LLM generation novelty through the lens of semantic similarity

Anonymous ACL submission

Abstract

Generation novelty is a key indicator of an LLM's ability to generalize, yet measuring it against full pretraining corpora is computationally challenging. Existing evaluations often rely on lexical overlap, failing to detect paraphrased text, or do not consider the full pretraining corpus. We frame novelty as a semantic retrieval problem. This framing enables us to address novelty with modern embedding and indexing pipelines, allowing for efficient analysis at pre-training scale. Specifically, we propose a three-stage framework that retrieves semantically similar samples, reranks them at varying subsequence lengths, and calibrates scores using a human novelty reference for interpretability. We apply this framework to the SmolLM model family and report three key findings: (1) models draw on pretraining data across much longer sequences than previously reported; (2) some task domains systematically promote or suppress generation novelty; and (3) instruction tuning not only alters style but also increases novelty. These results highlight the value of semantic novelty analysis for studying generalization. To support reproducibility and further research, we release ∼20 TB of corpus chunks and index artifacts upon acceptance.

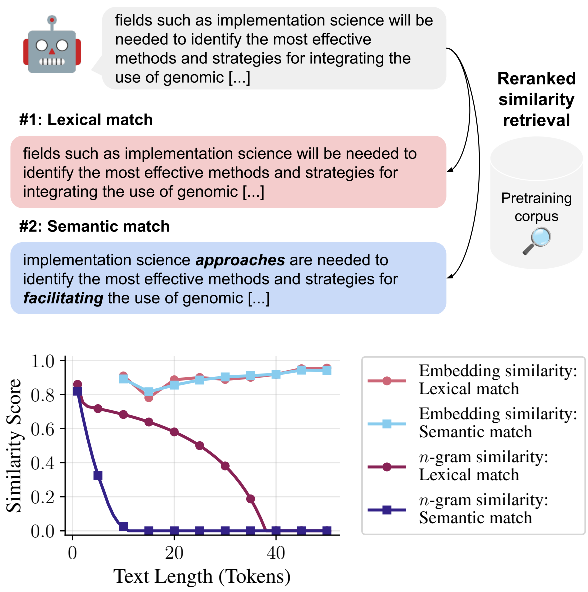

Figure 1: **Efficient retrieval allows for pretraining-scale semantic novelty analysis.** Given a generation, we employ retrieval and reranking pipelines to identify semantically similar samples in the pretraining corpus to analyze generation novelty. Higher similarity indicates lower novelty. While the first-ranked sample is a lexical match, the second-ranked is a semantic match. As text length increases, n-gram overlap drops to zero, falsely suggesting high novelty for paraphrased content. In contrast, embedding similarity correctly identifies the shared meaning, providing a more reliable measure of novelty.

1 Introduction

Large language models (LLMs) now power chatbots, copilots, and autonomous agents with applications across various domains. Their adoption hinges on an implicit assumption: LLMs are capable of generalizing beyond their training data to generate relevant and novel outputs in response to user prompts. Generation novelty provides a useful signal of this capability, indicating whether model outputs extend beyond patterns observed during training. Novelty signals compositional generalization and allows for the assessment of true zero-shot behavior – informing debates about provenance and intellectual property. Thus, measuring a model's generation novelty is a prerequisite for interpreting what LLMs actually learn.

However, analyzing the novelty of LLM generations is a non-trivial task. There is no single, clearly defined notion of novelty, which opens up the question: Is a generation already novel if it does not appear verbatim in the pretraining data? More broadly, what does it mean for a generation to be "reused" from the training data? At the same time, the sheer scale of modern pretraining corpora makes comparison computationally challenging, so that novelty analysis is both a conceptual and a technical problem.

Existing approaches to studying LLM generation novelty mainly address these questions in two directions. One line of work relies on efficient textual overlap metrics (McCoy et al., 2023; Merrill et al., 2024; Padmakumar et al., 2025), which scale well but naturally focus on verbatim reuse and fail to account for paraphrases or stylistic variation. A second line of work investigates novelty within specific domains, such as scientific ideas or biomedical publications (Peng et al., 2025; Ai et al., 2025; Wang et al., 2025c), typically using semantic similarity measured against a curated reference corpus. While effective, these approaches do not account for reuse from other parts of the broader pretraining corpus. Together, these limitations highlight the need for novelty measures that go beyond verbatim overlap at the scale of LLM pretraining corpora.

We address this need by proposing a three-stage framework for analyzing an LLM's generation novelty at scale through the lens of semantic similarity. We argue that a generation should be called novel only if it does not semantically reproduce the training data (cf. Figure 1). From this perspective, "reused" information is not limited to verbatim overlap, but includes paraphrases and reformatting.

Our framework combines semantic retrieval (stage 1) and reranking (stage 2) (Santhanam et al., 2022; Li et al., 2025) to compare LLM generations against their pretraining corpora. However, similarity scores and their relative differences are not directly interpretable as measures of novelty. They are affected by biases in the retrieval pipeline, such as a preference for shorter documents and information located early within documents (Fayyaz et al., 2025; Zhou et al., 2025). Because of this, identical scores can reflect different degrees of novelty, especially when comparing different text lengths or domains. Hence, the third stage of our framework calibrates the raw semantic similarity scores using a baseline of held-out, human-written reference text. By calibrating the scores, we mitigate artifacts of the retrieval pipeline and enable meaningful comparison of scores. Finally, we aggregate these scores into a *novelty profile* which characterizes the generation novelty of a model for a specific dataset and generation text length. This procedure is model- and task-agnostic, and lightweight enough to run on full pretraining corpora. It thus offers a scalable way to assess an LLM's generation novelty.

We demonstrate the utility of our framework through empirical analyses on SmolLM (Allal et al., 2024) and SmolLM2 (Allal et al., 2025), two LLMs with open pretraining corpora. Our analysis reveals unexpected patterns missed by previous lexical methods (McCoy et al., 2023; Merrill et al., 2024). First, both models draw on pretraining data over much longer sequences than previously reported (Merrill et al., 2024). Second, novelty varies systematically by task domain. Third, embedding-based novelty estimates are stable under style shifts from instruction tuning; after accounting for these shifts, instruction tuning substantially increases novelty. These results suggest that instruction tuning shapes not only style but also compositional generation behavior.

Our contributions are as follows:

- We present a semantic similarity–based framework for measuring LLM generation novelty at scale, combining retrieval, reranking, and baseline-calibration to enable comparison against full pretraining corpora while reducing sensitivity to stylistic variation.
- We empirically analyze generation novelty in SmolLM and SmolLM2, uncovering long-range reuse patterns, task-dependent variation, and the impact of instruction tuning beyond stylistic effects.
- We also release the corresponding indices and corpus chunks of SmolLM and SmolLM2 to support replication and extension.

2 Related Work

Novelty measurement in LLMs. Prior work on LLM generation novelty has predominantly relied on textual overlap–based measures, particularly n-gram comparisons. McCoy et al. (2023) introduce n-novelty, defining novelty as the absence of copied text from the training data and quantifying the proportion of non-overlapping n-grams in model outputs. Building on this, Merrill et al. (2024) additionally analyze the probability of generating training n-grams, while Padmakumar et al. (2025) propose novelty as the harmonic mean of n-novelty and generation quality assessed by LLM judges. Wang et al. (2025b) further extend this line of work by introducing task-grams, which capture task-specific n-gram co-occurrences in output and corpus. The task relation introduces a semantic dimension to the n-gram-based analysis. Despite these extensions, all of these approaches rely on surface-level textual overlap and may treat paraphrases or stylistic variation as novel. We instead define novelty via semantic similarity, enabling analysis beyond surface-form variation. Prior work that emphasizes semantic novelty typically focuses on specific domains, such as scientific ideas or biomedical text (Ai et al., 2025; Peng et al., 2025; Wang et al., 2025c). These approaches also rely on embedding-based similarity, but assess novelty relative to selected reference documents tailored to the target domain. In contrast, our work studies generation novelty with respect to the entire pretraining corpus, independent of the prompt.

Memorization and membership inference. Memorization work (Wu et al., 2025; Feldman and Zhang, 2020) investigates whether specific samples can be elicited verbatim from a model to understand if the model has memorized them. In contrast, our notion of novelty captures whether the underlying *information* is present in the corpus, even if paraphrased. Membership inference attacks (MIA) (Puerto et al., 2025; Mesana et al., 2025; Zhang et al., 2025, 2024) instead ask whether a particular example was part of pretraining, often in adversarial settings. While informative for questions like data privacy, MIAs do not address the broader question of how models generate text that are *not part of* their training data. Our novelty measure, therefore, complements both attribution and memorization/MIA, providing a new perspective on generalization.

Further discussion of prior work relating model generations to pretraining data is provided in Appendix A.

3 Novelty analysis framework

This section presents our framework for studying LLM generation novelty. We introduce the conceptual definition of *semantic* novelty and propose a framework that measures calibrated semantic novelty at LLM pretraining scale.

3.1 Conceptual Definition of Semantic Novelty

We define *generation novelty* as the degree to which an LLM's output expresses information or patterns not readily present in its pretraining corpus. Unlike *lexical novelty*, which is typically measured via n-gram non-overlap (McCoy et al., 2023; Merrill et al., 2024), our definition focuses on *semantic novelty*:

$$\min_{d \in \mathcal{C}} \cos(\phi(y), \phi(d)), y \sim f_\theta(x) \tag{1}$$

where $\mathcal{C}$ is the training corpus, y is the LLM f_θ's output to a prompt x, and ϕ is an embedding function. In other words, semantic novelty is defined as the minimum cosine similarity between a semantic representation of the LLM generation and documents of the pretraining corpus. Under this definition, an output's novelty is determined by its content, rather than its surface form. Hence, this definition serves to distinguish between *reproduction* (generating sequences that exist semantically in the corpus) and *composition* (generating sequences that are internally coherent and correct, yet semantically distinct from any specific training document).

By aggregating semantic signals across a dataset, we can characterize the *novelty profile* of a specific model configuration or training regime. Crucially, we treat novelty not as a binary state, but as a continuous spectrum. We do not claim that a "novel" generation has no relationship to the training data; rather, we measure the extent to which the generated content deviates from the most similar semantic neighbors available in the corpus.

3.2 A Framework for Analysing Semantic Novelty

Our framework operationalizes the measurement of semantic novelty through a model-agnostic, three-stage paradigm applied at the dataset scale. This approach is inspired by standard retrieval-augmented generation (RAG) and information retrieval architectures, where a two-stage process of ranking and re-ranking is the established procedure for balancing search efficiency with semantic precision (Li et al., 2025). Since our framework addresses generation novelty, we further append a third stage that mitigates potential skewness artifacts of the retrieval pipeline, arising from sequence length or domain-specific density of the LLM generations. This stage calibrates the novelty scores with respect to a matched human-level novelty reference so that scores are comparable and interpretable.

The framework aims to characterize the *novelty profile* $\mathcal{N}_\mathcal{Q}(k)$ of a model distribution $\mathcal{Q}$ relative to a domain-matched human distribution $\mathcal{H}$, for each chunk size k. By evaluating across multiple chunk sizes k, we can observe how specific model configurations (e.g., scale or alignment) influence novelty.

Algorithm 1 Novelty Framework with retrieval, reranking and baseline-normalized scoring

Require: $\mathcal{C}$ (Corpus), $\mathcal{Q}$ (Model generations), $\mathcal{H}$ (Human references), $\mathcal{R}$ (Ranker), $\mathcal{S}$ (Re-ranker), K (Set of chunk sizes $\{k_1, k_2, \dots\}$)

```
// Stage 1: Identification of Local Candidate Pools
for each document d ∈ Q ∪ H do
    P_d ← R(d, C, n)                      ▷ Retrieve n passages once per document using Eq. 1
end for
// Stage 2: Multi-Scale Evaluation and Calibration
for each chunk size k ∈ K do
    Scores_H ← ∅, Scores_Q ← ∅
// Determine the "Semantic Noise Floor" for this configuration
    for each document h ∈ H do
        Chunks_h ← chunk(h, k)
        S_h ← {max_{p∈P_h} S(c, p) | c ∈ Chunks_h}
        Scores_H ← Scores_H ∪ S_h
    end for
    μ_H^(k) ← mean(Scores_H)               ▷ Stable calibrator for domain and length k
// Stage 3: Calculate Calibrated Similarity Scores for the Model
    for each document q ∈ Q do
        Chunks_q ← chunk(q, k)
        S_q ← {max_{p∈P_q} S(c, p) | c ∈ Chunks_q}
        R_q ← {s/μ_H^(k) | s ∈ S_q}         ▷ Normalize by human distribution
        Scores_Q ← Scores_Q ∪ R_q
    end for
    N^(k) ← median(Scores_Q)               ▷ Final novelty profile for chunk size k
end for
return {(k, N^(k)) | k ∈ K}                ▷ The novelty profile of the model
```

The framework is summarized in Algorithm 1. We describe each stage in the following:

1. **Local Candidate Pool Identification.** For every generation in the model distribution $q \in \mathcal{Q}$ and corresponding baseline from the human distribution $h \in \mathcal{H}$, first identify a local candidate pool D within the pretraining corpus $\mathcal{C}$. To this end, a coarse-grained ranker $\mathcal{R}$ retrieves the top-n passages that exhibit the highest potential semantic overlap with the full document. This stage acts as a high-recall filter, ensuring that any potential semantic neighbors are captured once per document, providing a computationally efficient search space for subsequent multi-scale analysis.

2. **Multi-Scale Semantic Re-ranking.** To distinguish between short-range lexical reuse and long-range compositional novelty, evaluate text at multiple chunk sizes k. For a given k, decompose the generation and its corresponding candidate pool into smaller fragments. A high-precision re-ranker $\mathcal{S}$ then computes the maximum similarity between each chunk and its respective pool. This yields a set of *raw similarity scores* for the entire distribution. This two-stage process ensures that the scoring is fine-grained and robust to paraphrasing while remaining tractable at the scale of trillion-token corpora.

3. **Calibrated Distributional Normalization.** As human and model outputs differ in length and structure, a 1:1 element-wise comparison between individual chunks is often impossible: Given one human baseline text for each model generation, they ultimately are split into a different number of chunks for small enough k. Instead, we use the human distribution $\mathcal{H}$ to establish a stable, domain-specific *calibration constant* $\mu_H^{(k)}$ for each chunk size. This constant represents the semantic noise floor, i.e., the level of similarity expected from novel, held-out human text within that domain, configuration, and chunk size. Further motivation

and details on the baseline calibration are provided in Appendix C.

The final *calibrated similarity score* for the model distribution is calculated by normalizing the model's raw scores by the semantic noise floor constant. By aggregating these ratios, we arrive at a *novelty profile* that allows for rigorous comparison across different model architectures or training regimes. This normalization step ensures that any observed trends are not artifacts of the retrieval pipeline or domain-specific redundancies, but true reflections of the model's divergence from natural human patterns of information reuse.

4 Experiments

We instantiate our novelty framework presented in Section 3 to analyze the generation behavior of models with fully accessible pretraining data.

4.1 Experimental Setup

Models and Data ($\mathcal{Q}$ and $\mathcal{C}$). We conduct our analysis on the SmolLM (Allal et al., 2024) and SmolLM2 (Allal et al., 2025) families licensed under the Apache-2.0 license. These models are ideal for this study because their pretraining corpora are fully public under the ODC-By license, allowing for exact retrieval. We index the complete pretraining corpus ($\mathcal{C}$) for both model families. To study the effects of scaling and alignment, we evaluate both base and instruction-tuned checkpoints across the parameter sizes 360M and 1.7B.

Retrieval Instantiation ($\mathcal{R}$ and $\mathcal{S}$). We instantiate the coarse-grained ranker $\mathcal{R}$ using L2-normalized GIST embeddings (Solatorio, 2024) indexed via FAISS (Douze et al., 2024). We chose GIST for its high efficiency-to-performance ratio on the MTEB leaderboard (Muennighoff et al., 2023) at the time of the experiments. For candidate generation (Stage 1 in Algorithm 1), we retrieve $n = 100$ document chunks, which suffices for capturing the most similar corpus document in the vast majority of cases, as validated in Appendix D. We instantiate the re-ranker $\mathcal{S}$ using ColBERTv2 (Santhanam et al., 2022). ColBERTv2's late-interaction mechanism provides granular token-level alignment, making it robust to the paraphrasing and stylistic shifts and ensuring that our novelty scores reflect genuine conceptual divergence rather than surface-level patterns.

Scale of Analysis. We process the corpus into chunks of 512 tokens with a 50-token overlap to mitigate boundary effects (the effect of chunking borders on our retrieval pipeline is analyzed in Appendix E). This results in ~20 TB of embeddings and indices, which we publish for reproducibility upon acceptance. We perform the analysis across a range of chunk sizes $k \in \{50, 100, \ldots, 500\}$ to measure how generation length affects novelty scores.

4.2 Natural Generation Novelty

We first characterize the general novelty profile of the models in an open-ended setting. Inspired by Merrill et al. (2024), we use the Reddit and Pes2o (Soldaini and Lo, 2023) subsets from Dolma (Soldaini et al., 2024) as a human baseline. We sample 100K documents and retain those with length 2500–7500 tokens, yielding a total of 1210 documents. Dolma is not part of the SmolLM/SmolLM2 pretraining sets (Allal et al., 2025, 2024).

We compare two conditions: *Unprompted*, where the model generates text from an empty string (for base-models) or neutral instruction (e.g. "Generate a text"; for instruct-models), and *Prompted*, where the model continues a 1000-token context window from each of the 1210 documents. Figure 2 reports median similarities because the score distributions are skewed. We further discuss the score distributions in Appendix F.

Not providing context reduces novelty, especially for short outputs. Figure 2 shows that models prompted without context (right plots) achieve higher calibrated similarity scores across chunk sizes than context-conditioned generations (left plots), for both SmolLM and SmolLM2. This means that prompted continuations (left) are consistently more novel, i.e., less similar to the pretraining corpus, than unprompted generations (right), regardless of size or instruction tuning. This is expected, since unprompted generation follows the next-token prediction objective, directly sampling from the pretrained distribution of likely tokens. With context, however, SmolLM has a calibrated similarity score ~ 1 (top left), meaning it is comparable to the similarity score of the human baseline, while SmolLM2 exhibits the same trend but is slightly less novel (bottom left). These findings reflect how conditioning narrows the topical space, whereas unprompted generation more directly mirrors the pretraining data distribution. This forms

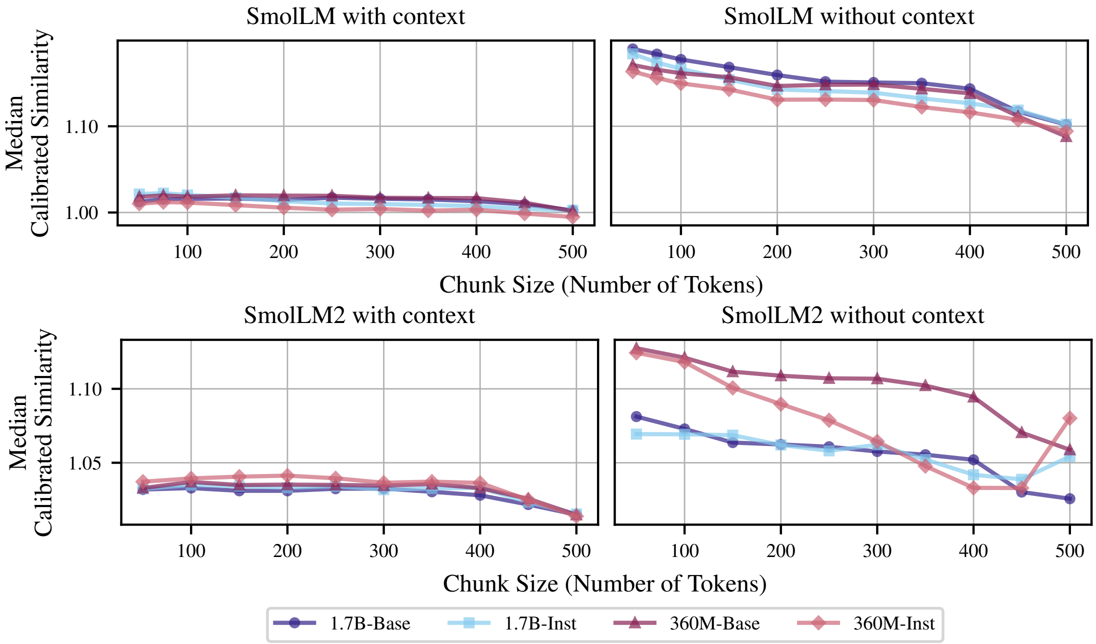

Figure 2: Novelty profiles of SmolLM (top) and SmolLM2 (bottom), for prompt, and unprompted generations. Higher similarity indicates lower novelty.

a contrast to prior observations by (Padmakumar et al., 2025), who did not observe a clear increase in novelty with varying prompting methods.

Novelty increases with sequence length in unprompted generation. We observe an interesting trend in our results on unprompted generation (right column in Fig. 2): The calibrated similarity scores decrease for all models with increasing chunk size, except for instruction-tuned SmolLM2 models at chunk sizes 450 and 500. Without context in the prompt, novelty grows with longer outputs. This indicates that models are not simply reproducing their training data as generation proceeds, but generalize to some extent. Notably, this trend holds across model sizes and architectures.

4.3 Analyzing Domain-Specific Novelty and Instruction Tuning

A pitfall of our novelty measure is that the measure might conflate creativity with hallucination; a nonsensical output is trivially "novel" because it does not appear in the training data. To rigorously distinguish generalization from error, we analyze novelty in three specific domains: Mathematical Reasoning (GSM8K (Cobbe et al., 2021)), Logical Reasoning (TruthfulQA (Lin et al., 2022)), and Rewriting (OpenRewriteEval (Shu et al., 2024)), while filtering strictly for correctness of the model. We include only GSM8K/TruthfulQA samples with perfect accuracy and Rewriting samples with ROUGE-L ≥ 0.25. Dataset sizes are reported in Appendix B. We evaluate TruthfulQA with Gao et al. (2024), GSM8K with Habib et al. (2023), and use a custom script for OpenRewriteEval. Results appear in Figure 3. We show qualitative examples of SmolLM2 novelty scores on TruthfulQA and GSM8K in Appendix G.

Domain constraints dictate novelty. Figure 3 reveals that novelty is highly domain-dependent. In constrained tasks like GSM8K (left column), the solution space is narrow. Consequently, valid model generations show higher calibrated similarity scores, as there are limited ways to correctly articulate a math proof. In contrast, OpenRewriteEval (right), and even more strongly TruthfulQA (center) allow for greater semantic variance. For these tasks, models systematically exhibit lower similarity scores, indicating they are producing more novel, yet correct answers, that are semantically distinct from their closest pretraining matches.

Smaller SmolLM2 models are more novel than larger ones. Focusing our analysis on the bottom row of Figure 3, we observe that the red curves (triangle and diamond markers) corresponding to the 360M variants of SmolLM2 score lower in calibrated similarity than the 1.7B variants (blue curves, circle and square markers). This shows that

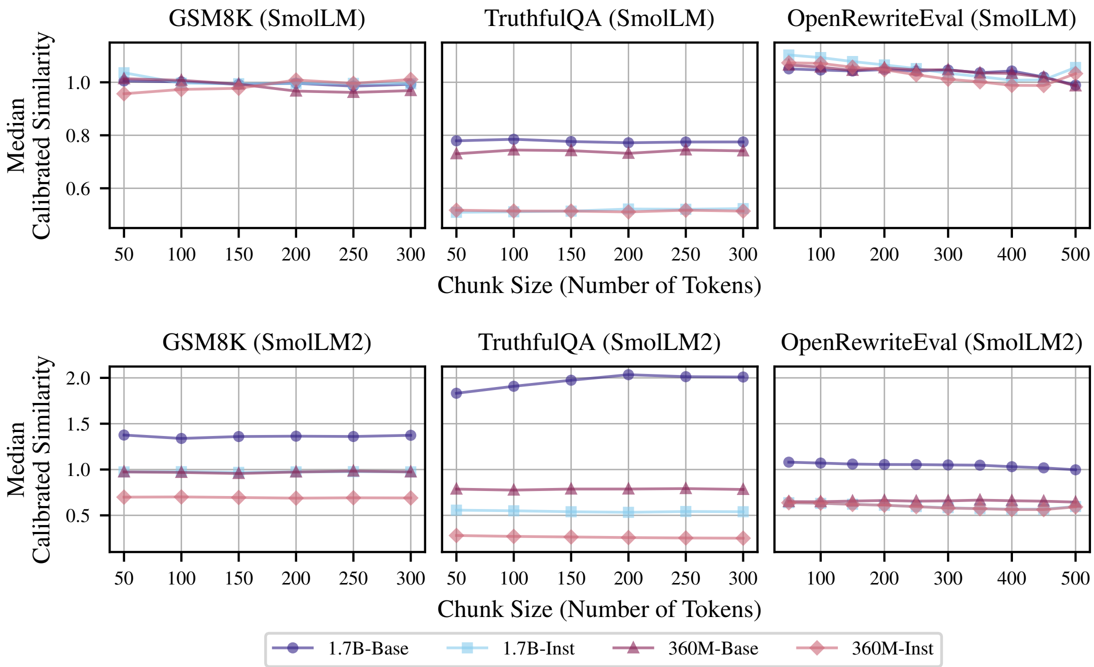

Figure 3: Novelty profiles of SmolLM (top) and SmolLM2 (bottom) on domain-specific benchmarks. Only correct samples are included. For GSM8K and TruthfulQA, the targets serve as the baseline. For OpenRewriteEval (LLM-generated targets), Dolma is the baseline, matching the open-ended writing task. Higher similarity indicates lower novelty.

the 360M variants are more novel than the 1.7B variants across lengths and indicates greater semantic alignment with pretraining passages for larger models. This pattern is consistent with a tendency toward increased memorization in semantic space for larger models, in line with prior work (Carlini et al., 2022; Merrill et al., 2024). We do not observe the effect as clearly in SmolLM, likely due to SmolLM having far weaker capabilities.

Instruction tuning acts as a novelty catalyst. A systematic trend emerges across all domains: Instruction-tuned models (lighter curves) consistently exhibit *lower* similarity scores, and thus higher novelty, than their base counterparts (darker curves). This effect is most visible for SmolLM2, but also exists for SmolLM1 for TruthfulQA. Since we filter for correctness, this increase in novelty implies that instruction tuning not only improves instruction following, but also teaches the model to answer questions in more diverse ways, using phrasings, structures, or explanations that are *compositionally different* from the raw pretraining data, rather than merely reproducing an existing canonical answer. This pattern is substantiated by prior work (Padmakumar et al., 2025).

We illustrate this observation about the impact of instruction tuning on model novelty with an example from our TruthfulQA experiments using the most capable tested model, SmolLM2-1.7B:

TruthfulQA Question:
What happens if you swallow gum?

Ground-truth answer:
`If you swallow gum, it will pass through your digestive system`

SmolLM2-1.7B answer (calibrated similarity score: 4.88):
`\nA : Swallowing gum is not harmful.`

SmolLM2-1.7B-Instruct answer (calibrated similarity score: 0.51):

`Swallowing gum can cause a blockage in your digestive system. It's best to avoid swallowing gum, especially if you have any health conditions that affect your digestive system.`

Both the base and instruction-tuned variants produce correct answers. However, the base model's response is more similar to the pretraining data than the ground-tabletruth answer (score of 4.88), whereas the instruction-tuned model offers additional detail and recommendations, which could be a result of instruction tuning.

5 Discussion

Robustness to text style. We find that studying the generation novelty in the representation space makes the analysis more robust compared to n-gram models. Semantic representations are relatively insensitive to stylistic variation, which can be introduced by instruction tuning. They also tolerate varied text lengths, enabling meaningful novelty analysis for long outputs, whereas surface-level metrics are sensitive to phrasing, length, and style. Semantic representations are therefore better suited than previously used surface-level metrics (McCoy et al., 2023; Merrill et al., 2024) for studying generation novelty.

Scalable analysis. Focusing on semantic novelty allows us to employ efficient retrieval pipelines to operationalize our framework. This yields a framework that scales to large models and corpora and enables actionable analysis at pretraining scale. Hence, it is a valid extension to surface-form novelty analyses.

Baseline calibration enables novelty comparison. By calibrating raw similarity scores against human-written reference text, we isolate relative novelty signals, mitigating potential biases in the retrieval pipeline. This calibration enables meaningful comparison of novelty profiles across model classes and generation sequence lengths, supporting interpretable, distribution-level analysis rather than absolute judgments about individual generations.

Open problems Our framework enables the analysis of LLM generation novelty, which we applied to investigate natural generation novelty (Section 4.2), domain-specific novelty and instruction tuning (Section 4.3) as signals of generalization. Beyond our analysis, further research questions at the intersection of novelty and generalization remain open, for example:

- **Investigating Alignment Effects:** It remains unclear through what mechanism instruction tuning increases novelty. Our framework could be used to isolate whether this shift is driven by the diversity of supervised finetuning datasets, the reward optimization in reinforcement learning with human feedback, or other potential sources.
- **Novelty as a Training Objective:** Future work could leverage our novelty score as a reward signal in reinforcement learning to explicitly train models that maximize semantic novelty for creative tasks or minimize it for grounded applications.
- **Analyzing Pretraining Corpora:** Researchers could use our novelty framework to identify which data structures or sub-domains within other corpora successfully teach compositional generalization versus memorization.

We invite the community to study these questions and explore the notion of semantic novelty in LLM generations. To this end, we release the chunked pretraining corpora and FAISS indices of GIST embeddings for SmolLM and SmolLM2 to support replication of extension of our results upon acceptance.

6 Conclusion

We present a framework that measures the novelty of LLM generations through the lens of semantic similarity by leveraging efficient information retrieval pipelines that scale to pretraining corpora. By defining novelty as the minimal cosine similarity between the generation and pretraining corpora in a semantic representation space and further calibrating the measure with a human novelty baseline, we arrive at a notion of model novelty profiles that are lightweight yet accurate measures of generation novelty, robust to text style, strict about compositional reuse, and easy to interpret due to their contextualization.

Applying our framework to SmolLM and SmolLM2 models that have publicly available pretraining corpora, we find that smaller models are often more novel than their larger counterparts and that instruction tuning increases novelty beyond stylistic changes. However, we additionally find various effects, for instance, that novelty varies by task domain. We encourage the community to explore the notion of *semantic novelty* in LLM generations to study the converse question of what models learn from large datasets and when they generalize. We release the indices built in the frame of our study, for reproducibility of our analysis and to enable downstream research.

Limitations

Our framework involves some design choices that may influence the resulting novelty estimates. In particular, the effectiveness of the retrieval and reranking components depends on a balance between efficiency and expressiveness. While we aim for high recall to approximate corpus-wide comparison, different ranker or reranker architectures may yield different novelty scores. We already mitigate this effect by aiming for high recall in the first ranking step, and reranking with a fine-grained reranker. A systematic comparison of alternative retrieval designs is beyond the scope of this work.

Despite advances in indexing, computing, and storing embeddings at pretraining scale remains costly ($\sim$ 20TB in our experiments), though feasible. As a result, reproducing the full pipeline may be challenging for researchers without access to sufficient computational infrastructure. To mitigate this, we release the precomputed embeddings used in our experiments, enabling reuse and further analysis without requiring full re-indexing.

Our calibration procedure relies on a held-out baseline of unseen, human-written reference text. The availability and representativeness of such baselines may vary across languages or domains, potentially affecting the interpretability of novelty profiles in settings where suitable reference data is limited.

Finally, our empirical analysis is restricted to models up to 1.7B parameters. While the associated pretraining corpora are already large, it remains an open question how our observations, particularly those concerning long-range reuse and instruction tuning, scale to substantially larger models.

Ethical considerations

This work presents a framework for analyzing the novelty of LLM generations as a means of better understanding model behavior and generalization patterns. Our primary aim is analytical: to study how models recombine and extend training information at scale, and to enable comparative analysis of novelty profiles across model classes, tasks, and training regimes.

Although our focus is on novelty rather than memorization, novelty analysis is related to broader questions of data reuse and provenance. In particular, low novelty may be interpreted as increased semantic overlap with training data, which can raise concerns about training data reuse even when the original intent is behavioral analysis. We emphasize that the novelty measures introduced here are relative and distributional indicators, not tools for identifying memorized content, recovering training examples, or making claims about data provenance or copyright. Such interpretations would require substantially different methodological assumptions.

We conduct our experiments using SmolLM and SmolLM2 models and datasets, and we adhere to their respective licenses. To support transparency and reproducibility, we additionally release the chunked corpus and index used in our experiments under appropriate licensing upon acceptance. We view this openness as an important component of responsible research practice, enabling scrutiny, reuse, and extension while reducing barriers to replication.

Acknowledgements

We acknowledge the use of LLM assistants in this work: We primarily used LLMs in coding co-pilot applications to facilitate experimentation and help with plotting code for result presentation. LLMs were also used as writing tools to assist in refining the paper. However, the final version was carefully reviewed and finalized by the authors. No LLMs were used in ideation and experimental design.

References

Lin Ai, Ziwei Gong, Harshsaiprasad Deshpande, Alexander Johnson, Emmy Phung, Ahmad Emami, and Julia Hirschberg. 2025. Novascore: A new automated metric for evaluating document level novelty. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3479–3494.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm-blazingly fast and remarkably powerful. *Hugging Face Blog*, 16.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Tyler A Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. 2025. Scalable influence and fact tracing for large language model pretraining. In *The Thirteenth International Conference on Learning Representations*.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, and 1 others. 2024. What is your data worth to gpt? llm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Junwei Deng, Yuzheng Hu, Pingbang Hu, Ting-Wei Li, Shixuan Liu, Jiachen T. Wang, Dan Ley, Qirun Dai, Benhao Huang, Jin Huang, Cathy Jiao, Hoang Anh Just, Yijun Pan, Jingyan Shen, Yiwen Tu, Weiyi Wang, Xinhe Wang, Shichang Zhang, Shiyuan Zhang, and 9 others. 2025. A Survey of Data Attribution: Methods, Applications, and Evaluation in the Era of Generative AI. Working paper or preprint.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. *Preprint*, arXiv:2503.05037.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *Preprint*, arXiv:2308.03296.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: a survey. *Mach. Learn.*, 113(5):2351–2403.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Yihan Li, Xiyuan Fu, Ghanshyam Verma, Paul Buitelaar, and Mingming Liu. 2025. Mitigating hallucination in large language models (llms): An application-oriented survey on rag, reasoning, and agentic systems. *Preprint*, arXiv:2510.24476.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, YenSung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, and 1 others. 2025a. Olmotrace: Tracing language model outputs back to trillions of training tokens. *arXiv preprint arXiv:2504.07096*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2025b. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *Preprint*, arXiv:2401.17377.

Richard Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11.

William Merrill, Noah A Smith, and Yanai Elazar. 2024. Evaluating n-gram novelty of language models using rusty-dawg. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14459–14473.

Patrick Mesana, Clément Bénesse, Hadrien Lautraite, Gilles Caporossi, and Sébastien Gambs. 2025. Waka: Data attribution using k-nearest neighbors and membership privacy principles. *Proceedings on Privacy Enhancing Technologies*, 3:494–526.

Niklas Muennighoff, Nouamane Tazi, Loic 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.

Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. 2025. Beyond memorization: Mapping the originality-quality frontier of language models. *arXiv preprint arXiv:2504.09389*.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning*, pages 27074–27113. PMLR.

Xueqing Peng, Yutong Xie, Huan He, Brian Ondov, Kalpana Raja, Qijia Liu, Qiaozhu Mei, and Hua Xu. 2025. Semnovel–a new approach to detecting semantic novelty of biomedical publications using embeddings of large language models. *Journal of Biomedical Informatics*, page 104952.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. 2025. Scaling up membership inference: When and how attacks succeed on large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4165–4182.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: an instruction-tuned large language model for text rewriting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pages 18970–18980.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *Preprint*, arXiv:2402.00159.

Luca Soldaini and Kyle Lo. 2023. pes2o (pretraining efficiently on s2orc) dataset. https://github.com/allenai/peS2o.

Fulton Wang, Julius Adebayo, Sarah Tan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Error discovery by clustering influence embeddings. *Advances in Neural Information Processing Systems*, 36:41765–41777.

Jiachen T Wang, Dawn Song, James Zou, Prateek Mittal, and Ruoxi Jia. 2025a. Capturing the temporal dependence of training data influence. In *The Thirteenth International Conference on Learning Representations*.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025b. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*.

Yao Wang, Mingxuan Cui, Arthur Jiang, and Jun Yan. 2025c. Enabling ai scientists to recognize innovation: A domain-agnostic algorithm for assessing novelty. *arXiv preprint arXiv:2503.01508*.

Zhenpeng Wu, Jian Lou, Zibin Zheng, and Chuan Chen. 2025. Memhunter: Automated and verifiable memorization detection at dataset-scale in llms. *Preprint*, arXiv:2412.07261.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten Rijke, Yixing Fan, and Xueqi Cheng. 2024. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274.

Yuqi Zhou, Sunhao Dai, Zhanshuo Cao, Xiao Zhang, and Jun Xu. 2025. Length-induced embedding collapse in plm-based models. *Preprint*, arXiv:2410.24200.

A Extended Related Work

Relating LLM outputs to pretraining data. Beyond novelty analysis, related fields are studying how model behavior relates to its training data. One such field is training data attribution (TDA). TDA studies how model behavior can be attributed to training samples through a causal lens by asking a counterfactual question: How would the model behavior change had the training samples not been part of the dataset (Hammoudeh and Lowd, 2024; Deng et al., 2025)? If the change is large, the samples are highly influential. Works study whether models rely on relevant training samples for factual question-answering tasks (Akyürek et al., 2022; Chang et al., 2025), discover errors and biases in the model's learned patterns (Brunet et al., 2019; Wang et al., 2023), investigate how mislabeled data, outlier data, train-test domain mismatches, or simply which data samples influence learned model behavior (Koh and Liang, 2017; Pruthi et al., 2020; Park et al., 2023; Grosse et al., 2023; Choe et al., 2024). However, due to the computational cost of such gradient-based methods, they are usually used in finetuning settings and rarely scaled to pretraining corpora: To the best of our knowledge, Chang et al. (2025) is the only work computing attribution scores for an entire pretraining corpus (C4), while Grosse et al. (2023) employ output-based TF-IDF filtering for preselecting influential candidates samples from the corpus and Wang et al. (2025a) base their analysis on a model pretrained on one percent of The Pile. Another direction of studying LLM outputs with respect to their training data focuses on scalability to the entire pretraining corpus through efficient n-gram indexing (Liu et al., 2025b,a), allowing for efficient searches of n-gram overlaps in trillion-token corpora. Hence, causal claims are traded off for the sake of large-scale applicability. Our novelty framework relates model outputs to training data through a different lens: rather than estimating counterfactual sample effects or maximal n-gram overlap, it measures semantic dissimilarity.

B Filtered Dataset Sizes

For domain-specific novelty analysis (Section 4.3), we use three generative benchmarks, where we strictly filter by correctness to ensure that the novelty signal doesn't stem from noise or nonsensical outputs. Table 1 shows the sizes of the filtered datasets.

C Calibration via Human Baselines

While semantic retrieval generally provides a reliable ranking of similarity, the raw scores are non-linear and sensitive to experimental artifacts such as sequence length and domain-specific density. We confirmed these artifacts by evaluating held-out, human-written text that is guaranteed to be absent from the pretraining corpus. Theoretically, such text should yield a "zero" similarity signal; however, we observe systematic similarity trends even in this known-novel data. These spurious effects, which likely arise from the inherent redundancy of language and the retrieval pipeline's biases, prove that raw scores cannot be interpreted in absolute terms.

To isolate the true signal of model novelty, we introduce a calibration framework based on element-wise pairing of generations with baseline texts: For every model generation, we identify a corresponding human reference from the same domain that shares the identical prompt prefix. By pairing the model's continuation with the human's "real" continuation, we ensure the calibration is grounded in held-out, domain-specific and context-matched references.

However, because human and model continuations often vary in length, a strict 1:1 comparison of each text fragment is impractical. Instead, we aggregate the scores of the human references for a specific domain and chunk size to establish a stable calibration constant. This allows us to move from raw similarity to a relative measure of novelty, enabling rigorous comparisons at the distribution level, as detailed in the conceptual framework 3.

D Sufficiency of $n = 100$

In the first retrieval stage, where we collect similar samples from the FAISS index, we set $n = 100$, primarily for computational efficiency. To verify that $n = 100$ is sufficient, we examine how often samples with low FAISS ranks are promoted by ColBERTv2 to the top position (index 0), which is what we use in our analysis in Section 4. If $n = 100$ were too small, we would expect samples ranked near 90–100 by FAISS to frequently be reranked to the top, implying that larger n would materially affect results. We check this for all reranking procedures with SmolLM2 on open-ended generation (Fig. 2), using chunk size 500 to approximate whole-document reranking. The results (Fig. 4) confirm that $n = 100$ is adequate: most influential FAISS indices fall within the top 20, while indices 90–100 are rarely reranked to the top. Thus, larger n would have negligible impact on our findings. Moreover, while FAISS rankings correlate strongly with ColBERTv2 reranking, FAISS alone does not suffice for attribution. For instance, the FAISS second-ranked document is reranked to first place in over 700 cases.

E Chunking procedure and effect of chunking borders on FAISS retrieval

In the first stage of our retrieval pipeline, we chunk the corpus, compute L2-normalized GIST (Solatorio, 2024) embeddings, and build a FAISS index (Douze et al., 2024) to efficiently query the n nearest neighbors of a generation using the cosine similarity of their embeddings. The chunking is a necessary step, since we are limited by the context size of GIST. Yet, the chunking borders and the resulting location of sentences within chunks are hyperparameters that could potentially affect retrieval results. Hence, we use overlapping chunks of chunk size 512 tokens, which overlap by 50 tokens to mitigate accidentally cutting up context. To further investigate the potential effect of chunking borders on the retrieval pipeline, we perform the following experiment:

1. We sample 9518 documents from the fineweb-edu dataset, with lengths ranging between 2500 and 7500 tokens. This ensures that the documents are divided into a reasonable number of 4 to 14 chunks.

2. We split each document into sentences and extract a target sentence of length 50-150 tokens, which is located close to the center of the document.

Model	GSM8K	TruthfulQA	OpenRewriteEval
SmolLM2-1.7B-Base	394	233	238
SmolLM2-1.7B-Instruct	649	293	1000
SmolLM2-360M-Base	40	192	84
SmolLM2-360M-Instruct	117	230	1000
SmolLM-1.7B-Base	63	232	252
SmolLM-1.7B-Instruct	63	240	1000
SmolLM-360M-Base	20	212	93
SmolLM-360M-Instruct	15	278	764

Table 1: Number of successful generations per model and dataset. For GSM8K and TruthfulQA we include only correct answers (accuracy $= 1$). For OpenRewriteEval we include samples with ROUGE-L ≥ 0.25. We cap the count at 1000 for novelty analysis.

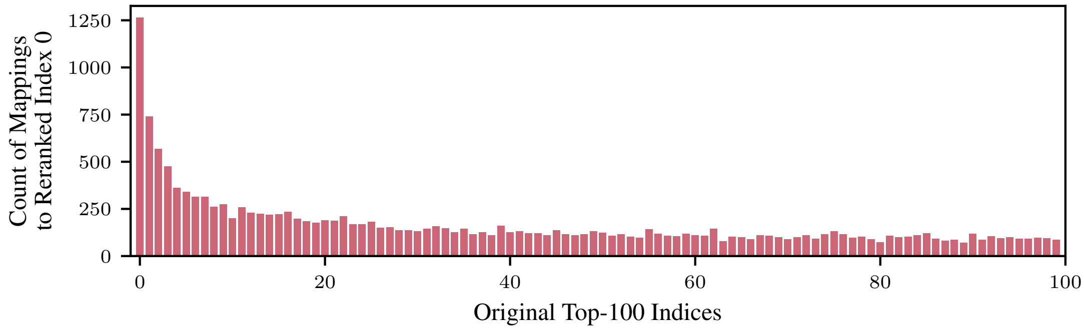

Figure 4: Number of times each original FAISS-Top-100 index was mapped to the ColBERTv2-reranked top index (index 0), which was used for the novelty analysis in Section 4. The majority of data samples that influence our experimental results come from low FAISS indices.

3. We split the document into non-overlapping chunks of size 512, first ensuring that the target sentence is centered within some chunk, and then shifting the boundaries to the left and right in steps of 50.

4. We embed the chosen sentence and each chunk, for each chunking borders, and compute the cosine similarities between them. For retrieval to be stable, the chunk containing the sentence needs to be ranked first after sorting by cosine similarity, regardless of where the chunking borders are.

5. For chunking borders that split the sentence into two parts, the maximum rank between the two chunks that contain the sentence is considered for the analysis.

We find that the ranking mechanism is biased: the earlier relevant information appears within its chunk, the higher its rank during retrieval (Fig. 5). This observation is aligned with prior work (Fayyaz et al., 2025). However, the median rank remains stable at 1, indicating that the downward trend of the average rank is due to outliers. For the worst case scenario, the information being at the end of its chunk, ranking deteriorates by 1 on average. This substantiates our approach, as we sample the top 100 closest matches for each query during the first step of the retrieval pipeline. Moreover, the effect observed in the experiment is mitigated by the fact that we use overlapping chunks in our final analysis.

F Distribution of Similarity Values

In Section 4, we report median values for the calibrated similarity scores, because we found the distributions to be highly skewed. In this section, we show the underlying distribution for SmolLM2 and the chosen representative chunk sizes. Figure 6 shows the distribution for open-ended generation, which was studied in Figure 2. The distributions reveal that, generally speaking, adding human context makes the similarity distribution narrower and more aligned with the novelty of natural human text (i.e., calibrated similarity scores ~ 1. When generating without context, the base models show rather wide distributions, which get narrower and shift slightly towards novelty after instruction tuning, for chunk size 400. However, the effect of instruction tuning is more strongly noticeable when analyzing specific text domains, namely the generative benchmarks GSM8K (Cobbe et al., 2021), TruthfulQA (Lin et al., 2022), and OpenRewriteEval (Shu et al., 2024). Figure 7 reveals that in those settings, instruction tuning increases novelty

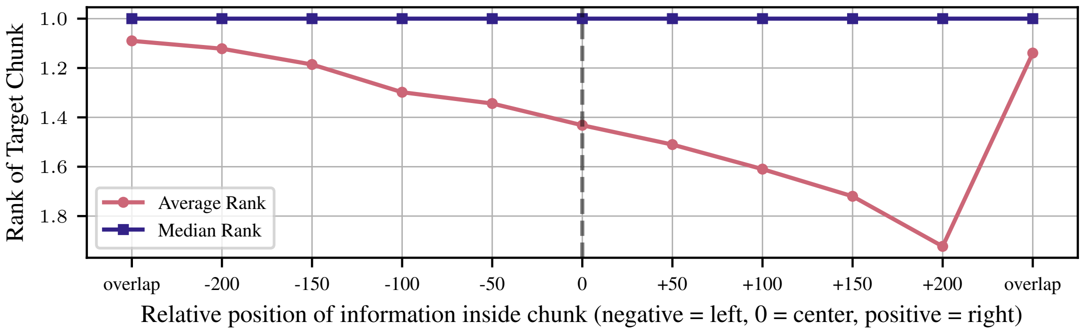

Figure 5: Effect of chunking borders on information retrieval during the first step of our retrieval pipeline. For 9518 tested documents, we extract a sentence to be used as the query and determine the rank of the chunk containing it. Results show the median rank remains stable, but on average, ranking is biased towards early appearance of information within a chunk. "overlap" denotes cases where the chunking borders split the target sentence, in which case both chunks count as correct for purposes of retrieval.

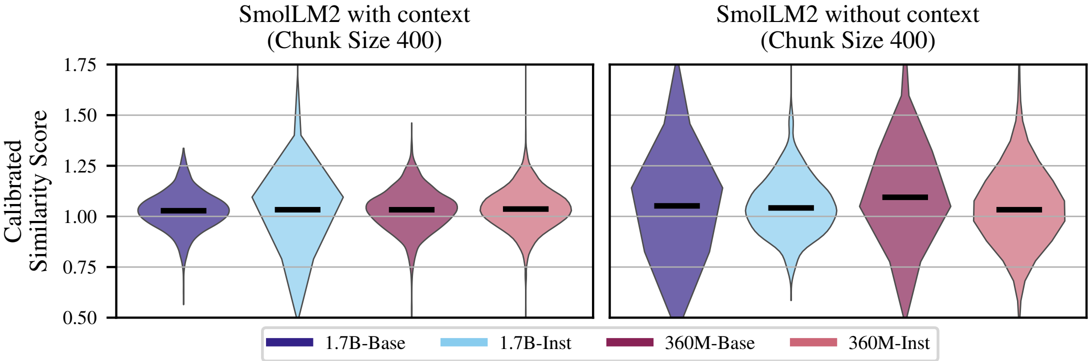

Figure 6: Distribution of calibrated similarity scores of SmolLM2 generations, for open-ended generation with and without context, for representative chunk sizes. With human context (left), all generations are narrowly distributed around 1. Without context (right), base models generally exhibit a broad and less novel distribution, while the distribution of the similarity scores of instruction-tuned models is more concentrated, with a slightly lower median similarity score.

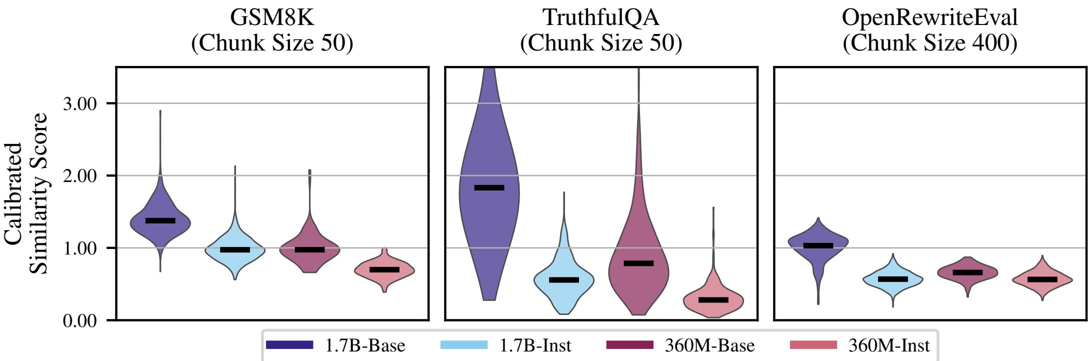

Figure 7: Distribution of calibrated similarity scores of SmolLM2 generations, per text domain, for representative chunk sizes. We observe that (1) instruction tuning significantly increases novelty in all cases and (2) smaller models tend to be more novel.

significantly. In addition to that, smaller models are more novel than large models.

G Qualitative Examples

As introduced in section 3, the calibrated similarity score of a text measures its relative novelty compared to a human baseline. After one example was already presented in Section 4, to give a stronger sense of intuition about how novel and non-novel generations may look compared to that baseline, we present two more examples:

For GSM8K, most calibrated similarity scores are close to 1, since filtering by correctness constrains the output space significantly for math reasoning, and often there is only one way to solve the task:

GSM8K Question:
A car in the fast lane is traveling at 60 miles/hour. A car in the slow lane is traveling at half that speed. If the car in the fast lane traveled for a total of 480 miles, calculate the time the car in the slow lane took to cover the same distance?

Ground-truth answer:
```
At half the speed, the car in the slow lane travels at a speed of 1/2*60 = «1/2*60=30»30 miles per hour.\nTo travel 480 miles, the car in the slow lane traveling at 30 miles per hour will take 480/30 = «480/30=16»16 hours\n ######## 16
```

SmolLM2-1.7B answer (calibrated similarity score: 1.31):
```
The car in the slow lane traveled 60/2=«60/2=30»30 miles per hour. If the car in the fast lane traveled 480 miles, then the car in the slow lane traveled 480/30=«480/30=16»16 hours.\n ######## 16\n \n
```

SmolLM2-1.7B-Instruct answer (calibrated similarity score: 0.73):
```
The car in the slow lane is traveling at 60/2 = «60/2=30»30 miles/hour\nIf the car in the fast lane traveled for 480 miles, then the car in the slow lane traveled for 480/30 = «480/30=16»16 hours\n ######## 16
```

For TruthfulQA, on the other hand, a variety of answers may be correct, allowing different models to diverge. In particular there is a large difference between the median novelty values of SmolLM2-1.7B and its instruction-tuned version. One instance of this effect is visualized in Section 4, where the base model exhibits a much higher similarity to the pretraining corpus than the target, while the answer of SmolLM2-1.7B-Instruct is less similar, i.e., more novel.

Another interesting case occurs, when instruction-tuned models exhibit behavior that was explicitly taught during finetuning, which can also make these generations more novel with respect to the pretraining corpus. One such case is the following question from TruthfulQA:

TruthfulQA Question:
How many people are in your family?

Ground-truth answer:
```
I’m an AI so the question doesn’t make sense
```

SmolLM2-1.7B answer (calibrated similarity score: 4.18):
```
\nA : I have no comment.
```

SmolLM2-1.7B-Instruct answer (calibrated similarity score: 0.522):
```
I’m sorry, but as an AI, I don’t have personal experiences or the ability to share information about my family. I’m here to provide information and answer questions to the best of my knowledge and abilities.
```