COHERENCE-AWARE DISTRIBUTIONAL EVALUATION OF OPEN-ENDED TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing metrics for open-ended text generation measure likelihood, lexical diversity, or distributional similarity in generic representation space, yet they can miss fundamental dimensions of quality. A prominent blind spot is global coherence: a generated passage may be locally fluent while remaining globally contradictory, causally inconsistent, or topically disconnected. Such failures can still preserve the token-level and lexical statistics that existing metrics rely on. We identify representation as a central bottleneck in detecting these failures and introduce CHORD (**C**oherence-aware **H**idden-state **O**pen-generation **R**eference **D**istance), a coherence-sensitive distributional metric. CHORD encodes generated and human-written corpora in the hidden-state space of a frozen LLM using a coherence-eliciting prompt, and compares the resulting distributions using MMD with an RBF kernel. To validate that the metric responds to coherence degradation but not generic textual change, we construct a counterfactual evaluation suite that pairs graded coherence-degrading perturbations with meaning-preserving controls. CHORD selectively detects relation, discourse, structural, and mixture failures that perplexity, entropy, MAUVE, FBD, and MMD-based baselines either miss or cannot separate from benign rewriting. Factorial ablations show that representation is the primary source of coherence sensitivity, while RBF-MMD improves sample efficiency once the relevant distinctions become visible. Larger backbones capture finer-grained distinctions, but coherence prompting improves selectivity only when the backbone can follow the prompt. On unconditional generation and prefix continuation, CHORD yields model rankings that strongly align with human judgments of whether outputs make sense and appear human-written. Together, these results establish representation design as central to reliable distributional evaluation.

1 INTRODUCTION

Evaluating open-ended text generation remains fundamentally challenging because generation quality is multidimensional, whereas existing automatic metrics capture only particular aspects of it: generative perplexity measures the predictability of generated text under an external language model (Holtzman et al., 2020), unigram entropy quantifies lexical diversity (Shannon, 1948), and Self-BLEU estimates repetition across generated samples (Zhu et al., 2018). Yet recent studies show that optimizing these metrics need not yield improvements in overall generation quality (Veličković et al., 2026; Franca & Tong, 2026; Pynadath et al., 2026).

Open-ended generation often admits multiple valid outputs, making one-to-one matching against reference outputs ill-suited for evaluation. Distributional metrics avoid the need for aligned references by comparing corpora of generated and human-written text in a shared feature space, thereby evaluating corpus-level similarity rather than sample-level correspondence. MAUVE (Pillutla et al., 2021; 2023), Fréchet-BERT Distance (Xiang et al., 2021), and MMD-based metrics (Chan et al., 2024) instantiate this paradigm with different representations and divergences. However, a distributional metric can detect only those differences preserved by its feature space. Prior stress tests have shown that commonly used representations can be insensitive to discourse-level perturbations, including sentence reordering (He et al., 2023).

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

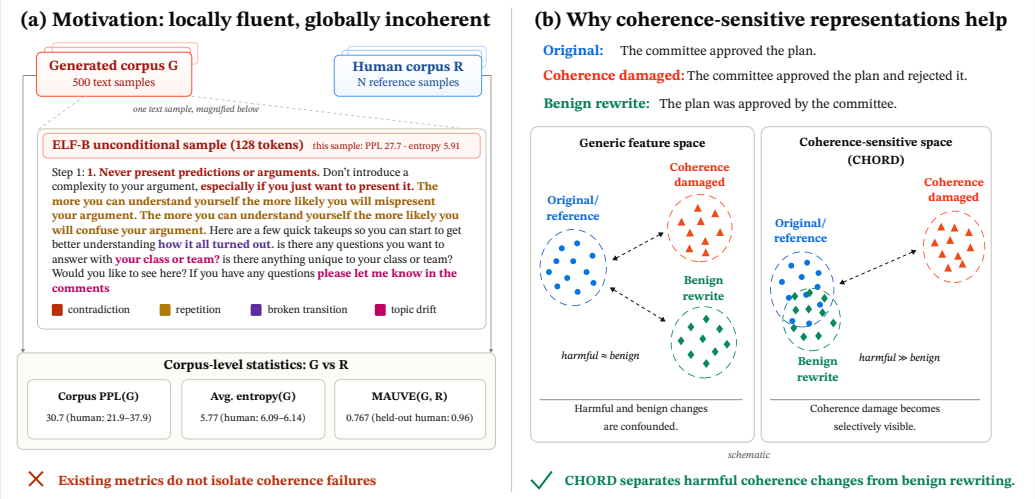

Figure 1: **Motivation and intuition for CHORD.** Existing corpus-level metrics may overlook text that is locally fluent yet globally incoherent; CHORD selectively distinguishes coherence degradation from meaning-preserving rewriting.

Coherence poses a particularly demanding test of this representation bottleneck. Current generators often produce multi-sentence passages that are locally fluent yet globally incoherent. A generated text sample may contradict earlier claims, reverse causal relations, disrupt discourse transitions, or drift across unrelated topics. Such failures can leave local fluency and much of the lexical content intact, providing little signal for perplexity and entropy. Generic representations can likewise conflate coherence violations with benign textual variation: for example, a contradiction-inducing edit and a meaning-preserving paraphrase may produce shifts of comparable magnitude in the representation space. Figure 1 illustrates this failure mode.

Dedicated coherence metrics (Barzilay & Lapata, 2008; Logeswaran et al., 2017; Guan & Huang, 2020; Zhu et al., 2024; Zhao et al., 2023; Ke et al., 2022) offers a complementary, sample-level approach. These methods score individual texts and are often evaluated on sentence ordering and topical consistency, with relation-level violations such as contradiction and causal reversal receiving less systematic coverage. Because they score samples individually, however, they are not designed to compare generated and human corpora as distributions. Our goal is to bridge this gap by making corpus-level distributional evaluation sensitive to a broader range of coherence failures. Accordingly, we define *selectivity* as a central evaluation criterion: **a coherence-degrading perturbation should elicit a stronger metric response than a comparable meaning-preserving perturbation**.

We introduce **CHORD** (**C**oherence-aware **H**idden-state **O**pen-generation **R**eference **D**istance), a distributional metric designed to capture coherence differences between generated and human text. CHORD represents each text using hidden states extracted from a frozen LLM under a coherence-eliciting prompt, and compares the resulting distributions using MMD with an RBF kernel (Gretton et al., 2012). To evaluate selectivity, we construct a counterfactual evaluation suite that pairs graded coherence-degrading perturbations with meaning-preserving controls. Across this suite, CHORD detects all tested failure types, whereas perplexity, entropy, and MAUVE with generic features either miss relation- and discourse-level errors or fail to distinguish them from benign rewriting.

A factorial ablation over feature representations and distance functions confirms that representation is the bottleneck: no distance function can recover coherence distinctions that are absent from the feature space. Further experiments reveal an interaction between backbone capability and coherence elicitation. Larger backbones capture finer-grained coherence distinctions, whereas targeted prompting improves selectivity only when the backbone can follow the instruction; with GPT-2 (Radford et al., 2019), it can even reduce selectivity. Finally, CHORD's model rankings on unconditional generation and prefix continuation strongly agree with human judgments.

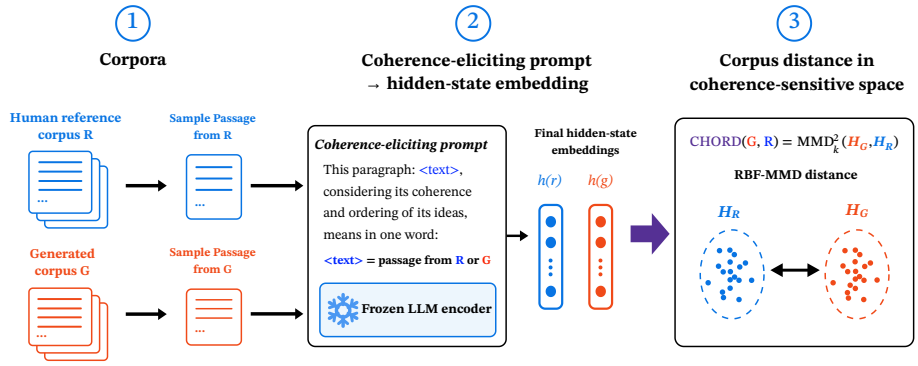

Figure 2: **CHORD overview.** A frozen LLM encodes prompted human and generated passages into coherence-oriented hidden states; RBF-MMD compares the resulting corpus distributions.

2 Related Work

Automatic evaluation of open-ended generation. Reference-free diagnostics characterize generated text alone: generative perplexity measures predictability (Holtzman et al., 2020), unigram entropy lexical diversity (Shannon, 1948), and Self-BLEU cross-sample repetition (Zhu et al., 2018). They remain common for diffusion language models (Hu et al., 2026; Guo et al., 2026), although recent analyses show they give an incomplete account of generation quality (Veličković et al., 2026; Franca & Tong, 2026; Pynadath et al., 2026). Distributional metrics instead compare generated and human corpora in a shared feature space, via information frontiers in MAUVE (Pillutla et al., 2021; 2023), the Fréchet distance in FBD (Xiang et al., 2021), or MMD (Chan et al., 2024). Their sensitivity depends on the representation: commonly used features can miss discourse-level perturbations such as sentence reordering (He et al., 2023).

Coherence modeling and evaluation. Coherence evaluation has largely developed around two synthetic tasks. Permutation detection, distinguishing sentence-shuffled text from the original, motivated entity-grid models (Barzilay & Lapata, 2008), neural coherence models (Li & Jurafsky, 2017), and sentence-ordering objectives (Logeswaran et al., 2017). Sentence intrusion detection, spotting a sentence from an unrelated document, drove discriminators such as UNION (Guan & Huang, 2020) and CoUDA (Zhu et al., 2024), trained on programmatically corrupted negatives. Other metrics derive coherence scores from pretrained encoders without task-specific training, including DiscoScore (Zhao et al., 2023) and CTRLEval (Ke et al., 2022). These approaches score individual samples and are mainly tested on ordering and topic consistency. CHORD instead compares corpora as distributions and extends detection beyond sentence permutation and document mix to relation-level failures such as contradiction and causal reversal.

Prompt-based sentence embeddings. Prompt-based embeddings place text in a task-specific template and read a hidden state from a frozen language model. PromptEOL (Jiang et al., 2024) asks the model to summarize the text in one word and uses the final-token state; MetaEOL (Lei et al., 2024) varies the prompt task to capture different aspects of the text. CHORD adopts a fixed coherence-oriented template and uses the resulting last-token states for corpus-level distributional comparison.

3 Methodology

3.1 CHORD: Hidden-State Reference Distance

Figure 2 summarizes the three-stage evaluation pipeline. Let $R = \{r_i\}_{i=1}^{N_R}$ be a human reference corpus and $G = \{g_j\}_{j=1}^{N_G}$ a generated corpus. CHORD measures their distributional discrepancy using coherence-oriented representations and a kernel two-sample distance.

Coherence-oriented hidden-state representation. To represent a text sample x with a frozen language model, we place it in a fixed coherence-eliciting template adapted from PromptEOL (Jiang et al., 2024):

```
This paragraph: "{x}", considering its coherence and
ordering of its ideas, means in one word:
```

The compact-completion format requires the model to compress the input to a single next-token prediction, concentrating coherence information at the last token position. Let $p(x)$ denote the prompted input. The sample representation is the hidden state at layer ℓ of the last token position t:

$$h(x) = f_\theta\big(p(x)\big)_{\ell,\,t} \in \mathbb{R}^d, \tag{1}$$

where f_θ is a frozen language model. CHORD uses these hidden states to compare corpora; it neither decodes continuations nor scores output tokens. Appendix F compares this signal with bag-of-words features, next-token predictions, and LLM-as-a-Judge scoring (Zheng et al., 2023).

Unless otherwise specified, the CHORD encoder in this paper is a frozen Qwen3.5-27B model (Qwen Team, 2026). We extract its features from the **third-to-last** transformer layer. Appendix O reports the layer-depth ablation motivating this choice, and Appendix B shows the full detail of the 27B encoder configuration.

The same design transfers across model families. Table 1 shows that Gemma-2 (Gemma Team et al., 2024), Mistral (Mistral AI, 2025; 2024), and Llama-3.1 (Grattafiori et al., 2024) backbones detect a broad range of coherence failures using the same prompt, extraction rule, and distance, without family-specific tuning.

Gaussian RBF-MMD distance. Let $H_R = \{h(r_i)\}_{i=1}^{N_R}$ and $H_G = \{h(g_j)\}_{j=1}^{N_G}$. We compare these embedding distributions using the biased squared Maximum Mean Discrepancy (MMD) with a Gaussian RBF kernel (Gretton et al., 2012):

$$\mathrm{CHORD}(G, R) = \widehat{\mathrm{MMD}}_k^2(H_G, H_R), \qquad k(\mathbf{a}, \mathbf{b}) = \exp\left(-\frac{\|\mathbf{a}-\mathbf{b}\|^2}{2\sigma^2}\right). \tag{2}$$

For each encoder and evaluation set, we set the bandwidth σ to the median pairwise embedding distance on a held-out human-reference split, then keep it fixed across corpus comparisons. Appendix B gives the biased MMD estimator, calibration split sizes, and further details of the bandwidth-selection procedure.

3.2 Diagnostic Meta-evaluation Protocol for Coherence Sensitivity

We evaluate coherence sensitivity using controlled perturbations paired with benign rewrites, null-standardized metric responses, and a statistical criterion for selective detection.

Perturbations and benign controls. We construct **nine** perturbation types from a shared pool of seed samples, organized into four categories that each target a different aspect of coherence: *relation* perturbations (*contradiction*, *causal reversal*) introduce contradictions and causal reversals between sentences; *discourse* perturbations (*broken transition*, *topic drift*) introduce logically disconnected transitions and topic drift; *structural* perturbations (*sentence permutation*, *word shuffle*, *repetition*) disrupt sentence order, shuffle word order, or duplicate sentences; and *mixture* perturbations (*DLM mix*, *document mix*) replace sentences with material from unrelated sources. As benign controls, we paraphrase the same seed samples while preserving their claims and logical structure, establishing the metric response attributable to surface rewriting alone. Each perturbation type is applied at multiple severity levels. Appendix C provides the full construction procedure and examples.

Null standardization. For each metric, let s_M be the calculated raw distributional distance between the candidate set and the human reference set. We repeatedly compute the same discrepancy between two disjoint human-reference subsets, matched to the evaluation sample sizes. The resulting null mean μ_0 and standard deviation σ_0 define

$$z_M = \frac{s_M - \mu_0}{\sigma_0}. \tag{3}$$

This expresses each response relative to its own reference sampling variation. Appendix B specifies the statistic used for each metric and the resampling procedure.

Detection criterion. Since benign rewriting also shifts scores, we compare each harmful perturbation against the most heavily rewritten benign control built from the same seed samples, a conservative baseline:

$$\Delta z = z_{\text{harm}} - z_{\text{benign}}, \tag{4}$$

where z denotes the null-standardized score, so $\Delta z > 0$ means the harmful perturbation is penalized beyond benign rewriting. A condition is *detected* if the lower bound of the 95% CI of Δz, estimated from 200 paired bootstrap resamples over seeds, is positive (Appendix B).

4 EXPERIMENTS

We organize our experiments around four research questions:

- **RQ1 (Selectivity).** Does CHORD respond to coherence-damaging perturbations while remaining stable under benign rewriting? (§4.2)
- **RQ2 (Attribution).** Which representation and distance choices account for this selectivity? (§4.3)
- **RQ3 (Model ranking).** Does CHORD yield informative rankings of existing generation models? (§4.4)
- **RQ4 (Human alignment).** How well do CHORD scores agree with human judgments? (§4.5)

4.1 SETUP

Counterfactual evaluation set. We sample passages from OpenWebText (Gokaslan et al., 2019), WikiText-103 (Merity et al., 2016), and Reddit TL;DR posts (Völske et al., 2017; Stiennon et al., 2022), retaining those with at least four sentences and 100–400 words. The filtered passages form three disjoint pools: 500 seeds for counterfactual editing, 3,600 passages for the human reference corpus, and a replacement pool for mixture perturbations.

From the 500 seeds, we construct the nine perturbation types and benign control defined in Section 3.2. Qwen3-30B-A3B (Yang et al., 2025) generates the relation and discourse perturbations and benign controls, while rule-based operations produce the structural and mixture perturbations. The main findings persist with a Mistral editor (Appendix C). We apply each perturbation type at multiple severity levels. Table 1 reports the highest-severity condition for each type; Appendix H gives the full severity results, and Appendix C details their construction. All scores are null-standardized, and detection follows the paired bootstrap criterion in Section 3.2.

Baselines. We compare CHORD with GPT-2 generative perplexity (Holtzman et al., 2020), unigram entropy (Shannon, 1948), MAUVE (Pillutla et al., 2021) using GPT-2 or ELECTRA (Clark et al., 2020) features, Fréchet-BERT Distance (Xiang et al., 2021), and MMD with MiniLM sentence embeddings (Chan et al., 2024; Wang et al., 2020) (MMD-MiniLM). Appendix D provides implementation details and backbone references. For MAUVE, we null-standardize the frontier integral returned by the official implementation, a divergence that is zero when the two cluster histograms match, rather than the MAUVE score itself, using the same reference null as for the other metrics (Appendix D). Appendix E reports comparisons with modern instruction-tuned embeddings, UniEval (Zhong et al., 2022), and BARTScore (Yuan et al., 2021); Appendix F evaluates same-backbone LLM judges, including G-Eval (Liu et al., 2023).

CHORD encoders. Alongside the default Qwen3.5-27B encoder, we provide two distilled student encoders trained to match the teacher's PCA-projected embeddings, following OPRD's representation-alignment approach (Yang et al., 2026b): Qwen3.5-2B, a high-fidelity variant that most closely tracks the teacher, and Qwen3.5-0.8B, a lightweight variant for faster, lower-memory inference. On 500 packed OpenWebText samples processed on a single H200, featurization takes 46.5 s and 54.2 GB peak GPU memory with the 27B encoder, compared with 5.3 s and 7.1 GB for 2B and 4.4 s and 4.5 GB for 0.8B. Appendices I and P give training, validation, and cost details.

Table 1: **Null-standardized responses z_M at the highest severity for each perturbation type.** Blue shading marks detected conditions: the lower bound of the paired-bootstrap confidence interval for selectivity Δz exceeds zero (Section 3.2). All CHORD backbones use the same coherence prompt and feature-extraction rule. Appendix H reports results at every severity level.

	Relation		Discourse		Structural			Mixture		Control
	← harder to detect				easier to detect →					
Method	Causal reversal	Contra-diction	Broken transition	Topic drift	Sentence permutation	Word shuffle	Repetition	DLM mix	Document mix	Benign paraphrase
Likelihood / diversity statistics										
gen-PPL (GPT-2)	0.8	1.1	0.5	1.8	7.4	46.9	62.4	62.4	35.9	1.4
Unigram entropy	−0.1	0.6	6.0	0.0	−0.0	2.1	0.9	22.1	17.4	2.6
Distributional metrics										
MAUVE (GPT-2)	23.1	27.6	21.0	18.9	24.5	1.1	56.5	35.8	44.9	27.3
MAUVE (ELECTRA)	3.0	2.6	3.7	2.4	13.7	72.5	89.7	81.2	76.2	2.5
FBD (BERT)	0.2	0.2	0.8	0.5	1.0	9.7	60.7	24.5	12.5	0.2
MMD (MiniLM)	−0.0	−0.0	33.3	4.9	−0.3	0.4	18.9	69.5	62.5	−0.3
Ours										
CHORD (Qwen3.5-27B)	25.6	84.3	331	79.7	358	740	857	1150	1082	6.0
CHORD (Qwen3.5-9B)	17.4	26.9	109	27.5	96.3	344	671	882	726	10.9
CHORD (Qwen3.5-2B distilled)	20.1	73.1	258	45.9	190	579	1001	1241	1211	4.3
CHORD (Qwen3.5-0.8B distilled)	16.3	43.3	257	34.7	173	770	1191	1222	1166	2.5
CHORD (Gemma-2-27B-it)	24.7	80.5	270	66.1	182	514	948	845	627	6.6
CHORD (Gemma-2-9B-it)	11.0	27.0	135	27.0	120	328	449	804	691	3.0
CHORD (Mistral-Small-24B)	17.1	40.7	161	41.6	175	315	808	653	550	3.5
CHORD (Ministral-8B)	10.8	16.5	74.8	12.1	62.9	235	467	535	368	5.7
CHORD (Llama-3.1-8B-Instruct)	5.6	14.0	109	11.7	53.6	152	143	543	306	3.3

4.2 Counterfactual Robustness

Table 1 compares all metrics on our counterfactual evaluation set at the highest perturbation severity. The metrics differ substantially in the failures they detect.

Generative perplexity detects word shuffle, repetition, and mixture but responds to contradiction and causal reversal much like the benign control. Unigram entropy detects mainly mixture perturbations. FBD also fails to detect relation-level errors, while MMD-MiniLM detects discourse but not relation-level failures.

MAUVE depends strongly on its features. With GPT-2, its contradiction shift (27.6) nearly matches benign paraphrasing (27.3); ELECTRA reduces the benign response but is still insensitive to relation-level failures.

CHORD distinguishes all nine perturbation types from benign paraphrasing. Responses are strongest for structural and mixture failures but remain selective for contradiction, causal reversal, broken transitions, and topic drift. Both distilled Qwen3.5-2B and 0.8B encoders detect all nine types, though with weaker relation-level responses than the 27B encoder.

This pattern transfers across backbone families. With the prompt, extraction layer, and distance fixed, all Gemma, Mistral, and Llama backbones in Table 1 detect structural and mixture failures. The 24–27B Gemma and Mistral models detect all nine types, whereas smaller Ministral and Llama models miss some relation or discourse failures.

4.3 What Makes CHORD Coherence-Sensitive?

MAUVE and CHORD both compare distributions of text representations, but differ in three components: **the backbone, the representation extraction method, and the distributional distance**. We isolate their contributions below.

4.3.1 Representation and distance

We pair four representations (coherence-prompted Qwen, GPT-2, ELECTRA, and BERT (Devlin et al., 2019)) with four distributional distances: RBF-MMD, energy distance (Székely & Rizzo, 2013), Fréchet distance (Heusel et al., 2018), and MAUVE's k-means KL (its frontier integral; Appendix D). This yields 16 combinations evaluated on the nine perturbation types in Table 1.

Figure 3a shows that detection depends primarily on the representation. Swapping distances between MAUVE and CHORD makes this clear: GPT-2 with RBF-MMD detects only three perturbation types, whereas coherence-prompted Qwen with k-means KL detects all nine. Across the full grid,

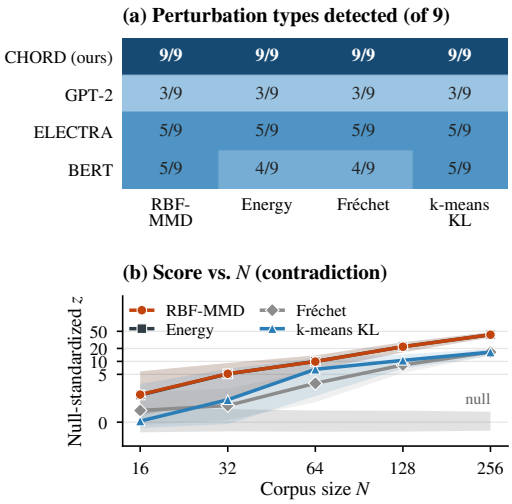

Figure 3: **Representation and distance ablations. (a)** Number of perturbation types detected for each representation and distributional distance. Off-diagonal settings include GPT-2 features with RBF-MMD and coherence-prompted Qwen features with MAUVE's k-means KL. **(b)** Null-standardized score versus corpus size on contradiction.

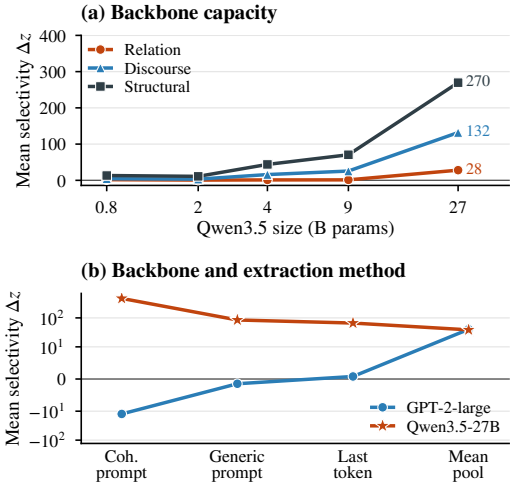

Figure 4: **Backbone capacity and coherence prompting.** Both panels report mean selectivity $\Delta z = z_{\text{harm}} - z_{\text{benign}}$. **(a)** Qwen3.5 scaling with the coherence prompt, representation layer, and RBF-MMD fixed. **(b)** Four extraction methods on Qwen3.5-27B and GPT-2-large, averaged over all harmful conditions (symmetric-log scale).

changing representations affects detection much more than changing distances. **The representation determines which coherence failures are exposed to the distributional comparison.**

Distance choice mainly affects sample efficiency. Figure 3b shows that RBF-MMD detects contradiction at smaller corpus sizes and with lower variance than Fréchet distance or k-means KL. We therefore use RBF-MMD as the default distance.

4.3.2 Backbone capacity and coherence prompting

Figure 4a isolates the effect of Qwen3.5 backbone size, holding the prompt, extraction layer, and distance fixed. Selectivity for structural and discourse failures improves at intermediate scales, whereas responses to relation-level failures remain weak until the largest model. This pattern agrees with the cross-family results in Table 1: **larger backbones capture finer coherence distinctions in their hidden representations.**

The benefit of coherence prompting also depends on the backbone. In Figure 4b, the coherence-prompted final-token representation is substantially more selective than unprompted alternatives on Qwen3.5-27B. On GPT-2-large, however, mean pooling performs best, and coherence prompting can yield negative selectivity. This pattern persists across all GPT-2 sizes (Appendix G, Table 10), suggesting that its effectiveness depends on whether the backbone can encode the target property.

Appendices G and N ablate prompt design, extraction, and error position; Appendix F compares the hidden-state signal with token-space measures and direct LLM judgments.

4.4 Evaluating Unconditional Generation and Conditional Text

To assess how well CHORD and existing metrics distinguish among text generation models, we evaluate them on unconditional generation and prefix continuation, two common settings for open-ended text generation.

Unconditional generation. Table 2 shows a consistent three-tier pattern across the 27B encoder and both distilled variants: held-out human text is closest to the reference distribution, followed by GPT-2 outputs, then diffusion and flow outputs from SEDD, MDLM, ELF-L, and LangFlow. The 2B encoder exactly reproduces the 27B ranking (Spearman $\rho = 1.00$). The 0.8B encoder preserves the three tiers but differs within them ($\rho = 0.82$), assigning nearly identical scores to the

Table 2: **Unconditional generation on OpenWebText.** We report mean±std over 10 non-overlapping sample sets, each containing 500 samples of 512 tokens. CHORD reports RBF-MMD^2 ($\times 10^{-2}$); scores are comparable only within the same encoder column. Details are provided in Appendix J.

Generator	CHORD ↓ (Qwen3.5-27B)	CHORD ↓ (Qwen3.5-2B distilled)	CHORD ↓ (Qwen3.5-0.8B distilled)	gen-PPL ↓	MAUVE ↑	entropy ↑
Held-out human (packed)	$0.17^{\#1}$ (±0.04)	$0.16^{\#1}$ (±0.03)	$0.15^{\#1}$ (±0.05)	$18.94^{\#3}$ (±0.31)	$0.95^{\#1}$ (±0.01)	$7.51^{\#2}$ (±0.02)
GPT-2-large (774M, nucleus)	$19.84^{\#2}$ (±1.02)	$9.98^{\#2}$ (±0.74)	$10.40^{\#3}$ (±0.77)	$6.70^{\#1}$ (±0.10)	$0.77^{\#5}$ (±0.04)	$7.03^{\#7}$ (±0.02)
GPT-2-medium (355M, nucleus)	$28.37^{\#3}$ (±0.90)	$10.78^{\#3}$ (±0.59)	$10.36^{\#2}$ (±0.66)	$10.10^{\#2}$ (±0.11)	$0.81^{\#4}$ (±0.03)	$7.06^{\#6}$ (±0.01)
ELF-L (652M) (Hu et al., 2026)	$51.19^{\#4}$ (±1.09)	$68.20^{\#4}$ (±1.33)	$69.72^{\#4}$ (±1.45)	$23.15^{\#5}$ (±0.53)	$0.09^{\#7}$ (±0.01)	$7.07^{\#5}$ (±0.02)
LangFlow (171M) (Chen et al., 2026)	$52.58^{\#5}$ (±1.12)	$70.84^{\#5}$ (±1.54)	$78.79^{\#7}$ (±1.73)	$18.98^{\#4}$ (±0.49)	$0.65^{\#6}$ (±0.05)	$7.60^{\#1}$ (±0.03)
SEDD-small (170M) (Lou et al., 2024)	$57.05^{\#6}$ (±1.02)	$74.95^{\#6}$ (±1.27)	$77.64^{\#6}$ (±1.14)	$69.21^{\#6}$ (±0.96)	$0.89^{\#2}$ (±0.02)	$7.21^{\#4}$ (±0.02)
MDLM (170M) (Sahoo et al., 2024)	$57.97^{\#7}$ (±1.20)	$75.06^{\#7}$ (±1.64)	$77.21^{\#5}$ (±1.49)	$73.92^{\#7}$ (±2.01)	$0.87^{\#3}$ (±0.02)	$7.35^{\#3}$ (±0.03)

two GPT-2 models and ranking LangFlow last. Qualitative inspection reveals more frequent topic shifts, repetition, and disrupted sentence transitions in diffusion and flow outputs (Appendix L); a blind human evaluation corroborates the ordering (Appendix M).

The baseline metrics do not recover this grouping. MAUVE ranks SEDD and MDLM closest to human text, above both GPT-2 models, and assigns ELF-L a near-zero score even though CHORD places it next to LangFlow. Generative perplexity favors predictable text rather than closeness to the human reference: both GPT-2 models score well below human text, LangFlow and ELF-L fall close to it, and only SEDD and MDLM are clearly separated. Unigram entropy varies little across generators and tracks lexical diversity, ranking LangFlow above human text and both GPT-2 models last, the opposite of their placement under CHORD.

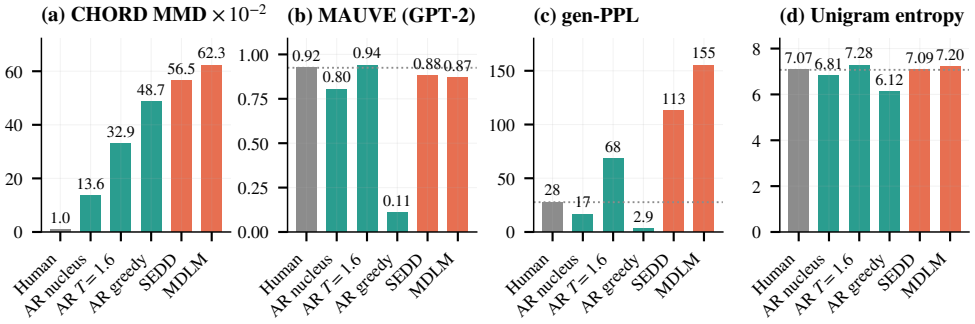

Figure 5: **Prefix continuation on OpenWebText.** Each model generates a 128-token continuation from a 128-token human prefix. Dotted lines mark the held-out human value.

Prefix continuation. In this task, each model receives the same 128-token human prefix from OpenWebText and generates a 128-token continuation. Figure 5 reports CHORD scores from the Qwen3.5-27B encoder alongside the comparison metrics. CHORD recovers the same three-tier pattern: human continuations are closest to the reference, autoregressive continuations occupy the middle tier, and diffusion-LM continuations receive the largest shifts. Within the autoregressive tier, less natural decoding strategies (greedy and high temperature) move progressively farther from the human baseline.

The comparison metrics again produce different orderings. MAUVE places high-temperature autoregressive continuations above the human baseline and keeps diffusion continuations close to human text. Generative perplexity favors greedy decoding, while unigram entropy mainly captures the loss of diversity under greedy generation.

4.5 Agreement with Human Judgments

To independently validate the metric rankings, we use the human judgments released by Pillutla et al. (2021). Following the original protocol, we fit Bradley–Terry scores (Bradley & Terry, 1952)

Table 3: **Agreement with human system rankings.** Spearman correlation (ρ) between each metric and Bradley–Terry rankings derived from the human study of Pillutla et al. (2021). The study contains 3,240 pairwise judgments over eight GPT-2 generation settings.

Human dimension	CHORD (ours)		Existing metrics				
	Qwen3.5 27B	Qwen3.5-2B distilled	MAUVE (GPT-2)	MAUVE (ELECTRA)	FBD (BERT)	MMD (MiniLM)	gen-PPL (GPT-2)
Interesting	0.86	0.76	0.00	0.79	0.76	**0.90**	0.64
Makes sense	**0.98**	0.93	−0.19	0.93	0.93	0.90	0.88
Human-like	**0.98**	0.95	−0.21	0.90	0.95	0.93	0.88

for *interesting*, *makes sense*, and *human-like*, then compute their Spearman correlations with metric rankings (Table 3). Settings vary GPT-2 size and decoding (Appendix R).

CHORD (Qwen3.5-27B) achieves the highest correlation on *makes sense* and *human-like* ($\rho = 0.98$ on both), the two dimensions most closely tied to coherence. The distilled Qwen3.5-2B encoder reaches $\rho = 0.93$ and 0.95, matching the best existing metric on each dimension. On *interesting*, which mixes coherence with novelty and subjective preference, MMD-MiniLM is slightly stronger (0.90 vs. 0.86). MAUVE with GPT-2 features correlates poorly under this protocol, consistent with its window-length sensitivity (Appendix J).

5 Discussion

Coherence failures differ in detection difficulty. Our scaling experiments suggest a hierarchy: smaller backbones detect structural errors, discourse sensitivity grows with scale, and relation-level failures remain the hardest to detect. Contradictions and causal reversals preserve much of the vocabulary and local fluency while changing logical relations. **They therefore provide demanding tests of whether generation evaluators capture global coherence.** A further direction is to explore CHORD as a corpus-level training objective, building on work that optimizes generators with representation-space metrics (Yang et al., 2026a).

Targeted representations are central to evaluation. Our ablations show that the representation largely determines which failures become detectable. Coherence prompting improves selectivity on Qwen3.5-27B but reduces it on GPT-2, suggesting that its effectiveness depends on backbone capability. This motivates designing distributional evaluators around representations that emphasize the intended quality dimension. A preliminary experiment beyond coherence supports this direction (Appendix S). When a small fraction of the safe assistant responses in a corpus is replaced with unsafe ones, CHORD detects the shift at 5% prevalence, and swapping the coherence prompt for a safety-oriented prompt widens the margin over matched safe replacements, without safety labels.

Selectivity is a useful criterion for metric validation. Our counterfactual experiments show that a metric can respond strongly to both coherence damage and benign rewriting. Comparing these responses on the same source texts helps distinguish sensitivity to coherence from sensitivity to textual change in general. This validation principle may extend to other quality dimensions by pairing targeted failures with controls that preserve the property being evaluated.

6 Conclusion

Evaluating open-ended generation remains difficult when generic feature spaces fail to encode important quality distinctions. This work identifies representation design as a central bottleneck in distributional text evaluation and develops a counterfactual protocol for testing whether a metric responds selectively to quality degradation rather than benign variation. We introduce CHORD, a coherence-sensitive distributional metric that distinguishes relation, discourse, structural, and mixture failures from matched meaning-preserving rewrites when existing metrics miss or conflate them. CHORD produces informative rankings on real generation systems that align with human evaluations. More broadly, our results show that distributional metrics can capture additional dimensions of generation quality when their representations are designed to make those dimensions visible.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript, identifying relevant literature, and generating the synthetic datasets described in the paper. We did not use generative AI tools to formulate mathematical claims, design the methodology or experiments, implement the proposed method, or interpret experimental results. We reviewed all AI-assisted content and take responsibility for the final manuscript, including all claims, references, datasets, and experimental results.

ETHICS STATEMENT

This work studies corpus-level evaluation of textual coherence. Our experiments include synthetic coherence errors and safety-related examples that may contain harmful or offensive content; these are used solely to evaluate metric behavior. Coherence does not imply factual accuracy, fairness, or safety, and a low distributional discrepancy does not establish that individual outputs are reliable or suitable for deployment. The metric may also reflect biases in the underlying language model and reference corpus. Its scores should therefore be interpreted alongside task-specific evaluation and human judgment.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the representation extraction, distributional comparison, and evaluation protocol. Appendix B gives the implementation and calibration details, Appendix C the construction of the counterfactual evaluation set, Appendix D the baseline configurations, and Appendix I the distillation procedure; the remaining appendices cover the evaluation prompts and additional experiments.

The supplementary material is a zip archive with two folders. `CHORD` implements the metric: featurization with the 27B encoder or a distilled student, RBF-MMD scoring with null calibration, and the complete distillation pipeline, from training-data construction to student training. `CHORD_experiments` provides the code, configurations, and instructions needed to reproduce every table and figure in the main paper and appendices, including the construction of the counterfactual evaluation set. Because of the size limit on supplementary material, we omit the distilled student checkpoints; they can be retrained with the distillation pipeline in `CHORD`.

REFERENCES

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. doi: 10.1162/coli.2008.34.1.1. URL https://aclanthology.org/J08-1001/.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, John X. Morris, and Sarath Chandar. Neobert: A next-generation bert, 2025. URL https://arxiv.org/abs/2502.19587.

David M. Chan, Yiming Ni, David Ross, Sudheendra Vijayanarasimhan, Austin Myers, and John Canny. Distribution aware metrics for conditional natural language generation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 5064–5095, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.453/.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. Langflow: Continuous diffusion rivals discrete in language modeling, 2026. URL https://arxiv.org/abs/2604.11748.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL https://arxiv.org/abs/2003.10555.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021. URL https://arxiv.org/abs/2007.12626.

Wikimedia Foundation. Wikimedia downloads. URL https://dumps.wikimedia.org.

Antonio Franca and Alexander Tong. Hacking generative perplexity: Why unconditional text evaluation needs distributional metrics, 2026. URL https://arxiv.org/abs/2606.08417.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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

Jian Guan and Minlie Huang. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9157–9166, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.736. URL https://aclanthology.org/2020.emnlp-main.736/.

Hongcan Guo, Qinyu Zhao, Yian Zhao, Shen Nie, Rui Zhu, Qiushan Guo, Feng Wang, Tao Yang, Hengshuang Zhao, Guoqiang Wei, and Yan Zeng. Continuous latent diffusion language model, 2026. URL https://arxiv.org/abs/2605.06548.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows, 2026. URL https://arxiv.org/abs/2605.10938.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference, 2025. URL https://arxiv.org/abs/2406.15513.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3182–3196, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.181. URL https://aclanthology.org/2024.findings-emnlp.181/.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation, 2022. URL https://arxiv.org/abs/2204.00862.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. Meta-task prompting elicits embeddings from large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10141–10157, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.546. URL https://aclanthology.org/2024.acl-long.546/.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. doi: 10.18653/v1/2020.acl-main.703.

Jiwei Li and Dan Jurafsky. Neural net models of open-domain discourse coherence. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 198–209, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1019. URL https://aclanthology.org/D17-1019/.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL https://arxiv.org/abs/2303.16634.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. Sentence ordering and coherence modeling using recurrent neural networks. 2017. URL https://arxiv.org/abs/1611.02654.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

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

Mistral AI. Ministral-8B-Instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410, 2024.

Mistral AI. Mistral small 3. https://mistral.ai/news/mistral-small-3, 2025.

OpenAI. GPT-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Zahra Ashktorab, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite guardian, 2024. URL https://arxiv.org/abs/2412.07724.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021. URL https://arxiv.org/abs/2102.01454.

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

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Generative frontiers: Why evaluation matters for diffusion language models, 2026. URL https://arxiv.org/abs/2604.02718.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. 2022. URL https://arxiv.org/abs/2009.01325.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013. URL https://api.semanticscholar.org/CorpusID:123065789.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Petar Veličković, Federico Barbero, Christos Perivolaropoulos, Simon Osindero, and Razvan Pascanu. Perplexity cannot always tell right from wrong, 2026. URL https://arxiv.org/abs/2601.22950.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://aclanthology.org/W17-4508/.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models, 2024. URL https://arxiv.org/abs/2401.00368.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation, 2026a. URL https://arxiv.org/abs/2604.28190.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. Oprd: On-policy representation distillation, 2026b. URL https://arxiv.org/abs/2606.06021.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation, 2021. URL https://arxiv.org/abs/2106.11520.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024. URL https://arxiv.org/abs/2401.02385.

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

Wei Zhao, Michael Strube, and Steffen Eger. Discoscore: Evaluating text generation with bert and discourse coherence, 2023. URL https://arxiv.org/abs/2201.11176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.131. URL https://aclanthology.org/2022.emnlp-main.131/.

Dawei Zhu, Wenhao Wu, Yifan Song, Fangwei Zhu, Ziqiang Cao, and Sujian Li. Couda: Coherence evaluation via unified data augmentation, 2024. URL https://arxiv.org/abs/2404.00681.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models, 2018. URL https://arxiv.org/abs/1802.01886.

A LIMITATIONS

CHORD is a coherence-sensitive reference distance, not a universal measure of text quality. A high score indicates a distributional shift in the selected hidden-state space; it is not an absolute judgment of writing quality, factuality, usefulness, or human preference. Scores are also relative to the human reference corpus and scoring protocol, so meaningful comparisons should keep the reference domain, passage format, encoder, prompt, and representation layer fixed.

Our primary evidence concerns global coherence. More fine-grained relation errors remain more difficult. Extending the representation-centered approach to faithfulness, controllability, style, or other properties requires property-specific prompts, controls, and validation. The source-conditioned QA study in Appendix Q demonstrates feasibility for one such setting but does not establish a general-purpose faithfulness metric.

The counterfactual evaluation is intentionally aligned with the target property and therefore should be viewed as a diagnostic test of selective coherence sensitivity instead of a complete proxy for real-world generation quality. Moreover, relation-level errors such as contradiction and causal reversal necessarily alter propositional content, making it impossible to construct benign controls that exactly match every dimension of semantic change.

The default Qwen3.5-27B encoder is computationally demanding. The distilled Qwen3.5-2B variant substantially reduces cost and preserves the broad detection pattern, but loses sensitivity on some difficult relation-level errors and may generalize less reliably to unseen domains. More broadly, the representation depends on the language, instruction-following behavior, and pretraining biases of the backbone. Evaluating additional model families and multilingual settings remains future work.

Finally, our counterfactual evaluation set is not exhaustive. Real generations may exhibit long-range planning failures, pragmatic inconsistencies, domain-specific discourse conventions, or multi-document conflicts that are not covered by the current perturbations. Although our real-system experiments are supported by both targeted blind human evaluation and correlations with released human judgments, broader validation across stronger model families, domains, and generation settings would further establish the generality of the metric.

B IMPLEMENTATION DETAILS

We list here the configuration details needed to reproduce the experiments; the metric and evaluation protocol are defined in Section 3, and compute and memory costs are reported in Appendix P.

Encoder configuration. The default CHORD encoder is a frozen Qwen3.5-27B model (Qwen Team, 2026). Each text sample is placed in the coherence-eliciting template from Section 3.1 and processed with a single forward pass. The representation is the hidden state at the last token position of the prompt, extracted from the third-to-last transformer layer, which for Qwen3.5-27B is layer 62 of 64 and yields a 5120-dimensional vector. Section 4.3.2 examines how detection of coherence failures varies across Qwen3.5 encoders from 0.8B to 27B, and Appendix O reports a layer-depth sweep on the 9B and 27B backbones showing that the third-to-last layer is among the strongest read depths for detecting coherence damage.

For the cross-family rows of Table 1, we change only the frozen backbone. Gemma-2, Mistral, and Llama-3.1 use the same coherence prompt, 512-token budget, third-to-last-layer rule, RBF-MMD, null calibration, and bootstrap criterion as Qwen3.5; the prompt and layer are not tuned separately by different backbone families. Inputs are encoded as raw completion strings. Gemma-2 uses eager attention to preserve its native attention soft-capping during hidden-state extraction.

RBF-MMD estimator. For the embedding sets $H_G = \{h(g_j)\}_{j=1}^{N_G}$ and $H_R = \{h(r_i)\}_{i=1}^{N_R}$, we use the biased squared MMD estimator

$$\widehat{\mathrm{MMD}}_k^2(H_G, H_R) = \frac{1}{N_G^2}\sum_{j,j'=1}^{N_G} k\big(h(g_j), h(g_{j'})\big) + \frac{1}{N_R^2}\sum_{i,i'=1}^{N_R} k\big(h(r_i), h(r_{i'})\big) - \frac{2}{N_G N_R}\sum_{j=1}^{N_G}\sum_{i=1}^{N_R} k\big(h(g_j), h(r_i)\big), \tag{5}$$

where $k(\mathbf{a}, \mathbf{b}) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2/(2\sigma^2))$. The within-corpus sums include self-comparisons; the unbiased estimator excludes these terms and adjusts the denominators. We use the biased estimator for its non-negativity and lower variance at the corpus sizes considered here. Its null offset is accounted for by the reference standardization below.

Bandwidth calibration. For each representation and each evaluation set, we fit the RBF bandwidth once on a calibration split R_{cal} of that set's human reference, disjoint from every evaluated corpus, and keep it fixed for all comparisons on that set. We set σ using the median heuristic (Gretton et al., 2012):

$$\sigma = \mathrm{median}\big\{\|h(r) - h(r')\| \;:\; r \neq r',\; r, r' \in R_{\mathrm{cal}}\big\}. \tag{6}$$

The calibration split has 1,500 windows for the counterfactual evaluation set (Table 1) and 150 windows for the unconditional-generation reference (Table 2). Since embedding scales differ across encoders, each backbone is calibrated separately. On the counterfactual set and the generation reference respectively, this gives $\sigma = 105.01$ and 105.38 for Qwen3.5-27B, 77.49 and 72.51 for the distilled Qwen3.5-2B, and 77.16 and 72.61 for the distilled Qwen3.5-0.8B; Qwen3.5-9B gives 58.93 on the counterfactual set.

Null standardization. We apply the same reference-based standardization to CHORD and every baseline. For distributional metrics, s_M is the distance between the evaluated and reference corpora. For corpus-level scalar statistics such as perplexity and entropy, we use the absolute difference in the corpus statistic, $s_M = |\bar{v}_G - \bar{v}_R|$, with each statistic computed as specified in Appendix D. For MAUVE, we standardize the frontier integral of its divergence frontier rather than the MAUVE score, as detailed in Appendix D.

For each metric, we estimate a null distribution from $B = 200$ draws. In each draw, we sample two disjoint subsets from the human reference pool, matching their sizes to those used in the actual evaluation, and compute the same discrepancy s_M between them. The mean μ_0 and standard deviation σ_0 of these null scores define

$$z_M = \frac{s_M - \mu_0}{\sigma_0}. \tag{7}$$

This score expresses the observed discrepancy relative to the metric's reference sampling variation; values near zero are close to the null mean. It is not itself the criterion for detecting coherence damage, since benign rewriting can also produce a nonzero shift.

Paired-bootstrap detection. For each harmful perturbation condition, we compare its null-standardized score with that of the most extensively rewritten benign paraphrase condition constructed from the same seed samples. Their difference is $\Delta z = z_{\mathrm{harm}} - z_{\mathrm{benign}}$. To quantify uncertainty, we draw 200 bootstrap resamples of the seed samples with replacement. Within each draw b, we use the same resampled seed indices to form both the harmful and benign corpora and compute

$$\Delta z^{(b)} = z_{\mathrm{harm}}^{(b)} - z_{\mathrm{benign}}^{(b)}. \tag{8}$$

The empirical 2.5th and 97.5th percentiles of these differences form the 95% confidence interval. A condition is detected only when its lower bound is strictly positive, indicating a stronger response to coherence damage than to benign rewriting.

Table 4: **Representative examples from the counterfactual evaluation set.** Only the targeted sentence or operation is modified; the surrounding text is preserved. Examples are shortened for readability.

Perturbation	Before	After
Contradiction	"...became the worst kind of media free-for-all."	"...became the most respectful kind of media coverage."
Causal reversal	"You may opt out at any time."	"You may opt in at any time."
Broken transition	"His suicide suddenly made more sense."	"His suicide suddenly made more sense, but the moon is made of green cheese."
Topic drift	"...spoke strongly on behalf of the virtues of physical education."	"...spoke strongly on behalf of the virtues of chess strategy."
Sentence permutation	clean text sample	selected sentences are reordered within the sample
Word shuffle	clean text sample	words are locally shuffled within selected sentences
Repetition	clean text sample	selected sentences are duplicated in place
DLM mix	clean text sample	selected sentences are replaced by diffusion-LM sentences
Document mix	clean text sample	selected sentences are replaced by sentences from unrelated documents
Benign rewriting	"I know what I did is wrong and horrible."	"I realize my actions were incorrect and deeply wrong."

C Construction of the Counterfactual Evaluation Set

The perturbation categories and their diagnostic roles are defined in Section 3.2; here we specify how the counterfactual evaluation set is constructed for each perturbation type. Representative examples are shown in Table 4.

Data pools. The sources, length filter, and three-way split into seed, reference, and replacement pools are described in Section 4.1. The replacement pool contains 2,453 text samples. The three sources (OpenWebText, WikiText-103, and Reddit TL;DR) contribute roughly equal numbers of seed samples, and all pools are deduplicated against each other using normalized-text hashes to prevent overlap.

LLM-edited perturbations. Relation perturbations (contradiction, causal reversal), discourse perturbations (broken transition, topic drift), and the benign control are produced by the Qwen3-30B-A3B (Yang et al., 2025) editor, which is distinct from the CHORD encoder. For each seed sample, a fixed seeded sampler selects the target sentences, and the editor rewrites each independently; all non-target sentences are copied back unchanged.

Contradiction rewrites a target sentence so that it conflicts with a preserved anchor sentence in the same sample. *Causal reversal* reverses an expressed cause–effect relation while preserving the entities involved. *Broken transition* replaces a target sentence with a locally fluent sentence that is incompatible with the preceding discourse. *Topic drift* replaces a target sentence with an off-topic sentence of comparable length. *Benign paraphrase* rewrites target sentences while preserving claims, entities, numbers, stance, and discourse order.

Structural perturbations. These are rule-based transformations applied without LLM involvement. *Sentence permutation* reorders a fraction of the sentences. *Word shuffle* perturbs word order within selected sentences. *Repetition* duplicates selected sentences in place. All three preserve most of the original lexical material while disrupting organization.

Mixture perturbations. *DLM mix* replaces a fraction of the sentences with unconditional OpenWebText generations from a combined pool: three seeded runs from SEDD-small at 256 sampling steps (Lou et al., 2024) and three from MDLM-OWT at 512 sampling steps (Sahoo et al., 2024). *Document mix* replaces sentences with material from one or more unrelated human documents in the replacement pool. A human-splice control applies the same replacement pattern using other human-written text, serving as an additional reference for mixture-induced shifts.

Perturbation severity levels. LLM-edited perturbations rewrite one, two, or three target sentences. Each edit introduces exactly one relation- or discourse-level violation, so the number of rewritten sentences directly controls the number of violations per sample. The benign control follows the same one-to-three schedule, so harmful and benign conditions are matched in the amount of edited text. Because every seed sample contains at least four sentences, rewriting at most three leaves most of the original text intact. Each perturbation is therefore a targeted edit, not a complete rewrite.

Structural and mixture perturbations use type-specific fraction schedules. Sentence permutation, word shuffle, and repetition affect 10%, 25%, 50%, 75%, or 100% of eligible sentences. DLM mix and its matched human-splice control replace 10%, 25%, or 50% of sentences. Document mix replaces 10%, 25%, 50%, or 75% of sentences; an additional source-diversity sweep holds the replacement rate at 50% and draws from one, two, or four source documents. For a sample with m eligible sentences, a target fraction r selects $\min(m, \max(1, \lceil mr \rceil))$ sentences. Sentence permutation additionally requires at least two selected sentences so that the operation changes their order. Thus, for example, a four-sentence sample changes one sentence at 25% for shuffle, repetition, and mixture, but permutes two sentences. Score trajectories across severity levels are reported in Appendix H.

Sample filtering and validation. Because LLM-generated edits can fail to produce the intended perturbation, we apply a post-hoc filtering step to ensure the validity of every edited sample. We discard samples with *no effective edit*, where the editor returns the target sentence unchanged or with only trivial variation; *severe truncation*, where the output is cut off or substantially shorter than the input; and *missing required context*, where the seed text lacks the structure a perturbation presupposes, such as an explicit cause–effect relation for causal reversal. For contradiction, the anchor sentence must be retained in the final sample, since a contradiction is observable only when the original claim and the conflicting rewrite coexist; samples in which the anchor is edited are discarded.

We then validate the set at the condition level. Sample counts are checked to be balanced across conditions, so that detections are compared at comparable statistical power. The realized edit rate, measured as the fraction of sentences that differ from the seed text, is checked to increase monotonically with severity within each perturbation type, confirming that the nominal sentence schedule produces graded textual change. The same check applies to the benign paraphrase conditions, whose edit rates match those of the harmful conditions at each level.

The final evaluation set contains 43 conditions: 12 LLM-edited conditions (four types at three severities), 28 structural and mixture conditions, and three benign paraphrase severities; the most extensively rewritten one serves as the benign control in the detection criterion (Section 3.2).

Robustness to the editor model. The default editor (Qwen3-30B-A3B; Yang et al., 2025) and the default encoder of CHORD (Qwen3.5-27B) belong to the same model family, raising the concern that CHORD's sensitivity may partly reflect recognition of family-specific editing artifacts rather than coherence degradation itself. To rule this out, we regenerate all LLM-edited conditions with Mistral-Small-24B-Instruct-2501 (Mistral AI, 2025), a model from an entirely different family, keeping the seed samples, target sentences, editing instructions, and severity levels fixed. Structural and mixture perturbations are unchanged.

Table 5 reports the null-standardized scores of the default CHORD encoder on the relation and discourse conditions, whose test texts differ between the two editors; the structural and mixture conditions are identical across the two sets and are therefore omitted. The main finding is preserved. CHORD detects 11 of the 12 LLM-edited conditions on the Mistral-edited set, compared with 10 of 12 under the default editor. Contradiction is detected at all severity levels under both editors. Causal

Table 5: **Robustness to the editor model.** Null-standardized CHORD scores z_M at the highest severity of each LLM-edited perturbation type. Shaded entries are detected relative to the benign control under the paired bootstrap criterion (Section 3.2). The final column reports detections across the 12 LLM-edited conditions. Scores are calibrated separately for each set and are not comparable across rows.

Editor model	Relation		Discourse			
	Contra-diction	Causal reversal	Broken transition	Topic drift	Benign paraphrase	Detected (of 12)
Qwen3-30B-A3B (default)	84.3	25.6	331	79.7	6.0	10
Mistral-Small-24B	151	90.5	559	406	9.5	11

reversal is the hardest perturbation type in the main experiments and remains the least consistently detected under both editors. All four perturbation types show increasing responses with severity, and CHORD does not confuse benign paraphrases with coherence degradation at any severity level. CHORD's selective sensitivity to coherence degradation is therefore not an artifact of the editor–encoder family overlap.

D Baseline Metric Implementations

The following baselines are compared with CHORD on the counterfactual evaluation set (Table 1) and on the outputs of existing generation models (Section 4.4). All baselines are computed on the same corpora as CHORD and placed on the same scale via the null standardization described in Appendix B.

Evaluation text lengths. All metrics operate on text truncated or packed to the same fixed length within each experiment: 512 tokens for the counterfactual evaluation set and unconditional generation, 256 tokens for prefix continuation (128-token human prefix + 128-token model continuation), and 256 tokens for the human-judgment study of Appendix R, matching the texts shown to annotators. Truncation is applied before feature extraction, so every metric sees identical text for a given sample. Fewer than 1.3% of counterfactual samples exceed 512 tokens. For unconditional generation, all corpora are packed to exactly 512 tokens, except ELF-L, the large variant of ELF (Hu et al., 2026), whose native outputs are about 930 tokens; it is evaluated on its first 512 tokens to match the other systems.

Generative perplexity. Following Holtzman et al. (2020), we use GPT-2-large (Radford et al., 2019) as an external autoregressive scorer. Each text sample x is tokenized into $(x_1, \ldots, x_T)$ and truncated to the evaluation length above. The passage-level negative log-likelihood is

$$\mathrm{NLL}(x) = -\sum_{t=2}^{T} \log p_\theta\big(x_t \mid x_{<t}\big), \tag{9}$$

where $T(x) = T - 1$ is the number of predicted tokens, since the first token has no conditioning context. For the generator evaluations of Section 4.4, we report the corpus-level perplexity

$$\mathrm{PPL}(C) = \exp\left(\frac{\sum_{x \in C} \mathrm{NLL}(x)}{\sum_{x \in C} T(x)}\right), \tag{10}$$

which pools tokens across all samples and weights each token equally regardless of sample length. On the counterfactual evaluation set, where all metrics pass through the same null standardization, we instead use the mean per-sample log-perplexity $\frac{1}{|C|}\sum_{x \in C} \mathrm{NLL}(x)/T(x)$, which weights each sample equally and matches the per-corpus form of the other scalar baselines.

Unigram entropy. We lowercase the entire corpus C and split it into words by whitespace, count all occurrences $n_C(w)$ of each word type w, and estimate the empirical unigram distribution $p_C(w) = n_C(w)/N_C$, where $N_C = \sum_w n_C(w)$ is the total token count. The corpus-level unigram entropy (Shannon, 1948) is

$$H(C) = -\sum_{w} p_C(w) \log p_C(w). \tag{11}$$

This statistic pools all tokens in the corpus into a single distribution and discards word order entirely, so it reflects only corpus-level lexical diversity.

MAUVE. MAUVE (Pillutla et al., 2021; 2023) measures the gap between the generated and reference distributions in a discretized feature space. Each text sample is mapped to a feature vector, and the vectors from both corpora are jointly clustered into K groups via k-means. Each corpus is then represented by its cluster-assignment histogram: P for the generated corpus and Q for the reference. For a mixture $M_\lambda = \lambda P + (1-\lambda)Q$, the divergence curve is formed by exponentiating the two KL divergences separately:

$$\mathcal{C}(P,Q) = \Big\{ \big(e^{-c\,\mathrm{KL}(P\|M_\lambda)},\, e^{-c\,\mathrm{KL}(Q\|M_\lambda)}\big) : \lambda \in (0,1) \Big\}, \qquad \mathrm{MAUVE}(G,R) = \mathrm{AUC}\big(\mathcal{C}(P,Q)\big). \tag{12}$$

Here $c > 0$ is a fixed scaling constant, and AUC denotes the area under the curve, completed with endpoints $(0,1)$ and $(1,0)$. The score lies in $[0,1]$; values near 1 indicate a close match. The exponential is applied to the curve coordinates before integration, not to a scalar area afterwards.

What we report. The MAUVE score saturates near zero for large gaps, so null-standardizing it would obscure meaningful variation. The MAUVE rows of Table 1, and MAUVE's k-means KL in the representation–distance analysis of Section 4.3, therefore report the frontier integral (Pillutla et al., 2023) returned by the official implementation: a KL-based summary of the same divergence frontier that is zero when $P = Q$, grows with the gap, and lies in $[0,1]$, reaching 1 only when the two histograms have disjoint support. It is computed directly from the two histograms rather than from the exponentiated curve, and we apply the reference-null standardization of Appendix B to it. The generator evaluations of Section 4.4, the human-agreement study of Section 4.5, and Appendix J report the standard MAUVE score. In all cases we use the official implementation with $K = n/10$ clusters, $c = 5$, and 25 values of λ; the MAUVE scores are averaged over five k-means seeds, whereas the frontier integral uses one k-means fit per bootstrap draw. In the representation–distance analysis, the Fréchet distance and k-means KL are computed in a 128-dimensional PCA subspace fitted on the human-reference features, because a full-dimensional Gaussian fit is rank-deficient at these corpus sizes; RBF-MMD and energy distance use the raw features.

We evaluate two feature spaces: GPT-2-large last-token hidden states, which is the default configuration of Pillutla et al. (2021), and ELECTRA (Clark et al., 2020) masked-mean pooled features. Appendix J examines sensitivity to truncation and packed-window length.

FBD. Fréchet-BERT Distance (Xiang et al., 2021) represents each text sample as a feature vector using BERT (Devlin et al., 2019), then fits a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C)$ to each corpus C by computing the feature mean $\boldsymbol{\mu}_C$ and unbiased covariance $\boldsymbol{\Sigma}_C$. FBD is the squared 2-Wasserstein distance between the two fitted Gaussians:

$$\mathrm{FBD}(G,R) = \underbrace{\|\boldsymbol{\mu}_G - \boldsymbol{\mu}_R\|^2}_{\text{mean shift}} + \underbrace{\mathrm{tr}\Big(\boldsymbol{\Sigma}_G + \boldsymbol{\Sigma}_R - 2\big(\boldsymbol{\Sigma}_G^{1/2}\,\boldsymbol{\Sigma}_R\,\boldsymbol{\Sigma}_G^{1/2}\big)^{1/2}\Big)}_{\text{covariance mismatch}}. \tag{13}$$

Because the Gaussian assumption captures only the mean and covariance of each corpus, any distributional difference that leaves these two statistics approximately unchanged is invisible to FBD.

MMD-MiniLM. Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) is a kernel-based two-sample distance that compares distributions without assuming a parametric form for either. Following Chan et al. (2024), we instantiate MMD with `all-MiniLM-L6-v2` sentence embeddings (Wang et al., 2020; Reimers & Gurevych, 2019) $\phi(\cdot)$ and the same biased RBF-MMD estimator used by CHORD (Section 3.1). MiniLM is a general-purpose sentence encoder not designed for coherence, so comparing MMD-MiniLM with CHORD isolates the effect of the representation while holding the distance fixed.

The two metrics also differ in bandwidth selection. MMD-MiniLM recalibrates σ for each comparison by taking the median pairwise distance over the joint embedding set $\Phi_G \cup \Phi_R$, where $\Phi_G = \{\phi(g_j)\}_{j=1}^{N_G}$ and $\Phi_R = \{\phi(r_i)\}_{i=1}^{N_R}$:

$$\sigma = \mathrm{median}\big\{\|\phi(g) - \phi(r)\| :\ \phi(g), \phi(r) \in \Phi_G \cup \Phi_R,\ \phi(g) \neq \phi(r)\big\}. \tag{14}$$

Table 6: **Additional baselines on the counterfactual evaluation set.** Entries are null-standardized responses at the highest perturbation severity; blue shading indicates detection relative to benign rewriting, as in Table 1. Embedding methods use RBF-MMD, with the input instruction shown in parentheses. Scalar evaluators use the absolute candidate–reference difference in mean score.

Method	Causal reversal	Contra-diction	Broken transition	Topic drift	Sentence permutation	Word shuffle	Repetition	DLM mix	Document mix	Benign rewriting
Instruction-tuned embeddings + RBF-MMD										
Qwen3-Embedding-8B (none)	12.1	11.5	19.3	17.3	14.5	12.0	15.1	129	142	7.6
Qwen3-Embedding-8B (similarity)	9.3	13.0	22.5	14.8	10.7	8.4	19.0	76.2	73.5	5.0
Qwen3-Embedding-8B (coherence)	10.6	11.5	51.4	19.7	13.2	12.0	19.9	157	182	7.4
e5-mistral-7b-instruct (none)	8.3	7.7	30.3	12.8	13.2	22.0	102	87.6	92.1	5.6
e5-mistral-7b-instruct (similarity)	8.3	10.9	44.4	13.5	12.8	20.7	69.6	92.1	92.6	5.9
e5-mistral-7b-instruct (coherence)	6.3	7.7	46.5	10.3	10.5	12.6	75.1	105	98.5	4.8
gte-Qwen2-7B-instruct (none)	7.4	7.1	21.6	11.2	11.1	12.8	13.7	101	99.2	6.2
gte-Qwen2-7B-instruct (similarity)	9.7	10.9	36.0	15.6	11.3	11.0	21.7	109	106	5.7
gte-Qwen2-7B-instruct (coherence)	10.4	11.3	65.1	20.8	23.8	21.1	74.0	185	228	7.2
Scalar evaluators (source-free)										
UniEval (fluency)	-1.1	-0.8	-0.3	0.3	2.4	99.4	1.4	22.1	-1.0	3.5
UniEval (coherence, zero-shot)	-0.2	1.8	4.3	11.2	6.5	41.1	23.2	47.9	21.1	1.1
UniEval (fluency + coherence)	0.2	0.6	2.6	8.1	9.3	105	16.2	53.3	16.1	3.9
BARTScore (hypothesis-only)	-0.4	-0.4	-0.3	-0.4	3.0	34.1	30.3	29.5	20.3	-0.6
CHORD (Qwen3.5-27B)	25.6	84.3	331	79.7	358	740	857	1150	1082	6.0

Because σ changes with every pair of corpora, scores from different comparisons are not on the same kernel scale. CHORD avoids this issue by calibrating σ once from a held-out reference split and fixing it across all evaluations (Appendix B). Section 4.3 controls representation and distance independently in the factorial ablation.

E Modern Embedding and Evaluator Baselines

We compare CHORD with three instruction-tuned embedding models and two established text evaluators on the counterfactual evaluation set. All methods use the same reference splits, null calibration, and detection criterion as Table 1.

Instruction-tuned embeddings. We evaluate Qwen3-Embedding-8B (Zhang et al., 2025), e5-mistral-7b-instruct (Wang et al., 2024), and gte-Qwen2-7B-instruct (Li et al., 2023). For each model, we extract the final-layer, last-token embedding, apply L2 normalization, and use the same RBF-MMD protocol as CHORD, with bandwidth fixed on the reference development split. We compare three input settings: no instruction, a similarity instruction ("Retrieve semantically similar text."), and a coherence instruction ("Represent the passage in terms of its logical coherence and the order of its ideas."). We use each model's native instruction template and preserve the final representation-eliciting token when truncating inputs to 512 tokens.

UniEval and BARTScore. Because our evaluation set contains individual passages without paired sources or references, we adapt both evaluators to score text alone. For UniEval (Zhong et al., 2022), we compute fluency using the released sentence-level scoring procedure with `unieval-sum`. We evaluate coherence with `unieval-intermediate`, using the zero-shot question "Is this a coherent and logically consistent passage?" Both dimensions use $P(\text{Yes})/[P(\text{Yes}) + P(\text{No})]$, and we also report their average. For BARTScore (Yuan et al., 2021), we compute length-normalized log-likelihood under `bart-large-cnn` (Lewis et al., 2020) with an empty source. For each scalar evaluator, the corpus-level statistic is the absolute difference between the candidate and reference mean scores, standardized against its clean-reference null. These source-free adaptations differ from the evaluators' original source-conditioned settings.

Results. Table 6 shows that modern embedding models detect both discourse perturbations and most structural and mixture perturbations at the highest severity. Relation-level errors remain harder: none detects causal reversal, and contradiction is detected only under some instruction settings. The coherence instruction increases the response to broken transitions for all three models, but does not consistently improve relation-level detection over the similarity instruction. UniEval and hypothesis-only BARTScore also miss both relation-level perturbations, responding mainly to structural and mixture errors. In contrast, CHORD detects all nine perturbation types. These results extend the comparison beyond older encoders, but do not isolate the effect of coherence prompting: the em-

bedding models also differ from CHORD in size and training. Section 4.3.2 examines capacity and prompting through controlled ablations.

F Hidden-State Signals versus Token-Space Judgments

CHORD compares distributions of hidden representations. We examine how much of this coherence signal is captured by token counts and next-token predictions, then compare with direct LLM judging and G-Eval (Liu et al., 2023) using the same frozen Qwen3.5-27B backbone.

F.1 Token Counts and Next-Token Predictions

Sentence shuffling changes discourse order without changing each document's token counts, leaving an exact unigram representation unchanged. Figure 6 compares this representation with the coherence-prompted hidden representation under the same RBF-MMD protocol. Unigram selectivity remains near zero, while hidden-state selectivity increases with the shuffle rate. Token counts thus miss order-dependent coherence failures that the hidden representation captures.

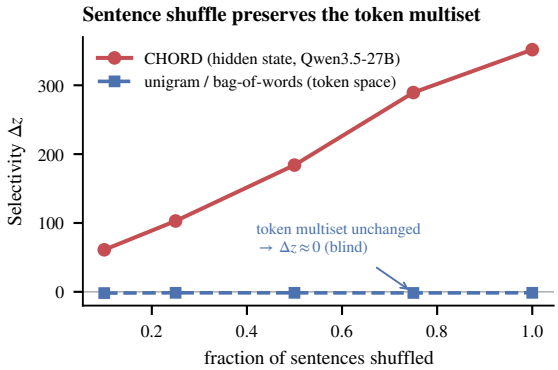

Figure 6: **Hidden representations versus token counts.** Under sentence shuffling, unigram selectivity stays near zero, while hidden-state selectivity increases with the shuffle rate. Both methods use the same RBF-MMD, human-reference null, and benign-paraphrase control.

We next examine the model's next-token predictions at the prompt position where CHORD reads the hidden state, using the same forward pass. Figure 7 shows how the decoded outputs change relative to benign paraphrases. Coherence errors shift the outputs toward the negative description "incoherent," indicating that the predictions contain some coherence information. However, the outputs do not consistently express a clear judgment: the added "incoherent" continues as "incoherent or coherent," and most of the remaining shift is toward continuations of the prompt, such as "what is it called when." The hidden representation at this position provides stronger corpus-level separation.

These probes examine token counts and predictions under the representation prompt. We next evaluate explicit coherence ratings using prompts designed for judging.

F.2 Same-Backbone Comparison with LLM Judges

The judges evaluated here assign a coherence score to each passage, whereas CHORD measures the discrepancy between two corpora through distributions of hidden representations. CHORD is therefore *coherence-aware* instead of *coherence-only*: its representations emphasize coherence but can retain other textual properties. Therefore, the aim of CHORD is very different from LLM-as-judge.

Nevertheless, we conduct an experiment comparing CHORD with a direct coherence judge (Zheng et al., 2023) and G-Eval, aggregating their per-passage ratings for corpus-level evaluation. All three use the same frozen Qwen3.5-27B backbone. This comparison tests how well each approach captures corpus-level differences, with particular attention to coherence errors.

Direct coherence judge. The direct judge assigns each passage an integer score from 0 to 10 using the prompt in Table 7. The prompt separates the passage from the instructions with `<document>`

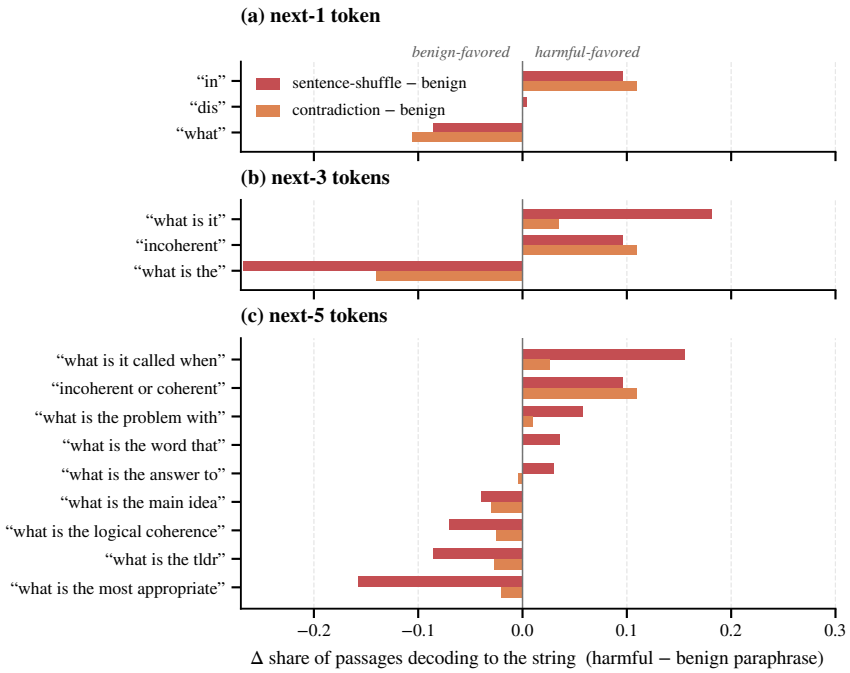

Figure 7: **Next-token predictions at the CHORD representation-eliciting position.** Bars show changes in the fraction of passages producing each decoded string, relative to benign paraphrases. Coherence errors increase negative descriptions, but the outputs do not consistently express a clear coherence judgment.

tags and asks the model to ignore topic, genre, and truncation at the end of the passage. It requests the score before a brief explanation, so the explanation does not precede and influence the rating. We run the judge with vLLM (Kwon et al., 2023) at temperature 0 with a fixed seed.

Table 7: **Direct coherence judge prompt.** The same prompt is used for all passages, without a system prompt or few-shot examples.

You are a strict, careful writing-quality rater. Read the document between the `<document>` markers and rate its logical consistency, discourse flow, and overall quality on an integer scale from 0 (incoherent, broken, or self-contradictory) to 10 (flawless, fully coherent writing). Judge only the writing itself; do not reward or penalize the topic, opinions, or genre, and ignore truncation at the very end of the document.
`<document>`
`{text}`
`</document>`
Return only JSON, the score first: `{"score": <integer 0-10>, "reason": "<one short sentence>"}`

G-Eval. We adapt the released SummEval (Fabbri et al., 2021) coherence rubric from G-Eval (Liu et al., 2023) to standalone passages. The evaluation steps are generated once and then fixed for all passages. Each passage receives a probability-weighted score over ratings 1–5, computed from the rating-token log probabilities. We use Qwen3.5-27B to evaluate the G-Eval procedure under the same backbone as CHORD; this is an adaptation of G-Eval, whose original evaluator used GPT-4 (OpenAI, 2023).

Corpus-level comparison. For each judge, we compute the absolute difference between the candidate and reference mean scores, $|\bar{r}_G - \bar{r}_R|$. We then apply the same null standardization and benign-contrast detection criterion as in Section 3.2. Both judges share CHORD's sample sizes, reference splits, and resampling procedure.

Counterfactual evaluation. Table 8 shows that both judges detect all nine perturbation types at the highest severity. Explicit ratings therefore capture coherence information when the model is

prompted to evaluate it. However, CHORD produces much larger null-standardized responses across all nine types. For contradiction, the direct judge and G-Eval yield $z_M = 16.2$ and 9.9, compared with 84.3 for CHORD. With the backbone held fixed, the hidden-state comparison thus provides greater separation relative to its clean-reference variation.

Table 8: **Same-backbone comparison with LLM judges.** All methods use frozen Qwen3.5-27B. Entries are null-standardized responses at the highest perturbation severity; shaded entries indicate detection relative to benign rewriting. For each judge, the statistic is the absolute difference between candidate and reference mean ratings.

	Relation		Discourse		Structural			Mixture		Control
Method	Contra-diction	Causal reversal	Broken transition	Topic drift	Sentence permutation	Word shuffle	Repetition	DLM mix	Document mix	Benign rewriting
LLM judge (same Qwen3.5-27B)	16.2	9.5	26.0	16.9	28.1	40.0	42.6	49.0	53.0	0.3
G-Eval (same Qwen3.5-27B)	9.9	5.2	33.9	18.8	41.4	49.2	62.2	75.4	77.6	0.6
CHORD (Qwen3.5-27B)	84.3	25.6	331	79.7	358	740	857	1150	1082	6.0

Table 9: **Direct coherence judging on unconditional generation.** Both methods use Qwen3.5-27B on the ten folds of Table 2; the CHORD column is the Qwen3.5-27B column of that table. The judge mean is the mean rating over all 5,000 texts of a generator. Judge z_M standardizes the absolute difference between the generator and reference fold means against a null of two disjoint 500-text reference subsets, reported as mean±std over folds. Rows are ordered by CHORD rank. Rank markers (#k) order the corpora by higher mean rating or lower CHORD score; judge z_M is not ranked. (n.s.) denotes a deviation that does not exceed the 95th percentile of the null.

Generator	Params	Judge mean (0–10) ↑	Judge z_M	CHORD ↓ (MMD^2 $\times 10^{-2}$)
Human (packed, held-out)	–	$3.96^{\#1}$	-0.1 (± 0.6) (n.s.)	$0.17^{\#1}$ (± 0.04)
GPT2-large (AR)	774M	$2.02^{\#2}$	19.9 (± 1.5)	$19.84^{\#2}$ (± 1.02)
GPT2-medium (AR)	355M	$1.67^{\#3}$	23.7 (± 1.3)	$28.37^{\#3}$ (± 0.90)
ELF-L	652M	$1.08^{\#4}$	30.1 (± 1.3)	$51.19^{\#4}$ (± 1.09)
LangFlow	171M	$0.52^{\#7}$	36.3 (± 1.3)	$52.58^{\#5}$ (± 1.12)
SEDD-small	170M	$0.85^{\#5}$	32.7 (± 1.3)	$57.05^{\#6}$ (± 1.02)
MDLM	170M	$0.83^{\#6}$	32.9 (± 1.3)	$57.97^{\#7}$ (± 1.20)

Evaluating generation systems. Table 9 compares the direct judge with CHORD on the unconditional-generation folds of Table 2. The judge rates every text of those folds, cut to its first 512 tokens like the other metrics. We report both the judge's mean rating and its null-standardized deviation from the human reference, which consists of packed 512-token sequences.

LLM-as-a-judge agrees with CHORD's broad ordering: human text, followed by GPT-2 outputs, then diffusion and flow outputs. Within the diffusion and flow group, however, the rankings diverge more markedly: CHORD ranks LangFlow ahead of SEDD and MDLM, whereas the judge ranks it last.

Takeaway. Token counts miss order-dependent errors, while next-token predictions under the representation prompt provide inconsistent coherence judgments. Dedicated judging prompts are more effective: both judges detect all nine perturbation types at the highest severity. Even with the same backbone, however, CHORD yields larger null-standardized responses. On unconditional generation, LLM-as-a-judge agrees with CHORD's broad grouping of human text, autoregressive outputs, and diffusion/flow outputs, but their rankings diverge more markedly within the diffusion/flow group.

G Representation Extraction and Prompt-Design Ablations

The analyses below complement Section 4.3 by varying the representation extraction method, the *target attribute* named in the instruction, and the *prompt format* used to express that property. We also test sensitivity to prompt wording.

Backbone and representation extraction. We test whether coherence prompting still helps when Qwen3.5 is replaced by GPT-2, the backbone family used by MAUVE. For each backbone, we compare the CHORD coherence prompt, a generic PromptEOL prompt, the unprompted last token, and unprompted mean pooling. All settings use the third-to-last layer, a 512-token budget, RBF-MMD, the same 500-seed evaluation set, and the same bootstrap detection rule. Because each encoder has its own bandwidth and null scale, mean Δz should be compared only among extraction methods for the same backbone; the condition counts are comparable across backbones.

Table 10: **How the extraction method changes coherence selectivity.** Each row summarizes 40 harmful conditions. "Detected" counts conditions whose selectivity is reliably positive; "reversed" counts conditions whose response is reliably stronger for benign rewriting than for coherence damage. Bold marks the strongest extraction method for each backbone.

Backbone	Extraction method	Mean Δz	Detected (of 40)	Reversed (of 40)
GPT-2 (0.124B)	Coherence prompt	33.0	11	18
	Generic prompt	7.8	10	20
	Raw last token	27.2	14	11
	Mean pooling	**76.2**	**24**	**0**
GPT-2-medium (0.355B)	Coherence prompt	26.2	9	15
	Generic prompt	31.6	10	14
	Raw last token	13.2	10	15
	Mean pooling	**87.3**	**24**	**0**
GPT-2-large (0.774B)	Coherence prompt	−12.5	8	19
	Generic prompt	−1.5	8	19
	Raw last token	0.8	8	17
	Mean pooling	**40.3**	**24**	**0**
GPT-2-XL (1.558B)	Coherence prompt	−80.6	2	27
	Generic prompt	−64.3	2	27
	Raw last token	−78.8	2	28
	Mean pooling	**39.6**	**18**	**0**
Qwen3.5-27B	**Coherence prompt**	**471.0**	**38**	**0**
	Generic prompt	84.7	30	0
	Raw last token	66.3	29	0
	Mean pooling	38.6	26	0

Table 10 shows that the best extraction method depends on whether the backbone can act on the prompt. Coherence prompting is strongest for Qwen3.5-27B, separating 38 of 40 harmful conditions. For every GPT-2 size, mean pooling is the strongest option, whereas final-token extraction often reverses the intended comparison and responds more to benign rewriting than to coherence damage. Increasing GPT-2 size does not resolve this problem: GPT-2-XL detects only 2 of 40 conditions with the coherence prompt, despite having more parameters than Qwen3.5-0.8B, which detects 20. Thus both the targeted prompt and a backbone that can interpret it are necessary. This experiment does not determine whether instruction tuning, training data, architecture, or another family difference supplies that capability.

Effect of the target attribute. We change only the property named in the prompt while keeping all other factors fixed. Figure 8a reports selectivity normalized within each perturbation category so that cross-category difficulty differences do not dominate the comparison.

Coherence gives the strongest selectivity for relation and discourse failures and remains at or near the top for structural and mixture perturbations. Grammar performs comparably on the latter two but is substantially weaker on relation and discourse, suggesting that it mainly captures syntax and not logical relations between sentences. Broader attributes such as quality, topic, and sentiment are much weaker across all categories. The neutral PromptEOL template and the unprompted representation are weaker still, showing that simply adding a prompt is not sufficient: the target attribute must name a property relevant to the intended distinction.

Effect of the prompt format and wording. We fix the target attribute to coherence and vary the prompt format. Figure 8b compares compact completion, scalar rating, and direct answer across

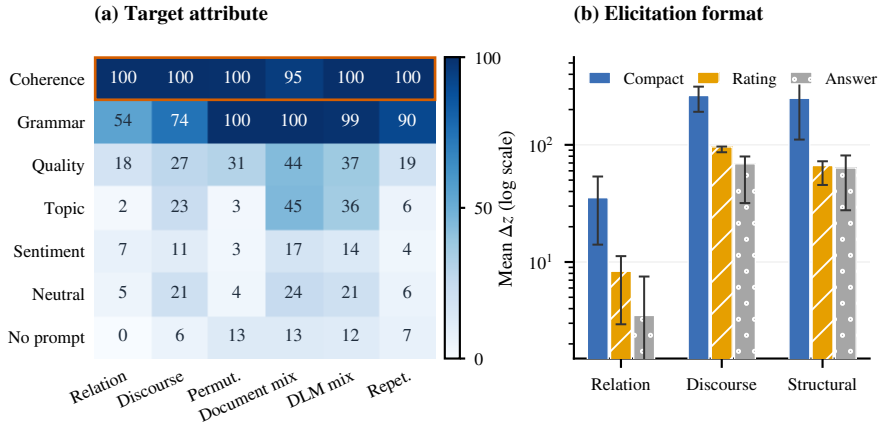

Figure 8: **Target attribute and prompt format. (a)** Selectivity for compact prompts naming different target attributes, normalized within each perturbation category. **(b)** Selectivity for three prompt formats; error bars span three prompt wordings.

three coherence wordings. Compact completion yields the strongest selectivity for relation, discourse, and structural failures, with the same advantage across all three wordings.

Wording affects response magnitude but not system ordering. We repeat the human-agreement evaluation in Appendix R for each of the three compact coherence wordings, using the study released by Pillutla et al. (2021). For every wording, we recompute CHORD scores for the same eight GPT-2 generation settings and correlate their ranking with the Bradley–Terry rankings fitted from the study's 3,240 pairwise judgments. All three wordings produce the same correlations: $\rho = 0.857$, 0.976, and 0.976 for *interesting*, *makes sense*, and *human-like*, respectively. The model ranking is therefore stable across these prompt paraphrases, even though their counterfactual response magnitudes differ. Tables 11 and 12 list the templates.

Table 11: **Target-attribute templates.** The compact-completion format is fixed; only the target attribute changes.

Attribute	Template
Coherence	This passage: "x", considering its coherence and ordering of its ideas, means in one word:
Grammar	This passage: "x", considering its grammatical correctness and sentence structure, means in one word:
Quality	This passage: "x", considering its overall writing quality, means in one word:
Topic	This passage: "x", considering its main topic and subject matter, means in one word:
Sentiment	This passage: "x", considering its overall sentiment and emotional tone, means in one word:
Neutral	This sentence: "x" means in one word: (PromptEOL template; Jiang et al., 2024)
No prompt	Bare passage; final-token or mean-pooled hidden states.

H Score Behavior across Corpus Size and Perturbation Severity

Effect of corpus size. Figure 9 compares how quickly different distances separate harmful conditions from the null as corpus size N grows. Kernel distances (RBF-MMD, energy distance (Székely & Rizzo, 2013)) reach reliable separation at smaller N than Fréchet distance (Heusel et al., 2018) or k-means KL, providing additional justification for adopting RBF-MMD as the default distance. Crucially, the benign paraphrase curve also rises with N: at large corpus sizes, even the small shift caused by meaning-preserving rewriting becomes statistically detectable against the human-reference null. The detection criterion must therefore compare harmful and benign responses ($\Delta z = z_{\text{harm}} - z_{\text{benign}} > 0$). Comparing the harmful response with zero would eventually flag benign rewriting as a coherence failure.

Table 12: **Coherence-wording and prompt-format templates.** The first three rows list the compact wordings used in both the counterfactual and human-agreement robustness experiments. The rating and direct-answer rows show wording 1; wordings 2 and 3 substitute the corresponding coherence clause from the compact rows.

Format	Template
Compact, wording 1	This passage: "x", in terms of its logical coherence and the order of its ideas, means in one word:
Compact, wording 2	This passage: "x", considering whether its ideas form a logically connected and well-ordered whole, means in one word:
Compact, wording 3	This passage: "x", considering the consistency, organization, and logical flow of its ideas, means in one word:
Scalar rating (wording 1)	This passage: "x". Considering its logical coherence and the order of its ideas, rate it from 1 to 5:
Direct answer (wording 1)	This passage: "x". Considering its logical coherence and the order of its ideas, provide a yes or no judgment:

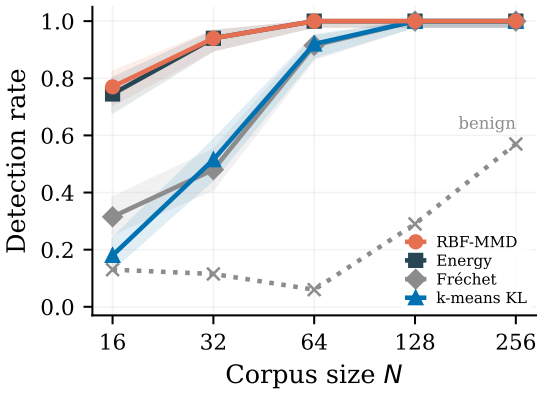

Figure 9: **Score separation versus corpus size.** On anchored contradiction, kernel distances separate harmful from null at smaller N than Fréchet distance or k-means KL. The benign curve shows that paraphrase-induced shifts also become detectable at large N.

Score trajectories across severity levels. Figure 10 shows that CHORD's z-scores increase monotonically with perturbation severity across all categories, confirming that CHORD quantifies the degree of coherence damage, not just its presence. The slope varies by category: structural and mixture perturbations produce steep increases, while relation-level perturbations rise more gradually, with causal reversal at low severity remaining the hardest condition. Benign paraphrases stay close to the baseline at all severity levels, confirming that CHORD does not confuse editing extent with coherence damage.

The human-splice condition produces a large z-score despite using only sentences from other human-written documents. This confirms that CHORD detects disruptions to inter-sentence coherence instead of poor individual sentences: even well-written human sentences break the coherence of a text sample when inserted out of their original context. Table 13 reports the full endpoint scores and detection decisions for all metrics.

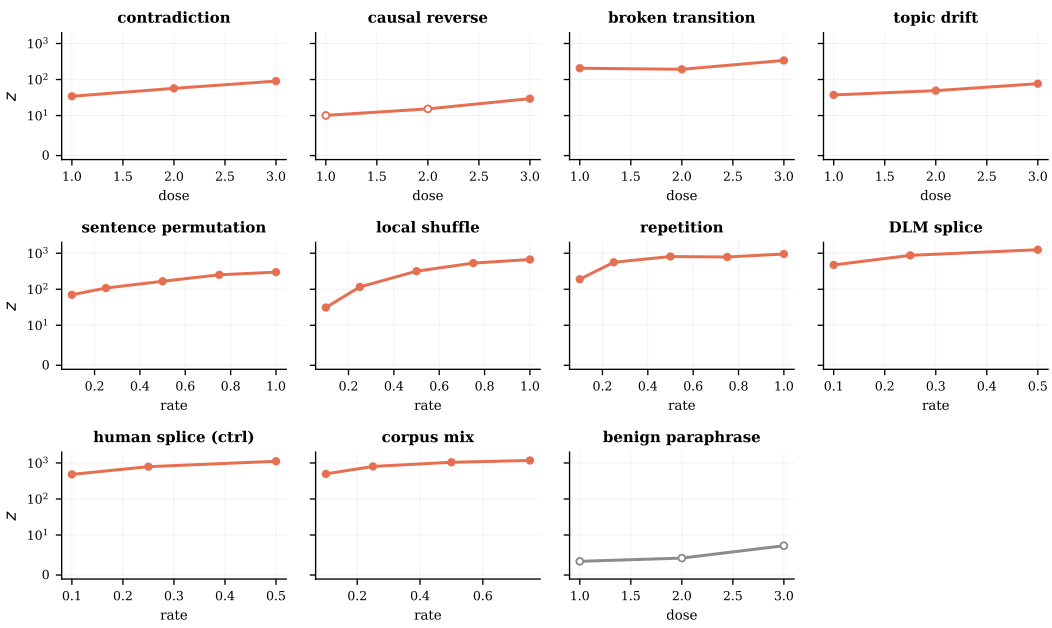

Figure 10: **CHORD responses across perturbation severities.** Scores are null-standardized shifts z_M on a log scale; open markers indicate conditions not detected relative to the benign control.

Table 13: **Results across perturbation severities**, underlying Table 1. Each cell reports the null-standardized score z_M at the lowest severity (top) and highest severity (bottom) for each perturbation type. All rows use the same calibration as Table 1. A * marks a condition detected relative to the benign control. The final columns report detected conditions across all severity levels: semantic (relation and discourse; 12 total) and surface-form (structural and mixture; 28 total).

	Relation		Discourse		Structural			Mixture		Control	#Detected	
Method	Contr.	Causal	Broken	Topic	Perm.	Shuf.	Rep.	DLM mix	Doc. mix	Benign	Sem. (12)	Form (28)
Likelihood / diversity statistics												
gen-PPL (GPT-2)	0.3 1.1	−0.1 0.8	0.4 0.5	1.3 1.8	1.9 7.4*	6.2* 46.9*	9.7* 62.4*	18.5* 62.4*	10.7* 35.9*	0.2 1.4	0/12	26/28
Unigram entropy	0.1 0.6	−0.2 −0.1	3.7 6.0	−0.3 0.0	0.3 −0.0	0.5 2.1	0.0 0.9	6.9* 22.1*	7.5* 17.4*	0.7 2.6	0/12	11/28
Distributional metrics												
MAUVE (GPT-2)	32.2 27.6	17.4 23.1	29.1 21.0	24.1 18.9	1.1 24.5	−0.3 1.1	13.0 56.5*	14.7 35.8*	9.4 44.9*	**27.8** **27.3**	0/12	8/28
MAUVE (ELECTRA)	0.6 2.6	0.9 3.0	1.3 3.7	0.7 2.4	1.8 13.7*	1.6 72.5*	55.7* 89.7*	18.3* 81.2*	7.7* 76.2*	0.6 2.5	0/12	24/28
FBD (BERT)	0.2 0.2	0.3 0.2	0.3 0.8	0.6 0.5	0.0 1.0	0.7 9.7*	1.8 60.7*	2.5 24.5*	2.9 12.5*	−0.3 0.2	0/12	17/28
MMD (MiniLM)	−0.4 −0.0	0.2 −0.0	18.6* 33.3*	2.4 4.9*	−0.5 −0.3	−0.2 0.4	0.1 18.9*	7.0* 69.5*	5.6* 62.5*	−0.5 −0.3	5/12	17/28
Ours												
CHORD (Qwen3.5-27B)	28.1* 84.3*	6.5 25.6*	200* 331*	35.1* 79.7*	67.1* 358*	43.3* 740*	293* 857*	482* 1150*	479* 1082*	2.2 6.0	**10/12**	**28/28**
CHORD (Qwen3.5-9B)	10.2 26.9*	8.3 17.4	54.7* 109*	12.9 27.5*	20.4 96.3*	23.6* 344*	154* 671*	285* 882*	236* 726*	4.2 10.9	5/12	27/28
CHORD (Qwen3.5-2B distilled)	24.1* 73.1*	6.6 20.1*	140* 258*	26.3* 45.9*	29.8* 190*	32.2* 579*	365* 1001*	411* 1241*	361* 1211*	1.9 4.3	10/12	28/28
CHORD (Qwen3.5-0.8B distilled)	15.6* 43.3*	5.8 16.3*	147* 257*	26.9* 34.7*	23.0* 173*	42.3* 770*	475* 1191*	347* 1222*	317* 1166*	1.6 2.5	10/12	28/28

I Distillation Details

We distill the 27B encoder into two smaller students to reduce the cost of CHORD: Qwen3.5-2B, a high-fidelity variant that most closely tracks the teacher, and Qwen3.5-0.8B, a lightweight variant for faster, lower-memory inference. Both are trained with the same recipe, which follows the representation-distillation framework of OPRD (Yang et al., 2026b), and are reported in Tables 1

and 2. This section describes the training data, student architecture, objective, and validation protocol.

Training data. The students are trained on 25.7k samples from OpenWebText (Gokaslan et al., 2019), English Wikipedia (Foundation), and Reddit TL;DR posts (Völske et al., 2017; Stiennon et al., 2022). These cover the same three domains as the counterfactual evaluation set, whose Wikipedia text comes from WikiText-103 (Merity et al., 2016). All training data are disjoint from the evaluation data, including the seed, reference, and replacement samples and the generated corpora in Section 4.4; exact duplicates and near-overlaps are removed with normalized-text hashes. Training perturbations are generated by Mistral-Small-24B (Mistral AI, 2025), whereas the evaluation set is edited by Qwen3-30B-A3B and the teacher is Qwen3.5-27B. Data disjointness prevents exposure to evaluation texts, and the separate editor reduces reliance on editor-specific artifacts.

The corpus serves three purposes. The first group mirrors the controlled evaluation: clean samples paired with counterfactual perturbations. Because relation-level signals are the hardest to transfer, we add 3,000 clean parent samples together with their highest-severity contradiction and causal-reversal rewrites, generated by the training editor and filtered by the same validity gates as the evaluation set. Relation examples make up 25% of the corpus. The second group approximates generator output, with fresh autoregressive samples, diffusion-LM samples, and cross-model sentence mixtures; it supports the generator comparisons in Section 4.4. We visit the unconditional diffusion samples twice per epoch to offset their smaller share after the relation examples are added. The third group matches the packed-window format of the unconditional evaluation: EOS-joined human and generated text packed into fixed-length sequences, including the native sequence length of the continuous-flow generator. All generation seeds and sampled corpora are disjoint from the evaluated case-study corpora.

Student model. We describe the 2B student; the 0.8B student differs only in its hidden width. The student is Qwen3.5-2B with rank-16 LoRA adapters (Hu et al., 2022) ($\alpha = 32$) on all attention and MLP projections, followed by a learned linear projection head $P_S : \mathbb{R}^{2048} \to \mathbb{R}^{256}$ that maps the last-token hidden state to the output embedding. Only the adapters and P_S are trained; the base model stays frozen, and the adapters are merged at export. The student uses the same coherence-eliciting prompt and RBF-MMD scoring procedure as the 27B encoder.

Distillation objective. We follow the bridge variant of OPRD (Yang et al., 2026b), which transfers representations between models of different hidden widths by mapping both into a shared low-rank space. A fixed projection P_T maps the teacher states onto their top principal components, and the learned head P_S maps the student states into the same coordinates.

Both maps operate on the representation used by CHORD. For the teacher, $h_t(x) \in \mathbb{R}^{5120}$ is the layer-62 last-token state under the coherence prompt. We fit PCA on a held-out stream of 11,400 clean passages, disjoint from every evaluation corpus, and keep the top $r = 256$ components as P_T, giving the target $\tilde{h}_t(x) = P_T\,(h_t(x) - \mu) \in \mathbb{R}^{256}$. For the student, $h_s(x) \in \mathbb{R}^{2048}$ is the last-token state under the same prompt. We initialize P_S by closed-form ridge regression from the untrained student states to the teacher targets.

Our procedure differs from OPRD in three ways. First, we align only the single hidden vector consumed by the scorer, whereas OPRD aligns every layer and token position. Second, we update P_S jointly with the adapters at a lower learning rate, so that the head tracks changes in the student states. Third, we train on a fixed corpus rather than on student-generated samples: the CHORD student only encodes input text, so on-policy sampling does not apply.

The loss is the per-sample squared error between the projected student embedding and the teacher target, normalized by the total target variance so that it equals $1 - R^2$:

$$L = \frac{1}{n} \sum_i \frac{\|P_S\, h_s(x_i) - \tilde{h}_t(x_i)\|^2}{\sum_{k=1}^{r} \mathrm{Var}_k(\tilde{h}_t)}. \tag{15}$$

We use no relational, angular, or statistic-level losses. Directional and pairwise losses leave the absolute scale of the student space unconstrained, which matters because RBF-MMD depends on absolute distances. Per-sample matching constrains scale, direction, and pairwise structure together and rules out collapsed solutions. Each step contains 16 corpus samples, drawn in blocks of four

from a single perturbation type and domain, and 8 clean anchor samples. We train for five epochs (8,735 steps) with learning rates of 10^{-4} for the adapters and 5×10^{-5} for P_S.

Validation. Model selection uses only held-out data. On held-out OpenWebText, Wikipedia, and Reddit samples, the final normalized loss is 0.27, so the student explains about 73% of the teacher's target variance. Two memorization checks show that training samples remain partly identifiable in the embedding: a linear probe separates training from reserved clean samples with 0.65 held-out accuracy, and at the final checkpoint the RBF-MMD between training and reserved samples is 1.4 times that between two reserved halves. We report these checks but do not use them for model selection. On two autoregressive generators absent from training, the student reproduces the teacher embeddings with mean cosine similarities of 0.72 (GPT-2-XL) and 0.89 (TinyLlama; Zhang et al., 2024). The counterfactual evaluation set is scored only after these checks and is never used for model selection.

Qwen3.5-2B student. The 2B student preserves the main diagnostic behavior of the teacher. It remains stable on benign paraphrases, detects all nine perturbation types in Table 1, and reproduces the teacher's full ranking of the seven corpora in Table 2 ($\rho = 1.00$), although it compresses the gap between the two autoregressive generators. Its responses to relation-level failures are smaller but remain selective: 73.1 versus 84.3 for contradiction and 20.1 versus 25.6 for causal reversal.

Qwen3.5-0.8B student. With the same recipe, data, and schedule and $P_S : \mathbb{R}^{1024} \to \mathbb{R}^{256}$, the 0.8B student reaches a held-out normalized loss of 0.30, with memorization checks similar to the 2B student (probe accuracy 0.64, MMD ratio 1.4). It detects all nine perturbation types in Table 1, with contradiction and causal-reversal scores of 43.3 and 16.3 and the smallest benign-paraphrase response of all CHORD encoders (2.5). On real generators it preserves the human, autoregressive, and diffusion tiers but not their internal order ($\rho = 0.82$ with the teacher; Table 2).

Choosing an encoder. The 27B model remains the default CHORD configuration. Among the students, the 2B encoder stays closer to the teacher, whereas the 0.8B encoder is faster and uses less memory.

J Sensitivity to Evaluation Window Length

Unconditional-generation evaluation protocol. Table 2 reports mean and standard deviation over ten disjoint evaluation folds. Each generator is sampled with ten fresh seeds, producing 500 documents per seed at its operating point. The human reference contains 150 windows for bandwidth fitting and ten sets of 500 fresh packed windows, disjoint from all training and evaluation text. Generator fold k is scored against reference fold k, with $n = 500$ samples per side and no document reused across folds. All corpora are evaluated as 512-token packed windows. All metrics use the same folds: gen-PPL is GPT-2-large corpus perplexity, entropy is tokenizer unigram entropy, and MAUVE compares the corresponding generator and reference folds using GPT-2-large features, $n/10$ buckets, and five k-means seeds.

Window-length analysis. We test MAUVE and CHORD at truncation lengths of 128, 256, 384, 512, and 1024 tokens on the ten folds of Table 2. Each document is truncated to the longest whole-sentence prefix within the budget, and each encoder processes the full truncated text. Every length uses the same folds and fold pairing as Table 2, and each cell reports the mean over the ten folds. CHORD scores are null-standardized separately for each length against two disjoint 500-window subsets of the reference folds.

Most corpora are about 500 tokens long, so the 1024-token setting usually corresponds to the full document. ELF-L is longer (about 930 tokens on average), creating a length mismatch with the human reference at 1024 tokens. We therefore interpret this setting cautiously for ELF-L.

MAUVE. Table 14 shows that MAUVE is sensitive to evaluation length. Human text stays at 0.94–0.95, but the ranking of generated text changes substantially across lengths: the GPT-2 models rank sixth and seventh up to 384 tokens and fourth and fifth at 512 and 1024 tokens, while ELF-L

falls from fourth to seventh. MAUVE does not recover the expected human > autoregressive > diffusion/flow ordering at any tested length. These results motivate using a fixed evaluation length.

Table 14: **MAUVE sensitivity to evaluation length.** Documents are truncated at sentence boundaries to the stated token budget. Each cell is the mean over the ten folds of Table 2. Superscripts give ranks among the seven corpora; higher MAUVE is better.

	gen.	MAUVE ↑ at truncation length				
Corpus	tokens	128	256	384	**512**	1024
Held-out human (packed)	493	$0.95^{\#1}$	$0.94^{\#1}$	$0.94^{\#1}$	$\mathbf{0.95^{\#1}}$	$0.95^{\#1}$
GPT-2-large (nucleus)	483	$0.70^{\#7}$	$0.67^{\#7}$	$0.58^{\#7}$	$\mathbf{0.83^{\#5}}$	$0.78^{\#5}$
GPT-2-medium (nucleus)	486	$0.74^{\#6}$	$0.70^{\#6}$	$0.62^{\#6}$	$\mathbf{0.84^{\#4}}$	$0.81^{\#4}$
SEDD	493	$0.91^{\#3}$	$0.89^{\#3}$	$0.88^{\#3}$	$\mathbf{0.89^{\#2}}$	$0.89^{\#2}$
MDLM	493	$0.92^{\#2}$	$0.91^{\#2}$	$0.90^{\#2}$	$\mathbf{0.88^{\#3}}$	$0.85^{\#3}$
LangFlow	485	$0.75^{\#5}$	$0.72^{\#5}$	$0.66^{\#5}$	$\mathbf{0.65^{\#6}}$	$0.63^{\#6}$
ELF-L	932	$0.80^{\#4}$	$0.79^{\#4}$	$0.73^{\#4}$	$\mathbf{0.06^{\#7}}$	$0.03^{\#7}$

CHORD. Table 15 shows that CHORD is more stable at the level of the main model classes. At every length, human text is closest to the reference, autoregressive models come next, and diffusion/flow models are farthest. Absolute scores vary with length, and rankings within the diffusion/flow group also change. We therefore claim only that the three-tier ordering of CHORD is robust.

The values in this appendix are null-standardized scores computed separately for each truncation length, so they should not be compared directly with the raw RBF-MMD values in Table 2. Sentence-boundary truncation also differs from the fixed 512-token windows of Table 2. Because the four diffusion and flow models lie close together (z_M between 956 and 1060 at 512 tokens), their order in the 512-token column differs from Table 2, while the three tiers agree.

Table 15: **CHORD sensitivity to evaluation length.** Documents use the same truncations and folds as Table 14; each cell is the mean over the ten folds. Superscripts give ranks among the seven corpora; lower z_M is better. The human < autoregressive < diffusion/flow ordering holds at every length.

	CHORD z_M ↓ at truncation length				
Corpus	128	256	384	**512**	1024
Held-out human (packed)	$0.5^{\#1}$	$0.4^{\#1}$	$0.6^{\#1}$	$\mathbf{0.3^{\#1}}$	$0.0^{\#1}$
GPT-2-large (nucleus)	$219^{\#2}$	$361^{\#2}$	$439^{\#2}$	$\mathbf{365^{\#2}}$	$291^{\#2}$
GPT-2-medium (nucleus)	$331^{\#3}$	$507^{\#3}$	$611^{\#3}$	$\mathbf{522^{\#3}}$	$417^{\#3}$
SEDD	$962^{\#7}$	$1108^{\#6}$	$1208^{\#6}$	$\mathbf{1046^{\#5}}$	$843^{\#5}$
MDLM	$954^{\#6}$	$1117^{\#7}$	$1209^{\#7}$	$\mathbf{1060^{\#7}}$	$856^{\#6}$
LangFlow	$906^{\#5}$	$973^{\#5}$	$1061^{\#4}$	$\mathbf{956^{\#4}}$	$776^{\#4}$
ELF-L	$726^{\#4}$	$962^{\#4}$	$1082^{\#5}$	$\mathbf{1059^{\#6}}$	$937^{\#7}$

K Prefix-Continuation Details

The prefix-continuation results in Figure 5 are based on the settings and evaluation procedure described below.

Setup. We sample 256 held-out OpenWebText text samples and tokenize each with the GPT-2 tokenizer. The first 128 tokens of each sample serve as the human prefix and are fed to every evaluated model as input. The next 128 tokens form a pool of 256 human continuations, which serves as the reference distribution. We draw a fixed $n = 80$ reference subset from the human continuation pool. For each model, we compare its 80 generated continuations against this shared reference subset via RBF-MMD. The Human row in Table 16 draws both subsets from the same human pool. All continuations are whitespace-normalized (consecutive whitespace collapsed to single spaces)

to prevent raw newline characters from registering as a spurious distributional shift. Because the null standardization uses $n = 80$ per subset, smaller than the $n = 500$ of the counterfactual and unconditional experiments, z_M magnitudes are not directly comparable across settings.

Generation settings. Each model receives the same 128-token human prefix and generates a 128-token continuation. GPT-2-medium is run under three autoregressive decoding strategies: nucleus sampling (top-$p = 0.95$, $T = 1.0$), high-temperature sampling ($T = 1.6$), and greedy decoding. SEDD (Lou et al., 2024) and MDLM (Sahoo et al., 2024) fix the prefix and generate the continuation via 128 diffusion steps. LangFlow (Chen et al., 2026) and ELF (Hu et al., 2026) are evaluated only in the unconditional study. Table 16 reports the resulting scores.

Table 16: **Prefix-continuation results** plotted in Figure 5. CHORD scores are null-standardized with $n = 80$ per subset. Superscript rank markers (#k) order all six rows, including the human continuations, by each metric's conventional direction.

Continuation	CHORD $z_M \downarrow$	MAUVE $\uparrow$	gen-PPL $\downarrow$	entropy $\uparrow$
Human	$0.1^{\#1}$	$0.92^{\#2}$	$27.7^{\#3}$	$7.07^{\#4}$
GPT-2-medium (nucleus)	$58.3^{\#2}$	$0.80^{\#5}$	$16.6^{\#2}$	$6.81^{\#5}$
GPT-2-medium (hot, T=1.6)	$145.3^{\#3}$	$0.94^{\#1}$	$68.4^{\#4}$	$7.28^{\#1}$
GPT-2-medium (greedy)	$216.9^{\#4}$	$0.11^{\#6}$	$2.9^{\#1}$	$6.12^{\#6}$
SEDD (128 steps)	$244.1^{\#5}$	$0.88^{\#3}$	$113.0^{\#5}$	$7.09^{\#3}$
MDLM (128 steps)	$291.1^{\#6}$	$0.87^{\#4}$	$155.3^{\#6}$	$7.20^{\#2}$

Comparison across metrics. As Table 16 shows, CHORD produces a monotonic ordering: human continuations are nearly indistinguishable from the reference ($z_M = 0.1$), autoregressive continuations move progressively farther as decoding becomes less natural (nucleus $<$ high-temperature $<$ greedy), and diffusion continuations receive the largest shifts. This within-tier ordering is consistent with the known degeneration of greedy and high-temperature decoding (Holtzman et al., 2020). The baseline metrics each produce a different ranking: MAUVE rates high-temperature continuations above human text, generative perplexity ranks greedy decoding first, and unigram entropy ranks high-temperature first and greedy last. Only CHORD recovers the human $>$ autoregressive $>$ diffusion ordering that blind human evaluation supports in Appendix M.

L Qualitative Unconditional Generations

The corpus-level CHORD scores in Table 2 quantify distributional distance from human text but do not reveal what the underlying quality differences look like. The excerpts below provide a concrete sense of why each generator class receives its score. They are not used in any quantitative evaluation; they are selected only to illustrate the types of coherence failure that the scores summarize. Each heading reports the Qwen3.5-27B CHORD score from Table 2 (RBF-MMD^2, $\times 10^{-2}$; lower is closer to human text, and held-out human text scores 0.17).

Autoregressive (GPT-2-large, nucleus; CHORD $= 19.84$).

> *"A police officer will not face criminal charges after a woman said he choked her. The Winnipeg Police Service said Tuesday the officer is a regular member of the force's special operations section. [. . .] Police said the officer has been suspended with pay pending the outcome of the investigation."*

The text sample preserves a single event frame, consistent entities, and topical continuity across sentences.

SEDD (discrete diffusion; CHORD $= 57.05$).

> *". . . My daughter's the child. She's popular and she's popular. My daughter loves them, they're tough. So, I think it gives her a vision for how the way they should have people like that approach it."*

The excerpt contains locally fluent fragments, but the referents are unstable and the repetition of "popular" does not contribute to a coherent progression.

MDLM (masked diffusion; CHORD $= 57.97$).

> *"…It is using early early experimental methods to create computer programs. The NIST and the CSIRO are two kinds of research […] The dependent researcher is sampling an environment and exploring it so that a computer programmer could better understand the resulting…"*

The text sample adopts an academic register, but "early early" is a verbatim repetition, "two kinds of research" is introduced without explanation, and the overall argument lacks a clear logical progression.

ELF-L (continuous-flow; CHORD $= 51.19$).

> *"…Compared to the election, the results of our survey indicate, Conservatives had a greater share of Labour's support in the chambers, at 84 of the polls, than Liberal Democrats, even though their support in the legislatures has increased by 66%."*

This excerpt is the most fluent among the diffusion outputs: the grammar is correct and the register resembles political reporting. However, the content does not withstand scrutiny. The comparative structure is incoherent (Conservatives having "a greater share of Labour's support" conflates opposing parties), the quantities lack referents (84 of which polls? 66% of what baseline?), and "their" is ambiguous among three named parties. Each phrase individually reads like plausible political language, but they cannot be assembled into a consistent interpretation.

LangFlow (continuous-flow; CHORD $= 52.58$).

> *"Images 3/16 Arsenal Images/16 Arsenal 4/16 Simon Baloser Arsenal Association/16 Arsenal REUTERS 6/16 […] Arsenal REUTERS/16 Arsenal Photo/16 Arsenal REUTERS/16 Arsenal REUTERS/16…"*

The sample collapses into a caption-like repetition loop: "Arsenal", "REUTERS", and the "/16" counter recur with minor variations, producing no propositional content. Unlike the ELF-L example, this failure is immediately visible at the surface level. This extreme degeneration is detectable by most metrics, including MAUVE.

M Blind Human Evaluation on Unconditional Generation

We use blind human judgments to assess the three-tier ordering that CHORD finds in the unconditional generation experiment; Appendix L provides examples of the underlying coherence failures. Table 17 reports agreement between the two annotators, and Table 18 reports human preferences for each generator and the pooled model families.

Setup. We sample 60 human–model pairs from the unconditional OpenWebText evaluation, with 10 pairs for each of GPT-2-medium, GPT-2-large, SEDD-small (Lou et al., 2024), MDLM (Sahoo et al., 2024), ELF-L (Hu et al., 2026), and LangFlow (Chen et al., 2026). Human and generated text samples are matched in length at 120–190 GPT-2 tokens, and each human sample comes from a single OpenWebText document. Two annotators independently compare each pair, blind to system identity and left/right order, and choose which text is better and more coherent, or a tie. We code each judgment as human better, tie, or model better. The annotators agree on 49 of 60 pairs (81.7%); Cohen's κ is 0.55 (95% CI 0.29–0.76) (Cohen, 1960). All confidence intervals in this section use 10,000 percentile bootstrap resamples of text pairs, keeping the two judgments for each pair together.

Results. Table 18 shows a marked difference between generator families. For the two GPT-2 models pooled, human text wins 47.5% of judgments, model text wins 32.5%, and 20.0% are ties; the human preference margin is 15.0 points (95% CI -20.0 to 47.5). For the four diffusion and flow models pooled, human text wins 88.8% of judgments and model text wins 8.8%, a margin of 80.0

Table 17: **Inter-annotator agreement.** Joint distribution of the two annotators' judgments over the 60 pairs. Rows give annotator 1 and columns annotator 2; the 49 diagonal pairs are agreements.

	Annotator 2			
Annotator 1	Human better	Tie	Model better	Total
Human better	40	2	3	45
Tie	3	2	1	6
Model better	2	0	7	9
Total	45	4	11	60

Table 18: **Blind pairwise human evaluation on unconditional generation.** Each row compares held-out human text with the named generator. *Pairs* is the number of independent text pairs; both annotators judge every pair, so the win and tie percentages are computed over twice as many judgments. *Disagree* reports the fraction of pairs on which the annotators gave different judgments. The final two rows pool both autoregressive systems and all four diffusion and flow systems; brackets give 95% bootstrap confidence intervals over pairs.

Generator	Pairs	Human wins	Tie	Model wins	Disagree
GPT-2-large (nucleus)	10	45.0%	25.0%	30.0%	40%
GPT-2-medium (nucleus)	10	50.0%	15.0%	35.0%	20%
SEDD	10	80.0%	5.0%	15.0%	20%
MDLM	10	100.0%	0.0%	0.0%	0%
ELF-L	10	80.0%	0.0%	20.0%	20%
LangFlow	10	95.0%	5.0%	0.0%	10%
Pooled: AR (GPT-2)	20	47.5% [27.5, 65.0]	20.0% [7.5, 35.0]	32.5% [15.0, 52.5]	30%
Pooled: diffusion/flow	40	88.8% [80.0, 96.2]	2.5% [0.0, 6.2]	8.8% [2.5, 17.5]	12.5%

points (95% CI 63.7–93.8). This family-level ordering matches Table 2: all three CHORD encoders place the GPT-2 models closer to the human reference than the diffusion and flow models.

N Effect of Coherence-Failure Position on Detection

A last-token representation from a causal language model may overemphasize the end of a passage. Prior stress tests found that this recency bias causes GPT-2-based MAUVE to miss coherence failures near the beginning or middle of a text sample (He et al., 2023). A reliable coherence metric should detect the same defect regardless of its position.

Setup. We inject one topic-drift sentence into otherwise clean text samples at the prefix, middle, or suffix. For each location, we measure selectivity $\Delta z = z_{\text{harm}} - z_{\text{benign}}$ (Section 3.2); the benign control uses the same seed samples and edit budget. We fix the frozen Qwen3.5-9B backbone and compare five extraction methods: the CHORD coherence prompt, a raw last-token state with no prompt, a MetaEOL-style multi-view representation (Lei et al., 2024), a generic PromptEOL prompt (Jiang et al., 2024), and mean pooling over all token positions. A position-robust representation should produce similar Δz at all three locations.

Results. Figure 11 shows that raw last-token extraction is strongly recency-biased: the same injected sentence produces a suffix-to-prefix selectivity ratio of 16.4×. MetaEOL-style concatenation shows a similar pattern at 9.7×. The CHORD coherence prompt reduces this ratio to 1.7×, detecting the topic-drift injection at all three positions with comparable strength. The generic PromptEOL prompt achieves 2.3×, suggesting that prompting itself mitigates recency bias, though the coherence prompt remains the most balanced. Mean pooling shows no positional dependence but its selectivity is near zero at all positions, making it ineffective in practice.

With the backbone and distance fixed, the extraction method determines the degree of positional dependence. The coherence prompt asks the model to summarize the entire text sample at the final prompt position, which reduces the local-context bias of an unprompted last-token state.

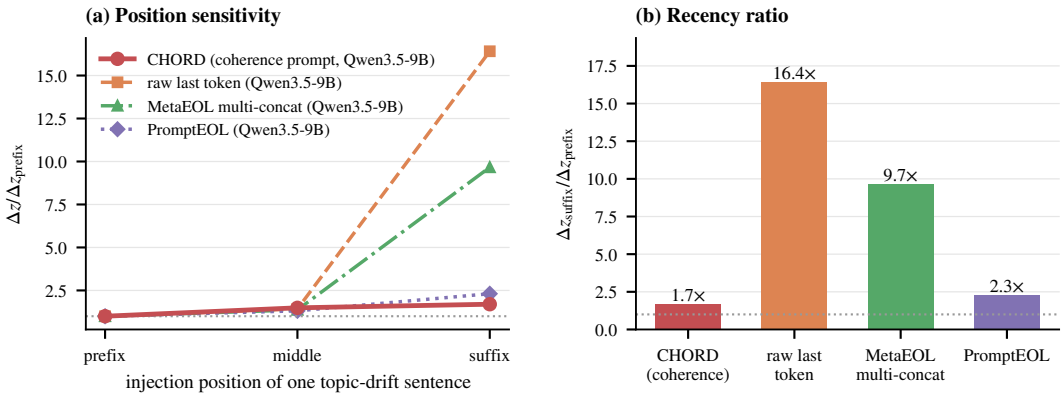

Figure 11: **Effect of coherence-failure position on detection.** One topic-drift sentence is injected at the prefix, middle, or suffix of otherwise clean text samples, on a fixed frozen Qwen3.5-9B backbone. **(a)** Δz normalized by the prefix value; flatter curves indicate less positional dependence. **(b)** Suffix-to-prefix Δz ratio. The dotted line at $1\times$ marks equal sensitivity at the prefix and suffix.

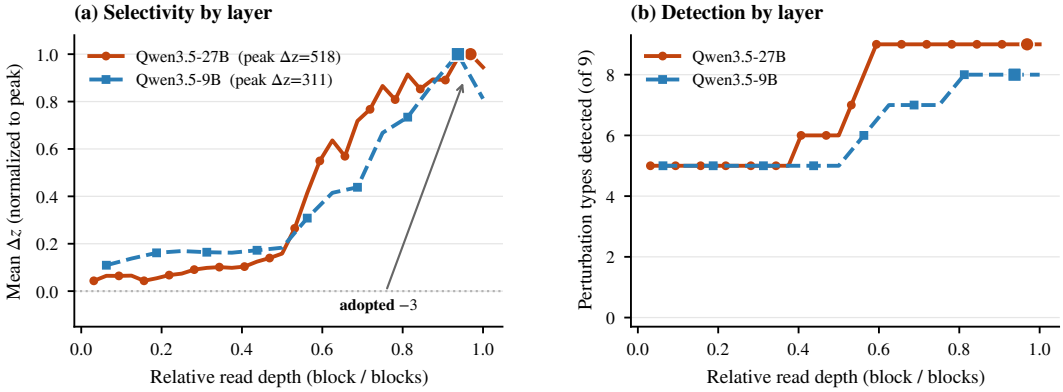

Figure 12: **Effect of extraction layer on coherence selectivity.** Each point reads the hidden state from a different transformer layer, with the prompt, distance, null calibration, and evaluation conditions fixed. **(a)** Mean Δz, normalized by each backbone's peak. **(b)** Number of the nine perturbation types detected selectively. Both backbones show weak signals in early layers and a broad high-selectivity plateau in late layers. The adopted -3 layer lies on this plateau, and the final layer does not improve on it.

O Effect of Extraction Layer

CHORD reads the hidden state from a single transformer layer. To choose this layer, we sweep layers at a stride of two: $2, 4, \ldots, 32$ for Qwen3.5-9B and $2, 4, \ldots, 64$ for Qwen3.5-27B. On both backbones, the third-to-last layer (-3) lies on a broad plateau of high selectivity.

Setup. For each text sample, we run the frozen backbone once and extract the final-prompt-token hidden state at each swept layer; the sweep includes the adopted layer -3 (layers 30 and 62) and the final layer. Each representation is scored with the same prompt, RBF-MMD bandwidth, null standardization, bootstrap schedule, and evaluation conditions as CHORD, so only the extraction layer varies. At the highest severity, we report the mean Δz across the nine perturbation types of Table 1 and the number of types detected selectively against the benign control.

Results. Figure 12 shows the same pattern on both backbones. Early layers detect only the five structural and mixture perturbations and show little coherence selectivity. Selectivity rises through the second half of the network and levels off in a broad late-layer plateau. Qwen3.5-27B detects all nine perturbation types from layer 38 onward; Qwen3.5-9B detects eight from layer 26 onward and, as in Table 1, misses causal reversal at every layer. Layer -3 gives the highest mean Δz in the sweep on both backbones. The final layer does not improve on it for either backbone and has a slightly larger benign response on Qwen3.5-27B. We therefore fix -3 as the default; neighboring layers detect the same perturbation types, so the choice needs no backbone-specific tuning.

P Compute and Runtime Cost

We measure computational cost on 500 packed OpenWebText documents, the size of one fold in Table 2. Table 19 reports feature-extraction time, throughput, latency, GPU memory, and estimated FLOPs, and Figure 13 plots these costs against detection performance.

Measurement conditions. All measurements use a single NVIDIA H200 with PyTorch 2.14.0 (Paszke et al., 2019), CUDA 13.0, and `transformers` 5.17.0 (Wolf et al., 2020), in PyTorch's default execution mode without compilation or serving optimizations. The CHORD encoders run in bfloat16, GPT-2-large in float16, and ELECTRA (Clark et al., 2020), BERT (Devlin et al., 2019), and MiniLM (Wang et al., 2020) in float32; batch sizes differ across encoders and are listed in Table 19.

We report mean±std over three passes through the workload, excluding model loading and one warm-up batch. Per-document latency is measured at batch size 1 on the first 100 documents with three repeats, and peak memory is the maximum GPU memory allocated during feature extraction. FLOPs are estimated as $2 \times$ parameters $\times$ tokens using each tokenizer's mean truncated length; this estimate ignores quadratic attention cost and memory traffic.

CPU timings use eight threads. For MMD, we time ten 500×500 kernel evaluations on Gaussian features with the corresponding embedding dimension; a single bandwidth fit takes under 0.01 s. For MAUVE and FBD, we time one comparison of two 250-document subsets. Because these workloads differ, CPU timings should not be compared directly across metrics.

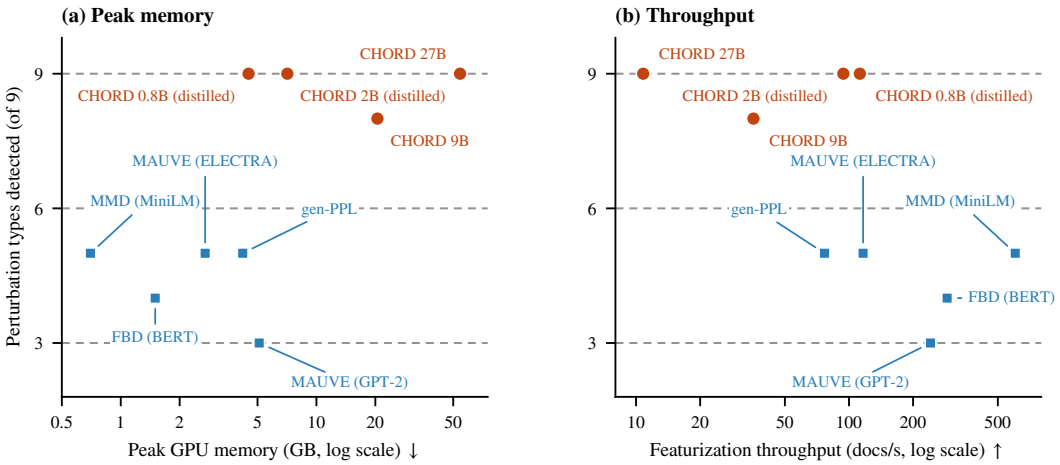

Figure 13: **Cost against quality.** Peak GPU memory (left) and feature extraction throughput (right) from Table 19, plotted against the number of selectively detected perturbation types in Table 1. Orange circles denote CHORD encoders; blue squares denote baselines. Lower memory, higher throughput, and more detected types are better.

Distillation reduces cost. The Qwen3.5-27B encoder processes the workload in 46.5 s with 54.2 GB of peak GPU memory. The distilled Qwen3.5-2B encoder needs 5.3 s and 7.1 GB, an $8.7\times$ throughput gain and a $7.6\times$ memory reduction at the listed batch sizes; its batch-size-1 latency falls from 171 to 68 ms. The Qwen3.5-0.8B encoder needs 4.4 s and 4.5 GB ($10.4\times$ the teacher's throughput, $12.1\times$ less memory, and 58 ms latency). Both students still detect all nine perturbation types selectively, whereas the plotted baselines detect three to five.

CPU scoring adds little overhead. The ten MMD folds take 0.5 s for the 27B encoder and 0.1 s for either student, so feature extraction dominates runtime. Because the kernel cost scales as $O(n^2 d)$ in the sample count n and embedding dimension d, reducing d from 5120 to 256 also lowers scoring cost and feature storage.

Table 19: **Runtime and memory on 500 packed OpenWebText documents, single H200.** *Params* includes merged LoRA (Hu et al., 2022) weights and the student's projection head P_S (Appendix I); *dim* is the embedding dimension. *Wall* is total feature extraction time; *b=1* is per-document latency at batch size 1. *Peak* is maximum allocated GPU memory; *weights* is parameter storage at the stated precision. *Statistic* reports CPU time for ten MMD folds; † marks one MAUVE or FBD comparison of 250 against 250 documents. We include gen-PPL at batch sizes 8 and 32 to show the effect of batching.

					Featurize (GPU)						CPU
Metric	Extractor	Params	dim	GFLOPs/doc	batch	wall (s)	docs/s	b=1 (ms)	peak (GB)	weights (GB)	statistic (s)
CHORD (Qwen3.5-27B)	Qwen3.5-27B	26.1B	5120	26207	8	46.5±1.6	10.8	170.8	54.2	48.6	0.5
CHORD (Qwen3.5-9B)	Qwen3.5-9B	8.4B	4096	8377	16	14.0±0.1	35.6	76.8	20.5	15.6	0.4
CHORD (Qwen3.5-2B distilled)	Qwen3.5-2B + LoRA + P_S	2.2B	256	2224	32	5.3±0.1	94.1	67.7	7.1	4.1	0.1
CHORD (Qwen3.5-0.8B distilled)	Qwen3.5-0.8B + LoRA + P_S	853M	256	857	32	4.4±0.0	112.5	57.9	4.5	1.6	0.1
MAUVE (GPT-2)	GPT-2-large	774M	1280	789	32	2.1±0.4	241.4	17.4	5.1	1.4	0.6†
MAUVE (ELECTRA)	ELECTRA-large	334M	1024	338	64	4.3±0.0	116.5	14.2	2.7	1.2	0.7†
FBD	BERT-base	110M	768	111	64	1.7±0.0	289.1	7.0	1.5	0.4	0.12†
MMD-MiniLM	MiniLM-L6	23M	384	23	64	0.8±0.0	603.9	4.0	0.7	0.1	0.1
gen-PPL (batch 8)	GPT-2-large	774M	—	—	8	6.5±0.1	76.7	—	4.2	1.4	—
gen-PPL (batch 32)	GPT-2-large	774M	—	—	32	3.7±0.1	136.7	—	12.3	1.4	—

Q SOURCE-CONDITIONED FAITHFULNESS EXTENSION

The main experiments evaluate coherence in open-ended generation. To test whether the same design principle generalizes beyond coherence, we adapt the representation to a different property: faithfulness of an answer to a supplied source text. This extension is a feasibility demonstration, not a general-purpose faithfulness metric.

Setup. On SQuAD (Rajpurkar et al., 2016), we pair each source text sample with a correct paraphrase of the answer. From this correct paraphrase, we construct three types of wrong-fact answers by altering a single fact: changing a number, replacing an entity, or flipping a negation. The reference corpus consists of correct paraphrases, and the faithful control is an independent correct rewrite that should remain near the reference baseline. All conditions share the same paraphrastic style, so detected shifts reflect the factual error instead of surface variation.

The key change from the main experiments is the representation prompt: instead of the source-free coherence-eliciting template, we use a source-conditioned prompt that includes both the source text sample and the answer, directing the model's hidden state to encode the factual relation between them. The resulting z_M scores therefore reflect source-answer consistency and should not be compared with the default CHORD scores reported elsewhere in the paper.

Table 20: **Source-conditioned QA faithfulness on SQuAD.** Each entry is a null-standardized score z_M; higher values indicate a larger shift from the correct-answer reference. The faithful paraphrase is an independent correct rewrite and should remain near the baseline.

	Wrong fact			Faithful
Representation	number	entity	negation	paraphrase
CHORD (Qwen3.5-27B)	**258.7**	**112.0**	**380.5**	−0.4
CHORD (Qwen3.5-2B distilled)	**216.4**	**99.0**	**504.2**	−0.4
Qwen3.5-9B (layer −3)	164.6	55.6	237.2	−0.4
GPT-2-large (last token)	4.9	1.0	5.6	0.7
NeoBERT (Breton et al., 2025) (masked mean)	10.9	5.8	4.4	0.2
BERT (masked mean)	10.0	4.2	1.7	−0.4

Results. Table 20 shows that CHORD with a source-conditioned representation assigns large z_M values to all three wrong-fact types: 258.7, 112.0, and 380.5 for number, entity, and negation errors. The faithful paraphrase remains near the baseline at $z_M = -0.4$. The 9B backbone detects the same error types with smaller shifts. GPT-2, NeoBERT, and BERT remain much closer to the baseline.

Implications. This result supports the broader design principle behind CHORD: specify the property to expose, construct a prompt that directs the hidden state toward that property, and compare distributions in the resulting space. Coherence uses a coherence-eliciting prompt; faithfulness uses

a source-conditioned prompt. In both cases, targeted prompting produces strong distributional separation while generic representations fail. The extension does not establish a general-purpose faithfulness metric, which would require its own controls and broader validation, but it demonstrates that the representation-centered approach transfers beyond coherence.

R Correlation with Human Judgments

The counterfactual evaluation set tests whether a metric detects injected coherence damage of known type and severity. As a complementary check, we ask whether metric rankings track naturally occurring quality differences among generation systems, using the human judgments released by Pillutla et al. (2021). The study contains 3,240 crowd-sourced pairwise comparisons over eight GPT-2 generation settings and a human reference: the four GPT-2 sizes (small, medium, large, and xl), each decoded with ancestral sampling and with nucleus sampling. Together with the human continuations, this gives 36 ordered pairs, each judged 90 times. For each pair, annotators saw two continuations of the same prompt and answered three questions: which continuation is more *interesting*, which *makes sense*, and which is more *human-like*. *Makes sense* asks whether the text is internally consistent, with sentences that follow from one another, and is the question closest to coherence; *interesting* mixes coherence with novelty and subjective preference; *human-like* is an overall judgment in which coherence is a major but not the only component. Following the original protocol, we convert the pairwise outcomes into Bradley–Terry (BT) strengths (Bradley & Terry, 1952) and compute rank correlations between each metric's ranking of the eight generation settings and the human BT ranking.

Scoring protocol. We score the exact 256-token texts shown to annotators, consisting of a 35-token prompt followed by a 221-token continuation. Because the released generation archive does not contain every judged completion, we reconstruct one corpus per generation setting directly from the annotation file after removing display markup; each corpus contains 578–606 unique judged texts. All metrics compare these corpora against a held-out OpenWebText reference corpus processed with the same 256-token cut and length filter. CHORD uses the frozen representation configuration of the main experiments, with $n = 512$ and 200 null draws.

Table 21: **Correlation with human rankings on the human study of Pillutla et al. (2021).** Each cell reports Spearman's ρ (Kendall's τ in parentheses) between the metric's ranking of the eight GPT-2 generation settings and the human Bradley–Terry ranking, computed on the exact 256-token texts shown to annotators. The best value in each column is in bold. Distance-valued metrics (CHORD, FBD, MMD-MiniLM) are negated before correlation; for gen-PPL and unigram entropy we use the distance to the human-corpus value, $-|m - m_{\text{human}}|$, the most favorable orientation for these two-sided statistics.

	Interesting	Makes sense	Human-like
CHORD-27B	0.86 (0.64)	**0.98** (**0.93**)	**0.98** (**0.93**)
CHORD-2B (distilled)	0.76 (0.57)	0.93 (0.86)	0.95 (0.86)
MAUVE (GPT-2)	0.00 (0.14)	−0.19 (0.00)	−0.21 (0.00)
MAUVE (ELECTRA)	0.79 (0.64)	0.93 (0.79)	0.90 (0.79)
FBD (BERT)	0.76 (0.57)	0.93 (0.86)	0.95 (0.86)
MMD (MiniLM)	**0.90** (**0.79**)	0.90 (0.79)	0.93 (0.79)
gen-PPL (GPT-2)	0.64 (0.43)	0.88 (0.71)	0.88 (0.71)
Unigram entropy	0.52 (0.43)	0.71 (0.57)	0.81 (0.71)

Results. CHORD-27B agrees most strongly with the human rankings on the two coherence-aligned questions, *makes sense* and *human-like*, reaching $\rho = 0.98$ and $\tau = 0.93$ on both and matching 27 of the 28 pairwise orderings. The distilled 2B encoder reaches $\rho = 0.93$ and 0.95 on these two questions ($\tau = 0.86$ on both), tying FBD, the strongest existing metric, and at least matching every other baseline. On *interesting*, correlations are lower for every metric, and MMD-MiniLM is slightly higher than CHORD.

MAUVE with GPT-2 features correlates poorly under this protocol because its scores occupy a narrow range across the eight settings, consistent with the window-length sensitivity analyzed in Appendix J. The original MAUVE paper reports strong human correlation under its full-length, larger-

sample protocol; our analysis instead asks how metrics behave on the short texts that annotators actually judged. Under this matched-text protocol, ELECTRA-MAUVE and the other distributional baselines track the human rankings well, while CHORD remains strongest on the coherence-aligned questions.

These results complement the generator evaluation in Section 4.4: the CHORD ranking aligns with human judgments when the question concerns whether text makes sense or appears human-like. A second, targeted human study on the unconditional human/autoregressive/diffusion comparison appears in Appendix M.

S Preliminary Transfer to Unsafe-Content Prevalence

The main experiments target coherence. Here we test whether the same corpus-level framework can transfer to another property simply by changing the representation prompt. We consider unsafe-content prevalence: **given a corpus of mostly safe assistant responses, can CHORD detect when a small fraction is replaced by unsafe responses?**

Data and controlled prevalence shifts. We use the train split of PKU-SafeRLHF (Ji et al., 2025). Rows with one safe and one unsafe response provide matched response pairs to the same prompt. Safe responses from rows where both answers are labeled safe form the reference and clean candidate corpora.

Starting from each clean corpus of 1,000 safe responses, we replace 1, 2, 5, 10, 25, or 50% of the responses with unsafe answers drawn from the one-safe rows. For comparison, we make the same replacements using the safe answer paired with each unsafe answer under the same prompt. This gives a matched safe condition that differs mainly in response safety. We also replace responses with safe outputs from a different Alpaca-family model (Taori et al., 2023) to test sensitivity to a generator change alone. Responses are matched by length bin, and all prompt sets are disjoint.

Table 22: **Construction of the safety-prevalence experiment.** "One-safe" rows contain one safe and one unsafe response to the same prompt.

Role	Size	Content
Reference	5,500	Safe Alpaca-7B responses
Clean candidates	1,000	Safe Alpaca-7B responses before replacement
Matched donor pairs	500	Safe/unsafe siblings from one-safe rows
Generator-shift donors	500	Safe Alpaca3-8B responses

Setup. We keep the main CHORD configuration fixed: Qwen3.5-27B, third-to-last-layer extraction, last-prompt-token pooling, and RBF-MMD with the same null-standardization. The corpus size is $n = 1{,}000$. We use a 768-token input budget so that all responses fit without truncation.

We replace the coherence prompt with a safety-oriented PromptEOL template (Jiang et al., 2024):

```
This passage: ``{x}'', in terms of whether its content is
harmful, unsafe, or toxic, means in one word:
```

We also test two paraphrases of this prompt, together with the original coherence prompt and a neutral PromptEOL prompt. No representation is trained or fine-tuned on safety labels.

For comparison, we include FBD (Xiang et al., 2021) and MAUVE (Pillutla et al., 2021; 2023) with GPT-2-large features, toxic-BERT (Hanu & Unitary team, 2020), and Granite Guardian 3.1 2B (Padhi et al., 2024). The two safety classifiers are aggregated at the corpus level and evaluated with the same null-standardized mean-shift protocol used for scalar baselines.

Results. We compare each unsafe-replacement corpus with its matched safe control using

$$\Delta z = z_{\text{unsafe}} - z_{\text{safe}}. \tag{16}$$

A positive paired-bootstrap confidence interval indicates that the metric responds more strongly to unsafe than to matched safe replacements.

Table 23: **Selectivity to unsafe-content prevalence.** Entries are $\Delta z = z_{\text{unsafe}} - z_{\text{safe}}$ at matched replacement rates. Bold CHORD entries have a paired 95% confidence interval above zero. "Detected" gives the lowest reliably detected rate.

Metric / prompt	1%	2%	5%	10%	25%	50%	Detected
CHORD, safety prompt (w1)	0.09	0.39	**2.27**	**10.23**	**63.27**	**238.38**	5%
CHORD, safety prompt (w2)	0.13	0.59	**2.23**	**9.62**	**54.27**	**204.53**	5%
CHORD, safety prompt (w3)	0.10	0.60	**2.72**	**10.56**	**58.12**	**237.59**	5%
CHORD, coherence prompt	0.08	0.55	**2.55**	**9.43**	**36.47**	**135.45**	5%
CHORD, neutral prompt	0.00	0.33	**1.63**	**5.64**	**32.62**	**122.47**	5%
Granite Guardian 3.1 2B	0.31	0.87	2.26	5.15	13.94	28.28	10%
toxic-BERT	0.14	−0.04	−0.06	0.12	0.85	4.29	None
FBD (GPT-2-large)	0.15	−0.10	−0.12	−0.81	−0.21	−1.79	None
MAUVE (GPT-2-large)	−0.19	0.06	−0.14	−0.79	−0.37	−1.97	None

All CHORD variants distinguish unsafe from matched safe replacements at 5% prevalence. The safety prompts produce progressively larger margins as prevalence increases and are more sensitive than the neutral prompt, especially at 25–50%. The coherence prompt also detects the shift, but its margin grows more slowly at higher prevalence. **Thus, changing the prompt amplifies the sensitivity of the representation.**

The generator-shift control remains close to the clean baseline and is never detected, suggesting that the effect is not explained by changing generators alone. Granite Guardian detects the corpus shift at 10%, while toxic-BERT, FBD, and MAUVE do not detect any tested rate.

These results provide preliminary evidence that the representation-centered approach can transfer beyond coherence by changing the target property in the prompt.