Evaluating Pluralism in LLMs through Latent Perspectives

Anonymous Authors[1]

Abstract

Pluralistic representation and generation in LLMs is becoming increasingly relevant because of the importance of showcasing the diversity of opinion. Models are known to reduce the diversity of training data and to exhibit homogeneity when generating. However, this issue has been demonstrated primarily on multiple-choice questionnaires or using aggregated, high-level characteristics for free-form text. Such approaches do not adequately capture the variety of human perspectives expressed in opinionated text, nor do they identify the missing aspects of human text driving the *pluralistic gap* in LLM-generated text. In this paper, we aim to analyze model pluralism by extracting and comparing latent perspectives from human and LLM-generated text. We propose a two-tiered framework which identifies which aspects of perspectives are underrepresented in LLM generation. We evaluate an instance of our framework on on highly opinionated data from book reviews and find that while LLMs reduce pluralism in human text across levels of abstraction, simple prompting techniques such as persona prompting can alleviate the pluralistic gap in subjective aspects.

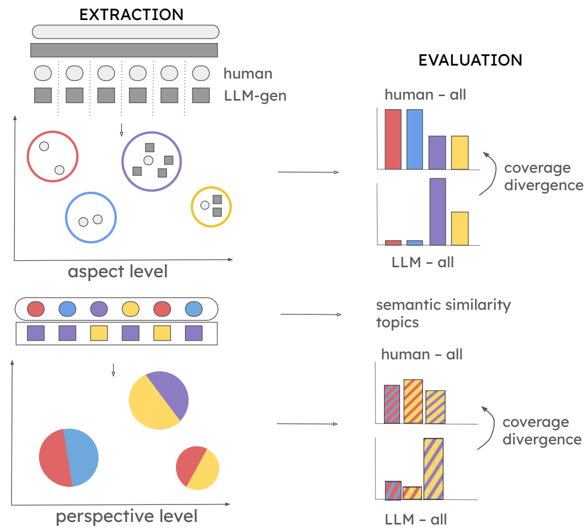


Figure 1. A depiction of our pluralistic evaluation framework, focusing on the extraction and comparison of aspects and latent perspectives between human and LLM-generated text

1. Introduction

As large language models scale, their competence across a broad range of tasks such as coding, math, and complex reasoning also improves (Jimenez et al., 2024; Phan et al., 2025). However, performance on tasks with objectively verifiable outputs does not necessarily indicate alignment to the inherent diversity of human perspectives in the myriad of subjective tasks where "correctness" is defined by specific cultural, social, or individual frameworks rather than universal facts (Frenda et al., 2025; Fleisig et al., 2024). To provide nuanced and balanced viewpoints on open-ended subjective tasks, language models should undergo *pluralistic alignment* (Sorensen et al., 2024), that is, they should be optimized to adequately represent the diversity of perspectives of a target population in their generated responses. Concerningly, current empirical evidence suggests that frontier models fall short of this ideal: frontier LMs exhibit a lower linguistic variety (Russo et al., 2025) and a high homogeneity in their generated texts – a phenomenon dubbed the *generative monoculture* (Wu et al., 2025). In general, models produce both inter- and intra-homogeneous outputs, generating repetitive text that is also highly similar across model families (Jiang et al., 2025).

One way to assess how puralistically aligned a model is is to analyze the diversity of *perspectives* represented in the outputs it generates. *Perspective*, in the most abstract sense defined as "a particular way of considering something" (Cambridge University Press, n.d.), is embedded in every communication act (Basile et al., 2022). In NLP, perspectives are generally used as an umbrella term for subjective language, a facet of pluralistic alignment (Sorensen et al., 2024). Although perspective is difficult to operationalize – serving as the latent driver behind overt manifestations like sentiment, opinions, and claims – it provides a powerful

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

framework for evaluating model pluralism by unifying these otherwise disparate constructs into a single, holistic lens.

Due the inherent complexity of modelling perspectives and analysing free-form text, prior work mainly evaluated the ability of LLMs to model human perspectives on MCQ surveys (Santurkar et al., 2023) and Likert scales (Meister et al., 2025). Where free-form text has been considered, homogeneity has been measured using aggregated, high-level features (Wu et al., 2025), with little attention to the range of perspectives expressed in human text. By reducing perspectives to aggregated labels, prior approaches fail to model the very complexity they seek to measure, ultimately masking the extent of the *pluralistic gap* in model generation.

To address these limitations, we introduce a framework that formalizes *latent perspective* as a composite of aspects extracted from text, offering a fine-grained alternative to surface-level homogeneity metrics. By using aspects as building blocks to characterize the perspectives expressed in free-form text, our method enables a high-resolution comparison of how perspective distributions vary between human and LLM-generated content.

We conduct our study on a book review dataset sourced from the Goodreads platform (Wan & McAuley, 2018), consisting of diverse and opinionated reviews. We evaluate pluralism across open-source (Llama 3 8B, OLMo2 1B & 8B) and proprietary (GPT 4.1 and Gemini 2.5) models of varying sizes. To elicit higher coverage of perspectives in LLM outputs, we utilize high temperature as well as persona prompting (Ge et al., 2025), a technique applied heavily in population simulation. We apply our framework to the generated reviews and assess diversity using various quantitative distributional measures as well as a qualitative analysis of the pluralistic gap between human and LLM-generated reviews.

While our results confirm previous findings regarding the surface-level homogeneity of LLM outputs, we demonstrate that when evaluating *plurality of perspective*, models vary significantly in both baseline performance and their sensitivity to prompting. Furthermore, while some models and configurations might come close to encompassing a spectrum of diverse aspects (overton pluralism), rarer aspects are still disproportionately underrepresented (distributional pluralism). Our contributions can be summarized as follows: **(i)** we propose a domain-agnostic framework for comparing latent perspectives expressed in two sources of text, suitable for analysing model pluralism, **(ii)** we implement the framework on the opinionated dataset of book reviews across different LLMs and prompting methods designed to elicit diverse outputs to identify gaps in perspectives, and **(iii)** we analyze perspective gaps in various layers of abstraction and differences across models and prompting techniques.

2. Related Work

The tendency of LLMs to flatten the diversity of human perspectives by converging toward a more uniform distribution of features and semantics in their outputs has been studied under several names, including *artificial hivemind* (Jiang et al., 2025), *generative monoculture* (Wu et al., 2025), and *distributional gap* (Peeperkorn et al., 2025), and reframed constructively as the effort for higher model *pluralism*. Sorensen et al. (2024) propose three pluralistic alignment modes: *Overton*, representing the set of possible answers, *Distributional*, representing the distribution of possible answers, and *Steerable*, representing output similar to a chosen group to steer towards. Evaluation of these modes in prior work is predominantly discrete: opinion alignment is operationalised through MCQ choices or Likert ratings (Santurkar et al., 2023; DURMUS et al., 2024; Meister et al., 2025), or reduced to classification accuracy over attribute profiles and preference rankings (Adams et al., 2025; Chen et al., 2025). These approaches share a common limitation by reporting aggregate statistics over a population, discarding fine-grained information indicating how opinions are framed and which aspects are pivotal.

Fewer works study output homogenisation directly on the features of free-form text, as opposed to the projected labels. Wu et al. (2025) demonstrate generative monoculture in book reviews by measuring diversity through extracted attributes (binary sentiment and coarse topic labels), while Peeperkorn et al. (2025) measure the diversity gap in narrative generation via aggregate scalar scores such as the Vendi Score. While both argue that LLM-generated text exhibits a more narrow distribution of features compared to human text on the matching tasks and topics, their metrics collapse the distributional structure of generated text into a single diversity score, making it difficult to identify which perspectives are systematically absent. A deeper, more structured evaluation is necessary, one that takes into account the multifaceted properties of perspective.

The limited evaluation on free-form text largely reflects the difficulty of working with it. Even within perspectivism (Frenda et al., 2025) – a paradigm that emphasizes preserving diverse perspectives across NLP tasks – perspective is typically operationalized as task-specific labels assigned to instances. These labels can then support various modeling and analysis approaches, such as creating annotator profiles using clustering (Vitsakis et al., 2024), but perspective is still effectively anchored to a label space, leaving the nuances of free-form expression largely unaddressed. Clustering has been used as a method of identifying common elements in various subjective tasks, both in argumentation (Reimers et al., 2019) and framing (Ajjour et al., 2019), though it has not been used for measuring and comparing the the homogeneity of LLMs, a gap which we fill with this paper.

3. A Framework for Perspective Analysis

3.1. Motivation

Perspectives in text are expressed through inherently subjective language, influenced by individual interpretation, emphasis, and context. A prerequisite for automatic extraction of perspective is operationalizing it, which is difficult because precise definitions and appropriate levels of granularity are context-sensitive. This results in two central challenges: (1) how to construct a meaningful representation of diverse perspectives on a given topic and (2) how to compare the distribution of perspectives originating from different data sources, such as human and LLM-generated text.

To address these challenges, we propose a two-tiered framework for identifying perspectives represented in a collection of texts and assessing their diversity. The first tier operates at the level of *aspects*, where the representations of individual discourse units from the texts are isolated and grouped to capture recurring semantic patterns. The second tier operates at the *perspective level*, using coarse labels of aspect groups produced at the previous level to characterize the perspective, both underlying and verbalized in the text.

The proposed framework enables a coarse-grained analysis of perspectives from free-form text, as well as a fine-grained comparison of the underlying aspects at the first-tier level. A concrete implementation of the framework requires operationalizing its core components: the first tier of aspect identification, the second tier where aspects are grouped into perspectives, and defining metrics that compare perspective diversity from different sources. In the following section, we provide the formal definition of the framework (§3.2), followed by a concrete implementation (§3.3) applied to a book corpus (§5) to rigorously verify the validity of its components (§6).

3.2. Definition

Let $D = \{d_1, d_2, \dots, d_n\}$ be a dataset of texts sharing a common topic, such as reviews of the same book.

Aspects. Each text d_i can be decomposed into a sequence of m_i aspect instances:

$$d_i = [a_1^i, a_2^i, \dots, a_{m_i}^i] \tag{1}$$

where each a_j^i represents the aspect expressed in unit j of text i. Common aspects should then be grouped across D, yielding a set of k aspect clusters $\mathcal{C} = \{C_1, \dots, C_k\}$. Each aspect instance a_j^i is assigned to exactly one cluster $C_l \in \mathcal{C}$, or designated as an outlier and excluded:

$$a_j^i \mapsto \begin{cases} C_l & \text{if } a_j^i \text{ belongs to a recognised cluster} \\ \varnothing & \text{if } a_j^i \text{ is an outlier} \end{cases} \tag{2}$$

Perspective. We define the perspective of a text d_i as the distribution over clusters induced by its non-outlier aspect instances. Let n_l^i denote the number of aspect instances in d_i assigned to cluster C_l, and let $N_i = \sum_{l=1}^{k} n_l^i$ be the total number of non-outlier aspect instances. The perspective of d_i is then the vector:

$$\mathbf{p}_i = \left(\frac{n_1^i}{N_i}, \frac{n_2^i}{N_i}, \dots, \frac{n_k^i}{N_i} \right) \in \Delta^{k-1} \tag{3}$$

where Δ^{k-1} denotes the $(k-1)$-dimensional probability simplex.

3.3. Implementation

3.3.1. Aspect level

In our work, we select a sentence as the target discourse unit representing an aspect, and use sentence embeddings as their semantic representation within the framework. We do not construct a structured representation of aspects, which keeps the approach context-independent and simplifies the construction of the framework. We provide a more structured analysis (§6.1) and validate the quality of this representation (§6.2) in later sections. To estimate the distribution of aspects present in book reviews, we (1) segment the original reviews into sentences, then (2) embed each sentence using a sentence encoder model, and finally (3) group the embeddings into aspect clusters.

Implementation. We use spaCy to segment free-form text into sentences.[1] We opt for the F2LLM 0.6B model (Zhang et al., 2025) as the sentence embedder, as it currently achieves the best clustering performance among sub-1B parameter English models on the MTEB leaderboard.[2] To cluster the aspects encoded within sentences, we build on the BERTopic (Grootendorst, 2022) library, and choose HDBSCAN (Campello et al., 2013) as the aspect-level clustering algorithm, as unlike centroid-based methods such as k-means, it does not require the number of clusters to be specified a priori.

This property is crucial in our setting, as the number of aspect clusters in a group of reviews is not known beforehand, and predefining it would impose unnecessary constraints. Prior to clustering, we reduce the dimensionality of sentence embeddings to 5 using UMAP (McInnes et al., 2018). We select 1500 human reviews per book for our analysis, and split them into a fixed train (R_{train}) and test (R_{test}) set, with $|R_{train}| = 1000$ and $|R_{test}| = 500$. We compute the clusters on the train set, and use the held-out test set for comparison with LLM-generated reviews. We opt for this split to ensure a larger set of reviews that serves both as a

[1] https://spacy.io/models/en
[2] https://huggingface.co/spaces/mteb/leaderboard

stable base for cluster computation and as a more accurate approximation of the full distribution, while the smaller number of held-out reviews allows for an apples-to-apples comparison with LLM-generated reviews. The sentencized and embedded reviews from R_{train} are used to fit a clustering model for each book (b), resulting in K_b clusters per book. Sentences from R_{test} and LLM-generated reviews are then assigned aspect cluster labels using the fit clustering models.

Evaluation. We use two metrics to evaluate aspect-level differences between human and LLM-generated texts. If human and LLM-generated reviews are properly aligned, the cluster assignment should be: (1) complete, covering all clusters to reflect *Overton pluralism*, and (2) proportionate, maintaining the same ratios to reflect *distributional pluralism*. We operationalize these using cluster coverage percentage and Jensen-Shannon Divergence (JSD), respectively, measured between the cluster distributions at a book level. Apart from these contrastive measures, we also include individual population statistics which measure aspect diversity within reviews originating from a single source. As these statistics, we opt for average semantic similarity, measuring sentence homogeneity, and aspect cluster entropy.

3.3.2. Perspective level

To aggregate the individual aspects to the perspective level, we opt for two approaches based on how we interpret aspect cluster assignment – either a *hard* or a *soft* clustering approach. Each review is then encoded as a set of aspects based on the cluster assignments of its constituent sentences. Since HDBSCAN is a probabilistic model, the output can be interpreted as the distribution across K_b aspect clusters, or as the most probable cluster. We construct the perspective vectors as either the count of aspects in a given cluster (*hard* vectors) or an aggregation of all individual aspect probabilities (*soft* vectors).

Implementation. After constructing the perspective vectors in the same human (R_{train}, R_{test}) and LLM-generated sets, we implement another layer of clustering models to determine the represented perspectives. In contrast to the aspect level, perspective vectors are sparser, and the previous density-based HDBSCAN model is no longer effective. Therefore, we experiment with other clustering models, including K-means and community detection.

Evaluation. Akin to the aspect layer, we evaluate the distribution of perspectives using total cluster coverage and the Jensen-Shannon Divergence (JSD) to explicitly compare Overton and distributional pluralism, complemented by cosine similarity of the perspective representation vectors to assess population diversity.

4. Dataset and Models

4.1. Dataset

We utilize the Goodreads dataset (Wan & McAuley, 2018) for all experiments. The dataset consists of book reviews and corresponding metadata across various genres. We choose this dataset because it contains opinionated texts not explicitly related to political stances. Furthermore, the 2017 data cut-off nullifies the risk of LLM-generated text.

For our analysis, we select from English-language books that contain more than 1500 reviews. To ensure a representative sample, we sample 20 books from each genre, including five from each of the following categories: highest average score, lowest average score, highest score deviation, and highest number of reviews. The represented genres are *Mystery, Thriller and Crime*, *Young Adult*, *History and Biography*, *Fantasy and Paranormal*, and *Romance*. From these, we select a preliminary *analysis subset* of 10 books sampled across genres used in initial testing and analyses. We filter out reviews not written in English or those shorter than 20 characters, using *langdetect* for language detection.[3]

4.2. Models and prompting configurations

We run our experiments across various closed- and open-source LLMs. For closed-source models, we select `gpt-4.1-mini-2025-04-14` (Achiam et al., 2023) and `gemini-flash-2.5-mini` (Comanici et al., 2025) (abbreviated hereafter as `GPT 4.1` and `Gemini 2.5`). For open-source models, we select `Llama-3.1-8B` (Grattafiori et al., 2024) as well as `OLMo-2-1B` and `OLMo-2-8B` (OLMo et al., 2025). The availability of pre-training corpora for OLMo models allows us to verify whether they were exposed to the review texts.

We measure and compare the diversity of perspectives in book reviews generated by humans and LLMs under different prompting setups. We generate 300 reviews for the 10-book subset and 100 reviews for the full 100-book set. We utilize three different prompting setups:

Baseline. *Baseline* prompts contain the vanilla instruction to write a book review, as well as rate it. We use a temperature $T = 0.7$ to obtain variety across samples.

High temperature. Higher temperature values are often used in creative tasks (Wu et al., 2025). To evaluate the influence of higher temperature on review diversity, we utilize the baseline prompt, but evaluate it under $T = 1.5$.

Persona prompts. Socio-demographic prompting and persona prompting are promising ways to increase the diversity

[3] `https://pypi.org/project/langdetect/`

of LLM output. We select a fixed batch of 300 personas from Ge et al. (2025) and use each persona to generate a single review.

We report prompts used across all setups in Appendix A.

5. Results

In this section, we first present aggregate measures (§5.1) and then fine-grained aspect- and perspective-level comparisons from our framework (§5.2).

5.1. High-level measures

We first perform a set of analyses focusing on high-level, aggregated measures to assess consistency with prior findings and to confirm differences in superficial features between the baseline and other setups. We report results on the analysis subset of 10 books, comparing 300 model-generated reviews with 300 held-out human reviews from R_{test}.

Semantic similarity. To measure the novelty of generated reviews across runs, we process reviews for each book in sequential batches of 10. For each new batch, we compute the cosine similarity between every sentence embedding in the incoming batch and all sentence embeddings from previous batches, recording the maximum value. We then average this score across books to obtain a mean-max similarity score. A high mean-max similarity indicates that new reviews are progressively more repetitive, while a low value suggests they introduce novel content.

Evaluating the original and generated reviews separately allows us to compare how quickly each source saturates. Figure 2 shows the mean-max similarity curves for human reviews and those generated by `GPT 4.1`. The results point to two major findings: (1) diversity saturates at around 100 reviews, leading us to opt for that sample size in further experiments, and (2) the semantic similarity of human reviews is lower than that of generated reviews, with *persona* prompting having a positive influence on review diversity. We report the results for other models in Appendix B.

Sentiment and rating. The review dataset is highly subjective, where attitudes toward a book are conveyed both implicitly through sentiment and explicitly through assigned ratings ranging from 1 to 5. Wu et al. (2025) show LLM-generated reviews are more positive in sentiment than original reviews, but do not compare the assigned rating. We estimate sentiment of book reviews using `distilbert-base-uncased-finetuned-sst-2-english` (Sanh et al., 2019). The sentiment range is $[-1, 1]$, with -1 denoting negative and $+1$ positive sentiment, respectively.

We report the average sentiment and assigned ratings across books for original and generated reviews in Table 1. Results

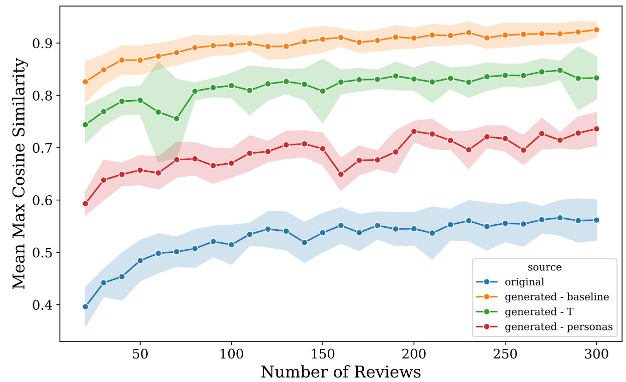


Figure 2. Mean maximum cosine similarity between embeddings for original reviews and across generation configurations for `GPT 4.1`. Results for other models are presented in Appendix B.

differ across models, with `Gemini 2.5` producing the lowest average rating and sentiment, and `GPT 4.1` the highest. Similar to our findings for semantic similarity, the baseline prompts are farthest from the original scores, scoring highly both in average sentiment (a trend most pronounced for `GPT 4.1`) and average assigned rating. Higher temperature sampling ($T = 1.5$) produces mixed results, usually reducing the average rating and sentiment while increasing the corresponding standard deviation. Persona prompting generates ratings closest to the original scores; notably, `Gemini 2.5` produces lower average ratings than the original reviews.

Table 1. Mean$_{\pm\text{std}}$ book review rating and detected sentiment per model and generation configuration.

		GPT 4.1	Gemini 2.5	Llama3 8B	OLMo2 1B	OLMo2 7B
Rating	base	$4.07_{\pm.57}$	$3.40_{\pm.86}$	$4.11_{\pm.35}$	$3.11_{\pm.73}$	$4.16_{\pm.41}$
	temp	$4.09_{\pm.56}$	$3.46_{\pm.88}$	$4.08_{\pm.47}$	$3.15_{\pm.99}$	$4.20_{\pm.58}$
	pers	$3.48_{\pm.70}$	$3.02_{\pm.98}$	$3.58_{\pm.69}$	$3.21_{\pm.74}$	$3.64_{\pm.76}$
	original			$3.52_{\pm1.30}$		
Sentiment	baseline	$0.92_{\pm.39}$	$0.45_{\pm.87}$	$0.67_{\pm.72}$	$0.34_{\pm.92}$	$0.91_{\pm.40}$
	temp	$0.89_{\pm.44}$	$0.46_{\pm.86}$	$0.67_{\pm.72}$	$0.28_{\pm.93}$	$0.84_{\pm.51}$
	personas	$0.66_{\pm.73}$	$0.26_{\pm.92}$	$0.37_{\pm.90}$	$0.41_{\pm.89}$	$0.70_{\pm.70}$
	original			$0.03_{\pm.95}$		

5.2. Framework results

Aspect level. To compare the topic distributions of human and LLM-generated reviews, we use two measures. First, we estimate alignment with Overton pluralism, reflecting the breadth of the generating corpus through *cluster coverage*, which measures the percentage of human-identified clusters that appear at least once in the generated reviews. Second, to estimate distributional pluralism – whether LLM-generated aspects are discussed in proportions similar to those of human reviewers – we compute the JSD, quantifying how much the generated corpus diverges from the human baseline in terms of aspect distributions.

We report results across models and configurations in Table 2. Persona prompting consistently yields higher topic coverage and lower divergence from the human distribution than temperature-based sampling across all models and metrics. Gemini-2.5 under persona prompting reaches the highest coverage ($98.0 \pm 2.6\%$) and lowest divergence ($\text{JSD} = 0.11 \pm 0.02$), suggesting it most closely approximates both Overton and distributional pluralism. Interestingly, GPT-4 lags behind both Gemini-2.5 and the smaller Llama-3.1-8B across all configurations, peaking at only $62.7 \pm 3.4\%$ coverage under persona prompting. Higher temperature ($T = 1.5$ vs. $T = 0.7$) yields modest but consistent gains on both metrics. Overall, we find that the choice of prompting strategy has a stronger effect on topical diversity than temperature scaling and that persona prompting can effectively narrow the pluralistic gap.

Table 2. Cluster coverage and JSD (mean$_{\pm\text{std}}$ across books). Cluster coverage estimates Overton pluralism, while JSD estimates distributional pluralism.

	Setup	GPT 4.1	Llama 3.1 8B	Gemini 2.5	OLMo2 1B	OLMo2 7B
Cover.	$t = 0.7$	$41.4_{\pm 9.5}$	$63.8_{\pm 9.2}$	$71.8_{\pm 8.9}$	$72.7_{\pm 10.4}$	$59.3_{\pm 10.9}$
	$t = 1.5$	$51.9_{\pm 11.4}$	$76.1_{\pm 7.7}$	$76.9_{\pm 8.6}$	$83.8_{\pm 8.0}$	$74.4_{\pm 9.0}$
	Personas	$65.5_{\pm 11.0}$	$92.2_{\pm 5.2}$	$97.9_{\pm 2.6}$	$75.6_{\pm 9.8}$	$69.1_{\pm 10.4}$
JSD^2	$t = 0.7$	$0.33_{\pm .07}$	$0.25_{\pm .05}$	$0.22_{\pm .05}$	$0.23_{\pm .06}$	$0.27_{\pm .06}$
	$t = 1.5$	$0.30_{\pm .06}$	$0.23_{\pm .04}$	$0.21_{\pm .05}$	$0.21_{\pm .05}$	$0.24_{\pm .05}$
	Personas	$0.26_{\pm .06}$	$0.17_{\pm .04}$	$0.12_{\pm .03}$	$0.23_{\pm .06}$	$0.24_{\pm .05}$

Table 3. Intra-cluster semantic similarity and entropy (mean$_{\pm\text{std}}$ across books).

	Setup	GPT 4.1	Llama 3.1 8B	Gemini 2.5	OLMo2 1B	OLMo2 7B
Cl. Sim.	$t = 0.7$	$0.68_{\pm .05}$	$0.56_{\pm .05}$	$0.51_{\pm .05}$	$0.46_{\pm .04}$	$0.56_{\pm .04}$
	$t = 1.5$	$0.61_{\pm .04}$	$0.47_{\pm .04}$	$0.49_{\pm .04}$	$0.38_{\pm .03}$	$0.47_{\pm .03}$
	Personas	$0.46_{\pm .04}$	$0.40_{\pm .04}$	$0.30_{\pm .04}$	$0.44_{\pm .03}$	$0.50_{\pm .05}$
	Original: $0.30_{\pm .04}$					
H (bits)	$t = 0.7$	$3.14_{\pm .54}$	$3.65_{\pm .65}$	$3.85_{\pm .71}$	$3.74_{\pm .64}$	$3.53_{\pm .61}$
	$t = 1.5$	$3.37_{\pm .56}$	$3.73_{\pm .69}$	$3.91_{\pm .71}$	$3.86_{\pm .65}$	$3.69_{\pm .64}$
	Personas	$3.59_{\pm .60}$	$4.13_{\pm .69}$	$4.49_{\pm .77}$	$3.77_{\pm .63}$	$3.68_{\pm .63}$
	Original: $5.02_{\pm .92}$					
$\hat{H}$	$t = 0.7$	$0.77_{\pm .10}$	$0.77_{\pm .10}$	$0.78_{\pm .09}$	$0.76_{\pm .09}$	$0.76_{\pm .09}$
	$t = 1.5$	$0.76_{\pm .09}$	$0.74_{\pm .09}$	$0.78_{\pm .09}$	$0.75_{\pm .08}$	$0.74_{\pm .08}$
	Personas	$0.75_{\pm .08}$	$0.78_{\pm .08}$	$0.83_{\pm .08}$	$0.75_{\pm .09}$	$0.76_{\pm .09}$
	Original: $0.92_{\pm .10}$					

Perspective level. To estimate perspective-level coverage, we utilize metrics similar to the aspect level: JSD, cluster coverage, and normalized entropy. We also include a novel perspective-level metric – *perspective diversity* – which measures the average number of perspective clusters covered in a set of reviews relative to those in the test set, where total number of test set perspectives identified by clustering is 5.

We report results of all metrics in Table 4. We find that perspective-level coverage is considerably lower across all models than aspect-level coverage, indicating that while constituent aspects are adequately represented in generated texts, the generated perspectives are still largely homogeneous. This homogeneity is best seen through the perspective diversity metric, which is considerably lower than in the human reviews, with LLM-generated reviews frequently mapping to only one or two perspective clusters. These results show that while LLMs are capable of generating texts corresponding to diverse aspects, this diversity is merely performative - with the overall perspectives present in the reviews still largely *monocultural*.

Table 4. JSD^2, normalised entropy, perspective diversity, and perspective coverage (mean$_{\pm\text{std}}$ across books).

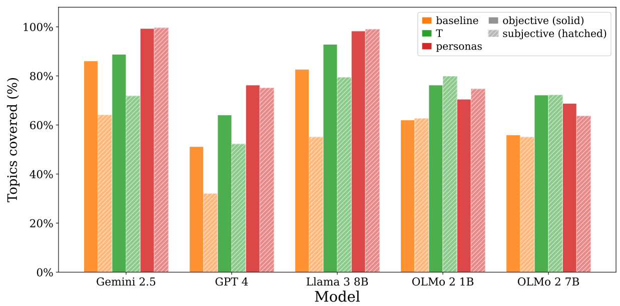
	Setup	GPT 4.1	Llama 3.1 8B	Gemini 2.5	OLMo2 1B	OLMo2 7B
Persp. Cov.	Baseline	$35.9_{\pm 16.7}$	$31.7_{\pm 14.5}$	$28.5_{\pm 13.1}$	$43.4_{\pm 19.6}$	$38.4_{\pm 18.0}$
	$t = 1.5$	$40.8_{\pm 18.2}$	$36.4_{\pm 16.2}$	$29.1_{\pm 12.8}$	$51.6_{\pm 20.4}$	$44.7_{\pm 18.2}$
	Personas	$50.4_{\pm 19.8}$	$38.4_{\pm 17.6}$	$33.8_{\pm 13.3}$	$39.5_{\pm 17.8}$	$44.4_{\pm 19.3}$
	Original: $97.6_{\pm 7.1}$					
JSD^2	$t = 0.7$	$0.085_{\pm .061}$	$0.080_{\pm .046}$	$0.087_{\pm .050}$	$0.070_{\pm .039}$	$0.075_{\pm .041}$
	$t = 1.5$	$0.075_{\pm .046}$	$0.077_{\pm .041}$	$0.087_{\pm .055}$	$0.063_{\pm .037}$	$0.070_{\pm .038}$
	Personas	$0.065_{\pm .034}$	$0.071_{\pm .036}$	$0.075_{\pm .036}$	$0.073_{\pm .038}$	$0.071_{\pm .039}$
	Original: —					
Norm. $\hat{H}$	$t = 0.7$	$0.218_{\pm .290}$	$0.204_{\pm .308}$	$0.146_{\pm .282}$	$0.247_{\pm .244}$	$0.233_{\pm .289}$
	$t = 1.5$	$0.251_{\pm .284}$	$0.244_{\pm .316}$	$0.145_{\pm .271}$	$0.299_{\pm .241}$	$0.283_{\pm .291}$
	Personas	$0.305_{\pm .257}$	$0.212_{\pm .271}$	$0.175_{\pm .253}$	$0.235_{\pm .259}$	$0.252_{\pm .268}$
	Original: $0.496_{\pm .134}$					
Persp. Div.	Baseline	$1.74_{\pm .79}$	$1.54_{\pm .70}$	$1.38_{\pm .62}$	$2.11_{\pm .96}$	$1.87_{\pm .88}$
	$t = 1.5$	$1.99_{\pm .88}$	$1.77_{\pm .79}$	$1.41_{\pm .60}$	$2.52_{\pm .99}$	$2.19_{\pm .90}$
	Personas	$2.46_{\pm .97}$	$1.86_{\pm .82}$	$1.64_{\pm .63}$	$1.92_{\pm .84}$	$2.15_{\pm .91}$
	Original: $4.88_{\pm .36}$					

6. Analysis

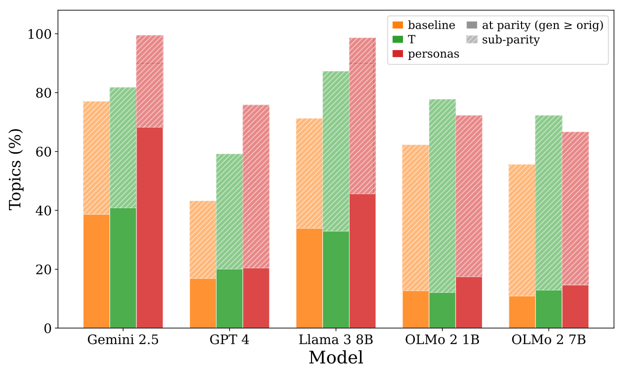
We now analyse the validity of our framework by evaluating cluster contents and quality and approximating the proportion of reviews present in model pre-training sets. We first categorize the aspect clusters into meaningful topic categories and analyse their coverage (§6.1), then evaluate the coherence and separability of the aspect clusters (§6.2), and finally analyze the pre-training corpus of the `OLMo` model family (§6.3).

6.1. Topic evaluation

Results of our experiments (§5) have identified the pluralistic gap, where the LLM-generated reviews lack diversity of perspectives. We now aim to identify which aspects, and perspectives, are consistently underrepresented. To do this, we label each inferred aspect cluster with a category and corresponding sentiment. For each of the aspect clusters across the analysis subset, we use five sentences closest to each aspect centroid to label clusters using `GPT-4.1-mini`. For labels, we use a predefined taxonomy of 24 categories proposed by Yang & Jin (2025) as a fixed set, making labels consistent and directly comparable across books. The categories span both more objective (plot, characters, writing style) and subjective categories (emotional impact, enjoyment, and expectation fulfillment). We also assign a sentiment label (positive, negative, neutral, or mixed) expressed in the aspect toward that category. The resulting labels allow us to pinpoint the reasons LLMs diverge from human reviews in the topic space, and exactly along which aspects.



(a) Objective vs. subjective coverage.



(b) Parity of coverage across models.

Figure 3. Topic coverage across generation modes for objective and subjective categories (3a), and parity of aspect coverage compared to original distribution (3b).

To jointly evaluate the coverage of aspects across books, as well as study the influence of prompting configurations, we aggregate aspects into subjective and objective categories, then measure which percentage of those aspects was covered across configurations and models. In Figure 3a we report the difference in coverage between objective (full bar) and subjective (shaded bar) aspects. We observe that across models, persona prompting generally affects the coverage of subjective aspects, even in cases where it does not reach maximum coverage (with `GPT 4.1`). In OLMo models, higher temperature has a higher influence in topic coverage than persona prompting. Finally, model size doesn't necessarily imply lower coverage, as `OLMo2 1B` generally achieves higher coverage than `GPT 4.1`.

However, Overton pluralism measured by topic coverage (1 review per topic at least) is not informative enough. In Figure 3b we compare whether the distributions of the LLM-generated topics correspond to human ones. These results indicate that Overton pluralism doesn't guarantee distributional pluralism, showing models report more spurious rather than systematic coverage of diverse topics.

6.2. Cluster coherence

We now aim to check whether the produced cluster assignments are valid. We evaluate aspect-cluster coherence using a leave-one-out intrinsic probe: for each cluster, we sample four representative sentences and one intruder sentence from a different cluster. Then, we use`GPT 4.1` as LLM-as-a-judge to determine the intruder. We vary this setup across two dimensions: (1) whether representative sentences are drawn from the centroid or sampled randomly, and (2) whether the outlier comes from a random or the closest neighbouring cluster. The prompt is provided in Appendix A.

We report the results in Table 5. Intruder detection accuracy ranges from 72.4% in the hardest condition (random–closest) to 94.6% in the easiest (centroid–random), against a 20% random baseline, confirming that the inferred topic structure is coherent and separable.

Table 5. Leave-one-out topic coherence accuracy across representative document and outlier selection strategies (baseline = 20.0%).

	$\text{outlier}_{\text{random}}$	$\text{outlier}_{\text{closest}}$
$\text{in-cluster}_{\text{centroid}}$	94.6%	94.3%
$\text{in-cluster}_{\text{random}}$	77.4%	72.4%

6.3. Memorization

In this section, we aim to estimate the proportion of review data present in model pre-training data. We conduct two complementary analyses of the `DCLM` corpus (Li et al., 2024), which is used in the pre-training pipeline of `OLMo` models (OLMo et al., 2025), and likely contributes to the training mixtures of other LLMs as well. In both analyses, we first retrieve all URLs corresponding to Goodreads from DCLM, as well as alternative book-review sources that contain user-generated literary reviews, book overviews, or related discussion content. The first analysis focuses specifically on Goodreads-derived content within DCLM, while the second examines alternative book-review and reading-community websites represented in the crawl.

Goodreads. The Goodreads subset of the DCLM crawl contains $25,831$ pages from two page types: $19,052$ individual review pages (`goodreads.com/review/show/`) and $6,779$ book-summary pages (`goodreads.com/book/show/`). Title extraction succeeded for $20,803$ pages (80.5%), yielding $11,993$ unique book titles. Because each `review/show` page contains one complete review, these pages contribute $19,052$ reviews directly. The `book/show` pages additionally embed up to 30 community reviews each; using truncation markers as a conservative estimate yields approximately $131,441$ further review texts, for a total estimate of $150,493$ Goodreads reviews. Linking crawl pages to the goodreads/books metadata via Goodreads IDs or normalized titles matched $14,873$ distinct books containing 16.1 million Goodreads reviews in total. The crawl captured approximately $131,971$ of these reviews, corresponding to an overall coverage rate of 0.82%. The coverage is highly skewed: 32.6% of matched books have more than 5% of their Goodreads reviews present in the crawl, while 12.5% have coverage below 0.1%. Among the 100 selected books used for downstream analysis, 77 appear in the crawl, yielding 612 matched review/show pages. The high number of Goodreads reviews present in the training data confirms the LLMs were exposed to community reviews, validating they could feasibly reproduce the diverse perspectives present in those reviews.

Other sources. To complement the Goodreads-focused analysis, we additionally examine review and discussion content originating from alternative reading-community and book-review platforms present in the DCLM crawl. We use LLM-as-a-judge on a random sample of content from 100 URLs of chosen book-related domains, estimating whether and how many book-reviews and book-overviews were present in the page. We then multiply the average number of review and overviews with the total number of pages for that domain. We report complete results in Appendix C. In total, the estimate comes to over $300k$ reviews and over $340k$ overviews, confirming that LLMs were exposed to an wide amount of evaluative book content.

7. Conclusion

In our work, we present a framework for evaluating pluralism in LLMs by first identifying constituent aspects of perspectives from free-form text and then grouping them into perspective representations. This two-level scheme extends beyond previously used aggregate statistics and allows for a fine-grained analysis of the *pluralistic gap* between human and LLM generations. We propose one concrete instance of our framework and apply it to a dataset of book reviews, extracting aspect and perspective distributions and allowing for comparison between authentic human reviews and LLM generated ones. We verify results from previous works which indicate that LLMs are part of a *generative monoculture*, but also show that the pluralistic gap can be bridged by utilizing prompting techniques such as persona prompting. Our subsequent analyses verify the validity of our method through cluster coherence and separability and identify the coverage of which topics was improved. Taken together, we offer a principled way of analysing fine-grained diversity of perspectives in free-form texts which can be used to identify concrete targets for pluralistic alignment.

Impact Statement

In this work, we study pluralism in LLMs through fine-grained analysis of aspects which are constituent of human perspectives. We identify the *pluralistic gap* and evaluate whether it can be bridged using simple prompting techniques. As LLM generated content becomes more prevalent, and humans interact with LLMs in an increasing degree, it is paramount to fairly represent individual voices, no matter how infrequent they may be. Concerningly, various works show that LLMs sharpen the distribution of perspectives from training data, either not encoding or not generating less frequent ones. Our work provides a fine-grained, structured framework for analysing this phenomenon, and paves a way towards effective mitigation strategies aimed at mitigating the pluralistic gap.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Adams, J., Hu, B., Veenhuis, E., Joy, D., Ravichandran, B., Bray, A., Hoogs, A., and Basharat, A. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. 2025.

Ajjour, Y., Alshomary, M., Wachsmuth, H., and Stein, B. Modeling frames in argumentation. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2922–2932, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1290. URL https://aclanthology.org/D19-1290/.

Basile, V., Caselli, T., Balahur, A., and Ku, L.-W. Editorial: Bias, subjectivity and perspectives in natural language processing. *Frontiers in Artificial Intelligence*, Volume 5 - 2022, 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.926435. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.926435.

Cambridge University Press. Perspective. https://dictionary.cambridge.org/dictionary/english/perspective, n.d. Accessed: 2026-05-05.

Campello, R. J. G. B., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates.

In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G. (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Chen, Q. Z., Feng, K., Park, C. Y., and Zhang, A. X. SPICA: Retrieving scenarios for pluralistic in-context alignment. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 748–765, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.41. URL https://aclanthology.org/2025.findings-acl.41/.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DURMUS, E., Nguyen, K., Liao, T., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

Fleisig, E., Blodgett, S. L., Klein, D., and Talat, Z. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2279–2292, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.126. URL https://aclanthology.org/2024.naacl-long.126/.

Frenda, S., Abercrombie, G., Basile, V., et al. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59:1719–1746, 2025. doi: 10.1007/s10579-024-09766-4.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. Scaling synthetic data creation with 1,000,000,000 personas, 2025. URL https://arxiv.org/abs/2406.20094.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., and Choi, Y. 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*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y., Nezhurina, M., Abbas, A., Hsieh, C.-Y., Ghosh, D., Gardner, J., Kilian, M., Zhang, H., Shao, R., Pratt, S., Sanyal, S., Ilharco, G., Daras, G., Marathe, K., Gokaslan, A., Zhang, J., Chandu, K., Nguyen, T., Vasiljevic, I., Kakade, S., Song, S., Sanghavi, S., Faghri, F., Oh, S., Zettlemoyer, L., Lo, K., El-Nouby, A., Pouransari, H., Toshev, A., Wang, S., Groeneveld, D., Soldaini, L., Koh, P. W., Jitsev, J., Kollar, T., Dimakis, A. G., Carmon, Y., Dave, A., Schmidt, L., and Shankar, V. Datacomp-lm: In search of the next generation of training sets for language models, 2024.

McInnes, L., Healy, J., Saul, N., and Großberger, L. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Meister, N., Guestrin, C., and Hashimoto, T. Benchmarking distributional alignment of large language models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *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)*, pp. 24–49, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.2. URL https://aclanthology.org/2025.naacl-long.2/.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., ..., and Hajishirzi, H. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models, 2025. URL https://arxiv.org/abs/2507.20956.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Yue, S., Wang, A., and Hendrycks, D. Humanity's last exam. *CoRR*, abs/2501.14249, 2025. doi: 10.48550/ARXIV.2501.14249. URL https://doi.org/10.48550/arXiv.2501.14249.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. Classification and clustering of arguments with contextualized word embeddings. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 567–578, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1054. URL https://aclanthology.org/P19-1054/.

Russo, G., Nozza, D., Röttger, P., and Hovy, D. The pluralistic moral gap: Understanding judgment and value differences between humans and large language models, 2025. URL https://arxiv.org/abs/2507.17216.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29971–30004, 2023.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Vitsakis, N., Parekh, A., and Konstas, I. Voices in a crowd: Searching for clusters of unique perspectives. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12517–12539, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.696. URL https://aclanthology.org/2024.emnlp-main.696/.

Wan, M. and McAuley, J. J. Item recommendation on monotonic behavior chains. In Pera, S., Ekstrand, M. D., Amatriain, X., and O'Donovan, J. (eds.), *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pp. 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

Wu, F., Black, E., and Chandrasekaran, V. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yZ7sn9pyqb.

Yang, D. and Jin, Q. What matters in evaluating book-length stories? a systematic study of long story evaluation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16375–16398, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.799. URL https://aclanthology.org/2025.acl-long.799/.

Zhang, Z., Liao, Z., Yu, H., Di, P., and Wang, R. F2llm technical report: Matching sota embedding performance with 6 million open-source data, 2025. URL https://arxiv.org/abs/2510.02294.

A. Prompts

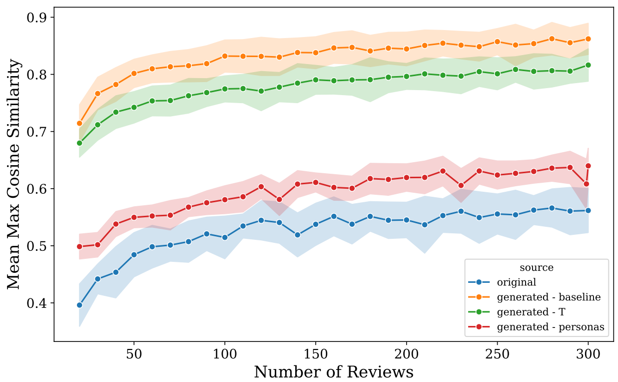
We report all prompts used to generate reviews across various setups and LLM-as-a-judge for cluster coherence, labelling and pretraining data analysis in Table 6.

Table 6. Prompts used for generating reviews (top), leave-one-out cluster coherence evaluation (second), cluster labelling (third), and web page classification (bottom).

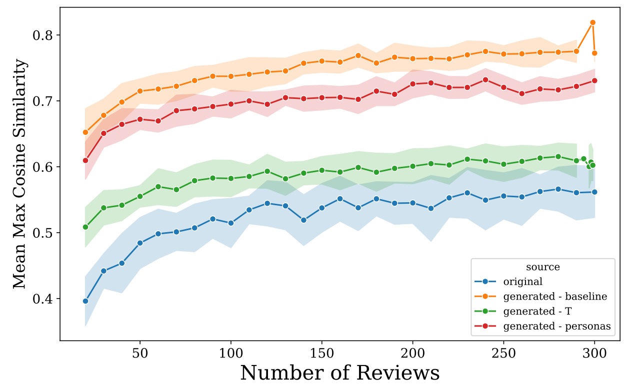
<table>
<tr><th>Usage</th><th>Prompt Type</th><th>Content</th></tr>
<tr><td rowspan="3">Review Generation</td><td>system</td><td>"You just finished reading a book."</td></tr>
<tr><td>baseline user</td><td>"Write a book review for [TITLE] by [AUTHOR]. Provide an integer rating from 1 to 5, and a written review. Format it like 'Rating: X/5, Review:...'"</td></tr>
<tr><td>persona user</td><td>"You are [PERSONA]. Write a book review for [TITLE] by [AUTHOR]. Provide an integer rating from 1 to 5, and a written review. Format it like 'Rating: X/5, Review:...'"</td></tr>
<tr><td rowspan="3">Topic Coherence Evaluation</td><td>system</td><td>"You are an expert at analyzing book reviews. You will receive 5 book review sentences. Four of them come from the same topic cluster and share a common theme or aspect. One is an outlier from a different cluster. Your task: identify which sentence (by number, 1–5) is the odd one out — the one that does NOT belong with the other four. Respond with a JSON object only."</td></tr>
<tr><td>user</td><td>"Five book review sentences: 1. [s1] ... 5. [s5]. Which sentence is the odd one out?"</td></tr>
<tr><td>response format</td><td>{"outlier": <int 1–5>, "reason": <string>}</td></tr>
<tr><td rowspan="3">Cluster Labelling</td><td>system</td><td>"You are an expert at analyzing book reviews. You will receive 5 representative sentences from a single review topic cluster. Your task is to identify: 1. The aspect of the book being discussed — pick the single best match from the provided enum. 2. The sentiment towards that aspect (positive, negative, or mixed). Base your judgment on what all 5 sentences have in common. Respond with a JSON object only."</td></tr>
<tr><td>aspects enum</td><td>plot_development, structure, ending, character_development, characterization, character_relationships, character_diversity, writing_style, language, readability, theme_exploration, theme_depth, world_building, setting, empathy, emotional_depth, enjoyment, engagement, genre_fulfillment, premise_fulfillment, originality, content_warnings, cover_design, personal_bias</td></tr>
<tr><td>response format</td><td>{"aspect": <enum>, "sentiment": "positive" | "negative" | "mixed"}</td></tr>
<tr><td rowspan="2">Page Classification</td><td>user</td><td>"You are a data quality analyst. I will give you the text content scraped from a web page. Your task: determine how many individual book reviews and book overviews are present. A 'book review' evaluates or discusses a specific book (professional critic, reader review, or blog post). A 'book overview OR excerpt' is a summary or recap of a book's plot/content without being evaluative — e.g. publisher blurbs, plot summaries, 'about the book' sections, editorial synopses, or excerpts. Do not flag movie or series reviews, or overviews on non-book topics. Page URL: [URL]. Page text (truncated): [TEXT]."</td></tr>
<tr><td>response format</td><td>{"review_count": <int>, "has_reviews": <bool>, "overview_count": <int>, "has_overviews": <bool>, "confidence": "high"|"medium"|"low", "notes": <one short sentence or empty string>}</td></tr>
</table>

B. Semantic Similarity

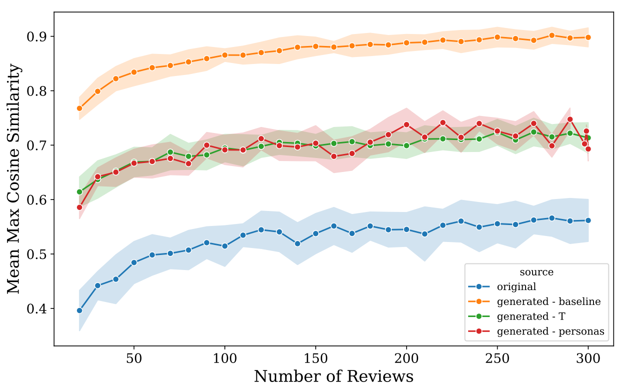
We report the mean max semantic similarity of novel batches of reviews for other models used throughout our experiments (Figure 2) in Figure 4. We find that similar findings to ones for `GPT-4.1` hold for other models. The most notable cases are Gemini 2.5 (Figure 4a), where persona prompting exhibits the strongest effect and LLaMA 3.1 (Figure 4c), where persona prompting introduces little novel content compared to sampling with higher temperature. Overall, we believe (and the results support) the fact that properly utilizing personas in generation is an emergent capability in LLMs, as evidenced by semantic similarity of novel content being much higher for Gemini and GPT, the two largest models we studied.



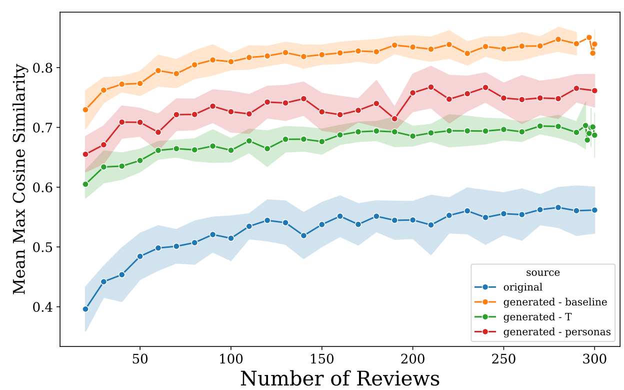
(a) Gemini 2.5



(b) OLMo2 1b



(c) Llama 3.1 8B



(d) OLMo2 7b

Figure 4. Similarity results across models.

C. Pre-training Dataset Analysis.

We report the full results of reviews and Goodreads overview pages identified in the DCLM component of the OLMo pretraining data mix in Table 7.

D. Book Corpus Details

Throughout our experiments, we used a set of 100 books, 20 representative books chosen from 5 categories. In Tables 8 and 9 we enumerate the books, their metadata, and the criterion we used to select them for the core book set.

Table 7. Estimated review and overview pages per source (sources with <10% in both categories omitted).

		Reviews			Overviews		
Source	Total Rows	% Pages	Avg Count	Est. Pages	% Pages	Avg Count	Est. Pages
`amazon.com`	155,879	26	1.03	40,529	13	0.13	20,264
`barnesandnoble.com`	293,727	26	0.68	76,369	62	0.66	182,111
`bookpage.com`	8,762	60	0.62	5,257	80	0.93	7,010
`electricliterature.com`	2,177	28	0.29	610	51	0.64	1,110
`kirkusreviews.com`	68,352	93	0.99	63,567	42	0.43	28,708
`publishersweekly.com`	97,337	78	0.78	75,923	73	0.77	71,056
`theguardian.com/books`	69,519	53	0.62	36,845	38	0.45	26,417
`washingtonpost.com/entertainment/books`	1,597	66	0.76	1,054	64	0.88	1,022
Total	720,505			301,222			346,558

Table 8. Metadata of books used in experiments (1–50), adapted from the Goodreads dataset (Wan & McAuley, 2018)

Title	Author	Reviews	Avg Score	Genre	Category
Cinder	Marissa Meyer	36235	4.11	fantasy	most reviews
Dead Ever After	Charlaine Harris	7415	2.92	fantasy	lowest average
New Moon	Stephenie Meyer	40992	3.26	young adult	highest std
Marked	P.C. Cast	12092	3.14	young adult	lowest average
The Secret History	Donna Tartt	11596	3.80	mystery	highest std
The Lost Symbol	Dan Brown	21569	3.13	mystery	lowest average
Me Before You	Jojo Moyes	51715	4.26	romance	most reviews
The Selection	Kiera Cass	31715	3.76	romance	most reviews
The Historian	Elizabeth Kostova	11613	3.32	history	lowest average
Wuthering Heights	Emily Bronte	17961	3.47	history	lowest average
Archer's Voice	Mia Sheridan	6985	4.62	romance	highest average
It Ends with Us	Colleen Hoover	14236	4.52	romance	highest average
Origin	Jennifer L. Armentrout	6017	4.42	romance	highest average
Point of Retreat	Colleen Hoover	6855	4.41	romance	highest average
Hopeless	Colleen Hoover	16800	4.38	romance	highest average
Fifty Shades of Grey	E.L. James	67203	2.73	romance	lowest average
Grey	E.L. James	8114	3.20	romance	lowest average
Fifty Shades Darker	E.L. James	25251	3.42	romance	lowest average
Fifty Shades Freed	E.L. James	13165	3.40	romance	lowest average
The Heir	Kiera Cass	11690	3.48	romance	lowest average
Beautiful Disaster	Jamie McGuire	21970	3.74	romance	highest std
The Notebook	Nicholas Sparks	14929	3.64	romance	highest std
Beautiful Bastard	Christina Lauren	6659	3.54	romance	highest std
Emma	Jane Austen	8855	3.60	romance	highest std
Romeo and Juliet	William Shakespeare	12510	3.51	romance	highest std
Pride and Prejudice	Jane Austen	35918	4.20	romance	most reviews
Everything, Everything	Nicola Yoon	22363	3.98	romance	most reviews
Anna and the French Kiss	Stephanie Perkins	20077	4.18	romance	most reviews
The Nightingale	Kristin Hannah	33361	4.47	history	highest average
Unbroken	Laura Hillenbrand	38878	4.50	history	highest average
To Kill a Mockingbird	Harper Lee	59827	4.43	history	highest average
When Breath Becomes Air	Paul Kalanithi	10779	4.40	history	highest average
The Book Thief	Markus Zusak	77448	4.35	history	highest average
Love in the Time of Cholera	Gabriel Garcia Marquez	11731	3.31	history	lowest average
Wolf Hall	Hilary Mantel	10254	3.48	history	lowest average
Abraham Lincoln: Vampire Hunter	Seth Grahame-Smith	10031	3.54	history	lowest average
Atonement	Ian McEwan	12609	3.61	history	highest std
Outlander	Diana Gabaldon	30463	3.78	history	highest std
A Tale of Two Cities	Charles Dickens	10362	3.86	history	highest std
The Pillars of the Earth	Ken Follett	23342	3.89	history	highest std
The Diary of a Young Girl	Anne Frank	17771	4.11	history	highest std
The Help	Kathryn Stockett	71318	4.40	history	most reviews
All the Light We Cannot See	Anthony Doerr	49416	4.27	history	most reviews
Water for Elephants	Sara Gruen	51195	3.90	history	most reviews
The Guernsey Literary Society	Mary Ann Shaffer	34373	4.14	history	most reviews
11/22/63	Stephen King	25838	4.18	history	most reviews
Big Little Lies	Liane Moriarty	27346	4.23	mystery	highest average
And Then There Were None	Agatha Christie	12618	4.24	mystery	highest average
Career of Evil	Robert Galbraith	7743	4.11	mystery	highest average
Doctor Sleep	Stephen King	11885	4.11	mystery	highest average

Table 9. Metadata of books used in experiments (51–100), adapted from the Goodreads dataset (Wan & McAuley, 2018)

Title	Author	Reviews	Avg Score	Genre	Category
NOS4A2	Joe Hill	6646	4.11	mystery	highest average
Luckiest Girl Alive	Jessica Knoll	8582	3.12	mystery	lowest average
The Dinner	Herman Koch	11691	3.15	mystery	lowest average
The Woman in Cabin 10	Ruth Ware	12871	3.47	mystery	lowest average
The Da Vinci Code	Dan Brown	33535	3.47	mystery	lowest average
Angels & Demons	Dan Brown	19683	3.70	mystery	highest std
The Girl with the Dragon Tattoo	Stieg Larsson	48632	3.75	mystery	highest std
The Good Girl	Mary Kubica	10018	3.52	mystery	highest std
Inferno	Dan Brown	29204	3.44	mystery	highest std
The Girl on the Train	Paula Hawkins	78438	3.56	mystery	most reviews
Gone Girl	Gillian Flynn	69096	3.80	mystery	most reviews
The Cuckoo's Calling	Robert Galbraith	22980	3.71	mystery	most reviews
Dark Places	Gillian Flynn	24433	3.78	mystery	most reviews
The Girl Who Played with Fire	Stieg Larsson	24868	4.01	mystery	most reviews
The Hate U Give	Angie Thomas	10729	4.66	young adult	highest average
Crooked Kingdom	Leigh Bardugo	8674	4.58	young adult	highest average
Clockwork Princess	Cassandra Clare	10423	4.57	young adult	highest average
Wonder	R.J. Palacio	31536	4.50	young adult	highest average
Cress	Marissa Meyer	18007	4.50	young adult	highest average
Evermore	Alyson Noel	8736	3.03	young adult	lowest average
Fallen	Lauren Kate	16254	3.06	young adult	lowest average
Crossed	Ally Condie	7994	3.11	young adult	lowest average
Reached	Ally Condie	9645	3.22	young adult	lowest average
Twilight	Stephenie Meyer	90766	3.29	young adult	highest std
Breaking Dawn	Stephenie Meyer	42587	3.38	young adult	highest std
Eclipse	Stephenie Meyer	32909	3.54	young adult	highest std
Midnight Sun	Stephenie Meyer	9247	3.96	young adult	highest std
The Fault in Our Stars	John Green	129572	4.37	young adult	most reviews
The Hunger Games	Suzanne Collins	142645	4.33	young adult	most reviews
Mockingjay	Suzanne Collins	86946	3.82	young adult	most reviews
Catching Fire	Suzanne Collins	80495	4.26	young adult	most reviews
Divergent	Veronica Roth	68482	3.96	young adult	most reviews
A Court of Mist and Fury	Sarah J. Maas	18765	4.63	fantasy	highest average
A Monster Calls	Patrick Ness	18601	4.55	fantasy	highest average
Harry Potter and the Deathly Hallows	J.K. Rowling	45748	4.54	fantasy	highest average
Harry Potter and the Goblet of Fire	J.K. Rowling	25258	4.50	fantasy	highest average
Clockwork Princess	Cassandra Clare	10861	4.55	fantasy	highest average
Pride and Prejudice and Zombies	Seth Grahame-Smith	11216	2.88	fantasy	lowest average
Wicked	Gregory Maguire	20367	2.92	fantasy	lowest average
The Magicians	Lev Grossman	16275	3.00	fantasy	lowest average
The Short Second Life of Bree Tanner	Stephenie Meyer	8939	3.29	fantasy	lowest average
Jonathan Strange & Mr Norrell	Susanna Clarke	8374	3.57	fantasy	highest std
Hush, Hush	Becca Fitzpatrick	20227	3.60	fantasy	highest std
Eragon	Christopher Paolini	16700	3.39	fantasy	highest std
A Discovery of Witches	Deborah Harkness	22410	3.53	fantasy	highest std
Shiver	Maggie Stiefvater	18330	3.47	fantasy	highest std
The Martian	Andy Weir	49988	4.25	fantasy	most reviews
Miss Peregrine's Home for Peculiar Children	Ransom Riggs	46665	3.66	fantasy	most reviews
Harry Potter and the Cursed Child	John Tiffany	36121	3.47	fantasy	most reviews
City of Bones	Cassandra Clare	43280	3.67	fantasy	most reviews