Revisiting the Homogeneity of Large Language Models with a Multi-dimensional Textual Similarity Framework

Anonymous ACL submission

Abstract

Text embedding similarity is the default proxy for measuring LLM response similarity, but reducing it to a single scalar hides differences across linguistic aspects. We propose a multi-dimensional framework that decomposes response similarity into six linguistic dimensions: surface, structural, lexical, syntactic, semantic, and stylistic. Each dimension is instantiated by interpretable metrics, yielding 24 in total. Across three validation studies, our metrics and embeddings are complementary, capturing different aspects of pair-level similarity and human indistinguishability. Applying the framework to 32 models spanning open and closed families, we find that homogenization depends strongly on the dimension: lexical and syntactic similarity is already high across nearly all models, while surface and structural similarity vary widely and depend on the model family. Tracking similarity across training stages further shows that lexical similarity is mostly inherited from pretraining, structural patterns are set at SFT, and later alignment stages add only small, gradual changes.

1 Introduction

Text Embedding-based similarity has long been used as a practical proxy for semantic relatedness across several NLP tasks, including dense retrieval and RAG (Karpukhin et al., 2020; Lewis et al., 2020), semantic textual similarity (Reimers and Gurevych, 2019), and creativity evaluation (Johnson et al., 2023; Wenger and Kenett, 2025). More Recently, motivated by the model-collapse hypothesis, which posits that LLM responses may converge due to shared training targets and overlapping training data (Shumailov et al., 2024; Seddik et al., 2024; Padmakumar and He, 2024), embeddings have also been adopted to assess homogeneity among model outputs (Kirk et al., 2024; Guo et al., 2025; Jiang et al., 2025).

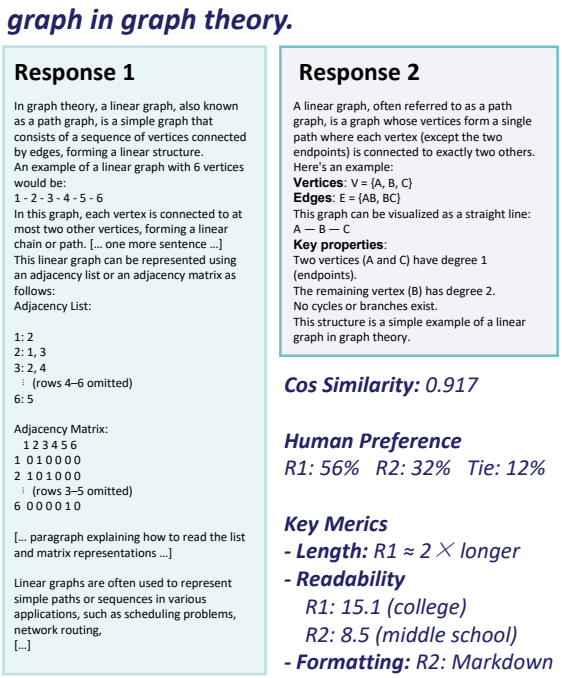

Figure 1: A motivating example from Infinity-Chat-Eval (Jiang et al., 2025), with embedding similarity (Qwen3-8B-Embedding) and human preference scores aggregated by 25 annotators. Although the two responses share a high embedding similarity, they differ markedly in presentation style and receive distinctly different human ratings (full example in Appendix B).

However, text embeddings alone provide insufficient evidence. By reducing response similarity to a single scalar score, they offer limited interpretability (Zhong et al., 2022; Leiter et al., 2024) and cannot reveal which linguistic dimensions drive homogenization, which aspects remain diverse, or at what stage such similarities are introduced during model training. More fundamentally, widely used embedding models such as GTE (Li et al., 2023), BGE (Xiao et al., 2024), and Qwen3-Embedding (Zhang et al., 2025a) are mainly designed for semantic matching and retrieval. While effective for assessing topical similarity (e.g., between a web page title and its body, or a question and related pas-

sages), they are less suited to capturing similarities in form, style, or linguistic structure (Patel et al., 2025). As a result, a high embedding similarity score often reflects shared topics or broad meaning, rather than true textual homogeneity. Figure 1 illustrates this gap: even when paired responses differ significantly in linguistic features and human evaluations, embedding-based similarity can still rate them as highly similar.

Thus, what is needed instead is an interpretable framework that decomposes similarity into different linguistic dimensions. By making each dimension separately observable, such a framework can identify not only whether responses are becoming more similar, but also how and where this similarity emerges. Beyond this, the framework also offers a general tool for analyzing the linguistic characteristics of LLMs, for instance, how AI reshapes human writing through interaction (Liang et al., 2024; Lu et al., 2025; Abdulhai et al., 2026).

Drawing on linguistic approaches to text analysis, we introduce a comprehensive framework for evaluating LLM responses, quantifying pairwise similarity with 24 interpretable metrics spanning six dimensions: surface, structural, lexical, syntactic, semantic, and stylistic. After validating the framework through three comparative studies, we use it to analyze 32 models across open- and closed-source families. Experimental results show that response homogenization is strongly dimension-dependent: lexical and syntactic similarity is already high at the base model stage, while surface and structural similarity varies widely and is largely set by supervised fine-tuning. Our metrics capture form, style, and structure signals that are complementary to embeddings, and stage-wise analysis reveals that the main convergence occurs at the Base→SFT/Instruct transition, with later alignment stages contributing only gradual changes. This decomposition clarifies how and when different aspects of model outputs become similar, providing a more nuanced diagnosis of LLM homogenization than embedding-based measures alone.

2 Framework: Multi-dimensional Similarity Metrics

In this section, we begin by introducing the design of the multi-dimensional evaluation framework, then compare it with embedding-based measurement to illustrate what may be overlooked when relying on embeddings alone.

Dimension	#	Metrics
Surface	3	Response Length, Punctuation Profile, Punctuation Density
Structural	8	# Paragraphs, # Bullet-like Lines, Bold Emphasis Ratio, Opening Template, Closing Template, Bullet Points, Headers, Bold Emphasis
Lexical	2	Lexical Diversity (MTLD), Word Frequency (Zipf)
Syntactic	3	POS Distribution, Dependency Distribution, Function Word Use
Semantic	4	BLEU, ROUGE-L, InfoMatch (Sent), Embedding Similarity
Stylistic	4	Readability, Concreteness, Emotion Profile, Emotion Intensity
Total	24	

Table 1: Summary of the six dimensions and their constituent metrics.

2.1 Multi-dimensional Similarity Metrics

The framework (see Table 1) is inspired by traditional linguistic studies, which decompose language into form, structure, and meaning—a decomposition that maps naturally onto **lexical, syntactic, and semantic** layers. We further extend this scheme with two dimensions that capture the visible formatting character of modern LLM outputs: a **surface** dimension covering low-level textual signals (response length and punctuation patterns), and a **structural** dimension covering the adoption and intensity of Markdown formatting conventions (e.g., bullets, headers, bold emphasis). Together these two dimensions constitute a major characteristic of model language and a key source of divergence between human and model outputs. Stylistic features (e.g., affect) are also considered: because these models are designed primarily for chat, the manner in which information is conveyed matters substantially.

Within each dimension, metrics are selected according to an explicit principle: **surface** metrics capture coarse-grained textual volume and punctuation distributions; **structural** metrics combine binary indicators (whether a formatting pattern is adopted) with continuous ones (intensity of use); **lexical** metrics target type-level word-choice statistics; **syntactic** metrics are uniformly distributional, computed as Jensen–Shannon divergences over grammatical categories; **stylistic** metrics draw on affect, concreteness, and readability as the three most consistently validated psycholinguistic axes; and **semantic** metrics combine reference-based overlap (BLEU, ROUGE, information-match F1)

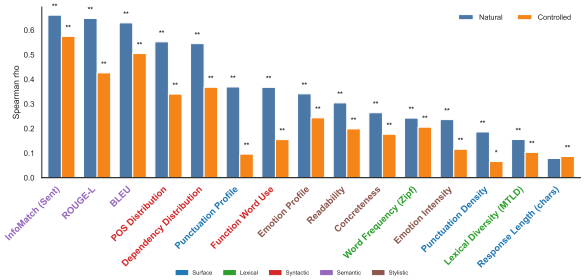

Figure 2: Spearman correlation (ρ) between embedding similarity and each metric, on the natural dataset vs. controlled writing prompts. Structural metrics (e.g., Bold Emphasis Ratio, are omitted under control since they do not vary across writing-task responses. Significance: * $p < 0.05$, ** $p < 0.01$.

with embedding-based similarity, enabling us to triangulate where embedding measurement succeeds and where it systematically fails.

Appendix A describes how metrics are extracted, how similarities are computed, and how dimension scores are aggregated. We also compute pairwise correlations between metrics to verify that no two metrics are empirically redundant (Appendix A.3).

2.2 What Embeddings Capture and What They Miss

Experimental Settings. We use Spearman's rank correlation to measure how embedding-based similarity correlates with each metric in our multidimensional framework across response pairs.

For the embedding model, we use `Qwen3-Embedding-8B` (Zhang et al., 2025a), one of the leading open-source embedding models, which achieves strong performance on the MTEB (English) leaderboard and ranks first on the STS task (Muennighoff et al., 2023). In addition to Qwen3-Embedding, we include `StyleDistance` (Patel et al., 2025), a content-independent style-embedding model trained with synthetic parallel examples to capture writing-style similarity while reducing content leakage.

For the datasets, we employ two datasets with complementary emphases.

(1) Natural AI responses. Jiang et al. (2025) introduce a large collection of open-ended questions (referred to as "INFINITY"), drawn from the real-world human–AI interactions in WildChat (Zhao et al., 2024). They further propose a open-ended question taxonomy with which the questions are annotated. They also provide an annotated subset (referred to as "INFINITY-EVAL") of 50 questions, each paired with two model-generated responses and evaluated by 25 independent human raters, yielding 500 annotated pairs.

Form	n	Semantic		Style	
		Mean	Std	Mean	Std
opening_template	300	0.9294	0.0362	0.9555	0.0366
closing_template	300	0.9534	0.0244	0.9776	0.0225
bullet_points	300	0.8901	0.0874	0.8998	0.0844
headers	300	0.9399	0.0429	0.8941	0.0649
emphasis	300	0.9426	0.0495	0.9100	0.0518

Table 2: Cosine and StyleDistance similarity between form-bearing responses and content-matched bases. Both embeddings assign uniformly high similarity once content is fixed.

(2) Controlled AI responses. To examine whether embeddings' explanatory power decreases when content is held fixed, we sample 100 creative-writing and 100 creative-ideation questions from INFINITY, targeting linguistic and structural metrics respectively. For each question, we generate a base response and controlled variants with GPT-5.4-mini: writing-task variants are stylistic rewrites of a sentence-level template (preserving sentence-level content while broadening linguistic features), and ideation-task variants are produced by adding targeted form elements to a form-free base response. Full generation prompts are in Appendix C.

Relationships between Metrics and Embeddings. Figure 2 compares, for each metric, the Spearman ρ between embedding similarity and metric similarity on the natural dataset (blue bars) versus the controlled writing prompts (orange bars), from which two observations emerge. **(1) Embeddings correlate strongly only with semantic metrics.** The three semantic metrics retain the tallest bars under both settings, with ρ remaining in 0.43–0.57 even after content is controlled. The two distributional syntactic metrics (POS and Dependency) also reach high correlations on the natural dataset ($\sim$0.55) but shrink to $\sim$0.35 under control, and Function Word Use drops further (from 0.37 to 0.15), indicating that a substantial part of the apparent syntactic alignment is driven by content rather than by syntactic structure itself. **(2) Embeddings largely fail to capture the surface, lexical and stylistic dimensions.** The corresponding bars are uniformly short already on the natural dataset (all $\rho < 0.4$) and most shrink further once content is controlled, confirming that these non-semantic dimensions are often overlooked by embeddings even when they vary across responses.

For the **binary form metrics** in solution ques-

Data	Feature setting	Macro-F1
Solution	Qwen3 Emb.	0.815
Solution	StyleDistance	0.492
Solution	Metrics	0.888
Solution	Metrics + Qwen3 Emb.	**0.914**
Solution	StyleDistance + Qwen3 Emb.	0.867
Writing	Qwen3 Emb.	0.546
Writing	StyleDistance	0.341
Writing	Metrics	0.843
Writing	Metrics + Qwen3 Emb.	**0.914**
Writing	StyleDistance + Qwen3 Emb.	0.635

Table 3: Macro-F1 of response-pair classification on Solution (cat1–cat3) and Writing (cat1–cat4). We compare Qwen3 embedding features, StyleDistance, our multidimensional metrics, and their combinations with Qwen3 embeddings. Robustness results with additional embedding backbones are reported in Appendix E.

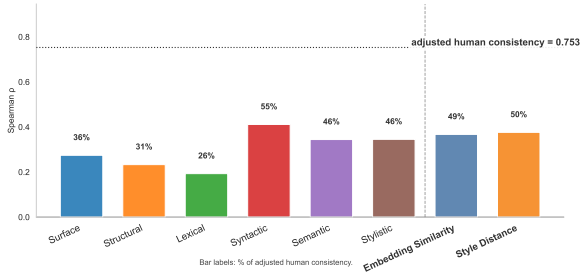

Figure 3: Spearman correlation between dimension-level similarity and the pairwise tie rate ($n = 500$). Bar labels report ρ as a percentage of the split-half human consistency upper bound ($\rho_{\text{SB}} = 0.753$, dotted line). Semantic Match excludes embedding similarity; Bootstrap validation results can be found in Table 12. Significance: $^{*}p < 0.05$, $^{**}p < 0.001$.

tions, where ρ is uninformative, we instead inspect the similarity scores directly. Table 2 reports both semantic and style similarities between each form-bearing response and its content-matched base. Both measures yield uniformly high similarity, indicating that neither generic semantic embeddings nor style-specialized embeddings capture these surface differences.

3 Validation of Multi-dimensional Metrics

We further validate the framework's effectiveness through two experiments: its ability to distinguish response pairs, and its relationship with human quality scores.

3.1 Classifying Response-Pair Types

Experimental Settings. We evaluate whether our metrics capture linguistic similarity beyond existing similarity representations by formulating response-pair comparison as a classification task. We construct four pair types: multiple samples from the same model and prompt (cat1), responses to the same prompt from different models (cat2), responses from the same model to different prompts (cat3), and style-controlled rewrites of the same prompt and model response (cat4). We train logistic-regression classifiers with five feature settings: Qwen3 embedding-only features, StyleDistance, our multidimensional metrics without embeddings, our metrics combined with Qwen3 embeddings, and StyleDistance combined with Qwen3 embeddings. Robustness results with additional embedding backbones and per-category results are reported in Appendix E.

Metrics and Embeddings Are Complementary for Response-Level Similarity. Table 3 reports the Macro-F1 results. Metrics-only consistently outperforms embedding-only, especially on Writing task (0.82–0.88 vs. 0.55–0.59), where response content is held roughly constant. Combining metrics with embeddings achieves the best performance in nearly every setting, reaching up to 0.915, which supports the complementary role of semantic embeddings and our multi-dimensional linguistic metrics.

The per-category breakdown in Table 11 clarifies this complementarity. Embedding-only excels on cat3 ($F_1 \approx 0.97$), where prompts differ and content cues are informative, but collapses to 0.21 on cat4, where stylistic rewrites preserve semantic content. Metrics-only remains strong across all categories (≥ 0.77), capturing model-specific linguistic patterns beyond content. The combined setting reaches at least 0.84 in nearly every category: embeddings dominate when content varies, while our metrics are more reliable when content is fixed.

3.2 Predicting Human Indistinguishability

Experimental Settings. To test whether our metrics capture human perception of similarity, we use the INFINITY-EVAL data introduced in Section 2.2, in which 25 annotators provide pairwise preference judgments for each response pair. We adopt the *tie rate*—the proportion of zero labels across the 25 annotators—as a behavioral proxy for perceived similarity, since indistinguishable pairs more often elicit ties than clear preferences. For each pair, we compute similarity under every metric and aggregate the values into dimension-level scores, then

compute Spearman correlations against the tie rate ($n = 500$). Details can be found in Appendix E.2.

Linguistic metrics complement embedding-based similarity. Figure 3 reports the Spearman correlation between each dimension-level similarity score and the pairwise human tie rate. Syntactic similarity achieves the strongest metric-based correlation ($\rho = 0.41$, 95% CI $[0.34, 0.48]$), comparable to Embedding Similarity ($\rho = 0.37$) and Style Distance ($\rho = 0.38$), despite relying only on POS, dependency, and function-word distributions. Semantic ($\rho = 0.34$) and Stylistic ($\rho = 0.35$) similarities are also competitive with the embedding baselines, while Surface, Structural, and Lexical dimensions remain weaker but significant.

Together, these results suggest that human indistinguishability depends on multiple linguistic cues beyond semantic closeness, including syntax, style, surface form, and lexical choice. Our metrics therefore complement embedding-based similarity while remaining largely orthogonal to embeddings by design: textbfembedding similarity is not incorrect, but incomplete.

4 Where Homogenization Emerges: A Dimension-Wise Diagnosis

We next ask where response homogenization enters the model pipeline and whether it affects all linguistic dimensions equally. To this end, we apply our multi-dimensional framework to a shared set of open-ended questions and compare pairwise *output* similarity across models and training stages.

4.1 Models and Stage-Wise Setup

Our analysis covers 32 models, including 29 open-source models and 3 closed-source models. For several open-source families, including Tülu, OLMo, Qwen, and Gemma, we also include intermediate checkpoints so that changes can be traced across training stages. We sample 300 open-ended questions from INFINITY, balanced across six question categories, and remove prompts with explicit style or format constraints, such as "write in bullet points" or "in Hemingway's style." After filtering invalid or noisy outputs, we retain 225 questions for which every model produces at least 10 valid responses. All models are sampled with temperature 1.0 and top-$p = 0.9$; base models are evaluated under both *pure* and *form* few-shot prompts. Appendix F provides the full model list, sampling details, few-shot prompts, and filtering procedure.

Dimension	Typical $\widetilde{S}_{ij}$	Min	Residual variance
Lexical	0.88–1.00	0.78	Tülu series
Syntactic	0.86–0.96	0.77	`gpt-5.4-nano`
stylistic	0.84–0.97	0.84	Thinking, `gpt-5.4-nano`
Semantic	0.80–0.93	0.77	Closed vs open source
Surface	0.65–0.97	0.50	Thinking vs non-thinking
Structural	0.65–0.93	0.41	Thinking vs non-thinking

Table 4: Per-dimension summary of pair similarity across the 32 final-stage models.

We first examine the final-stage models to characterize the end result of post-training. Figure 4 reports pairwise similarity across all 32 final-stage checkpoints, separately for each linguistic dimension. Each cell compares two models' responses to the same questions and self-normalizes this cross-model similarity by the models' own within-model variation across samples. Thus, values close to 1 indicate that two models are nearly as similar to each other as repeated samples from the same model. More details can be found in in Appendix F.5.

4.2 Homogenization Is Strongly Dimension-Dependent

Figure 4 reveals that homogenization is not a single global property of model outputs. The six heatmaps differ sharply: model pairs that are nearly saturated in one dimension can be only moderately similar in another. Table 4 provides a compact view of these ranges, which fall into three broad bands.

High homogenization (Lexical, Syntactic). Lexical similarity is uniformly very high: nearly every off-diagonal cell sits in $[0.88, 1.00]$, and the lowest values (around 0.78–0.81) are confined to pairs involving the Tülu series. Syntactic similarity is almost as uniform, mostly $[0.86, 0.96]$, with the only consistent dip coming from `gpt-5.4-nano`, whose syntactic similarity to open-source models falls to 0.86–0.89. In other words, when measured at the word-distribution and part-of-speech level, today's models look almost interchangeable, with only modest residual differences attributable to closed-source training.

Mid homogenization (stylistic, Semantic). stylistic similarity occupies an intermediate band of $[0.84, 0.97]$, but with a visible long tail: `Qwen3-30B-A3B-Thinking` and `gpt-5.4-nano` each fall below 0.88 against several mainstream instruct models. Semantic similarity shows the most disperse mid-range pattern: most cells sit in $[0.80, 0.93]$, but `gpt-5.4-nano` consistently drops to 0.77–0.81 when paired with open-source models. Semantic content is therefore the dimension

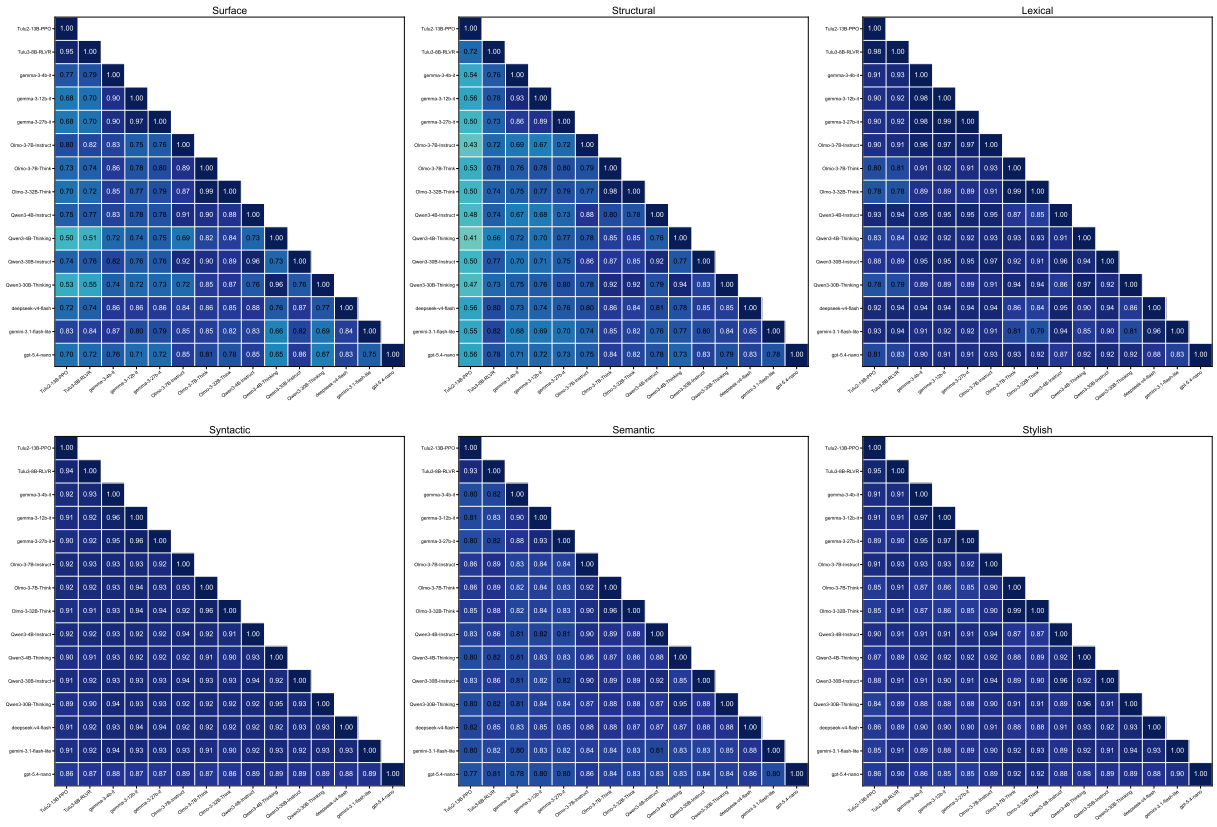

Figure 4: Self-normalized pairwise similarity across 32 final-stage models, shown separately for each of the six linguistic dimensions (Surface, Structural, Lexical, Syntactic, Semantic, Stylistic). Each cell compares two models' responses on the same questions, normalized by their within-model variation across samples; values close to 1 indicate that two models are nearly as similar to each other as repeated samples from the same model. Lexical and Syntactic dimensions are near-saturated, while Surface and Structural exhibit the widest dynamic range, with Thinking models forming a visibly distinct block.

where the open-vs-closed gap is most visible.

Low homogenization (Surface, Structural). Surface and Structural similarity show the widest dynamic ranges: Surface spans $[0.50, 1.00]$, while Structural spans $[0.41, 0.99]$. The lowest cell in the entire figure is `Tulu2-13B-PPO` $\leftrightarrow$ `Qwen3-4B-Thinking` at 0.41 on Structural, indicating that their output shapes remain highly distinct. Taken together with Table 4, these results show that asking whether models are "similar" in the singular is misleading: lexical and syntactic forms have largely converged, while surface formatting and structural fingerprints remain strongly model-family dependent.

4.3 Sub-Structure: Family, Thinking, and Closed-Source Effects

The heatmaps also reveal consistent cluster patterns across dimensions. Although the absolute similarity level changes from one dimension to another, the same types of model pairs repeatedly stand out. Three patterns are most visible.

Same-family models are almost identical. Pairs from the same family and post-training setup produce nearly indistinguishable outputs. The three Gemma-3 instruct checkpoints (4B, 12B, 27B) reach 0.97–0.99 on Surface, Structural, Syntactic, and Stylistic similarity; OLMo-3-7B-Think and OLMo-3-32B-Think reach $0.99/0.98$ on Surface/Structural, and Qwen3-4B-Thinking and Qwen3-30B-A3B-Thinking reach $0.96/0.94$. Scaling within a fixed setup changes the output signature only marginally.

Training regime leaves a dimension-specific fingerprint. Two regimes pull outputs away from the main cluster along different axes. The four Thinking-style models (`Olmo-3-{7B,32B}-Think`, `Qwen3-{4B,30B-A3B}-Thinking`) are tightly grouped on Surface and Structural (0.76–0.99) yet far from non-Thinking models—the most extreme cells pair the two types (e.g., Tülu2-13B-PPO vs. Qwen3-4B-Thinking: 0.41 Structural, 0.50 Surface). The closed-source `gpt-5.4-nano`, by contrast, sits at the lower end of Syntactic, Seman-

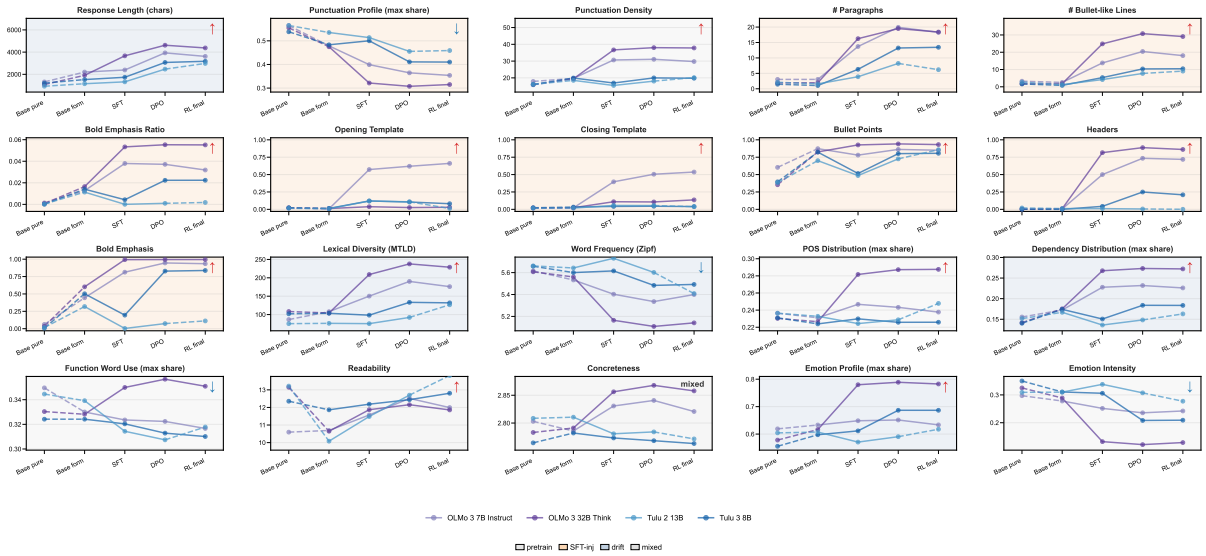

Figure 5: Linguistic metric trajectories across checkpointed SFT → DPO → RLHF/RLVR sub-stages on three models. Top-right annotations report directional consistency (e.g., 6/6 ↑). Subplots are color-framed by linguistic dimension. Post-training induces highly consistent directional shifts, concentrated at the SFT step. Figure 7 shows results from broad-stage transitions (Base pure → Base form → Instruct → Thinking)

tic, and Stylistic while its Lexical similarity stays in range (0.83–0.94): Thinking training leaves a surface/structural signature, whereas closed-source training preserves semantic and stylistic distinctiveness even after lexical convergence.

5 Stage-Wise Trajectories

The snapshot in Section 4.2 shows *what* has converged, but not *when* these similarities emerge. We therefore turn to a stage-wise analysis, tracking outputs from base models through post-training to distinguish similarities inherited from pretraining from those introduced by SFT, DPO, RL, or Thinking-style training.

5.1 Checkpointed Trajectory Setup

We use two checkpoint panels on the same 225-question set. Figure 7 compares Gemma and Qwen3 models across Base pure, Base form, Instruct, and Thinking stages, where applicable, to expose paradigm-level effects. Figure 5 follows four full pipelines (OLMo-3-7B-Instruct, OLMo-3-32B-Think, Tülu 2 13B, and Tülu 3 8B) through Base, SFT, DPO, and RL checkpoints to locate when each effect appears.

5.2 Where Each Dimension Converges

Across the 20 non-pairwise metrics, four recurring patterns explain the dimension-level results, and they correspond to three distinct stages of the training pipeline.

Already similar at base (pretraining layer). Function Word Use, POS Distribution, Punctuation Density, and Emotion Intensity sit within a narrow band at Base (e.g., POS Distribution within 0.22–0.24) and barely move through SFT, DPO, or RL. This explains the near-saturated similarity of the *Lexical* and *Syntactic* dimensions: models look similar there because they were already similar before instruction tuning.

Installed at SFT, one-shot (SFT template layer). Bold, Headers, Opening/Closing Template, Bullet Points, # Paragraphs, and # Bullet-like Lines are near zero at Base, then make their entire jump at SFT (OLMo-3 Headers: 0 → 0.78); DPO and RL barely move them. Final values vary widely across models (Headers spans 0–1.0), so *Surface* and *Structural* similarity is high within a family but low across families, with these differences locked in at SFT.

Slow drift through DPO and RL (preference-drift layer). Word Frequency (Zipf), Lexical Diversity (MTLD), and Dependency Distribution drift gradually across SFT, DPO, and RL in the same direction across models, reinforcing the lexical convergence already present at base.

Thinking branches off on surface metrics. For Response Length, Punctuation Density, Emotion Intensity, and Concreteness, Thinking checkpoints depart from the Instruct trajectory (e.g., ∼5,500 vs. ∼4,000 characters). The branch appears only

on surface-shaped metrics, which is why Thinking models are outliers on *Surface* and *Structural* in Figure 4 but not on *Lexical* or *Syntactic*. A few metrics (e.g., Readability) match none of these patterns and are marked *mixed*.

Together, these patterns show that lexical similarity is inherited from pretraining, structural variance is paradigm-driven (the carrying features only exist after SFT), and within the pipeline structural divergence is locked in at SFT while lexical convergence accumulates through DPO and RL. Appendix H.3 traces the SFT-template and preference-drift layers to two upstream sources: distillation-driven formatting in post-2024 SFT corpora, and the uncalibrated form preferences of generative LLM judges.

6 Related Work

Concerns about model collapse have raised the worry that LLMs may generate increasingly similar and stereotyped content (Shumailov et al., 2024; Seddik et al., 2024), with implications for downstream tasks (Padmakumar and He, 2024; West and Potts, 2025; Gu et al., 2026) and training pipelines that reuse model-generated data (Gerstgrasser et al., 2024; Shumailov et al., 2024; Guo et al., 2024; Dohmatob et al., 2025). These concerns have motivated recent work on response diversity and homogeneity (Zhang et al., 2025b; Guo et al., 2025; Jiang et al., 2025). Because such claims depend critically on how similarity is measured, we focus on the measurement methodology itself.

Existing approaches fall into two broad groups. Embedding-based methods provide scalable similarity estimates and have been applied to large model collections, for example by measuring pairwise similarity across representative models on open-ended prompts from real-world human–AI interactions (Jiang et al., 2025; Zhao et al., 2024). Dimension-specific approaches instead measure lexical, semantic, or syntactic variation using representative metrics. For instance, Kirk et al. (2024) and Guo et al. (2025) combine unique-N with SBERT, and further add entailment- or dependency-based signals. However, these studies typically instantiate each linguistic dimension with a single metric and do not systematically validate the effectiveness. Our framework uses multiple interpretable metrics per dimension and validates them against several experiments (§3).

A related line of work identifies stable linguistic traits of LLM outputs, such as longer words, more fluent prose, and lower lexical variability relative to human writing (Urban et al., 2025; Lee et al., 2025). Although these studies do not directly measure model similarity, the traits they identify suggest a complementary form of homogenization: models converge toward a shared linguistic profile even when individual responses differ. This motivates our use of both pairwise and scalar features.

Finally, several studies examine how diversity changes across training stages, often in the context of the alignment tax (Lin et al., 2024; Padmakumar and He, 2024; Murthy et al., 2025). Some work attributes part of the base–instruct diversity gap to improved quality and the suppression of irrelevant tokens rather than genuine loss of variation (Lake et al., 2025; Shypula et al., 2025). Checkpoint-level analyses further suggest that SFT → DPO → RL/RLVR can progressively compress diversity, with DPO often producing the sharpest drop (Zhang et al., 2025b; Shypula et al., 2025). Our work extends this stage-wise view to 32 models spanning Base, SFT, DPO, RL/RLVR, Instruct, and Thinking checkpoints, and localizes where homogenization emerges across specific linguistic dimensions.

7 Conclusion

We introduced a six dimension linguistic framework with 24 interpretable metrics for diagnosing similarity and homogenization in LLM responses, complementing embedding-based measures with signals that capture form, style, and structure. Applying the framework to 32 models, we find that response homogenization is not a single phenomenon but the joint footprint of three mechanisms operating on disjoint subsets of metrics: (i) a *pretraining layer*, where lexical and syntactic distributions are shared at base and account for the near-saturated similarity on those dimensions at the final stage; (ii) an *SFT template layer*, where formatting features (bullets, headers, bold, opening/closing templates) are injected discontinuously at supervised fine-tuning with family-specific strength, producing high within-family but low across-family structural similarity; and (iii) a *preference-drift layer*, where DPO and RL induce slow co-directional drift that further compresses lexical variation, with Thinking-style training deflecting this drift on a subset of surface metrics. This decomposition explains why the Base→SFT transition dominates homogenization while later alignment stages and parameter scale contribute little.

Limitations

Language, decoding, and model coverage. All experiments are conducted on English prompts drawn from WildChat-derived data, and responses are generated under a single decoding configuration (temperature 1.0, top-p 0.9). Our closed-source coverage is further limited to three lightweight API systems, without access to intermediate checkpoints. Whether the same dimension-level patterns (particularly the SFT-centered structural shift and the form-bias asymmetry between BT and generative judges) generalize to other languages, broader prompt distributions, alternative decoding settings, or larger proprietary models remains an open question.

Coverage and reliability of individual metrics. Although the 24 metrics are grounded in established linguistic theory and span complementary aspects within each dimension, they are not exhaustive; features such as discourse-level coherence and rhetorical structure are not captured. Several metrics additionally depend on external tools (spaCy parsers, the GoEmotions classifier, a FastText concreteness regressor, and Qwen3-Embedding-8B), whose errors propagate into our scores and bound the precision of individual metrics.

Tie rate as an approximate proxy for indistinguishability. In Section 3.2 we adopt the tie rate, i.e., the proportion of zero labels among the 25 annotators' pairwise judgments, as a behavioral proxy for perceived similarity. The underlying annotation protocol in INFINITY-EVAL, however, asks annotators which response is of higher *quality*, not whether the two responses look the same. A tie can therefore arise from at least two routes that we cannot disentangle post hoc: (a) the two responses are genuinely difficult to distinguish along the linguistic dimensions we measure, or (b) the responses differ along visible dimensions yet are judged to be of comparable overall quality. The mapping from "tie" to "indistinguishable" is consequently approximate, and our correlations between dimension-level metric similarity and the tie rate (Figure 3, Table 12) should be interpreted as bounded by this proxy rather than as a direct measurement of perceptual similarity. A dedicated annotation that explicitly asks raters to judge similarity, rather than quality, along each linguistic dimension would tighten this validation, and we leave this to future work.

Reference baseline for cross-metric aggregation. Combining metrics within a dimension requires a common scale; we standardize each metric using its intra-model similarity as the reference. A more principled alternative would be to estimate each metric's distribution over a large reference corpus of model responses spanning diverse models, prompts, and decoding settings, so that normalization reflects population-level rather than within-model variance. Constructing such a corpus exceeds our current data and computational budget, and we leave this to future work.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Leibo, Max Kleiman-Weiner, and Natasha Jaques. 2026. How llms distort our written language. *arXiv preprint arXiv:2603.18161*.

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. 2026. jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*.

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, Toby Drane, and Young Sun Park. 2025. Atla selene mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf. Preview release, April 24, 2026.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. 2025. Strong model collapse. In *International Conference on Learning Representations*, volume 2025, pages 15656–15691.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, and 1 others. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiangming Gu, Soham De, Michalis Titsias, Larisa Markeeva, Petar Veličković, and Razvan Pascanu. 2026. The illusion of stochasticity in llms. *arXiv preprint arXiv:2604.06543*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dan R Johnson, James C Kaufman, Brendan S Baker, John D Patterson, Baptiste Barbot, Adam E Green, Janet van Hell, Evan Kennedy, Grace F Sullivan, Christa L Taylor, and 1 others. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior research methods*, 55(7):3726–3759.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for Navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. In *International Conference on Learning Representations*, volume 2024, pages 20620–20653.

Thom Lake, Eunsol Choi, and Greg Durrett. 2025. From distributional to overton pluralism: Investigating large language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6794–6814.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024a. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024b. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024c. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Jinsook Lee, AJ Alvero, Thorsten Joachims, and Rene Kizilcec. 2025. Poor alignment and steerability of large language models: Evidence from college admission essays. *arXiv preprint arXiv:2503.20062*.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, and 1 others. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

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

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and 1 others. 2025. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. In *International Conference on Learning Representations*, volume 2025, pages 91005–91056.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258.

OpenAI. 2026. Introducing GPT-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *International Conference on Learning Representations*, volume 2024, pages 642–669.

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

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin.

Nils Reimers and Iryna Gurevych. 2019. 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.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025.

Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*.

Team Olmo. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

The Gemini Team. 2026. Gemini 3.1 Flash-Lite: Built for intelligence at scale. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-lite/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Marek Urban, Petra Kmoníčková, and Kamila Urban. 2025. Ai is misled by genai: Stylistic bias in automated assessment of creativity in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity*.

Walter J. B. van Heuven, Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*.

Peter West and Christopher Potts. 2025. Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Peng Yu, En Xu, Bin Chen, Haibiao Chen, and Yinfei Xu. 2025. Qzhou-embedding technical report. *Preprint*, arXiv:2508.21632.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

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. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025b. Novelybench: Evaluating language models for humanlike diversity. In *Proceedings of the Conference on Language Modeling (COLM)*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*.

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

A Similarity Computation and Aggregation

This section specifies how the metrics are defined and how similarity is computed. Table 15 summarizes the dimensions, metrics, metric descriptions, similarity computation methods, and direction-alignment rules. The conversion from individual metrics to dimension-level scores consists of three steps: (i) computing a similarity value for each metric on a pair of responses (§A.1); (ii) aggregating across response pairs to obtain a single value for each metric under each setting; and (iii) combining metrics into group-level homogeneity scores (§A.2).

A.1 Per-metric pairwise similarity

Each metric in our framework produces one of four output types, and we apply a similarity formulation matched to that type:

- **Scalar metrics** (e.g., MTLD, readability, average concreteness) are compared via the absolute difference $|m(r_i) - m(r_j)|$. Larger values indicate greater dissimilarity.
- **Distributional metrics** (e.g., POS distribution, function-word use, emotion profile) are compared via Jensen–Shannon divergence with log base 2, yielding a distance in $[0, 1]$.
- **Native pairwise metrics** (BLEU, ROUGE-L, InfoMatch, embedding cosine similarity) are defined on response pairs by construction; we use their native score directly.
- **Binary form metrics** (the five *form_** indicators) are compared via an exact-match indicator $\mathbb{I}[f(r_i) = f(r_j)]$, yielding 1 for agreement and 0 for disagreement.

For consistency in downstream aggregation, we sign-align all per-metric similarities so that higher values uniformly indicate *greater similarity* between two responses (i.e., absolute differences and JS divergences are negated prior to aggregation).

In section 3, we use pairwise similarity to evaluate relationship between embedding and those metrics. In section 4, we use following aggregation methods to gain the dimension scores.

A.2 Aggregation across responses and metrics

For settings that compare across questions or models, we first summarize the N responses in each (model, question) cell. Scalar metrics are averaged into a single scalar, distributional metrics into a mean distribution, and binary form metrics into a five-dimensional usage-rate vector $\mathbf{p} \in [0, 1]^5$, where each entry denotes the empirical use rate of one template feature. Native pairwise metrics, including BLEU, ROUGE-L, InfoMatch, and embedding cosine similarity, cannot be reduced to a single response-level representation and are therefore averaged directly over response pairs.

Metric A	Metric B	ρ	n
POS dist.	Dependency dist.	0.91	500
Bullet count	Bullet indicator	0.83	500
Emphasis ratio	Emphasis indicator	0.77	500
BLEU	ROUGE-L	0.77	500
Dependency dist.	Function-word JS	0.76	475
POS dist.	Function-word JS	0.75	475
#Paragraphs	Response length	0.66	500
BLEU	Dependency dist.	0.65	500
Emotion profile	Emotion intensity	0.63	500
Punctuation dist.	Dependency dist.	0.62	487
ROUGE-L	InfoMatch (Sent)	0.61	500
BLEU	POS dist.	0.61	500
Punctuation dist.	POS dist.	0.61	487

Table 5: Metric pairs with $|\rho| \geq 0.6$ on INFINITY-EVAL.

We then aggregate metrics within each group in Table 15. Because raw metric scales differ substantially, for example JS divergence lies in $[0, 1]$ whereas length differences can be much larger, we standardize each metric before averaging it with other metrics in the same group.

In the main dimension-level analysis in Section 4.2, we use intra-similarity as the reference baseline for z-score normalization, since the inter-vs-intra contrast more intuitively conveys how close two models are to each other relative to their own internal variation. In Section 3.2, by contrast, z-scores are computed against corpus-level statistics over all pairs in INFINITY-EVAL, which is the appropriate reference when comparing against the Embedding Similarity baseline. We keep the six dimensions separate rather than collapsing them into a single scalar, because Section 3 shows that no single dimension fully substitutes for the others.

A.3 Pairwise Metric Correlations

Experimental Setup. We computed Spearman rank correlations between every pair of continuous metrics on INFINITY-EVAL ($n = 500$ pairs). For punctuation and function-word distributions, responses with too few tokens of the relevant class yield undefined JS divergences and are excluded pairwise (e.g., the prompt *"Write a movie title, and*

write a literal opposite of that movie title" elicits short responses with no usable punctuation distribution), reducing n to 487 and 475 respectively. Table 5 lists all pairs with $|\rho| \geq 0.6$.

Justification for retention. POS–Dependency–Function-word JS (ρ=0.91, 0.76, 0.75): the three syntactic metrics tap different grammatical levels—spaCy coarse POS tags, spaCy dependency labels, and a focused subset of seven LIWC-style function-word classes. Pairwise correlation is expected because function-word classes overlap with both POS and dependency categories, but each metric still carries distinct signal (e.g., POS distinguishes adjectives from adverbs without reference to attachment, which dependency does not). All three are retained. **Bullet count–indicator and Emphasis ratio–indicator** (ρ=0.83, 0.77): adoption vs. intensity. The binary indicator is zero whenever the count is zero, which mechanically inflates ρ, but among responses with the feature on, intensity varies independently. **BLEU–ROUGE-L** (ρ=0.77): both are surface overlap measures, but BLEU is n-gram precision (1- to 4-grams; lexical-choice agreement) while ROUGE-L is longest-common-subsequence F1 without contiguity (overall argument-ordering agreement). The two facets are both of interest in the semantic-match dimension and are retained.

B Two Contrasting Case Studies

To illustrate the two failure modes of embedding similarity, we walk through two response pairs from INFINITY-EVAL that lie on opposite sides of the embedding scale but on the *same* side of the metric and human-judgment scales. For each pair we report the embedding cosine, the human pairwise distribution over 25 annotators, the absolute quality scores, and the linguistic metrics that drive the disagreement.

B.1 Case 1: High Embedding Similarity, Low Metric and Human Similarity

Setup. For the prompt *"Give an example of a linear graph in graph theory."*, the two responses receive an embedding cosine of 0.917, placing the pair in the top decile of INFINITY-EVAL. Despite this, the pair is far from indistinguishable: 56% of annotators prefer Response 1, 32% prefer Response 2, and only 12% tie. The absolute quality scores diverge by 1.38 points on the 5-point scale (4.17 vs. 2.79), well outside the typical within-pair noise band.

Metric	Response 1	Response 2
Structural		
Bullet points (binary)	False	True
Bold emphasis (binary)	False	True
# Bullet-like lines	0	3
Bold emphasis count	0	4
Surface		
Response length (chars)	1,193	576
Avg. sentence length (words)	42.7	23.0
Commas / Parentheses	17 / 0	5 / 24
Lexical		
MTLD	34.8	65.7
Stylistic		
Flesch–Kincaid grade	15.1	8.5
Semantic (reference)		
Embedding cosine	0.917	
BLEU	0.033	
InfoMatch F1 (sent)	0.726	

Table 6: Case 1 metric breakdown. The two responses agree on semantic content but diverge on structural, surface, lexical, and stylistic dimensions.

Where the responses actually differ. Table 6 reports every metric on which the two responses diverge by more than within-pair noise. Three patterns stand out:

- **Structural (form flips).** Response 2 uses bullet points and Markdown bold emphasis; Response 1 uses neither. Two of the five binary form indicators therefore flip.
- **Surface.** Response 1 is 2.07× longer (1,193 vs. 576 characters) and has an average sentence length of 42.7 words against 23.0. The punctuation profile is dominated by commas in Response 1 (17 commas, 0 parentheses) but by parentheses in Response 2 (5 commas, 24 parentheses, used to delimit LaTeX-style math).
- **Stylistic and lexical.** Flesch–Kincaid grade level drops from 15.1 to 8.5, and MTLD rises from 34.8 to 65.7. Response 2 is markedly easier to read and lexically more varied per token.

The semantic dimension is the only one on which the two responses are genuinely close: both define a path graph, give a small concrete example, and discuss adjacency. The embedding cosine of 0.917 reflects this semantic agreement. The structural, surface, lexical, and stylistic divergence is invisible

to the embedding but is precisely what the human annotators react to: Response 2 packages the same concept in a more readable, more structured form, while Response 1 wins the majority through informational coverage rather than presentation.

Prompt

Give an example of a linear graph in graph theory.

Response 1

preferred by 56% **of annotators; mean quality** 4.17

In graph theory, a linear graph, also known as a path graph, is a simple graph that consists of a sequence of vertices connected by edges, forming a linear structure. An example of a linear graph with 6 vertices would be:
`1 - 2 - 3 - 4 - 5 - 6`
In this graph, each vertex is connected to at most two other vertices, forming a linear chain or path. The vertices are labeled from 1 to 6, and each vertex is connected to its adjacent vertices by an edge.
This linear graph can be represented using an adjacency list or an adjacency matrix as follows:
Adjacency List:
```
1: 2
2: 1, 3
3: 2, 4
4: 3, 5
5: 4, 6
6: 5
```
Adjacency Matrix:
```
1 2 3 4 5 6
1  0 1 0 0 0 0
2  1 0 1 0 0 0
3  0 1 0 1 0 0
4  0 0 1 0 1 0
5  0 0 0 1 0 1
6  0 0 0 0 1 0
```
In the adjacency list representation, each vertex is listed along with its adjacent vertices. In the adjacency matrix representation, a 1 is placed at the intersection of row i and column j if there is an edge between vertices i and j, and a 0 otherwise.
Linear graphs are often used to represent simple paths or sequences in various applications, such as scheduling problems, network routing, and certain types of data structures like linked lists.

Response 2

preferred by 32% **of annotators; mean quality** 2.79

A linear graph, often referred to as a **path graph**, is a graph whose vertices form a single path where each vertex (except the two endpoints) is connected to exactly two others. Here's an example:
Vertices: $V = \{A, B, C\}$
Edges: $E = \{AB, BC\}$
This graph can be visualized as a straight line:
$A — B — C$
Key properties:

- Two vertices (A and C) have degree 1 (endpoints).
- The remaining vertex (B) has degree 2.
- No cycles or branches exist.

This structure is a simple example of a linear graph in graph theory.

B.2 Case 2: Low Embedding Similarity, High Metric and Human Similarity

Setup. For the prompt *"Write me 3 short tips for self-development."*, the two responses receive an embedding cosine of 0.553, well below the dataset median. Yet the pair is close to indistinguishable on both human axes: 40% of annotators prefer Response 1, 48% prefer Response 2, and 12% tie, with mean absolute quality scores of 4.28 and 4.32, a difference well within annotator noise.

Metric	Response 1	Response 2
Structural		
Bullet points (binary)	True	True
Headers / Emphasis / Opening / Closing	all False	all False
# Bullet-like lines	3	3
Surface		
Response length (chars)	990	404
Avg. sentence length (words)	11.6	11.5
Lexical		
MTLD	159.6	166.6
Zipf frequency	5.62	5.84
Stylistic		
Flesch–Kincaid grade	9.05	7.61
Concreteness	2.62	2.80
Syntactic		
Function-word ratio	0.290	0.256
Semantic (reference)		
Embedding cosine	0.553	
InfoMatch F1 (sent)	0.771	

Table 7: Case 2 metric breakdown. The two responses agree on every structural, syntactic, lexical, and stylistic axis; the only meaningful divergence is response length.

Where the responses actually agree. Table 7 reports the relevant metrics. The two responses align almost perfectly on the dimensions that the framework treats as fingerprints of model behaviour:

- **Structural (form flags identical).** All 5 binary form indicators match: both responses use bullet points; neither uses headers, bold emphasis, opening templates, or closing templates. Both contain exactly 3 bullet-like lines.
- **Surface (sentence length matches).** Average sentence length is 11.6 words in Response 1 and 11.5 in Response 2.
- **Lexical and stylistic.** MTLD is 159.6 vs. 166.6; Zipf frequency is 5.62 vs. 5.84; Flesch–Kincaid grade is 9.05 vs. 7.61; concreteness is 2.62 vs. 2.80. All differences are small relative to the within-model spread.

Variant	n	Mean	Std	Min	Max
Detailed	300	0.9174	0.0450	0.5978	0.9821
Reordered	300	0.8996	0.0532	0.6768	0.9815
Single	300	0.7932	0.0876	0.4135	0.9535

Table 8: Embedding similarity between the base response and each controlled variant. **Detailed** responses are nearly indistinguishable from the base in embedding space, while **single**-point responses are noticeably more distant.

The one dimension on which the responses diverge is **length**: Response 1 elaborates each tip with 3–4 supporting sentences and reaches 990 characters, whereas Response 2 keeps each tip to a single sentence and totals 404 characters, a ratio of $2.45\times$. This single-axis difference is sufficient to drive the embedding cosine down to 0.553, consistent with Table 8, which shows that embeddings systematically register elaborated and condensed versions of the same content as dissimilar.

Prompt

Write me 3 short tips for self-development.

Response 1

preferred by 40% **of annotators; mean quality** 4.28

1. Set Specific and Measurable Goals: Identify what you want to achieve and make sure your goals are specific, measurable, and realistic. This helps you track your progress and stay motivated. Break down larger goals into smaller, manageable steps to make the process less overwhelming.
2. Practice Mindfulness and Reflection: Take time daily to reflect on your thoughts, emotions, and actions. Mindfulness exercises like meditation or journaling can help you understand yourself better, reduce stress, and make more intentional decisions. Reflect on what you've learned and how you can improve, fostering personal growth.
3. Cultivate a Learning Mindset: Embrace the idea that learning is a lifelong process. Regularly read books, listen to podcasts, take online courses, or engage in new experiences to expand your knowledge and skills. Curiosity and a willingness to learn from mistakes are key to self-development. Surround yourself with people who inspire and challenge you to be better.

Response 2

preferred by 48% **of annotators; mean quality** 4.32

1. Start each day with a purpose: Set one clear goal every morning and make it your priority to accomplish it, no matter how small.
2. Practice the "two-minute rule": If a task takes less than two minutes to complete, do it immediately instead of putting it off.
3. Read for 15 minutes daily: Whether it's a book, article, or industry publication, consistent reading expands your knowledge and perspective.

C Controlled Datasets

As described in Section 2.2 and illustrated in Figure 9, we construct controlled datasets that hold content fixed while systematically varying linguistic form. We sample 100 creative-writing questions and 100 creative-ideation (solution) questions from INFINITY, targeting linguistic and structural metrics respectively. For each question, we generate a base response and a set of controlled variants with GPT-5.4-mini. Writing-task variants are stylistic rewrites of a sentence-level template that preserve sentence-level content while broadening linguistic features, and ideation-task variants are produced by adding targeted form elements to a form-free base response. In this section, we take the creative-writing dataset as a running example to illustrate the full generation pipeline.

C.1 Generating the Base Response

For each creative-writing prompt, we first generate a content-level skeleton (a sentence-by-sentence functional plan) rather than the prose itself. This decouples *what* the response says from *how* it is written, so that later stylistic rewrites can vary form without altering content. The full generation prompt is shown below.

Prompt: Base Response (Sentence-Level Framework)

You will be given a writing prompt. Your task is plan how to write sentence by sentence. For each sentence you envision, write a brief note describing what that sentence should do or say. Only indicates its function in the story, not its content. Do not write the real sentences themselves. The framework must contain between 8 and 15 sentences.

Example
Example Question: Write a short story about a lighthouse keeper who hasn't seen another person in six months.
Example Response:
Sentence 1: He is climbing the stairs to the light — describe the physical routine, no emotion yet
Sentence 2: He checks the time, it's the same time he always does this
Sentence 3: A small odd detail — he says "evening" out loud, to no one
Sentence 4: He looks out the window at the sea, it's empty as usual
Sentence 5: He notices a ship on the horizon, closer than normal
Sentence 6: He leans forward — his body reacts before his mind does

Sentence 7: He watches the ship, thinking maybe it will signal, or slow down
Sentence 8: The ship keeps moving, it doesn't stop
Sentence 9: He stays at the window a little longer than necessary
Sentence 10: He goes back downstairs, picks up the second mug on the shelf, starts to wash it, then puts it back still dirty

Input
writing prompt: {question}

C.2 Re-writing with Different Styles

Given the sentence-level framework, we rewrite the response in several controlled styles. The styles are intentionally designed so that, taken together, they cover the full set of linguistic and structural metrics used in our analysis (pos_dist, emotion, readability, word_frequency, dep_dist, avg_concreteness, lexical diversity, punctuation, function words, and avg_sent_length). Each prompt below is annotated with the metrics it primarily targets.

Style 1 — Plain (*target: pos_dist, emotion*)

Write in a plain, neutral style. Use short declarative sentences. Minimize adjectives, metaphors, and decorative language. State facts directly and avoid any emotional language or personal stance.

Style 2 — Formal Academic (*target: readability, word_frequency, dep_dist*)

Write in a formal academic style. Use complex, multi-clause sentences. Employ hedging language (e.g., 'it can be argued', 'evidence suggests'). Use passive voice where appropriate. Maintain an impersonal, third-person perspective throughout.

Style 3 — Childlike (*target: readability, avg_concreteness, word_frequency, emotion, dep_dist*)

Write in a way that a 5-year-old could easily understand, using simple, everyday words. Keep the tone warm, enthusiastic, and encouraging throughout.

Style 4 — Creative Literary (*target: lexical diversity, word_frequency*)

Write in a creative literary style. Use vivid sensory imagery, metaphors. Choose unusual, unexpected words and actively avoid repeating the same word. Prioritize rhythm and atmosphere over information delivery.

Style 5 — Stream-of-Consciousness (*target: punctuations, avg_concreteness*)

Write in a loose, stream-of-consciousness style. Let thoughts drift and digress into abstract reflections. Use fragments and self-interruptions. Avoid concrete examples or specific details.

Style 6 — Critical/Complaining (*target: function words, emotion*)

Write in a critical, complaining tone. Use lots of conjunctions and connectors (but, although, however, yet, because) to chain skeptical thoughts together. Express doubt and dissatisfaction throughout.

Style 7 — Elaborate/Long Sentences (*target: dep_dist, punctuation, avg_sent_length*)

Write in long, elaborate sentences. Connect ideas using semicolons, colons, and em-dashes rather than full stops. Embed qualifications and parenthetical remarks throughout.

C.3 A Worked Example

To make the pipeline concrete, we walk through a single creative-writing prompt and three of its stylistic variants.

Creative-Writing Question

Could you write an 'Escape from Planet Earth' franchise?

Sentence-Level Framework (Base Response)

Sentence 1: Introduce the franchise premise by establishing the alien world, the main cast, and the basic sci-fi adventure tone
Sentence 2: Present the central threat or problem that forces the characters into motion
Sentence 3: Show the first major escape attempt and why it fails or goes wrong
Sentence 4: Introduce the protagonist's defining trait or flaw that will carry through the series
Sentence 5: Bring in the key supporting characters and clarify how they fit into the escape mission
Sentence 6: Reveal a larger conspiracy or hidden truth that expands the story beyond one planet
Sentence 7: Escalate the stakes by showing the danger spreading to multiple worlds or civilizations
Sentence 8: Set up the first major victory, but make it incomplete so the franchise can continue
Sentence 9: Introduce the sequel hook by hinting at a new destination, enemy, or mystery
Sentence 10: End with a strong franchise-style tease that signals the adventure is just beginning

Variant: Plain Style

On Planet Earth, a group of aliens and their human allies face a series of space adventures.
A threat begins when a force puts their freedom and survival at risk.
Their first escape plan fails because the situation changes faster than they expect.
The main character is cautious and often makes mistakes under pressure.
Other allies join the mission and each one takes a role in the effort to escape.
They learn that the threat is part of a larger plan that reaches beyond one planet.
The danger spreads and affects more worlds and more groups of people.
They win an important battle, but the problem is not fully solved.
A new clue points them toward another planet and another enemy.
The story ends by showing that the escape is only the start of a larger journey.

Variant: Childlike Style

On the bright blue planet of Baab, funny little aliens like Scorch, Gary, and their friends go on big space adventures that feel exciting, silly, and full of wonder.
But a scary bad place called Area 51 traps aliens and makes the friends want to get away fast.
Their first try to escape goes wrong, and they end up in even more trouble than before.
Gary is very careful and a little worried, and that helps him notice things others miss.
His brave friends and family join the plan, and each one helps in a different way to get everyone free.
They learn that the trouble is bigger than one bad place, because a hidden plan reaches far past Baab.
Soon, the danger grows and starts hurting more than one world, so many places need help.
The friends win a happy first step and save some of the day, but the problem is not all gone yet.
Then a new journey begins to a strange new place, with more questions and a new mystery to solve.
And so the fun space adventure is only starting, with lots more courage, laughs, and surprises ahead.

Variant: Creative Literary Style

On the iridescent, storm-licked planet of Baab, a chorus of blue-skinned beings, led by the stout-hearted astronaut Scorch and his cautious brother Gary, tumble through a bright, humorous sci-fi odyssey that crackles like static under a candy-colored sky.
When a shadowy menace rises from the black beyond Baab, their carefree orbit is shattered, and the world begins to feel like a lantern cupped in a predator's hands.
Their first desperate bid to flee spirals into chaos, the plan buckling like wet paper as alarms howl and the escape route twists into calamity.
At the center of it all is Gary, whose nervous caution clings to him like a second skin, a flaw and a gift that shape every step he takes.
Around him gather sharp-edged allies and oddball companions, each one threading their talents into the delicate machinery of the escape, like mismatched gears in a dazzling clock.
Beneath the scramble, a buried conspiracy unfurls, revealing that Baab is only one bright bead on a much larger, darker necklace of hidden control.
Soon the peril swells outward, licking across distant worlds and civilizations as if the whole galaxy has caught fire at the edges.
One hard-won triumph finally lands in their hands, but it arrives incomplete, a half-open door that still rattles in its frame.
Then a new horizon glimmers ahead, hinting at a stranger destination, a fiercer enemy, and a mystery with teeth.
And so the tale hangs in the air like a comet's tail, wild and unfinished, promising that the voyage has only just begun.

As the example illustrates, all three variants preserve the same ten-step narrative content dictated by the framework, while differing sharply along the linguistic dimensions we measure — sentence length, lexical diversity, concreteness, emotional valence, and so on. This property is what makes the dataset suitable for isolating form effects from content effects in our analysis.

D Single vs. Detailed Version

Our metrics cover a range of linguistic and form features, and earlier sections probe each of them individually to test whether embeddings capture them. Here we take a complementary angle: rather than asking which fine-grained features embeddings encode, we ask how embedding-based similarity behaves when the *informativeness* of a response is varied while its underlying content set is held fixed. The goal of this section is diagnostic — we show that embeddings can miss certain kinds of differences — though we do not yet have a satisfactory remedy for the failure mode we surface.

D.1 Pipeline

Following the same pipeline as Appendix C, we start from the base responses generated for the solution (ideation) split and prompt the model to regenerate three controlled variants: **detailed**, **reordered**, and **single**. The three rewriting prompts are shown below.

Prompt: Base → Detailed

You will be given a question with an example response.
Your task is to rewrite the example response in a more detailed and elaborated way — expand on each point with more explanation, evidence, or reasoning. Do

not introduce new points or topics that are not present in the original. Do not use bold, italics, or headers. Preserve the original structure and order of ideas.

Prompt: Base → Reordered

You will be given a question with an example response. Your task is to rewrite the example response by presenting all the same points in a different order — rearrange the points so that the sequence differs from the original, and write a complete, natural response. Elaborate on each point fully and in depth so that your response is approximately the same length as the original. Do not reference, allude to, or imply that the order has changed. Do not use bold, italics, or headers. Do not add or remove any points from the original.

Prompt: Base → Single

You will be given a question with an example response. Your task is to rewrite the example response by focusing on only one point from the original. Select the single most important point and write a complete, natural response based solely on that point — elaborate on it fully, as though this were a standalone response and no other points exist. Do not reference, allude to, or imply that other points exist. Do not use bold, italics, or headers. Do not add new points not present in the original.

Note that both the **reordered** and **single** prompts explicitly require the rewritten response to be expanded to roughly the same length as the base response. This is deliberate: we want each variant to look like a natural, fully-formed model output rather than a truncated stub, so that any drop in similarity reflects a real change in informational coverage rather than a trivial length mismatch.

D.2 Results

For each of the 100 solution questions we compute the embedding similarity between the base response and each of its three variants ($n = 300$ per variant). The aggregate statistics are reported in Table 8.

Two patterns stand out. First, reordering the points in a response has little effect on its embedding similarity to the base response (mean 0.8996, only $\sim!0.02$ below *detailed*), confirming that the embeddings considered here are largely insensitive to point order. Second, the *detailed* variant, which expands the base response by adding explanation, evidence, and reasoning around the same set of points, is even closer to the base response (0.9174) than a response that simply restates the same content in a different order. Conversely, a *single point* response, even when elaborated to a comparable length, drops sharply to 0.7932 and exhibits substantially higher variance.

This observation has a concrete implication for embedding based evaluation of large scale model outputs. If response quality, diversity, or agreement is evaluated using embedding similarity alone, models that produce highly detailed and comprehensive responses may be systematically advantaged. Such responses semantically subsume the content space of their shorter counterparts and therefore obtain high similarity scores against almost any reasonable response to the same question. As a result, the similarity score conflates "covering the same ideas" with "covering strictly more ideas," a mismatch that practitioners should recognize when using embeddings as a proxy for response level evaluation. We leave a principled remedy, such as a coverage asymmetric similarity measure that distinguishes containment from equivalence, to future work.

D.3 A Worked Example

To make the three variants concrete, Table 9 reports the cosine similarities for a single solution-type question from our dataset, and the boxes below show the corresponding base response together with each rewritten variant.

Variant	Cosine similarity to base
Detailed	0.8819
Reordered	0.8851
Single	0.7738

Table 9: Per-variant cosine similarity to the base response for the worked example. As in the aggregate results (Table 8), the *detailed* and *reordered* variants sit very close to the base, while the *single* variant drops noticeably.

Solution Question

Give me fun date ideas to do in Toronto.

Base Response (Framework)

For fun date ideas in Toronto, you could plan a relaxed walk and food stop through Kensington Market and Chinatown, a scenic Harbourfront or Toronto Islands outing with a ferry ride and picnic, a playful evening at a Board Game Cafe or arcade bar, a culture-focused date at the Art Gallery of Ontario or a live show in the Distillery District, or an active adventure like skating at Nathan Phillips Square in winter, biking along the waterfront, or trying a new neighborhood restaurant together.

Variant: Detailed (*cosine = 0.8819*)

For fun date ideas in Toronto, you could plan a relaxed walk and food stop through Kensington Market and Chinatown, where you can wander through colorful streets, browse little shops, and try different snacks or meals along the way, making it easy to keep the date casual and interactive. You could also do a scenic Harbourfront or Toronto Islands outing with a ferry ride and picnic, which gives you a nice mix of city views, fresh air, and time to talk while enjoying a slower pace away from the busier parts of downtown. Another fun option would be a playful evening at a Board Game Cafe or arcade bar, since those places naturally create conversation and friendly competition, which can make the date feel relaxed and entertaining at the same time. If you want something more culture-focused, you could go to the Art Gallery of Ontario or catch a live show in the Distillery District, both of which offer a more memorable shared experience and give you something interesting to discuss afterward. Or, for an active adventure, you could go skating at Nathan Phillips Square in winter, biking along the waterfront, or trying a new neighborhood restaurant together, all of which add a sense of energy and variety while still leaving room to enjoy each other's company.

Variant: Reordered (*cosine = 0.8851*)

For fun date ideas in Toronto, you could start with an active adventure like skating at Nathan Phillips Square in the winter, biking along the waterfront, or trying a new neighborhood restaurant together. Those kinds of dates are great if you want something a little more interactive and relaxed, and they give you a chance to explore the city while doing something memorable together.
You could also plan a scenic Harbourfront or Toronto Islands outing with a ferry ride and picnic. That makes for a really charming date because you get the water views, a slower pace, and a chance to spend time together away from the busiest parts of the city. A picnic adds an easy, romantic touch, and the ferry ride itself can feel like part of the experience.
If you're looking for something more playful, a Board Game Cafe or arcade bar can be a fun choice. These kinds of dates naturally create conversation and a bit of friendly competition, which can make things feel lively and low-pressure at the same time. It's an especially good option if you want to laugh a lot and keep the energy up.
Another great option is a culture-focused date at the Art Gallery of Ontario or a live show in the Distillery District. That works well if you want something more artistic or memorable, and it gives you a shared experience you can talk about afterward. Whether you're wandering through exhibits or enjoying a performance, it can make for a thoughtful and engaging date.
You could also do a relaxed walk and food stop through Kensington Market and Chinatown. That kind of date is perfect if you enjoy exploring neighborhoods, trying different snacks or meals, and keeping things casual while still feeling interesting. The mix of food, atmosphere, and people-watching makes it easy to spend a fun afternoon or evening together.

Variant: Single (*cosine = 0.7738*)

A great fun date idea in Toronto is to spend the day at the Toronto Islands. Take the ferry over together, enjoy the short ride across the water, and use the trip as part of the experience. Once you arrive, you can explore the island at an easy pace, walk along the waterfront, relax in the green spaces, and find a quiet spot for a picnic. The setting feels peaceful and a little special, which makes it ideal for a date because you get to talk, take in the scenery, and enjoy a break from the city without having to plan anything too complicated.

Qualitatively, the *detailed* variant keeps all five date ideas from the base response and pads each one with reasoning about why it would work, leaving its embedding very close to the base (0.8819). The *reordered* variant likewise preserves all five ideas but rearranges and re-paraphrases them; despite this restructuring, its cosine (0.8851) is essentially indistinguishable from the detailed variant, confirming that the embedding is largely insensitive to the order in which content is presented. More strikingly, the *single* variant retains only one of the five ideas, discarding roughly four-fifths of the semantic content of the base response, yet still reaches a cosine of 0.7738. In other words, embedding similarity does not just over-estimate agreement between a detailed answer and its content-preserving rewrites; it continues to register a response that has dropped most of its content as similar to the original. This is precisely the failure mode that motivates our caution against treating semantic embeddings as a sufficient evaluator of model responses when stylistic, structural, or coverage-level differences matter.

E Evaluation Between Embeddings and Metrics

E.1 Response-Pair Classification Details

E.1.1 Experimental Settings.

We construct response pairs from four categories: (cat1) multiple samples from the same model on the same prompt; (cat2) responses to the same prompt from different models; (cat3) responses from the same model to different prompts; and (cat4) responses from the same model to the same prompt but written in different styles. For cat1–cat3, we sample 50 creative ideation and 50 creative writing questions from the controlled datasets

Data	Embedding	Emb.	Met.	Met.+Emb.
Solution	Qwen3	0.815	0.888	**0.914**
Solution	Qzhou	0.705	0.837	0.867
Solution	Jina	0.791	0.878	0.912
Writing	Qwen3	0.546	0.843	0.914
Writing	Qzhou	0.507	0.821	0.853
Writing	Jina	0.588	0.882	**0.915**

Table 10: Robustness results for response-pair classification with additional embedding backbones. Across both datasets, our multidimensional metrics outperform embedding-only features, and combining metrics with embeddings generally yields the strongest performance.

and prompt five models to generate responses. For cat4, we reuse the style-controlled responses from Section 2.2. Since cat4 requires stylistic rewrites, it is available only for writing prompts. We therefore evaluate three-way classification on Solution prompts (cat1–cat3) and four-way classification on Writing prompts (cat1–cat4).

We train logistic-regression classifiers under different feature settings. To avoid prompt-level leakage in cat3, response pairs are constructed separately within each of five folds. In the main text, we report results for Qwen3 embeddings and StyleDistance. For robustness, we additionally evaluate two embedding backbones: `Jina-embeddings-v5-text-small` (Akram et al., 2026) and `Qzhou` (Yu et al., 2025). We report StyleDistance separately because it produces a single aggregate style score that overlaps with several of our individual style metrics.

E.1.2 Per-Category Classification Results

This appendix expands the aggregate Macro-F1 results of Table 3 into per-category precision, recall, and F1 for the writing-task four-way classification (Section 3.1). For convenience we restate the four categories:

- **cat1**: multiple samples from the same model on the same prompt;
- **cat2**: responses to the same prompt from different models;
- **cat3**: responses from the same model to different prompts;
- **cat4**: responses from the same model to the same prompt but written in different styles.

Roughly, cat3 maximizes content variation while cat4 minimizes it, and cat1/cat2 lie in between. The three feature configurations and the five embedding choices follow Section 3.1; metric columns are not applicable to StyleDist.-based configurations and are therefore omitted from the "Met. only" and "Met.+Emb." blocks.

Additional Per-Category Observations. Three finer-grained patterns complement the summary in Section 3.1. *(i) StyleDist. and content embeddings have inverted profiles.* StyleDist. is the strongest embedding-only model on cat1 ($F_1 = 0.603$) but the weakest on cat3 ($F_1 = 0.055$), while content embeddings (Qwen3, Jina) show the opposite pattern. Combining the two (StyleD.+Qwen3) yields gains on cat1 and cat4 over Qwen3 alone ($0.564 \rightarrow 0.654$ on cat1; $0.208 \rightarrow 0.464$ on cat4), confirming that style and content signals are complementary rather than redundant in the embedding space. *(ii) cat2 is the hardest category for embeddings.* Across all embedding-only encoders, cat2 (same prompt, different models) sits below cat3 and often below cat1: e.g. Qwen3 reaches $F_1 = 0.436$ on cat2 versus 0.976 on cat3. Because prompts are shared within a cat2 pair, content embeddings have little surface to exploit, and the remaining signal – model-specific linguistic style – is exactly what our metrics target. Adding metrics raises cat2 F1 from $0.436/0.496/0.558$ (embedding only) to $0.845/0.788/0.852$ (Met.+Emb.) for Qwen3, Qzhou, and Jina respectively, the largest absolute jump among the four categories. *(iii) Metrics are not uniformly weakest on cat3.* While the gap between Met. only and Emb. only is widest on cat4, metrics still reach $F_1 \in [0.738, 0.880]$ on cat3, where content cues should in principle make embeddings dominant. This indicates that inter-prompt differences leave detectable traces in surface-level linguistic structure as well, not only in semantic content. Across rows, Jina is the most consistent embedding backbone in the combined configuration, achieving the best F1 on three of four categories and matching Qwen3 on the fourth.

E.2 Human Tie-Rate Predictions

E.2.1 Dimension-Level Aggregation

Each pair is first scored under every individual metric using the unified per-metric formulations summarized in Appendix A.1. The resulting raw values are heterogeneous in scale and sign (Jensen–Shannon divergences, absolute differences, $0/1$ form indicators, BLEU, ROUGE-L, and embedding cosines all live on different ranges) so we

Config	Embedding	cat1			cat2			cat3			cat4		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Emb. only	Qwen3	0.466	0.715	0.564	0.459	0.415	0.436	0.952	1.000	0.976	0.299	0.160	0.208
	Qzhou	0.307	0.155	0.206	0.475	0.520	0.496	0.761	0.860	0.808	0.464	0.620	0.531
	Jina	0.373	0.220	0.277	0.571	0.545	0.558	0.935	1.000	0.966	0.475	0.675	0.558
	StyleDist.	0.494	0.775	0.603	0.354	0.255	0.297	0.087	0.040	0.055	0.429	0.420	0.424
	StyleD.+Qwen3	0.571	0.765	0.654	0.516	0.395	0.448	0.962	1.000	0.980	0.497	0.435	0.464
Met. only	Qwen3	0.856	0.860	0.858	0.796	0.740	0.767	0.769	0.800	0.784	0.938	0.980	0.958
	Qzhou	0.867	0.845	0.856	0.768	0.710	0.738	0.717	0.760	0.738	0.921	0.985	0.952
	Jina	0.849	0.870	0.859	0.847	0.800	0.823	0.880	0.880	0.880	0.951	0.980	0.966
Met.+Emb.	Qwen3	0.860	0.860	0.860	0.860	0.830	0.845	0.980	0.990	**0.985**	0.951	0.980	0.966
	Qzhou	0.876	0.845	0.860	0.806	0.770	0.788	0.798	0.830	0.814	0.925	0.980	0.951
	Jina	0.869	0.865	**0.867**	0.870	0.835	**0.852**	0.952	0.990	0.971	0.956	0.980	**0.968**

Table 11: Per-category precision, recall, and F1 on the writing-task four-way classification (cat1–cat4) under three feature configurations: Embedding only, Metrics only, and Metrics + Embedding. Metric columns are not applicable to StyleDist. and StyleD.+Qwen3 and are therefore omitted. Best F1 per category is in **bold**.

standardize each metric before averaging within a dimension.

Sign convention. Distance-style metrics (divergences and absolute differences) are sign-flipped so that, for every metric, larger values indicate greater similarity.

Per-metric standardization. For each metric m, we fit a corpus-level mean μ_m and standard deviation σ_m over all pairs in INFINITY-EVAL and convert each raw value v to $z_m = (v - \mu_m)/\sigma_m$.

Dimension score. The score of dimension d on a given pair is the unweighted mean of the z-scores of the metrics belonging to d, ignoring entries that are undefined on that pair (e.g. form indicators that are inactive on both responses):

$$S_d = \frac{1}{|\mathcal{M}_d^{\text{def}}|} \sum_{m \in \mathcal{M}_d^{\text{def}}} z_m,$$

where $\mathcal{M}_d^{\text{def}}$ is the subset of metrics in dimension d with a finite value on the pair. For the tie-rate analysis in Section 3.2 and Appendix E.2.3, the Semantic dimension excludes its embedding-similarity component to keep it orthogonal to the Embedding Similarity baseline.

E.2.2 Per-metric Prediction Results

The main text reports correlations at the dimension level (Figure 3). This section provides the per-metric breakdown.

Figure 6 shows per-metric Spearman correlations with the pairwise tie rate. The strongest predictors within each dimension are *POS Distribution* ($\rho = 0.45$) and *Dependency Distribution* ($\rho = 0.42$) for Syntactic; *ROUGE-L* ($\rho = 0.35$) and *BLEU* ($\rho = 0.34$) for Semantic; *Punctuation Profile* ($\rho = 0.34$) for Surface; *Emotion Profile* ($\rho = 0.33$) for Stylistic; and *Bullet Points* ($\rho = 0.32$) and *# Bullet-like Lines* ($\rho = 0.27$) for Structural. The Lexical dimension is the weakest at the metric level, with both *Lexical Diversity (MTLD)* and *Word Frequency (Zipf)* falling below the leaders above. Most remaining metrics across the other dimensions are smaller but still significant. The notable exceptions are all structural: *Opening Template* ($\rho = 0.03$) is not significant, and *Headers* shows a small negative correlation ($\rho = -0.10$, $p < 0.05$), consistent with the larger within-dimension spread observed for structural features.

E.2.3 Robustness of Dimension-Level Predictions

To quantify the uncertainty in the dimension-level correlations reported in Figure 3, we resample the 500 response pairs with replacement $B = 1{,}000$ times and recompute Spearman's ρ against the pairwise tie rate on each resample. Table 12 reports the observed correlation, bootstrap mean, standard deviation, and 95% percentile confidence interval for each dimension and for the two embedding baselines.

The eight rows fall into three loose clusters. Syntactic stands at the top with the highest point estimate ($\rho = 0.41$); its lower bound (0.34) sits at or above the point estimate of every other dimension. A middle cluster — Style Distance (0.38), Embedding Similarity (0.37), Stylistic (0.35), and Semantic (0.34) — has heavily overlapping intervals, with point estimates within a 0.04 window. The remaining three dimensions form a lower clus-

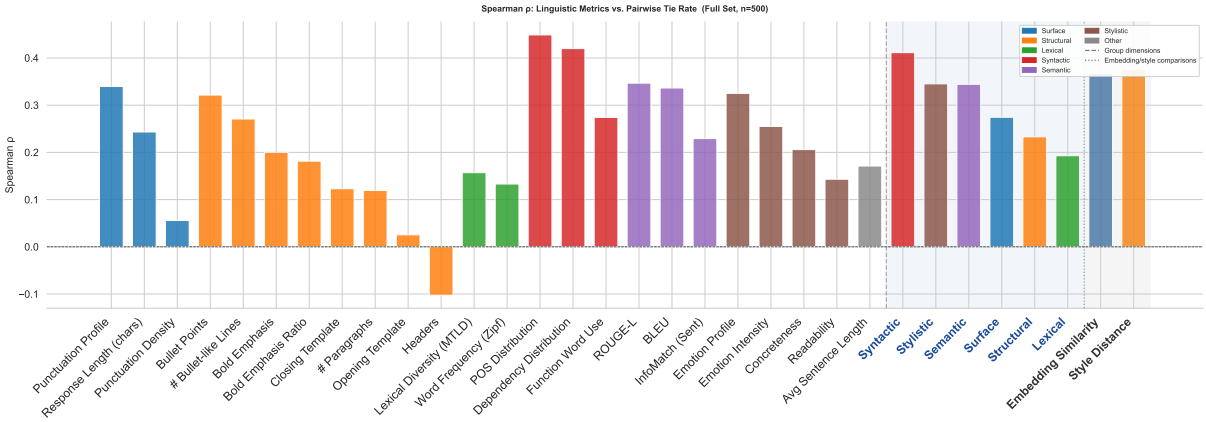

Figure 6: Spearman correlations between individual linguistic metrics and the pairwise tie rate on INFINITY-EVAL ($n = 500$). Bars are colored by dimension. The shaded panel on the right reproduces the dimension-level summary from Figure 3 for comparison. Significance markers: $*\,p < 0.05$, $**\,p < 0.001$.

Dimension	ρ	mean	sd	95% CI
Syntactic	0.41	0.41	0.038	[0.34, 0.48]
* Style Distance	0.38	0.38	0.039	[0.30, 0.45]
* Embedding Similarity	0.37	0.37	0.042	[0.28, 0.44]
Stylistic	0.35	0.34	0.038	[0.26, 0.42]
Semantic	0.34	0.35	0.040	[0.27, 0.42]
Surface	0.27	0.27	0.042	[0.19, 0.35]
Structural	0.23	0.23	0.044	[0.14, 0.32]
Lexical	0.19	0.19	0.043	[0.11, 0.27]

Table 12: Bootstrap statistics ($B = 1{,}000$ resamples) for the Spearman correlation between each dimension-level similarity score and the pairwise tie rate on INFINITY-EVAL ($n = 500$). The Semantic dimension here aggregates BLEU, ROUGE-L, and InfoMatch only, with its embedding-similarity component held out to keep it orthogonal to the Embedding Similarity baseline. Confidence intervals are 2.5/97.5 percentiles.

ter (Surface 0.27, Structural 0.23, Lexical 0.19); their point estimates lie below those of every dimension in the middle cluster, though their upper CIs extend into its lower edge.

While Structural yields one of the weaker dimension-level correlations on the full set, one likely reason is that INFINITY-EVAL contains a large share of open-ended creative generation prompts (e.g. story writing, poetry, role-play), for which neither response naturally contains those features. On such pairs the structural metrics are constant or near-constant, contributing noise rather than signal to the correlation.

To check this, we restrict the analysis to pairs in which at least one response activates a form indicator (a non-empty value on any of *Bullet Points*, *Bold Emphasis*, *Headers*, *Opening Template*, or *Closing Template*), reducing the sample from $n = 500$ to $n = 268$. On this subset, Structural rises from $\rho = 0.23$ to $\rho = 0.33$ ($p < 10^{-7}$). The remaining dimensions are not the focus of this analysis and are omitted.

These numbers indicate that structural metrics carry meaningful signal when the task elicits the corresponding formatting, and that the lower full-set correlation partly reflects coverage rather than predictive validity.

F Models, Responses, Prompts, and Filtering

F.1 Model List

Table 13 lists every model evaluated in this work, organized by series and training stage. The set spans (i) three closed-source API models, which we select three state-of-the-art lightweight systems, and (ii) several open-weight families covering the full post-training pipeline (base → SFT → DPO → RLHF/RLVR/PPO) together with, where available, separate *thinking* variants.

F.2 Prompts and Question Types

We sample open-ended questions from INFINITY across six top-level categories: Creative Content Generation, Brainstorming and Ideation, Open-Endedness, Alternative Perspectives, Alternative Styles, and Information-Seeking. Since not every item in the dataset is suitable for homogeneity evaluation, we manually review the sample and further use GPT-5.4-Mini to strip stylistic constraints (e.g., "in Hemingway's style") and explicit format requirements (e.g., "write in bullet points") so as to avoid evaluation bias. In total we retain 300 questions, 50 per category. The sample list of questions is shown in Table 14.

Model	Series	Stage	Ref.
Closed-source API models			
deepseek-v4-flash	–	Instruct	(DeepSeek-AI, 2026)
gemini-3.1-flash-lite	–	Instruct	(The Gemini Team, 2026)
gpt-5.4-nano	–	Instruct	(OpenAI, 2026)
Gemma 3			
gemma-3-4b-pt	gemma-3-4b	Base	(Gemma Team et al., 2025)
gemma-3-4b-it	gemma-3-4b	Instruct	
gemma-3-12b-pt	gemma-3-12b	Base	
gemma-3-12b-it	gemma-3-12b	Instruct	
gemma-3-27b-pt	gemma-3-27b	Base	
gemma-3-27b-it	gemma-3-27b	Instruct	
Olmo 3			
Olmo-3-1025-7B	olmo-3-7b-Instruct	Base	(Team Olmo, 2025)
Olmo-3-7B-Instruct-SFT	olmo-3-7b-Instruct	SFT	
Olmo-3-7B-Instruct-DPO	olmo-3-7b-Instruct	DPO	
Olmo-3-7B-Instruct	olmo-3-7b-Instruct	RLHF	
Olmo-3-7B-Think	olmo-3-7b-Instruct	Thinking	
Olmo-3-1125-32B	olmo-3-32b-Think	Base	
Olmo-3-32B-Think-SFT	olmo-3-32b-Think	SFT	
Olmo-3-32B-Think-DPO	olmo-3-32b-Think	DPO	
Olmo-3-32B-Think	olmo-3-32b-Think	RLVR	
Qwen 3			
Qwen3-4B-Base	qwen3-4b	Base	(Yang et al., 2025)
Qwen3-4B-Instruct-2507	qwen3-4b	Instruct	
Qwen3-4B-Thinking-2507	qwen3-4b	Thinking	
Qwen3-30B-A3B-Base	qwen3-30b	Base	
Qwen3-30B-A3B-Instruct-2507	qwen3-30b	Instruct	
Qwen3-30B-A3B-Thinking-2507	qwen3-30b	Thinking	
Tülu 3 (Llama-3.1-8B base)			
Llama-3.1-8B	llama-3.1-Tülu	Base	(Grattafiori et al., 2024)
Llama-3.1-Tülu-3-8B-SFT	llama-3.1-Tülu	SFT	(Lambert et al., 2024a)
Llama-3.1-Tülu-3-8B-DPO	llama-3.1-Tülu	DPO	
Llama-3.1-Tülu-3-8B-RLVR	llama-3.1-Tülu	RLVR	
Tülu 2 / 2.5 (Llama-2-13B base)			
Llama-2-13b-hf	llama-2-Tülu	Base	(Touvron et al., 2023)
Tülu-2-13b	llama-2-Tülu	SFT	(Ivison et al., 2023)
Tülu-v2.5-dpo-13b-uf-mean	llama-2-Tülu	DPO	(Ivison et al., 2024)
Tülu-v2.5-ppo-13b-uf-mean-70b-uf-rm	llama-2-Tülu	PPO	

Table 13: Models evaluated in this work, grouped by series and training stage. **Base**: pretrained checkpoint; **SFT**: supervised fine-tuning; **DPO**: direct preference optimization; **RLHF/RLVR/PPO**: reinforcement-learning stage (RLVR for Tülu 3, PPO for Tülu 2.5); **Instruct**: official instruction-tuned release; **Thinking**: release with explicit reasoning traces.

F.3 Few-shot for Base Models

Since base models tend to be unstable when answering questions directly, we provide three in-context examples drawn from the highest-rated human responses in INFINITY-EVAL. The three examples are deliberately chosen at varying response lengths (short, medium, long) to avoid biasing the base model's output length. The *form* variant retains all formatting markers from the original responses; the *pure* variant strips these markers while preserving the content, ensuring that any output-quality difference between the two conditions is attributable to formatting rather than to content. A short example illustrating the form/pure contrast is shown below; the full set of few-shot prompts will be released with our code.

Example (*form* variant)

Input: Write a movie title, and write a *literal* opposite of that movie title.

Response: **Movie Title:** "Fast and Furious" **Literal Opposite:** "Slow and Calm".

Table 14: Exampled Questions by Category

ID	Question / Prompt
Creative Content Generation	
1	Create a character named Bill Honda.
2	Show me a joke?
3	Write a sentence that begins with ‘Nuclear war’, contains ‘pizza’, and ends with ‘Rubik’s cube’.
4	Write some FOMO titles for a video with ocean waves.
5	Write a sentence that begins with ‘A blue shirt’ and ends with ‘5 pelicans’.
6	Write a story about Swapna and Hrushi.
7	Write about puns with the word "solana" and animal names.
8	Write a poem about a man.
9	Write an essay about Adolf Hitler.
10	Make a football game.
Brainstorm & Ideation	
11	Suggest a theme for shoe care.
12	Hypothetical TV advert from 1987 for deep-fried carrot crisps sold under the name ‘Bunny Chips!’ — might use some interesting advertising ideas.
13	Questions for an OSEA brand interviewee.
14	Give me an Erasmus+ KA154 project idea about digital entrepreneurship skills that would be more creative.
15	Design an electronics kit for teenagers that lets them build a simple low-band TV circuit.
16	For an RPG moment, how exactly could a ‘power move’ happen without it feeling too much like a ‘power fantasy’?
17	Can you imagine fictional cable TV networks inspired by Bravo in 1982, with details including viewership?
18	How can I make a Breaking Bad-type Roblox game?
19	Sci-fi version of Little Red Riding Hood.
20	For a 3D RPG medieval-style game, how could I implement things to do if a player becomes a leader of a faction?
Open-Endedness	
21	How could a found footage film work with a creature that can’t turn up on camera?
22	I am traveling around the Earth in a straight line. After 40,075 km, I am back where I started. On my travels, I encountered a tiger. Which species was the tiger?
23	What are the things people focus on or spend their lives doing, in detail and according to each age group?
24	What would a condemned Australian home in 1967 look like in detail?
25	If humans had immunity to alcohol, what would happen?
26	Describe an anti-economic world.
27	What are some things that the 1980s had in common with the 1960s, but the 1990s did not have in common?
28	Evaluate the analogy made by a non-materialist between human consciousness and a TV/TV signal, where consciousness is compared to the TV signal.
29	What are some reasons why a consulting firm might fail on a project to help an automotive manufacturing company expand into the EV market for SUVs and trucks?
30	What did the Fyre Festival look like during the night?
Alternative Perspectives	
31	Is being a game QA tester a good job?
32	Do you agree that athletes’ salaries should be proportional to the sport’s popularity?
33	What is the best feeling in the world?
34	I think, if a software engineer says something like "it’s not a bug, it’s a feature," they’re a bad software engineer.
35	Unethical life hacks to earn money.
36	Why is it important to name your eldest son after you?
37	I know this might be a controversial topic, but what do you think about abortion?
38	Show a typical Liverpool fan.
39	What to think when everything goes the wrong way?
40	What do you think is scary?
Alternative Styles	
41	How could an invasion of alien beasts being stopped by one man with a gun who simply refused to give up be written in a believable and interesting way?
42	What are good informal and formal responses to ‘see you when I see you’?
43	How to say thank you for your time in the subject line of an email?
44	Write an email to Angela Wu, thank her for her guidance, and tell her I need to inform my CEO.
45	Write a professional greeting message for my friend, Mr. Qubow.
46	Reply to an email that once I receive feedback from the store, I will reach back to you.
47	Please write a message to my friend telling him that if he doesn’t find the test, I will buy it from the internet.
48	Write an email to inform and seek approval for leave to visit the doctor due to a recurring medical condition on Saturday, 22nd March 2023.
49	Write an email asking for an update that is expected within 48 hours.
50	Draft a follow-up email to an interviewer about setting up a time to speak.
Information-Seeking	
51	What is a way to structure a general chemistry course so that every topic is built upon, starting with the periodic table?
52	You are a college intern at a small financial planning company. Give ideas on what to do to make a good impression.
53	Please create a detailed plan for what to cook for the Christmas holidays, starting on Friday, the 22nd, and ending on the following Wednesday.
54	What are some innovations in science education?
55	Give me lesson plan ideas for a high-level ESL middle school student that include creative writing.
56	I need to know the current online problems, as I want to develop an AI service to solve them. Please provide information related to web design.
57	What should I see in Oregon and Washington states to make sure I miss nothing during my travels?
58	What 3D features could be incorporated into a public safety worker self-care app to enhance user experience?
59	Generate more comprehensive details and examples on the Core Principles of ZenVitaFuel.
60	Can you propose some possible and easy math IA topics?

Same example (*pure* variant)

Input: Write a movie title, and write a literal opposite of that movie title.

Response: Movie Title: "Fast and Furious". Literal Opposite: "Slow and Calm".

F.4 Filtering Prompts

Shypula et al. (2025) argue that small models often generate incoherent outputs, thereby inducing a bias toward homogeneity in evaluation, since random tokens are meaningless and would not be produced by more sophisticated models. While they use a code generation task, in which the validity of outputs can be checked directly, we use a text generation task, which is easier for smaller models. The notion of quality we refer to here is that the model does not generate meaningless content, which we further expand into three criteria and evaluated based on GPT-OSS-120b: **relevance** (directly addresses the question; refusals fail), **completeness** (coherent and finished, with no truncation or repeated junk tokens), and **focus** (contains only the answer, with no metadata, debug text, or reasoning traces). A response is discarded if it fails any one of these. The exact prompt is reproduced below.

Filtering prompt given to GPT-OSS-120B

Task Description
You will be given a QUESTION and its corresponding RESPONSE. Decide whether the RESPONSE passes all of the checks below. You are not evaluating answer quality or factual correctness, only compliance with these criteria.

The RESPONSE must pass all of the checks:

- **Relevance**: it directly addresses the QUESTION. A refusal (e.g., "I'm sorry, I can't help with that", "As an AI, I cannot...") fails this check.
- **Completeness**: it is coherent and finished. Fails if truncated mid-sentence or mid-thought, if it contains only a stub (e.g., the QUESTION asks for a letter and the RESPONSE is just "Dear recipient,"), or if it contains repeated junk tokens or artifacts such as `<br><br><br>`, long runs of newlines or whitespace, or the same sentence repeated.
- **Focus**: it contains only the intended answer. Fails if it includes model metadata, system or debug text, token usage logs, tool-invocation traces, or reasoning markers not meant for the user.

F.5 Dimension Similarity Aggregation

Each cell of Figure 4 is produced by four aggregation steps. We sample $K = 10$ responses per model per question and retain the $|\mathcal{Q}| = 225$ questions on which every model produces at least K valid responses.

Steps 1–3: Metric → dimension → question. Let $s_{ij}^{(m)}(q)$ be the similarity between model i and model j on metric m and question q. Write $\mathcal{M}_d^{q,ij}$ for the finite-valued metrics in dimension d. We average within the dimension, then across questions:

$$S_{ij}(d) = \operatorname*{mean}_{q\in\mathcal{Q}} \underbrace{\operatorname*{mean}_{m\in\mathcal{M}_d^{q,ij}} s_{ij}^{(m)}(q)}_{S_{ij}(q,d)} . \qquad (1)$$

Non-finite metric values are dropped before averaging; metrics are averaged (not summed) so dimensions with different metric counts remain comparable.

Step 4: Intra-model baseline. $S_{ii}(d)$ follows the same formula with $s_{ii}^{(m)}(q)$ replaced by the mean pairwise similarity over the $\binom{K}{2}$ same-prompt sample pairs of model i.

Step 5: Self-normalized score. The displayed value is

$$\widetilde{S}_{ij}(d) = \frac{\frac{1}{2}(S_{ij}(d) + S_{ji}(d))}{\sqrt{S_{ii}(d)\, S_{jj}(d)}}. \qquad (2)$$

The numerator symmetrizes over the two directions; the denominator rescales each pair by the geometric mean of the two models' self-similarities, ensuring that low-diversity models (high S_{ii}) are not penalized when compared against diverse ones. By construction $\widetilde{S}_{ii}(d) = 1$ on the diagonal.

G Panal A Trajectories Values

H Form Features and Reward Model Preference

This appendix details the reward models, prompts, and scoring procedure used to produce Figure 9.

H.1 Selected Reward Models

We compare two paradigms of preference modelling. The three Bradley–Terry (BT) reward models are the top-ranked open 8B-scale BT models on RewardBench (Lambert et al., 2024c) and RewardBench 2 (Malik et al., 2025) at the time of writing.

Bradley–Terry reward models.

- **Skywork-Reward-V2-Llama-3.1-8B** and **Skywork-Reward-V2-Qwen3-8B** (Liu et al., 2025), a BT pair trained on the curated SynPref-40M mixture whose paper explicitly advertises improved resistance to stylistic biases via data curation.

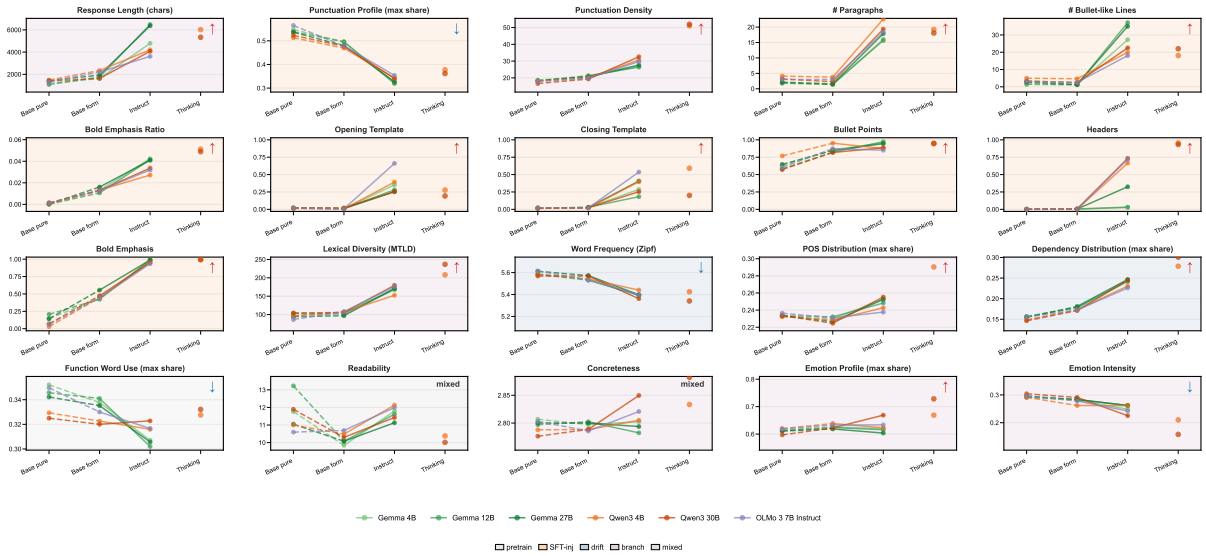

Figure 7: Linguistic metric trajectories across broad-stage transitions (Base pure → Base form → Instruct → Thinking) for the Gemma and Qwen3 families, with OLMo-3-7B-Instruct included as a reference. Subplots are color-framed by linguistic dimension, and top-right annotations report directional consistency (e.g., $6/6 \uparrow$). Most form-bearing metrics (bullets, headers, bold, opening/closing templates) make their entire jump at the Base→Instruct transition, while Thinking checkpoints branch off on a subset of surface metrics (Response Length, Punctuation Density, Emotion Intensity, Concreteness) rather than continuing the Instruct trajectory. Figure 5 shows the complementary fine-grained SFT → DPO → RLHF/RLVR view.

- **Llama-3.1-Tulu-3-8B-DPO-RM-RB2** (Malik et al., 2025), the RewardBench 2 BT reward model fine-tuned from Llama-3.1-Tulu-3-8B-DPO (Lambert et al., 2024b), included to pair directly with its generative counterpart below.

Including two Skywork backbones (Llama and Qwen) and a Tülu RM lets us check whether residual form preferences track the training recipe or the underlying base model.

Generative (LLM-as-judge) reward signals.

- **Llama-3.1-Tulu-3-8B-DPO** (Lambert et al., 2024b), the DPO checkpoint of the Tülu 3 recipe, paired with its BT RM above.
- **Qwen3-30B-A3B-Thinking-2507** (Yang et al., 2025), a recent reasoning-trained MoE judge.
- **Gemma-3-27B-IT** (Gemma Team, 2025), a 27B instruction-tuned dense model from a different family, included to guard against family-specific artefacts.

Generative-judge prompt. For every generative judge we use the pairwise `Atla_PROMPT` template shipped with RewardBench (Lambert et al., 2024c) (in `rewardbench/generative.py`), which was originally introduced by Alexandru et al. (2025) for training Atla Selene Mini and elicits a short reasoning trace followed by a single token `A` or `B`. We caution that this prompt, like most LLM-as-judge templates, is explicitly comparative, and the two responses we feed it are nearly identical except for one formatting feature. A capable judge may notice this near-duplication and infer that the experimenter is probing a specific axis of variation, which could itself amplify or suppress the measured form effect. We therefore read the absolute numbers from the generative paradigm as indicative rather than calibrated, and rely primarily on the qualitative gap between the two paradigms.

The dataset of paired responses is the one introduced in Section 2.2: every pair shares the same prompt and semantically equivalent content, differing only in a single formatting feature (one of `base`, `opening_template`, `closing_template`, `headers`, `bullet_points`, `emphasis`).

H.2 Computing Win Rates

For a reward model R and a form f, we report the average pairwise win rate of f against the other five forms, which approximates the BT strength induced by R on the form space:

$$\overline{w}_f^R = \frac{1}{|Q|\,(|\mathcal{F}|-1)} \sum_{q \in Q} \sum_{j \in \mathcal{F} \setminus \{f\}} w_{f,j}^{R,q}, \quad (3)$$

where $\mathcal{F}$ is the set of forms, Q is the evaluation set, and $w_{i,j}^{R,q} \in [0,1]$ is the win probability that R assigns to form i over form j on question q. A value of 0.5 corresponds to chance.

For scalar BT models with reward $s_f^{R,q}$ we use the standard logistic link

$$w_{i,j}^{R,q} = \sigma\big(s_i^{R,q} - s_j^{R,q}\big), \tag{4}$$

which is the maximum-likelihood implication of the BT assumption under which these models are trained. For generative judges, which return only discrete verdicts, we query each pair in both presentation orders to neutralise position bias (Zheng et al., 2023; Wang et al., 2023) and set

$$w_{i,j}^{R,q} = \begin{cases} 1, & \text{both orders prefer } i, \\ 0, & \text{both orders prefer } j, \\ 0.5, & \text{otherwise (disagreement or tie).} \end{cases} \tag{5}$$

This conservatively absorbs ties and position-bias-induced flips into chance. We then plug the resulting $w_{i,j}^{R,q}$ into Equation 3 for both paradigms, so that scalar and generative numbers are on the same scale.

H.3 Sources of Form Homogenisation

This appendix collects three pieces of evidence supporting the claim that structural homogenisation is inherited from SFT data and reinforced by generative LLM judges.

H.3.1 Form Style Is a Model Property, Not a Question Property

Figure 8 reports the mean usage rate of each form feature across six question types, averaged over instruct-tuned models. Within-feature variation across question types is typically 0.10–0.20, much smaller than the across-feature range, which spans most of $[0,1]$. Question type does modulate rates — Information-Seeking and Brainstorming prompts elicit more Bullets and Headers than Creative Content Generation — but the form style of a typical instruct response is largely a property of the post-trained model, with question type playing only a secondary role.

H.4 Form Homogenisation Is Inherited from Modern SFT Data

The checkpoint-level view in Figure 5 confirms the aggregate stage attribution: most changes in form features occur during the BASE→SFT transition, while later DPO and RLVR stages introduce

Figure 8: Mean usage rate of each form feature across six question types, averaged over instruct-tuned models. The across-feature range (∼0.1–0.9) is much wider than the within-feature, across-question range (∼0.1–0.2), indicating that form style is determined primarily by the model rather than the question.

only small additional shifts, mainly in visual format cues such as Bullets and Emphasis. A striking exception to this SFT-installation pattern is Llama-2-Tülu, released in November 2023, whose trajectory remains nearly flat on every low-prior form feature: Headers and Emphasis stay below 0.20 throughout, and Opening and Closing never exceed 0.10 even after RLVR. This contrast is informative because Llama-2-Tülu shares the same broad post-training pipeline as Llama-3.1-Tülu and differs from it by less than 1B parameters. The divergence is therefore unlikely to be explained by the optimisation procedure or by scale alone. Instead, it points to the SFT corpus as the key source of the formatting shift. The Tülu-2 mixture predates the widespread use of GPT-4- and Claude-distilled responses in post-2024 instruction datasets; as these frontier-model outputs became common training targets, their markdown-heavy formatting was likely transferred into student models. These results suggest a more qualified interpretation of the familiar claim that "alignment causes homogenisation." Rather than implying that the alignment objective itself necessarily induces structural homogenisation, the evidence here points to the formatting patterns present in modern alignment data as an important contributing factor. This distinction highlights data curation as a key locus for understanding, and potentially mitigating, form-level homogenisation.

H.4.1 Uncalibrated Generative Judges Create a Second Channel

The SFT corpus is not the only entry point through which formatting bias can enter an alignment pipeline; the reward signal used for preference optimisation constitutes a second potential channel. We compare two paradigms of preference modelling on

response pairs that differ only in their formatting features: purpose-trained Bradley–Terry reward models and generative LLM-based judges. We use the dataset introduced in Section 2.2, in which each response applies a single formatting feature and "plain" denotes responses without any such feature; the variants are semantically equivalent.

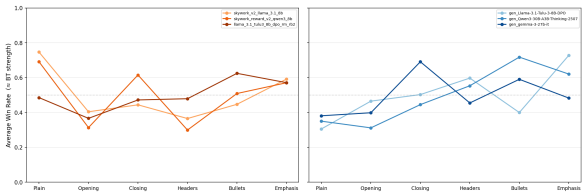

Figure 9: Average win rate of form variants under Bradley–Terry reward models (left) and generative LLM judges (right). The two paradigms give nearly opposite preference orderings: BT RMs favour Plain (0.69–0.75) and penalise Headers/Opening, while generative judges penalise Plain (0.30–0.38) and favour Bullets and Emphasis (0.59–0.73).

As shown in Figure 9, the two paradigms diverge sharply. Bradley–Terry RMs perform near chance and sometimes even prefer the plain variant (win rates up to 0.74), whereas generative LLM judges consistently penalise the plain variant (0.27–0.40) and favour Emphasis (≈ 0.70) and Bullets (≈ 0.60–0.72).

Since generative judges are widely used to produce synthetic preference data, RLAIF rewards, or constitutional feedback, their uncalibrated formatting preferences can be written back into the supervision signal. This creates a feedback loop in which form-preferring judges select form-heavy responses, which in turn train future models and future judges with similar biases. The choice between a calibrated reward model and an uncalibrated generative judge is therefore itself a curation decision, with measurable consequences for structural homogeneity.

Group	Metric	Description	Output	Similarity (aligned)	Ref.
Surface	Response Length (chars)	Total length in characters	diff	$-\lvert\Delta\rvert$	
	Punctuation Profile	Distribution over 11 regex-defined punctuation classes: comma, period, em dash, hyphen, semicolon, colon, exclamation, question mark, ellipsis, parentheses, and quotation marks	dist	$-$JS	
	Punctuation Density	Total punctuation count per 100 words	diff	$-\lvert\Delta\rvert$	
Structural	# Paragraphs	Number of non-empty paragraphs split by blank-line separators	diff	$-\lvert\Delta\rvert$	
	# Bullet-like Lines	Number of lines beginning with bullet or enumeration markers: *, -, •, en/em dash, or digit+./)	diff	$-\lvert\Delta\rvert$	
	Bold Emphasis Ratio	Count of Markdown bold spans matched by **...**, normalized by word count	diff	$-\lvert\Delta\rvert$	
	Opening Template	Phatic or meta opener before substantive content, e.g., "Sure!" or "Great question!"	binary	$1 - \text{Hamming}$	
	Closing Template	Formulaic closing offer or sign-off, e.g., "I hope this helps" or "Let me know if you have questions"	binary	$1 - \text{Hamming}$	
	Bullet Points	Use of bulleted or numbered lists, e.g., lines beginning with "-" or "1."	binary	$1 - \text{Hamming}$	
	Headers	Use of Markdown section headers, e.g., "## Key Points"	binary	$1 - \text{Hamming}$	
	Bold Emphasis	Use of Markdown bold spans, e.g., "`**important**`"	binary	$1 - \text{Hamming}$	
Lexical	Lexical Diversity (MTLD)	Length-robust type-token diversity index	diff	$-\lvert\Delta\rvert$	(McCarthy and Jarvis, 2010)
	Word Frequency (Zipf)	Mean log word frequency (Zipf scale)	diff	$-\lvert\Delta\rvert$	(van Heuven et al., 2014)
Syntactic	POS Distribution	Distribution over spaCy coarse POS tags, including noun, verb, adjective, adverb, pronoun, determiner, adposition, conjunction, auxiliary, particle, numeral, punctuation, symbol, and other	dist	$-$JS	(Honnibal et al., 2020)
	Dependency Distribution	Distribution over spaCy dependency labels, including subject, object, modifier, determiner, auxiliary, conjunction, marker, preposition, root, and punctuation relations	dist	$-$JS	(Honnibal et al., 2020)
	Function Word Use	Proportions of seven LIWC-style function-word classes approximated with spaCy POS/dependency rules: articles (DET/det), prepositions (ADP/prep), coordinating conjunctions (CCONJ/cc), subordinating conjunctions (SCONJ/mark), pronouns (PRON), auxiliaries (AUX/aux), and particles (PART)	dist	$-$JS	(Pennebaker et al., 2015; Honnibal et al., 2020)
Semantic	BLEU	Symmetric 4-gram overlap with brevity penalty	pair	native	(Papineni et al., 2002)
	ROUGE-L	F1 of longest common subsequence	pair	native	(Lin, 2004)
	InfoMatch (Sent)	BERTScore-style F1 over sentence-level Qwen3-Embedding-8B embeddings, computed by bidirectional best-match cosine similarity between sentence segments	pair	native	(Zhang et al., 2020, 2025a)
	Embedding Similarity	Cosine of Qwen3-Embedding-8B response embeddings	pair	native	(Zhang et al., 2025a)
Stylistic	Readability	Flesch–Kincaid grade level	diff	$-\lvert\Delta\rvert$	(Kincaid et al., 1975)
	Concreteness	Mean word concreteness (0–5; FastText regression)	diff	$-\lvert\Delta\rvert$	(Brysbaert et al., 2014; Bojanowski et al., 2017)
	Emotion Profile	28-way GoEmotions probability profile	dist	$-$JS	(Demszky et al., 2020)
	Emotion Intensity	Max non-neutral GoEmotions score	diff	$-\lvert\Delta\rvert$	(Demszky et al., 2020)

Table 15: Full metric inventory across the six groups. "Output" indicates the raw output type of the metric: *diff* (scalar difference), *dist* (distribution), *binary* (indicator), or *pair* (natively defined on response pairs). "Similarity (aligned)" gives the sign-aligned per-pair similarity used downstream, where higher values indicate greater similarity.