

Learning a Discriminator for Conceptual Diversity in Language Model Outputs

Stanford CS224N Custom Project

Hangoo Kang

Department of Computer Science
Stanford University
hangook@stanford.edu

James Liu

Department of Computer Science
Stanford University
jihaoliu@stanford.edu

Abstract

Existing diversity metrics for language model outputs largely measure surface variation, such as lexical overlap or embedding distance, and therefore often fail to capture *conceptual diversity*: whether two responses to the same prompt differ meaningfully in framing, reasoning, or semantic content. We propose a lightweight pairwise discriminator that predicts whether two responses are conceptually distinct. To construct supervision, we generate multiple responses per prompt under controlled stylistic modes and assign labels using bidirectional Natural Language Inference (NLI), treating mutually entailing pairs as non-diverse and all others as diverse. This labeling strategy targets semantic equivalence rather than geometric similarity in embedding space. Our discriminator operates on frozen sentence embeddings with simple pairwise feature interactions, including raw embeddings, element-wise differences, Hadamard products, and length features. On a balanced benchmark of 5,816 response pairs, the proposed model achieves an AUC of 0.843, outperforming all baselines, including Self-ROUGE-L (0.812), Self-BLEU (0.804), Embedding Distance (0.772), and Distinct- n metrics (0.703–0.788). Ablation studies show that element-wise embedding differences provide the strongest individual signal, while the full feature set yields the best AUC. These results suggest that conceptual diversity is better captured by a learned pairwise model than by fixed lexical or geometric heuristics.

1 Project Information

- Mentor: Caroline Choi.
- External collaborators: none.
- Sharing project: no.

2 Introduction

Large language models can often produce many plausible responses to a single open-ended prompt. In practice, this multiplicity is only useful when the candidate answers are not just different on the surface, but genuinely different in content, framing, or reasoning. A system that produces several paraphrases of the same answer is far less useful than one that offers multiple valid perspectives or solution strategies. This makes diversity evaluation important for applications such as candidate reranking, response set construction, and the analysis of model uncertainty. Yet most automatic diversity metrics in natural language generation were designed to capture surface variation. Metrics such as Distinct- n [1] and Self-BLEU [2] are simple and widely used, but prior work has shown that n -gram-based metrics track form diversity more reliably than content diversity [3, 4].

Measuring *conceptual* diversity is difficult because lexical variation and semantic variation do not align cleanly. Two responses may use very different words while making essentially the same claims,

or they may remain lexically similar while differing in argument structure, emphasis, or underlying assumptions. Embedding-based metrics and contextual evaluation methods move closer to semantics [5, 6], but they still mostly produce generic similarity signals rather than directly answering the question we care about: given two responses to the same prompt, are they conceptually different?

In this paper, we cast conceptual diversity as a pairwise classification problem over responses to the same prompt. Our key idea is to train a lightweight discriminator using synthetic supervision rather than costly human annotation. We begin with open-ended instruction prompts from Stanford Alpaca [7], generate multiple responses per prompt under controlled styles and random seeds, and use bidirectional natural language inference to assign weak labels indicating whether two responses mutually entail one another. This setup is inspired by the broader success of synthetic data generation in instruction-following work [8] and by the use of NLI as a formal framework for semantic equivalence and divergence [9, 10].

Our contributions are threefold. First, we build a same-prompt response-pair benchmark targeted at conceptual diversity rather than lexical novelty. Second, we introduce an NLI-based weak-labeling pipeline for this task. Third, we show that a small pairwise discriminator over frozen sentence embeddings outperforms lexical, overlap-based, and embedding-distance baselines, achieving 0.843 AUC, 0.753 Accuracy, and 0.773 F1 on the held-out test set.

3 Related Work

Automatic diversity evaluation in text generation has traditionally been dominated by surface-form metrics. Distinct- n measures the proportion of unique n -grams in model outputs [1], while Self-BLEU evaluates diversity by measuring overlap among generations, with lower overlap indicating greater diversity [2]. Related signals can also be derived from BLEU and ROUGE [11, 12]. Although these metrics are useful for detecting repetition, they operate mainly on wording rather than meaning. Prior work shows that such metrics capture form-level variation better than genuine content diversity [3]. More semantic approaches use sentence embeddings or contextual token matching, such as Sentence-BERT, BERTScore, and Sem-Ent [5, 6, 4], but these are generally designed as similarity scores or set-level diversity measures.

Our work instead focuses on a same-prompt pairwise decision problem: determining whether two specific responses are conceptually different. To build supervision, we draw on synthetic data generation and natural language inference. Self-Instruct and Stanford Alpaca show that language models can produce useful instruction-following data at scale [8, 7], while NLI benchmarks such as SNLI and MultiNLI formalize semantic relations between text pairs [9, 10]. We combine these ideas by generating controlled response pairs and labeling them with bidirectional NLI. The result is a lightweight learned metric tailored to conceptual diversity, rather than lexical variation or general semantic similarity.

4 Approach

Our goal is to learn a metric for conceptual diversity between two responses to the same prompt. Existing automatic diversity metrics primarily operate on surface form: they measure n -gram novelty, lexical overlap, or geometric distance in embedding space. While useful, such metrics do not directly capture whether two responses differ in their underlying framing, reasoning strategy, or semantic content. We instead train a lightweight pairwise discriminator that predicts whether two responses are conceptually different.

4.1 Data Generation

We construct a controlled response pool from open-ended prompts. We sample 2,000 prompts from Stanford Alpaca after filtering out prompts where diversity is either ill-defined or unlikely to be meaningful, such as factoid, yes/no, code-generation, and math questions. This filtering focuses the dataset on prompts that admit multiple valid response strategies, tones, or lines of reasoning.

For each prompt, we generate responses using Qwen3-30B-A3B [13] served through vLLM [14], with temperature 0.7, top-p 0.9, and a maximum generation length of 512 tokens. We prompt the model under five generation modes intended to induce distinct but still high-quality answer styles:

formal-analytical, casual-conversational, critical-skeptical, creative-analogical, and step-by-step. For each mode, we sample two responses with different random seeds. This yields 10 responses per prompt and 20,000 responses in total.

This controlled generation procedure produces naturally varying pairs that share the same prompt semantics while differing in tone, framing, and structure. At the same time, the use of multiple seeds within the same mode provides a source of paraphrase-like or near-paraphrase examples, which are valuable negatives for training a classifier that should distinguish conceptual diversity from mere lexical variation.

4.2 Pair Construction

From the generated response pool, we construct candidate response pairs in two categories.

First, we form cross-mode pairs, consisting of responses to the same prompt generated under different modes. To reduce combinatorial explosion and keep pair construction consistent, we use the first seed for each mode when constructing these pairs. For a prompt with $K = 5$ modes, this yields up to $\binom{5}{2} = 10$ cross-mode pairs, or $10P$ candidate pairs over P prompts.

Second, we form same-mode pairs, consisting of two responses generated under the same mode but with different random seeds. These pairs are often paraphrastic or only weakly divergent, and they serve as a useful source of non-diverse examples. Since there are five modes, this yields up to $5P$ same-mode candidate pairs.

This design provides two qualitatively different sources of supervision. Cross-mode pairs tend to contain meaningful differences in framing or reasoning, while same-mode pairs provide harder negatives that prevent the classifier from simply using topic mismatch or prompt identity as a shortcut.

4.3 NLI-Based Label Assignment

A central design choice in our label construction pipeline is the use of a Natural Language Inference model rather than embedding similarity thresholds. Our objective is to measure whether two responses are conceptually different, so the labeling procedure should reflect semantic equivalence rather than geometric proximity in a sentence-embedding space.

To assign labels, we use bidirectional textual entailment with the pretrained model `cross-encoder/nli-deberta-v3-large` [15]. For each candidate pair (y_a, y_b) , we truncate each response to its first 200 tokens. This truncation is motivated by the observation that high-level framing usually appears early in the response, and it ensures that the pair fits within the NLI model’s context window.

We then run the NLI model in both directions:

$$\text{NLI}(y_a, y_b), \quad \text{NLI}(y_b, y_a), \quad (1)$$

where each inference predicts one of entailment, neutral, or contradiction.

We assign the binary diversity label as follows:

$$\ell(y_a, y_b) = \begin{cases} 0 & \text{if both directions predict entailment} \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

If both responses mutually entail one another, they express the same claims and are treated as non-diverse. If either direction is neutral or contradictory, then the two responses differ in at least part of their semantic content or framing and are treated as diverse.

To reduce label noise, we apply confidence filtering. If the maximum softmax probability is below 0.7 in either NLI direction, we drop the pair from the dataset. This removes ambiguous cases where the entailment model is uncertain.

After labeling, we retain both positive and negative examples from both pair categories. The resulting dataset includes cross-mode pairs labeled diverse, cross-mode hard negatives labeled non-diverse because they converge to the same content despite stylistic differences, same-mode paraphrase negatives, and same-mode diverse pairs where seed variation leads to genuinely different arguments.

We then balance the dataset and split by prompt to avoid leakage. The final dataset contains 4,064 training pairs, 890 validation pairs, and 862 test pairs.

4.4 Pairwise Discriminator

Given two responses y_a and y_b , we first encode each response using a frozen sentence encoder, `all-mpnet-base-v2` [16], producing embeddings

$$u = e(y_a), \quad v = e(y_b), \quad u, v \in \mathbb{R}^{768}. \quad (3)$$

We then construct a feature vector

$$z = [u \parallel v \parallel |u - v| \parallel u \odot v \parallel \phi(y_a, y_b)], \quad (4)$$

where $\parallel$ denotes concatenation, $|u - v|$ is the element-wise absolute difference, $u \odot v$ is the element-wise Hadamard product, and $\phi(y_a, y_b)$ is a small set of scalar length features. Specifically, ϕ includes the token length of each response and the absolute difference in their lengths. This yields a final input dimension of 3075.

This representation combines several complementary signals. The raw embeddings u and v preserve information about each response individually. The absolute difference emphasizes divergence between responses, while the Hadamard product captures agreement patterns. The length features allow the model to account for stylistic differences such as verbosity, which may correlate weakly with response mode.

The discriminator is a multilayer perceptron:

$$\text{MLP}(3075 \rightarrow 512 \rightarrow 128 \rightarrow 1), \quad (5)$$

with ReLU activations, dropout 0.2 after each hidden layer, and a sigmoid output.

4.5 Training

We train the discriminator as a binary classifier using binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -[y \log p + (1 - y) \log(1 - p)], \quad (6)$$

where $y \in \{0, 1\}$ is the ground-truth label and p is the predicted probability.

We optimize with AdamW using learning rate 1×10^{-3} , weight decay 10^{-4} , and batch size 256. We train for up to 20 epochs with early stopping (patience 3) based on validation AUC and use `ReduceLROnPlateau` with factor 0.5 and patience 2.

5 Experiments

5.1 Data

Our evaluation dataset is constructed using the controlled response-generation pipeline described in Section 4.1, the pair-construction procedure in Section 4.2, and the NLI-based labeling method in Section 4.3. Starting from 2,000 filtered open-ended prompts from Stanford Alpaca, we generate 20,000 total responses and form same-prompt response pairs for conceptual-diversity classification. Given $P = 2,000$ prompts, this construction yields up to $10P$ cross-mode pairs and $5P$ same-mode pairs, for a total of 30,000 candidate pairs before NLI filtering and class balancing.

After confidence filtering and class balancing, the final dataset contains 5,816 labeled pairs (2,908 diverse, 2,908 non-diverse). We split the dataset by prompt into train, validation, and test partitions in a 70/15/15 ratio, yielding 4,064 training pairs, 890 validation pairs, and 862 test pairs. Splitting by prompt rather than by pair prevents leakage from the same response pool into multiple splits.

The resulting dataset contains both relatively easy and difficult examples. Easy positives are typically cross-mode pairs with clearly different semantic framing, while easy negatives are paraphrastic same-mode pairs. More challenging cases include cross-mode hard negatives, where stylistically different responses are still logically equivalent, and same-mode diverse pairs, where two responses generated under the same mode nevertheless diverge conceptually.

5.2 Evaluation method

We formulate conceptual diversity prediction as a binary classification task over pairs of responses to the same prompt. For each method, we compute a scalar diversity score for every response pair and evaluate performance in both threshold-free and thresholded settings.

Our primary metric is **AUC**, which measures ranking quality independent of any classification threshold. We emphasize AUC because conceptual diversity is naturally graded rather than strictly binary, and ranking quality is especially relevant for downstream applications such as reranking candidate responses.

We also report **Accuracy** and **F1** after converting each scalar score into a binary prediction. For every baseline and for our discriminator, we select the decision threshold on the validation set by maximizing Youden’s J statistic (sensitivity + specificity -1), then apply that threshold to the held-out test set.

We compare our model against six baselines spanning lexical, overlap-based, and embedding-based diversity signals. Distinct-1, Distinct-2, and Distinct-3 measure the proportion of unique n -grams in the concatenated response pair. Self-BLEU measures symmetrized BLEU overlap between the two responses, where lower overlap indicates greater diversity. Self-ROUGE-L measures ROUGE-L overlap in the same way. Embedding Distance (EmbDist) is defined as

$$1 - \cos(u, v), \tag{7}$$

where u and v are frozen sentence embeddings of the two responses. Higher values indicate greater separation in embedding space.

Beyond topline evaluation, we also perform stratified analysis by pair type, stratified analysis by cosine-similarity range, correlation analysis with baseline metrics, a length-confound analysis, and feature ablations.

5.3 Experimental Details

The discriminator is implemented in PyTorch and trained using the setup described in Section 4.5. We use `all-mpnet-base-v2` as the sentence embedding model and optimize the classifier with AdamW. Full training hyperparameters are provided in Appendix B.5.

The classification threshold used for accuracy and F1 is selected on the validation set by maximizing Youden’s J statistic. Final evaluation reports AUC, accuracy, and F1 on the held-out test set.

5.4 Results

Table 1 reports the main test-set results.

Method	AUC	Accuracy	F1
Distinct-1	0.703	0.591	0.706
Distinct-2	0.777	0.688	0.752
Distinct-3	0.788	0.712	0.761
Self-BLEU	0.804	0.733	0.760
Self-ROUGE-L	0.812	0.745	0.772
EmbDist	0.772	0.707	0.745
Discriminator (ours)	0.843	0.753	0.773

Table 1: Main results on the NLI-labeled test set. Our discriminator achieves the best AUC, accuracy, and F1 among all methods.

The learned discriminator achieves the strongest overall performance, with an AUC of 0.843 (Figure 1). This exceeds the best baseline, Self-ROUGE-L, by 3.1 points and EmbDist by 7.1 points. The discriminator also achieves the highest Accuracy and F1, indicating that it is not only a better ranker but also competitive as a thresholded classifier.

A notable finding is that EmbDist is not the strongest baseline under our evaluation setup. Its AUC of 0.772 falls below both Self-BLEU and Self-ROUGE-L, suggesting that cosine distance alone is not sufficient to capture conceptual differences between responses. In particular, two responses may remain close in embedding space while differing in framing, reasoning style, or semantic emphasis.

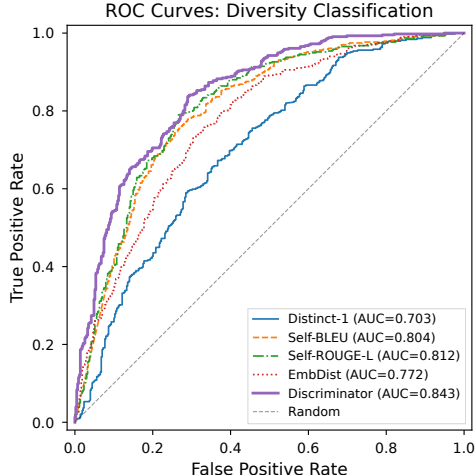


Figure 1: ROC curves for the discriminator and selected baselines on the test set. The discriminator (AUC = 0.843) consistently dominates all baselines across operating points.

The Distinct- n baselines perform worst overall, reinforcing the intuition that lexical novelty is a weak proxy for conceptual novelty. Overlap-based metrics such as Self-BLEU and Self-ROUGE-L are more competitive, likely because they better capture redundancy between paired responses, but they still fall short of the learned discriminator.

We next investigate which parts of the discriminator’s feature representation are responsible for these gains through an ablation study.

5.5 Ablation study

To understand which components of the pairwise feature representation drive performance, we train several ablated versions of the discriminator while keeping the overall training setup fixed. In each case, we use the same optimization procedure described in Section 4.5, scaling the input layer to match the feature dimension of the ablated representation.

Table 2 reports the results. The full model achieves the best overall AUC at 0.843. Removing the length features causes only a small drop to 0.834, indicating that response length carries a modest but non-dominant signal. Using only the raw embeddings and their absolute difference remains strong at 0.831, while the variant with raw embeddings and Hadamard-product features performs notably worse at 0.791.

Condition	Dim	AUC	Accuracy	F1
Full	3075	0.843	0.753	0.773
No length	3072	0.834	0.754	0.771
With diff	2304	0.831	0.759	0.766
With hadamard	2304	0.791	0.718	0.706
Diff only	768	0.793	0.719	0.720
Hadamard only	768	0.783	0.689	0.639
Embeddings only	1536	0.784	0.722	0.730

Table 2: Ablation results for the discriminator feature representation.

The ablation results show that the element-wise absolute difference $|u - v|$ is the most informative individual feature family. Adding $|u - v|$ to the raw embeddings improves AUC from 0.784 to 0.831, a gain of 4.7 points, whereas adding the Hadamard product $u \odot v$ improves AUC only from 0.784 to 0.791. This suggests that explicit divergence information is substantially more useful than agreement information for conceptual-diversity prediction.

The *diff only* model is also informative. Since it uses only $|u - v|$, it can be viewed as a nonlinear analog of embedding-distance-based scoring. Its AUC of 0.793 already exceeds the EmbDist baseline

(0.772), indicating that a learned nonlinear decision rule over embedding differences is more effective than cosine distance alone.

At the same time, the full model still outperforms the diff-only variant by 5.0 AUC points. This gap shows that the best performance comes from combining multiple pairwise signals rather than relying on any single feature family. Raw embeddings preserve information about each response individually, the Hadamard product contributes complementary agreement structure, and the length features provide a small but measurable gain. Together, these results support the view that conceptual diversity is captured by feature interactions that fixed heuristic metrics cannot express.

6 Analysis

We further analyze what the discriminator captures beyond existing metrics and where its gains appear most clearly.

Across pair types, the model assigns the highest scores to clearly diverse cross-mode pairs and the lowest scores to paraphrase negatives, with harder cases such as same-mode diverse pairs and cross-mode hard negatives occupying the middle range (Table 3). This suggests that the model is not simply responding to stylistic differences, but is learning a more nuanced notion of conceptual divergence. Figure 2 visualizes the overall score distributions by true label.

Pair Type	Count	Mean Disc Score
Cross-mode diverse	321	0.620
Same-mode diverse	113	0.523
Paraphrase negative	276	0.373
Hard negative	152	0.447

Table 3: Mean discriminator scores by pair type on the test set.

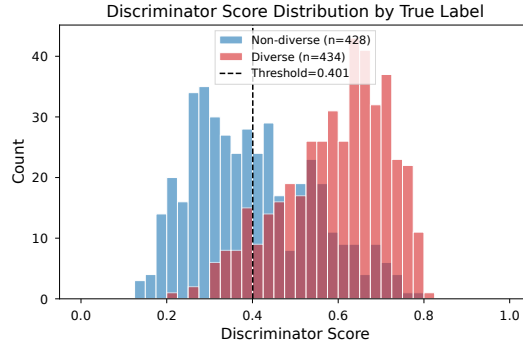


Figure 2: Distribution of discriminator scores by true label on the test set. Diverse pairs (label = 1) cluster toward higher scores while non-diverse pairs cluster toward lower values, with moderate separation around the decision threshold.

We also stratify performance by cosine-similarity range and compare against EmbDist within each bucket. The results are shown in Table 4.

Similarity Range	Count	Disc AUC	EmbDist AUC
Low sim (< 0.747)	216	0.758	0.729
Mid sim $[0.747, 0.941]$	430	0.780	0.601
High sim $[0.941, 1.0]$	216	0.817	0.537

Table 4: Discriminator vs. EmbDist AUC by cosine-similarity range. In the high-similarity bucket, EmbDist values are confined to a ~ 0.06 -wide interval and approach chance-level discrimination.

Within the high-similarity bucket, all EmbDist values lie in $(0, 0.059]$, leaving almost no variance with which to rank pairs. The discriminator, by contrast, achieves its best AUC (0.817) precisely in

this regime, confirming that it captures framing and reasoning differences that geometric distance cannot detect.

Correlation analysis shows that the discriminator is moderately correlated with existing baselines, especially EmbDist and Self-ROUGE-L, but is far from redundant with any single metric (see Appendix B.3). The strongest correlations are with EmbDist ($\rho = 0.693$) and negated Self-ROUGE-L ($\rho = 0.690$), while correlations with lexical metrics such as Distinct-1 are substantially lower ($\rho = 0.469$).

Finally, length-controlled evaluation suggests that response length is not the primary driver of performance. Although discriminator scores are weakly correlated with absolute length difference ($\rho = 0.267$, $p < 10^{-6}$), performance on a length-controlled subset of 257 pairs remains strong, achieving AUC 0.850 (see Appendix B.4).

To illustrate where the discriminator succeeds and embedding distance fails, consider a representative disagreement from the test set. Given the prompt “*Describe the benefits of exercising regularly,*” one response (formal-analytical mode) presents standard physiological benefits, cardiovascular health, endorphin release, cognitive improvement, while the other challenges the premise, arguing that benefits depend heavily on exercise type, intensity, and individual context. These responses share enough topical vocabulary to remain close in embedding space (EmbDist = 0.031), yet they differ fundamentally in framing: one endorses conventional wisdom, the other interrogates it. The discriminator assigns a score of 0.72, correctly identifying diversity, while embedding distance places the pair well below its decision threshold. In contrast, two casual-mode responses to the same prompt that use very different language but make identical claims receive a low discriminator score (0.29) despite a moderate embedding distance (EmbDist = 0.14). These cases show that the discriminator learns to distinguish conceptual content from surface variation in both directions.

These results suggest that the model captures many framing-level differences that heuristic metrics miss, though it still struggles with some discourse-level mismatches such as direct-answer versus meta-explanation pairs.

7 Conclusion

This paper presented a lightweight method for measuring conceptual diversity between two responses to the same prompt. By pairing controlled generation with bidirectional NLI labeling, we convert an otherwise subjective evaluation problem into a scalable supervised learning task. The resulting discriminator outperforms lexical, overlap-based, and embedding-distance baselines on this NLI-derived benchmark, indicating that conceptual diversity can be captured more effectively by a learned pairwise model than by surface-form heuristics alone.

At the same time, the study has clear limitations. Our labels are derived from an NLI model rather than direct human annotation, so the benchmark inherits some model bias and label noise. In addition, our data is restricted to English and open-ended prompts. Future work should validate the benchmark against human judgments, extend the setting to multilingual and multi-turn responses. Another promising direction is to move beyond binary classification toward a calibrated diversity score that jointly captures quality and conceptual distinctness.

Team contributions

James Liu helped formulate the project direction and experimental design, including the controlled response-generation setup, pair-construction pipeline, and NLI-based weak-labeling procedure. He also contributed to dataset preparation, baseline setup, and writing the report.

Hangoo Kang implemented the pairwise discriminator, training pipeline, and evaluation framework, including the main experiments, ablation study, and analysis code. He also contributed to interpreting the results and writing the report.

References

- [1] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the*

- North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics.
- [2] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Tegygen: A benchmarking platform for text generation models. *arXiv preprint arXiv:1802.01886*, 2018.
 - [3] Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online, April 2021. Association for Computational Linguistics.
 - [4] Seungju Han, Beomsu Kim, and Buru Chang. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
 - [5] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.
 - [6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.
 - [7] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models, 2023.
 - [8] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.
 - [9] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
 - [10] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
 - [11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.
 - [12] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
 - [13] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
 - [14] 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. *arXiv preprint arXiv:2309.06180*, 2023.

- [15] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023.
- [16] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, 2020.

A Appendix

B Extended Analysis

B.1 Analysis by Pair Type

See Table 3 in the main text.

B.2 Analysis by Cosine-Similarity Range

Similarity Range	Count	Disc AUC	EmbDist AUC
Low sim (< 0.747)	216	0.758	0.729
Mid sim $[0.747, 0.941)$	430	0.780	0.601
High sim $[0.941, 1.0)$	216	0.817	0.537

Table 5: Discriminator vs. EmbDist performance across cosine-similarity ranges. See also Table 4 in the main text.

B.3 Correlation with Baseline Metrics

Metric	Spearman ρ
Distinct-1	0.469
Distinct-2	0.653
Self-BLEU (negated)	0.660
Self-ROUGE-L (negated)	0.690
EmbDist	0.693

Table 6: Correlation between discriminator scores and baseline metrics.

B.4 Length Confound Analysis

Metric	Value
Disc score vs. length diff (Spearman ρ)	0.267
Disc score vs. length diff (p-value)	$< 10^{-6}$
Length-controlled subset size	257
Length-controlled AUC	0.850

Table 7: Length confound analysis.

B.5 Training Configuration

Embedding model	all-mpnet-base-v2
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Batch size	256
Max epochs	20
Early stopping	Validation AUC (patience 3)
LR scheduler	ReduceLROnPlateau (factor 0.5, patience 2)

Table 8: Training configuration for the discriminator.