$

[1] We verified on a SummEval subset that this approximation has negligible effect on document-level rankings (Spearman(exact, approximate) ≥ 0.9969 across all tested rates).

where the intervals form a partition of $[0, 1)$. Each coordinate captures reconstruction loss within a different range of masking rates, yielding a structured evaluation profile rather than a single scalar score.[2]

3.2 Interpreting the Masking-Rate Axis

The masking-rate axis is operational rather than manually designed: each coordinate in Eq. 4 corresponds to a reconstruction regime already present in the diffusion objective. Different masking rates expose different reconstruction behaviors because the amount of visible context changes with s. For example, reconstruction under lower masking rates may better reflect fine-grained consistency with nearby context, while higher masking rates can become more sensitive to broader contextual coherence when local evidence is sparse. In our experiments, different masking regimes favor different evaluation behaviors, including summary-level ranking and localized error detection.

We therefore treat masking rate primarily as a structured evaluation axis rather than assigning fixed semantic meanings to individual bands. While the precise behavior of individual masking regimes may depend on the diffusion evaluator and evaluation setting, the underlying masking-rate axis remains operationally consistent: all generated texts are evaluated through the same family of reconstruction conditions, yielding comparable reconstruction profiles across systems and outputs. Unlike prompt-based decompositions or manually specified evaluation dimensions, the masking-rate structure arises directly from the evaluator's reconstruction process itself.

3.3 Autoregressive Likelihood as a Scalar Baseline

For AR baselines, we use the standard full-sequence negative log-likelihood

$$\mathcal{L}_{\mathrm{AR}}(x) = -\sum_{i=1}^{N} \log p_\phi(x_i \mid x_{<i}), \qquad (5)$$

which provides a single likelihood-based evaluation score for the entire text. We compare this scalar baseline with the masking-rate reconstruction profile in Eq. 4 to test whether different masking regimes provide complementary evaluative information beyond standard autoregressive likelihood.

3.4 Roundtrip Reconstruction for Token-Level Localization

The masking-rate axis can also be used for token-level evaluation. Instead of measuring reconstruction difficulty for an entire text, we ask which tokens are repeatedly revised under reconstruction. For a fixed masking rate s, we repeatedly mask and reconstruct the same text under independently sampled mask patterns. Let $\hat{x}^{(k)}$ be the k-th reconstruction under mask pattern $m^{(k)}$. We define the roundtrip conditional revision frequency for token i as

$$\mathrm{RTP}_s(i, x) \triangleq \frac{\sum_{k=1}^{K} \mathbb{1}[m_i^{(k)} = 1]\, \mathbb{1}[\hat{x}_i^{(k)} \neq x_i]}{\max\left\{1, \sum_{k=1}^{K} \mathbb{1}[m_i^{(k)} = 1]\right\}}. \qquad (6)$$

High values indicate tokens that are likely to change when exposed to reconstruction. We normalize by the number of times a token is actually masked, so the score estimates conditional revision frequency rather than marginal exposure. Masking rate still controls the reconstruction context available to the model. We evaluate these token scores with AUROC and thresholded F1 against human span annotations. Varying s therefore produces a localization operating curve across masking regimes, allowing different masking rates to favor different localization behaviors. Unlike the band-conditioned losses, which are derived from the diffusion objective, this score is a reconstruction-based heuristic designed for token-level localization.

3.5 Evaluation Hypotheses

The masking-rate framework suggests three empirical questions. (i) If reconstruction profiles capture evaluative structure beyond scalar metrics, then preserving masking-rate information should provide complementary evaluation behavior beyond standard scalar evaluators. (ii) If masking rate changes what reconstruction evidence is available to the evaluator, then different masking regimes may support different forms of token-level error localization. Lower masking rates may expose a broader set of local inconsistencies, while higher masking rates may produce more selective error signals under sparse context. (iii) if masking rate changes

[2] In experiments, we use equal-width masking-rate bands, with the final band covering rates close to $s = 1$ while excluding the degenerate endpoint itself. The derivation is specific to absorbing-mask diffusion models; extending the same decomposition to other non-autoregressive generators would require different assumptions.

Role	Model
Main diffusion	LLaDA-8B-Instruct
Replication diffusion	Dream-7B-Instruct
AR / judge	Qwen2.5-7B-Instruct
Static cloze control	XLM-R-XXL

Table 1: Model roles in the main experiments. Exact checkpoints and inference details are in Appendix A.3.

the reconstruction evidence available to the evaluator, then the masking regimes that are most useful should depend on the evaluation objective. We study this through both summary-level evaluation and token-level localization.

4 Experimental Setup

Benchmark roles. The experiments assign each benchmark a specific role. **SummEval** (Fabbri et al., 2021) tests whether masking-rate reconstruction profiles provide complementary evaluative information beyond AR likelihood and prompted judges for multidimensional summary quality. **WMT MQM** zh-en, following the machine translation evaluation setting of Kocmi and Federmann (2023), tests token-level localization behavior across masking regimes.

Datasets. For WMT MQM, we use the released newstest2021 Chinese-to-English MQM TSV from the WMT21 metrics shared task and public MQM annotation release (Freitag et al., 2021; Google Research, 2024). We retain the 8,431 segment-system hypotheses whose system id begins with `hyp.`, covering 13 MT systems after excluding non-hypothesis entries. Marked MQM spans are stripped from the target text and projected to character-level gold masks. We evaluate all-error, major-error, and minor-only severity slices.

Models. The main diffusion models are LLaDA-8B-Instruct (Nie et al., 2025) and Dream-7B-Instruct (Ye et al., 2025). Qwen2.5-7B-Instruct (Yang et al., 2025) serves as both the AR likelihood baseline and prompted judge.

Localization baseline. We use XLM-R-XXL (Conneau et al., 2020) as a non-diffusion masked-language-model baseline. Each token is independently masked and scored by reconstruction cross-entropy under full left and right context. This tests whether a strong static cloze-style evaluator can match the localization behavior of roundtrip diffusion reconstruction.

Evaluation protocols. For summary-level evaluation, we report band-conditioned correlation analyses. For localization, we report area under the receiver operating characteristic curve (AUROC) and token-level F1. Fusion experiments use grouped cross-validation by source document, with diffusion-band selection and linear fusion weights fit only on training folds.

Implementation details. Unless otherwise stated, band-loss experiments use 10 equal-width masking-rate bands. Roundtrip localization experiments use $s \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with 100 reconstructions per rate. Token-level F1 uses a fixed prediction threshold of 0.2 on the conditional revision frequency in Eq. 6, chosen before test-set analysis. Appendix A.3 gives inference settings, runtime notes, fusion selection, and prompted-judge parsing details; Appendix A.4 reports Monte Carlo budget, band-boundary, and threshold sensitivity controls.

5 Results

5.1 Preserving Masking-Rate Structure Adds Evaluative Signal

We first test the central evaluation hypothesis of the paper: whether preserving masking-rate structure provides evaluative information beyond standard scalar evaluators. SummEval provides a controlled setting for this comparison because it contains multidimensional human quality judgments together with strong AR and prompted-judge baselines. We evaluate this hypothesis in three steps: first by testing complementarity with scalar evaluators, then through controls that isolate masking-rate structure from generic diffusion features, and finally through replication across diffusion evaluators.

Complementarity with scalar evaluators. The strongest and most consistent complementarity effect appears on the relevance dimension. Figure 2 summarizes these relevance results. For both LLaDA and Dream, combining AR likelihood with the best masking-rate band substantially improves over either signal alone. Smaller gains also appear for consistency and fluency, although the effect is weaker and less consistent across models. Coherence is the main exception: AR likelihood remains dominant there, and diffusion bands provide little additional benefit. Table 6 in Appendix gives the full per-dimension breakdown.

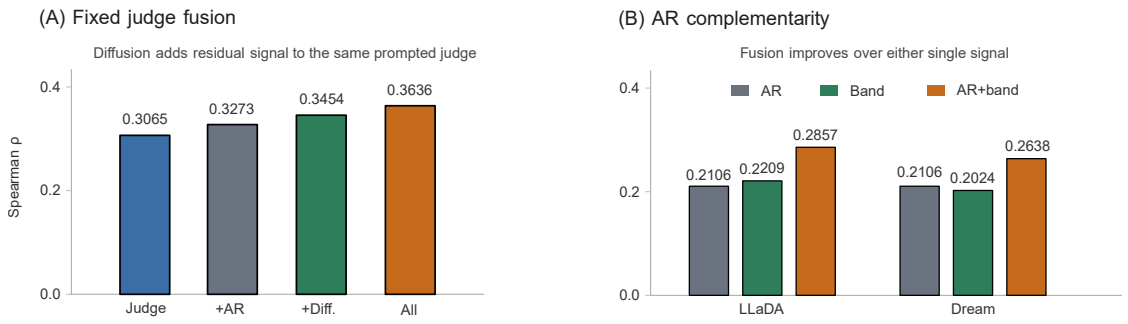

Figure 2: SummEval relevance complementarity. Left: LLaDA masking-rate reconstruction bands remain complementary even to a strong prompted Qwen judge, with judge+AR+diffusion achieving the strongest relevance correlation. Right: for both LLaDA and Dream, combining AR likelihood with the best masking-rate band consistently outperforms either signal alone, indicating that evaluative signal is carried by reconstruction behavior across masking regimes rather than by a single scalar diffusion score.

Evaluator	SummEval relevance ρ
Qwen judge	0.3065
Judge + AR	0.3273
Judge + diffusion	0.3454
Judge + AR + diffusion	0.3636

Table 2: Fixed judge fusion on SummEval relevance. Diffusion bands add residual signal to the same prompted judge, and the strongest grouped-CV result uses judge, AR, and diffusion together.

These gains persist even when the scalar evaluator itself is already strong. Adding LLaDA diffusion reconstruction signals to a fixed prompted Qwen judge improves relevance correlation from 0.3065 to 0.3454, while combining judge, AR, and diffusion signals reaches the strongest overall performance at **0.3636** (Table 2; Figure 2, left). The judge+diffusion improvement over the judge alone remains statistically stable under grouped bootstrap ($p = 0.0080$), and the three-way judge+AR+diffusion combination further improves over the judge baseline ($p < 0.001$).

These patterns suggest that masking-rate reconstruction may be especially sensitive to evaluation dimensions that depend on broader semantic support. High-mask reconstruction requires the evaluator to recover text under sparse contextual evidence, which may make reconstruction difficulty more dependent on whether summary content remains semantically recoverable from surrounding context. In contrast, coherence and fluency are more closely tied to local sequential compatibility, where autoregressive likelihood already provides a strong signal. Consistency lies between these extremes, producing smaller but still positive gains.

The gain depends on preserving masking-rate structure. We next test whether the SummEval gains come specifically from preserving masking-rate structure rather than simply adding another diffusion-derived score. If the improvement came only from introducing an additional model signal, collapsing diffusion reconstruction into a single scalar should perform as well as the best masking-rate band. Appendix Table 7 gives the full scalarization controls for LLaDA and Dream. On LLaDA relevance, the scalar control falls short: when fused with AR likelihood, the best masking-rate band reaches 0.2795, compared with 0.2610 for the corresponding full diffusion scalar ($\Delta = +0.0185$, $p = 0.0015$). The same LLaDA comparison is positive on consistency (0.1705 vs. 0.1428, $\Delta = +0.0276$, $p = 0.0085$). For Dream, the scalarization control is weaker and dimension-dependent, with a positive effect on consistency and little gain on relevance. These controls place the clearest evidence for masking-rate structure in LLaDA, with Dream providing a weaker replication.

A complementary sample-shuffle control preserves the same 200 reconstruction samples while destroying masking-rate order. Over 100 shuffles, AR+pseudo-band fusion averages $\rho = 0.2098$ on relevance and never matches the true AR+rate-band result of 0.2795. Together, these controls suggest that evaluative signal is carried specifically by how reconstruction behavior changes across masking regimes, rather than by arbitrary multidimensional diffusion features.

Replication across diffusion evaluators. To test whether this complementarity is specific to LLaDA, we repeat the analysis with Dream-7B-Instruct. Dream reproduces the same qualitative pattern: the

best masking-rate band remains complementary to AR likelihood on relevance, and AR+band fusion again outperforms either signal alone. The effect is weaker for fluency and coherence, where single-band diffusion signals remain limited. Table 6 in Appendix gives the full Dream breakdown.

Masking-rate profiles. Figure 3 shows the corresponding masking-rate profiles. Both diffusion evaluators concentrate their strongest relevance and consistency signals in high masking-rate bands, although the peak location and magnitude differ across models. Fusion gains are largest in the same regions where diffusion bands diverge most strongly from AR rankings, suggesting that high-mask reconstruction exposes evaluative behavior that scalar AR likelihood does not fully capture. Appendix A.5 gives qualitative examples illustrating where judge, AR, and diffusion evaluators fail differently.

5.2 Masking Rate Provides a Localization Operating Curve

We next ask whether masking rate exposes different localization behaviors. In the SummEval experiments, masking-rate reconstruction profiles revealed different summary-level evaluation signals. Here, we test whether the same control axis also changes how the evaluator identifies localized translation errors.

WMT MQM provides a useful setting for this analysis because it contains dense human error annotations at token-span level. Rather than assigning a single quality score to an entire output, the evaluator must determine which regions of the translation are likely to be problematic. We therefore use the conditional roundtrip revision frequency from Eq. 6 as a token-level localization signal and examine how localization behavior changes across masking regimes.

From an evaluation perspective, the masking-rate axis defines a localization operating curve rather than a single detection setting. Different masking regimes may favor different evaluation objectives: some may better support ranking suspicious regions, while others may favor broader thresholded error coverage. The WMT experiments test whether these tradeoffs emerge systematically across masking rates.

Figure 4 shows that masking rate defines a clear localization operating curve. Different masking regimes favor different evaluation objectives. Ranking-oriented AUROC peaks near $s = 0.5$, while fixed-threshold F1 peaks earlier near $s = 0.3$. The same qualitative pattern appears across major and minor-only error slices, corresponding to severe versus lower-severity MQM annotations.

This separation is consistent with the interpretation that masking rate changes the type of reconstruction evidence available to the evaluator. Intermediate masking rates appear to produce more selective revision behavior, improving the ability to rank suspicious regions reliably and yielding stronger AUROC. Lower masking rates retain denser local context, so when a token is selected for reconstruction, the model can make more locally grounded revisions. This produces broader thresholded error coverage and leads to higher F1. In this view, masking rate acts as an operational tradeoff between selective ranking behavior and broader error detection.

The observed shifts are stable under bootstrap resampling. Overall AUROC peaks at 0.6705 near $s = 0.5$, while overall F1 peaks at 0.2311 near $s = 0.3$. The AUROC shift is small but consistent under 2,000 paired bootstrap resamples ($\Delta = 0.0029$, 95% CI $[0.0015, 0.0044]$, $p = 0.0005$), while the F1 shift is larger ($\Delta = 0.0127$, 95% CI $[0.0108, 0.0147]$, $p = 0.0005$).

Against AR token surprisal, roundtrip diffusion reconstruction is strongest for ranking behavior. Overall AUROC reaches 0.6705 at $s = 0.5$, compared with 0.6330 for AR surprisal. The fixed-threshold F1 comparison is more mixed: AR performs better on pooled overall F1, while diffusion performs better on major and minor-only error slices. Under held-out-system threshold selection, diffusion at $s = 0.3$ also remains ahead of AR on macro-F1 for error-containing, major-error, and minor-only slices (Appendix A.4).

A bidirectional cloze baseline does not close the gap. XLM-R-XXL reaches 0.6236 overall AUROC, below both LLaDA roundtrip probing at $s = 0.5$ and AR surprisal. Although the comparison is not perfectly matched in model scale, training data, or objective, the result suggests that the localization behavior is not explained simply by access to bidirectional context. Appendix A.7 reports the weaker Dream replication, and Appendix A.5 gives qualitative localization examples and error-position stress tests.

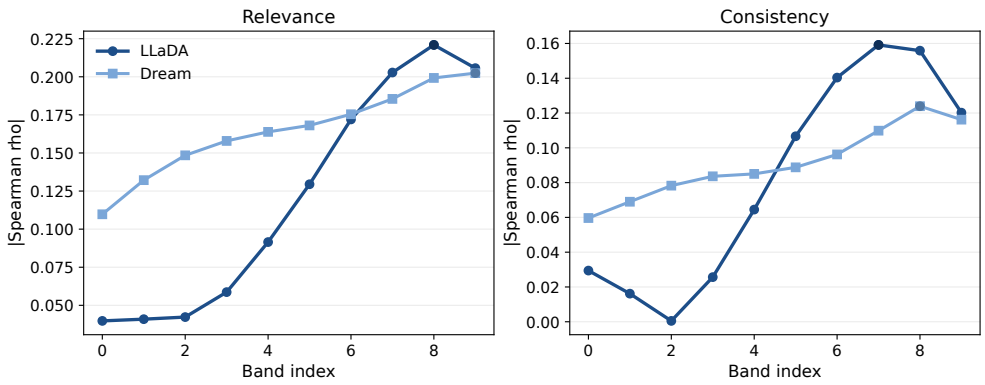

Figure 3: SummEval masking-rate profiles for relevance and consistency. Both diffusion evaluators concentrate their strongest evaluative signal in high masking-rate bands, although the peak location and magnitude differ across models. Larger values indicate stronger alignment between diffusion reconstruction loss and human quality judgments.

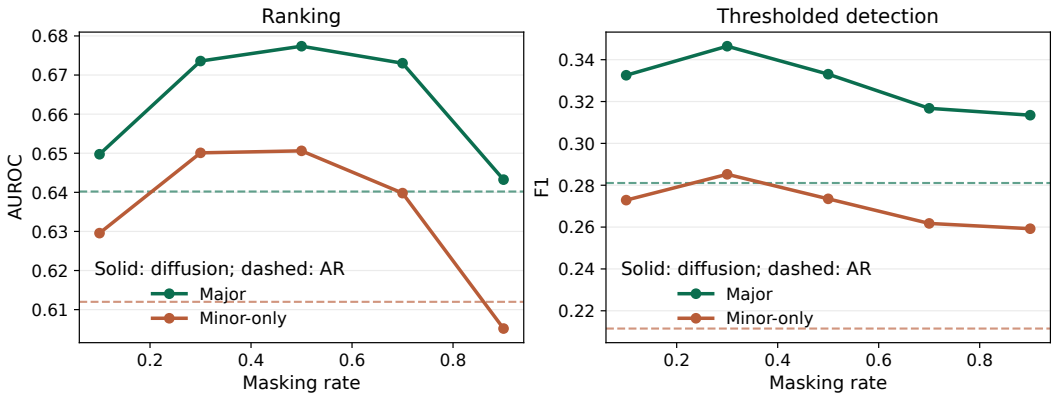

Figure 4: Masking-rate-dependent localization behavior on WMT MQM zh-en. Different masking regimes favor different localization objectives: intermediate masking rates produce the strongest ranking behavior (AUROC), while lower masking rates favor broader thresholded error detection (F1). Dashed horizontal lines show AR token-surprisal baselines for the same severity slice.

6 Conclusion

This work introduces masked diffusion language models as evaluators for text generation. We show that diffusion reconstruction exposes useful evaluation structure through the masking-rate axis: preserving reconstruction behavior across masking regimes yields complementary information beyond autoregressive likelihood and prompted judges, while also defining a localization operating curve for token-level error detection. The results suggest that, for the tested masked diffusion evaluators, evaluation behavior can depend not only on model scale or instruction tuning, but also on the reconstruction process used by the evaluator. Reconstruction under partial observation probes generated text differently from left-to-right prediction, and the resulting masking-rate profiles expose evaluative structure not preserved in standard scalar metrics.

In our experiments, high masking rates are particularly informative for relevance-oriented evaluation, while varying masking rate reveals systematic tradeoffs between selective ranking behavior and broader localization coverage. More broadly, modern evaluation increasingly relies on the same foundation models used for generation itself. Prompted judges and autoregressive likelihood evaluators often inherit closely related modeling assumptions and inductive biases. Our findings suggest that diffusion models offer a distinct evaluation perspective through iterative reconstruction under controlled corruption. Masking rate is one example of how model-native generation processes can expose useful evaluation structure beyond single scalar scores. Future work may identify additional evaluative behavior emerging from other reconstruction, corruption, or sampling processes across different generative paradigms.

Limitations

Model-family generalization. The current evidence is strongest for masked diffusion models, especially LLaDA. Dream reproduces the SummEval relevance complementarity more weakly and gives weaker WMT localization. The supported scope is therefore the tested masked diffusion models. Flow matching and continuous score-based text diffusion require separate evaluation.

Semantic interpretation. Some low-band local-quality effects appear relatively stable, but factuality-related signals are less consistent across datasets and model families. We therefore limit the semantic claim: masking rate changes the reconstruction setting, while local/global or factuality interpretations require task-specific validation.

Baseline strength. The judge comparison is intentionally scoped. Prompted scalar judges can remain competitive or stronger on narrow scalar tasks, and more optimized judge prompts could raise the baseline. A point-in-time stronger-judge check showed the same directional pattern but was not intended as a primary comparison.

Practical cost. Band-loss estimates require Monte Carlo samples over both masking rates and mask patterns, so the method is heavier than AR likelihood or a single prompted judge call. This cost is acceptable for analysis, but it limits immediate use as a production metric. Our sample-efficiency controls suggest that the main SummEval fusion signal survives substantially smaller Monte Carlo budgets, but estimator optimization remains future work. WMT roundtrip probing is also compute-asymmetric relative to AR surprisal because it uses repeated reconstruction across several masking rates.

Ethical Considerations

Evaluation metrics influence which generation systems are trusted, iterated on, and ultimately deployed. In that sense, even a diagnostic method has downstream impact: emphasizing only a single scalar score can hide failure modes that are important to users and practitioners. Our goal is to encourage more transparent evaluation by exposing rate-dependent behavior that scalar metrics may miss.

At the same time, diffusion-based diagnostics inherit biases from the underlying language models and from the benchmarks used to assess them. A specific misuse risk is treating high-mask behavior as a guarantee of factual correctness or reliability. We therefore present the method as a complementary analysis tool and keep it separate from human oversight or stronger task-specific evaluation protocols.

References

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

1307–1323, Online. Association for Computational Linguistics.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations*.

Google Research. 2024. Expert-based human evaluations for the submissions of WMT 2020, WMT 2021, WMT 2022 and WMT 2023. GitHub repository. Accessed 2026-05-25.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *Advances in Neural Information Processing Systems*.

OpenAI. 2026. Introducing GPT-5.4 mini and nano. Accessed 2026-05-25.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2023. T5Score: Discriminative fine-tuning of generative evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15178–15195, Singapore. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. 2023. Self-conditioned embedding diffusion for text generation. In *The Eleventh International Conference on Learning Representations*.

Zecheng Tang, Pinzheng Wang, Keyan Zhou, Juntao Li, Ziqiang Cao, and Min Zhang. 2023. Can diffusion model achieve better performance in text generation? bridging the gap between training and inference! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11359–11386, Toronto, Canada. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. 2023.

Ar-diffusion: Auto-regressive diffusion model for text generation. In *Advances in Neural Information Processing Systems*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *Preprint*, arXiv:2508.15487.

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

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, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Additional Details

A.1 Masked-Diffusion Decomposition

We use the absorbing-mask forward process common in masked diffusion language models. For a sequence $x = (x_1, \ldots, x_N)$, continuous time t, and mask symbol M, each token is independently kept with probability $1 - \sigma(t)$:

$$\begin{aligned} q(z_{t,i} \mid x_i) =& (1-\sigma(t))\mathbf{1}[z_{t,i} = x_i] \\ &+ \sigma(t)\mathbf{1}[z_{t,i} = M]. \end{aligned} \tag{7}$$

For this absorbing-state process, the continuous-time variational objective can be written as

$$\mathcal{L}(x) = -\int_0^{1^-} \frac{\dot{\sigma}(t)}{1-\sigma(t)} \sum_{i=1}^{N} \mathbb{E}_{q(z_t|x)}[\mathbf{1}[z_{t,i} = M] \log p_\theta(x_i \mid z_t)]\, dt. \tag{8}$$

Changing variables to $s = \sigma(t)$ gives $ds = \dot{\sigma}(t)dt$ and turns the time integral into a masking-rate integral:

$$\mathcal{L}(x) = \int_0^{1^-} \frac{1}{1-s} \sum_{i=1}^{N} \mathbb{E}_m\left[m_i r_i(x, m)\right] ds, \tag{9}$$

where $r_i(x, m) = -\log p_\theta(x_i \mid x_{\bar{S}})$ and $S = \{i : m_i = 1\}$. The factor $1/(1-s)$ is the continuous-time absorbing-state weight: as fewer tokens remain visible, each masked-token reconstruction term receives higher ELBO weight.

To obtain the band loss used in the main paper, write

$$\sum_i \mathbb{E}[m_i r_i(x, m)] = \mathbb{E}\left[|S| \frac{1}{|S|} \sum_{i \in S} r_i(x, m)\right]. \tag{10}$$

More explicitly, letting $\bar{r}_S(x, m) = \frac{1}{|S|}\sum_{i \in S} r_i(x, m)$ on the event $|S| > 0$, the conditional identity is

$$\begin{aligned} \mathbb{E}[|S|\bar{r}_S \mid |S| > 0] =& \mathbb{E}[|S| \mid |S| > 0]\ell(s, x) \\ &+ \mathrm{Cov}(|S|, \bar{r}_S \mid |S| > 0). \end{aligned} \tag{11}$$

Our estimator replaces the conditional mean mask count by sN and drops the covariance term, yielding

$$\mathcal{L}(x) \approx \int_0^{1^-} \frac{s}{1-s} N\ell(s, x) ds. \tag{12}$$

The approximation is therefore neither claimed as an upper bound nor as a lower bound; it is a pragmatic decomposition whose accuracy depends on both mask-count concentration and the residual covariance term. The first part is easy to quantify: for $|S| \sim \mathrm{Binom}(N, s)$, the coefficient of variation is $\sqrt{(1-s)/(Ns)}$. At the sequence lengths used in the main paper ($N \leq 1024$ with most examples near that cap), this gives relative mask-count fluctuations of about 0.094 at $s = 0.1$, 0.031 at $s = 0.5$, and 0.010 at $s = 0.9$. What remains uncontrolled is the covariance between mask count and average masked-token loss. We therefore interpret Eq. (11) as a useful decomposition with an explicit omitted term, and we treat exact semantic band placement as an empirical question.

Empirical magnitude of the omitted covariance. To estimate the practical size of this omitted term, we run a small length-stratified check on a 90-document SummEval subset spanning the observed summary lengths in that benchmark (21

to 147 tokens). For each document and each $s \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, we draw 200 independent masks and compute both the exact Monte Carlo quantity $\mathbb{E}[|S|\bar{r}_S]$ and the approximation $sN\ell(s, x)$. Table 3 shows that the omitted covariance is positive in the median, so the approximation underestimates the exact Monte Carlo term by a small positive bias: the median relative error ranges from 2.46% at $s = 0.1$ to 0.30% at $s = 0.9$. The empirical mask-count coefficient of variation closely matches the Binomial prediction at every tested rate, and the empirical gap between $\mathbb{E}[|S|]$ and sN remains below 0.3% in the median across all conditions. Most importantly for the downstream use in this paper, replacing the exact Monte Carlo term by the approximation preserves document-level ranking almost perfectly: $\text{Spearman}(\text{exact}, \text{approx}) \geq 0.996$ for all tested masking rates. The worst-case relative error stays below 10.7% on this subset.

Within the observed SummEval length range, median relative error decreases monotonically with document length across all tested masking rates, consistent with the $|S|$-concentration prediction. A split-half sanity check also shows stable covariance estimates: the median relative gap between two independent halves of the 200 mask samples drops from 2.6% at $s = 0.1$ to 0.05% at $s = 0.9$. We therefore use Eq. (11) as an empirically calibrated approximation.

s	Med. rel. err.	IQR	Emp./theory CV	Rank ρ
0.1	0.0246	0.0317	0.344 / 0.338	0.9969
0.3	0.0207	0.0133	0.170 / 0.172	0.9986
0.5	0.0167	0.0078	0.112 / 0.113	0.9996
0.7	0.0102	0.0053	0.073 / 0.074	0.9994
0.9	0.0030	0.0016	0.037 / 0.038	0.9997

Table 3: Empirical covariance probe for Eq. (11) on a 90-document SummEval subset (21 to 147 tokens). Relative error is the signed covariance term divided by the leading approximation $sN\ell(s, x)$. The empirical mask-count coefficient of variation closely matches the Binomial prediction, and document-level rankings induced by the exact and approximate quantities are nearly identical.

A.2 SummEval Positioning

Table 4 places the fixed-evaluator SummEval fusion results against published scalar references. It is included to calibrate scope; the main comparison in the paper is residual signal under a fixed evaluator family.

A.3 Reproducibility and Computational Cost

The main band-loss configuration uses 10 equal-width bands, 20 masking-rate samples per band, and 20 mask samples per rate. Thus a single example uses up to $10 \times 20 \times 20 = 4000$ masked reconstructions before batching. The main roundtrip configuration uses five masking rates and 100 reconstructions per rate. For token-level localization, AUROC and thresholded F1 use the conditional revision frequency in Eq. 6. The implementation also records an unconditional change-frequency field divided by the total number of reconstructions, but that diagnostic field is not used for the reported WMT AUROC or F1 results. All stochastic experiments use seed 13, bfloat16 inference, maximum sequence length 1024, and single A100-80G jobs unless otherwise stated.

The cost is substantial. On our cluster, Qwen AR scoring on SummEval completed in 4 minutes 54 seconds, while LLaDA band-loss scoring on the same 1,600 examples took 2 days 23 hours 21 minutes under the high-fidelity Monte Carlo configuration. WMT MQM full roundtrip over 8,431 segments took 5 hours 10 minutes. These numbers motivate our positioning: masked-diffusion scoring is an analysis tool for exposing diagnostic structure and should be treated separately from low-cost AR likelihood.

Our prompted judges use the same instruction: "Rate only whether the summary covers the important content of the source article." The model receives the source article and summary, then must return only JSON of the form `{"score": 4}`, where the score is an integer from 1 to 5. We parse only this JSON score and discard unparsable outputs. The Qwen judge is `Qwen2.5-7B-Instruct`; the stronger-judge check uses `gpt-5.4-mini` (OpenAI, 2026) through an API endpoint in April 2026 with deterministic decoding. Because commercial API models may change over time, we treat this experiment as a point-in-time robustness check. The static bidirectional cloze baseline uses `facebook/xlm-roberta-xxl`.

Fusion uses a grid over $\alpha \in \{0, 0.1, \ldots, 1\}$ after z-scoring the diffusion and AR signals. To avoid interpreting test-set tuning as a deployed metric, we also run 5-fold cross-validation on SummEval grouped by source document: all system outputs for the same article remain in the same fold. For each fold, the diffusion band and fusion weight are selected on the training split and evaluated on

A. No-judge signals		
Signal	**Relevance ρ**	**Gain**
AR likelihood	0.2106	n/a
Diffusion best band	0.2209	n/a
AR + full diffusion scalar	0.2610	ref.
AR + best diffusion band	**0.2795**	**+0.0185**

B. Fixed-judge fusion		
Added signal	**Relevance ρ**	**Gain**
None	0.3065	n/a
AR	0.3273	+0.0208
Diffusion	0.3454	+0.0389
AR + diffusion	**0.3636**	**+0.0571**

Table 4: SummEval relevance complementarity under the fixed open-weight setting. Panel A tests whether preserving the diffusion rate axis helps beyond a collapsed diffusion scalar; Panel B tests whether diffusion adds residual signal to the same fixed prompted judge. Published SummEval scalar references from Liu et al. (2023) range from 0.312 to 0.547 and are used only to calibrate scale.

Item	Reproducibility detail
Code and scripts	The paper source package contains the LaTeX sources, rendered figures, and rendered tables. Experimental code and configuration files will be released after de-anonymization.
Core models	LLaDA uses `GSAI-ML/LLaDA-8B-Instruct`; Dream uses `Dream-org/Dream-v0-Instruct-7B`; AR and the fixed judge use `Qwen/Qwen2.5-7B-Instruct`; the static cloze baseline uses `facebook/xlm-roberta-xxl`.
Pinned model refs	The local Qwen cache resolves `main` to `a09a35458c702b33eeacc393d103063234e8bc28`; diffusion checkpoints are recorded by release name in the current artifact and should be archived with commit hashes before de-anonymized release.
Environment	Python 3.11 with package lower bounds `torch>=2.1`, `transformers>=4.46`, `accelerate>=0.34`, `datasets>=2.18`, `scipy>=1.11`, and `pandas>=2.1`; LLaDA and Dream configs record Transformers 4.46.3 and 4.46.2 respectively.
Benchmarks and licenses	SummEval is distributed with MIT licenses in the local artifact; WMT MQM is distributed under Apache-2.0. All are used through their released benchmark files and cited in the main text.
WMT preprocessing	The loader keeps `hyp.*` systems from newstest2021 zh-en, groups rows by system/document/segment, strips MQM markup, projects spans to character-level masks, derives severity slices from released segment-level counts, and aggregates rater penalties with the released severity weights.
Randomness and hardware	Seed 13, bfloat16 inference, max length 1024, and single A100-80G jobs unless otherwise stated; WMT significance uses 2,000 segment-level bootstrap resamples with paired comparisons matched by segment id.

Table 5: Reproducibility summary for the main experiments. The table separates what is already fixed in the anonymous artifact from what should be pinned again before public release, especially diffusion checkpoint commit hashes.

the held-out split. For relevance, grouped cross-validated fusion reaches $\rho = 0.2795$, close to the full-data diagnostic value of 0.2857; all folds select band 8 and weights near $\alpha = 0.5$.

A.4 Efficiency and Sensitivity Controls

Table 8 reports a direct Monte Carlo budget check on SummEval relevance. Reducing the estimator from 20×20 masking-rate and mask samples to 5×5 preserves the main diffusion and AR fusion result, while reducing runtime from the original high-fidelity multi-day run to under four hours on the same cluster setting. The complementarity signal persists with far fewer than 4,000 reconstructions per example.

Table 9 checks whether the relevance finding depends on the arbitrary choice of 10 equal-width bands. The exact selected index changes with the discretization, but cross-validated fusion remains positive and the selected region stays in the high-mask part of the axis.

Table 10 breaks the rate order while preserving the LLaDA reconstruction sample budget. For each SummEval example, the 200 stored reconstruction samples are randomly repartitioned into ten pseudo-bands. The same grouped-CV selection and fusion protocol is then applied to each shuffled run.

Table 11 summarizes the WMT operating-point comparison discussed in Section 5.2.

Table 12 reports cross-fitted threshold selection for WMT MQM token-level F1. Thresholds are selected on held-out folds and evaluated on separate folds. Diffusion remains ahead of both AR surprisal and the XLM-R bidirectional cloze baseline on macro-F1 for error-containing, major-error, and minor-only slices.

Model	Dim.	Band	AR	Band ρ	AR+band	Gain
LLaDA	Rel.	8	0.2106	0.2209	0.2857	+0.0648
	Cons.	7	0.1022	0.1592	0.1849	+0.0257
	Flu.	0	0.1699	0.1709	0.1887	+0.0177
	Coh.	1	0.1695	0.0869	0.1695	+0.0000
Dream	Rel.	9	0.2106	0.2024	0.2638	+0.0532
	Cons.	8	0.1022	0.1239	0.1465	+0.0226
	Flu.	0	0.1699	0.0287	0.1794	+0.0095
	Coh.	0	0.1695	0.0388	0.1784	+0.0088

Table 6: Full-data SummEval per-dimension AR and diffusion-band complementarity. Band correlations are sign-aligned Spearman correlations for the selected diffusion band. Gain is AR+band minus the stronger single signal. This table supplies the detailed relevance, consistency, fluency, and coherence breakdown summarized in Section 5.1.

Model	Dim.	AR	AR+scalar	AR+band	Δ	p
LLaDA	Rel.	0.2106	0.2610	0.2795	+0.0185	.0015
	Cons.	0.1022	0.1428	0.1705	+0.0276	.0085
	Flu.	0.1699	0.1692	0.1852	+0.0160	.0895
	Coh.	0.1695	0.1590	0.1530	-0.0059	.9675
Dream	Rel.	0.2106	0.2524	0.2517	-0.0006	.6577
	Cons.	0.1022	0.1361	0.1423	+0.0062	.0055
	Flu.	0.1699	0.1692	0.1693	+0.0001	.4848
	Coh.	0.1695	0.1609	0.1517	-0.0094	.8161

Table 7: SummEval scalarization controls. AR+scalar fuses AR likelihood with the full diffusion scalar from the same diffusion model, while AR+band fuses AR with the best masking-rate band selected within grouped-CV folds. Δ is AR+band minus AR+scalar.

Budget	Runtime	Band ρ	Fusion ρ	Gain
5×5	3h43m	0.2263	0.2837	0.0574
10×10	9h06m	0.2204	0.2786	0.0582
20×20	2d23h21m	0.2209	0.2795	0.0586

Table 8: Monte Carlo budget sensitivity on SummEval relevance. The 20×20 runtime is the original high-fidelity run; lower-budget runs were executed separately and preserve the same fusion gain.

Bands	Band ρ	Fusion ρ	Gain	Selected region
5	0.1668	0.2521	0.0414	mostly band 4
10	0.2209	0.2795	0.0586	band 8 in all folds
20	0.1799	0.2778	0.0672	bands 16 and 17

Table 9: Band-boundary sensitivity for SummEval relevance. The high-mask region remains selected under 5, 10, and 20 equal-width partitions.

A.5 Inductive-Bias Slices

Table 13 groups WMT MQM segments by where the gold error span falls in the hypothesis. This is a descriptive stress test for the localization claim. If the roundtrip gain were mainly an artifact of one position bucket, the advantage over AR token surprisal would likely collapse in at least one bucket. Instead, LLaDA stays ahead for beginning, middle, and end errors. The gap is largest for end-position errors, but we treat this as a descriptive pattern; position alone fails to explain the roundtrip advantage.

Table 14 gives three qualitative SummEval relevance slices selected by rank-error agreement among the prompted judge, AR likelihood, and diffusion. These cases illustrate the inductive-bias interpretation; the quantitative tables carry the main evidence. The judge+AR case is a high-quality factual news summary that diffusion ranks too low. The judge+diffusion case is locally repetitive and off-quality; AR ranks it differently from the human ranking, while the judge and diffusion assign it near-human rank. The AR+diffusion case is broadly on-topic and informative, but the judge assigns a noticeably lower rank than the human ordering. This is the one place where we explicitly surface a judge-side hypothesis: prompted judges may prefer more elaborated summaries, so a terse factual summary can be penalized even when probability-based signals stay close to the human order.

A.6 Representation Probe

Table 15 reports a small hidden-state probing sanity check on LLaDA roundtrip states for WMT MQM. We extract masked-token hidden states at $s \in \{0.1, 0.5, 0.9\}$ and train grouped-split probes, so the reported values are held-out values. Token-surface labels are local lexical classes of the masked token, while major-error labels are segment-level MQM indicators predicted from mean masked-token states. The main value of this table is robustness of ordering: both linear and nonlinear probes preserve the same broad pattern that low-mask states are easier to decode than high-mask states. The linear probe supports the local side of the information-resolution interpretation, since token-surface information is much more linearly decodable at low masking than at high masking. A nonlinear MLP probe raises major-error AUROC at $s = 0.9$ from 0.523 to 0.568, but it also raises low-mask decoding from 0.659 to 0.737 at $s = 0.1$. We therefore avoid reading this as evidence that high-mask states have a special nonlinear privilege. Instead, we take it as a calibration

Signal structure	AR fusion ρ	Notes
Observed rate bands	0.2795	grouped-CV result
Full diffusion scalar	0.2610	scalarized LLaDA
Shuffled pseudo-bands	0.2098	mean over 100 shuffles
Best shuffled run	0.2431	maximum over 100 shuffles

Table 10: Sample-shuffle control for SummEval relevance. Shuffled pseudo-bands preserve the number of reconstruction samples while removing the masking-rate ordering; the 95% shuffle range for AR+pseudo-band fusion is $[0.1911, 0.2362]$.

Metric	Diffusion best	AR	Winner
Overall AUROC	0.6705	0.6330	Diff.
Pooled F1	0.2311	0.2636	AR
Major F1	0.3464	0.2811	Diff.
Minor F1	0.2852	0.2115	Diff.

Table 11: WMT localization comparison at the reported operating points. Results are mixed: diffusion leads on ranking AUROC and severity-specific F1, while AR leads on pooled fixed-threshold F1.

check that the mask-rate ordering remains stable under a stronger decoder. Stronger mechanistic evidence would require nonlinear probes beyond this small check, or causal methods such as activation patching and controlled representation interventions.

Exploratory patching check. As a minimal follow-up, we patch low-mask hidden states into high-mask forward passes on a small WMT MQM subset. The check is qualitative and underpowered for effect-size claims, so we use it only as a design note: reconstruction loss moves differently under structured patching than under random patching. It leaves open whether higher-mask states encode globally distinct error semantics.

A.7 Dream WMT Replication

Table 16 reports the same WMT MQM roundtrip aggregation with Dream-7B-Instruct. This is a negative replication of the strongest LLaDA localization result: Dream retains a masking-rate-dependent curve, but the absolute AUROC values are much lower and the best masking rates shift upward. We therefore treat WMT localization as strong evidence for LLaDA and as an explicit model-dependence boundary for masked diffusion evaluation.

Slice	Diff. .3	Diff. .5	AR	XLM-R
Err.	0.3251	0.3246	0.2987	0.2874
Major	0.3385	0.3381	0.3116	0.3018
Minor	0.2832	0.2838	0.2602	0.2443

Table 12: Cross-fitted macro-F1 under threshold sensitivity analysis on WMT MQM. Diff. .3 and Diff. .5 denote diffusion at $s = 0.3$ and $s = 0.5$; Err. denotes error-containing segments. Threshold selection leaves the comparison with AR and the bidirectional cloze baseline unchanged.

Pos.	Rec.	Diff. AUC	AR AUC	Diff. F1	AR F1
Beginning	1,215	0.6587	0.6053	0.2763	0.2344
Middle	3,168	0.6365	0.5579	0.3582	0.2437
End	1,504	0.6478	0.5263	0.2809	0.1652

Table 13: WMT MQM error-position stress test. Diffusion AUROC uses roundtrip scores at $s = 0.5$; diffusion F1 uses $s = 0.3$. AR F1 uses the same per-segment top-20% token-surprisal rule as the AR token baseline.

Slice	Failure mode and excerpt
Judge + AR *w/o Diff.* *rank err.*↓ J 73.5 AR 29.0 Diff. 1329.0	**Long-range ordering sensitivity** Eight Iranian border guards have been killed in clashes with militants near the border with Pakistan. A militant group called Jaish al Adal claimed responsibility for the attack.
Judge + Diff. *w/o AR* *rank err.*↓ J 15.0 AR 1513.0 Diff. 3.0	**Local fluency trap under repetition** Liverpool beat Newcastle United 2 to 0. "I thought it was a penalty," Allen said. "I thought it was a penalty. I thought it was a penalty."
AR + Diff. *w/o Judge* *rank err.*↓ J 1425.5 AR 32.5 Diff. 181.5	**Possible judge preference mismatch for terse factual summaries** Wales' crunch Euro 2016 qualifier with Belgium has been declared a 33,000 sell-out. The top two sides in Group B meet in Cardiff on June 12, with Wales in their best position to qualify for a major tournament since 1958.

Table 14: Qualitative SummEval relevance slices. Each row shows a case where one evaluator disagrees with the other two; rank errors are measured against the human ranking, with lower values better.

Probe	$s = 0.1$	$s = 0.5$	$s = 0.9$
Token surface F1 (linear)	0.830	0.611	0.244
Major-error AUROC (linear)	0.659	0.652	0.523
Major-error F1 (linear)	0.567	0.546	0.479
Token surface F1 (MLP)	0.849	0.638	0.331
Major-error AUROC (MLP)	0.737	0.691	0.568
Major-error F1 (MLP)	0.685	0.592	0.570

Table 15: Hidden-state probing sanity check on WMT MQM. Low masking preserves more decodable local token-surface information. Nonlinear probes partially recover major-error signal at all mask rates, but the low-to-high decodability ordering remains intact, so this table is best read as an ordering-robustness check without claiming a white-box proof of a strict information-resolution law.

Model	Maj. AUC	Maj. F1	Min. AUC	Min. F1
LLaDA	0.6774 @ .5	0.3464 @ .3	0.6506 @ .5	0.2852 @ .3
Dream	0.5871 @ .7	0.3134 @ .9	0.5418 @ .7	0.2597 @ .9

Table 16: Dream replication on WMT MQM zh-en. Values report each model's best masking rate for the metric. Dream is weaker and shifts toward higher masking rates.