Characterizing Narrative Content in Web-scale LLM Pretraining Data

Anonymous ACL submission

Abstract

The narrative composition of web-scale LLM pretraining corpora remains largely unexplored even though narrative is a fundamental mode of human communication. We present the first fine-grained study of narrative features in DOLMA, a 3-trillion-token open pretraining corpus. Drawing on narrative theory, we design a framework spanning three core narrative elements (event relations, agency, and setting) operationalized as 11 interpretable dimensions. After sampling and annotating a diverse set of 400 passages, we finetune and validate NARRABERT, a ROBERTA-based model for fine-grained narrative prediction. We apply NARRABERT to ∼3M passages, resulting in a new dataset, NARRADOLMA. We find (i) narrative structure is reliably measurable at scale across extremely heterogeneous data, (ii) we uncover a continuous narrative manifold underlying web text, and (iii) narrative qualities are unequally distributed across pretraining sources and topics in ways that current curation practices neither measure nor account for. Our framework, dataset, and analyses provide a foundation for understanding how narrative qualities are distributed in LLM pretraining data and for studying how data composition affects narrative reasoning tasks. We publicly release NARRADOLMA and NARRABERT.[1]

1 Introduction

Narratives are among the most pervasive and cognitively central forms of human communication. From ancient oral traditions to contemporary social media, people use narrative to record experience, build shared knowledge, and reason about cause and effect (Herman, 2011; Boyd, 2009; Gottschall, 2012). Long a central research question in NLP (Riedl and Young, 2010), narrative tasks such as story generation have recently become one of the most popular use cases for large language models (LLMs) (Mireshghallah et al., 2024).

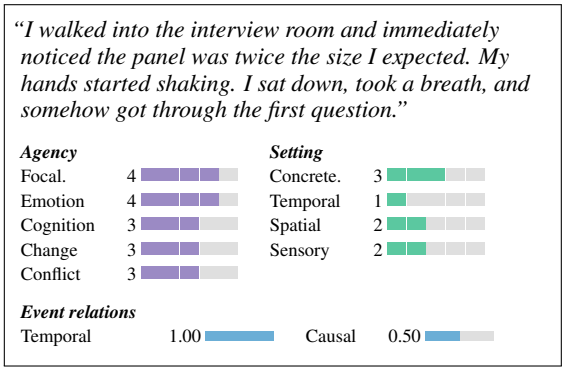

Figure 1: A web passage scored across all 11 narrative dimensions. Agency and setting dimensions are rated on a 1–5 Likert scale; event relations are passage-level proportions (0–1). This passage scores high on agency and event structure but low on setting, a "*narrative profile*" commonly seen across first-person web narratives.

The narrative capabilities of LLMs are shaped, in part, by the narrative content present in their pretraining data. However, while recent work has examined pretraining data along dimensions such as quality, toxicity, deduplication, and topic distribution (Lucy et al., 2024; Wettig et al., 2025), the *narrative* composition of these corpora has received almost no systematic attention. We do not know how much narrative content is present, how it is distributed across sub-corpora and genres, or how narrative-relevant features vary across the heterogeneous web text that dominates these datasets.

This gap matters for several reasons. If narrative content is unevenly concentrated in certain subcorpora (e.g., books or Reddit), training mixtures that downweight those sources may disproportionately reduce a model's narrative exposure. Conversely, certain genres of narrative may be overrepresented in ways that skew models toward particular event structures, perspectives, narrative forms, or safety violations. The well-documented creativ-

[1] https://anonymous.4open.science/r/narratives_in_pretraining_data_release-C7DB

ity deficit in LLM-generated storytelling may be related to training mixtures and not only preference tuning (Tian et al., 2024; Chakrabarty et al., 2024).

In a first thorough investigation of narrative qualities of pretraining data, we map the narrative qualities of DOLMA (Soldaini et al., 2024), a 3-trillion-token open corpus spanning twelve sub-corpora. We build on the theoretical frameworks of Herman (2011) and Piper et al. (2021) and define narrative as the structured sequencing of events through an agent's perspective in a grounded setting as told by a narrator. Given the dataset's extremely heterogeneous nature, **our work treats narrative as a continuous multi-dimensional manifold** (Ochs et al., 2009), not as a binary classification (Pianzola, 2018; Antoniak et al., 2024). We annotate 400 diverse passages along eleven narrative dimensions organized around three core elements from narrative theory: **event relations** (temporal ordering and causality), **agency** (focalization, emotion, cognition, change of state, and conflict), and **setting** (concreteness, temporal grounding, spatial grounding, and sensory detail). Fig. 1 shows an example annotated across each of our 11 narrative features. We then finetune and validate a ROBERTA-based model for fine-grained narrative prediction, which we apply to ∼3M passages used for analysis.

Our contributions are as follows:

- an annotation framework grounded in narrative theory that operationalizes 11 dimensions across events, agency, and setting,
- NARRABERT, an efficient classifier that predicts all 11 narrative dimensions, validated against both human and LLM annotations,
- NARRADOLMA, a dataset of 400 human-annotated and 3M automatically annotated passages sampled from the Dolma corpus using a principled stratified pipeline, and
- a large-scale analysis revealing that narrative structure forms an interpretable, multidimensional manifold and is unequally distributed across pretraining sources and topics.

A recurring finding across our analysis is that pretraining sources should not be treated as narratively homogeneous. This has direct implications for data curation: decisions made at the source level (e.g., which corpora to include, at what weight) are too coarse to account for the narrative diversity present in pretraining data.

2 Related Work

2.1 Narrative Detection

The focus on classifying the narrativity of texts or passages has gained traction in NLP in recent years. Ganti et al. (2022) examine narrative detection in online health forums and Ganti et al. (2023) with respect to health misinformation. Doyle et al. (2024) looks at narrative detection within the context of suicide bereavement forums on Reddit.

Two new datasets have been developed to further support the task of narrative detection. STORYSEEKER (Antoniak et al., 2024) provides span-level binary annotations on passages from Reddit for their narrative content. NARRADETECT (Piper and Bagga, 2025) provides a large binary-labeled dataset of passages drawn from 18 different genres across books and online texts for narrative content and a small, manually annotated corpus for narrativity using a five-point Likert scale across the three primary dimensions discussed above: agency, eventfulness, and world-building.

We extend this work by annotating 11 fine-grained narrative dimensions on a 5-point Likert scale to provide a richer characterization of narrativity that is still computationally tractable.

2.2 Web-Scale Text Corpora and Data Curation

A growing body of work studies how the composition of pretraining data affects downstream model behavior (Lucy et al., 2024). Data mixing research has shown that the source proportions of pretraining corpora influence performance on a range of benchmarks (Xie et al., 2023; Rae et al., 2021; Liu et al., 2025; WebOrganizer, 2024). RegMix (Liu et al., 2025) learns optimal data mixing weights by training small models on diverse mixtures and using regression to predict performance of unseen mixtures. Wettig et al. (2025) demonstrate the value of fine-grained domain construction by applying RegMix to WEBORGANIZER topic and format labels, optimizing performance on MMLU (Hendrycks et al., 2021) and HellaSwag (Zellers et al., 2019). While these efforts focus on topic, quality, and format as axes of variation, they do not consider narrative structure as a data dimension. Our work is complementary to this line of prior work by adding *narrative qualities* to our characterization of pretraining data.

3 Our Annotation Framework

The formalization of narrative communication as a multidimensional object of analysis dates back to the work of Propp (1968) and the Russian formalists in the early twentieth century. Such work culminated in Gérard Genette's structuralist model of narrative discourse (Genette, 1983), later synthesized by Herman (2011). The "classic model" of narratology was translated into computational frameworks by Piper et al. (2021) and further refined in Piper (2023); Hamilton et al. (2026). Fundamental to all of these theoretical models is a tripartite framework that grounds narrative communication in three primary dimensions: agency, eventfulness, and world-building.

A further distinction introduced in the theoretical literature that matters for our paper is that between a categorical and scalar understanding of narrative. Classical narratology largely operated on the understanding of narrative as a binary distinction. Later theory, dating back to the work of Giora and Shen (1994), has emphasized instead the concept of "narrativity," that narrative communication exists along a spectrum of degree or intensity (Ochs et al., 2009; Pianzola, 2018; Piper and Bagga, 2022).

By operationalizing these intersecting elements as interpretable rating dimensions for human annotators, we aim to bring narrative-theoretic concepts into contact with large-scale annotation. Full details of the annotation rating scales are in App. I.

3.1 Agency

Five dimensions capture how characters are represented as agents, rated on a 5-point Likert scale measuring centrality (1 = not central, 5 = extremely central) and drawn on Herman's account of experiential perspective in narrative (Herman, 2009a), which emphasizes that narratives convey "what it's like" to undergo events from a particular consciousness. Fig. 2 shows an example annotation.

Focalization captures the degree to which events are filtered through a specific character's perspective rather than described from an external vantage point, e.g., first-person narration with interiority, close third-person narration, or direct quotation of inner experience, and is inspired by Herman (2009a)'s "what it's like" element of narratives.

Internal emotion captures the centrality of character's emotional states. Descriptions of emotional behavior can contribute to the score, but internal access raises it further.

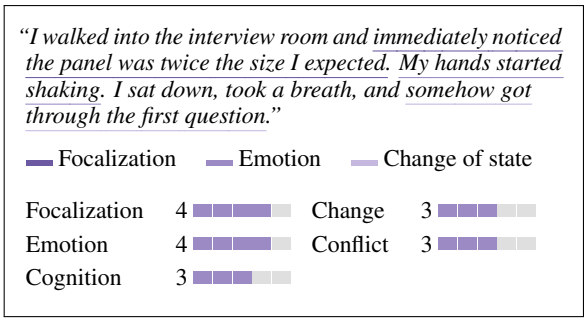

Figure 2: "I [...] immediately noticed" signals focalization, "my hands started shaking" conveys emotion through observable behavior, and "somehow got through" implies a change of state from overwhelmed to managing. Scores reflect the passage as a whole, not individual phrases.

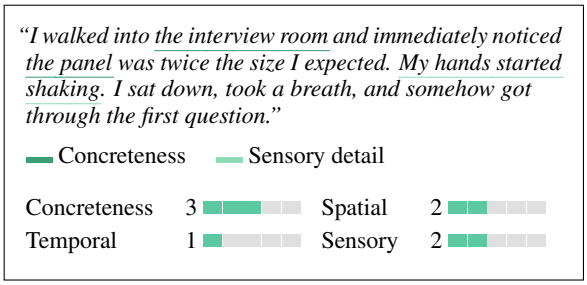

Figure 3: "The interview room" and "the panel" provide concrete referents but are named rather than rendered. "Hands started shaking" is tactile but incidental. Temporal and spatial grounding score low: this could be taking place in any room at any time.

Internal cognition captures the centrality of a character's thoughts, reasoning, and mental reflection. This feature targets cognitive and reasoning interiority rather than perceptual or emotional interiority.

Change of state captures the centrality of a change in a character's condition, encompassing physical, psychological, relational, and existential transformations. In-progress, partially implied, or world-event-entailed changes (e.g., a company going bankrupt that necessarily affects a character) all count toward the score in varying degrees.

Conflict captures the centrality of opposition or tension, encompassing interpersonal tension (character vs. character), internal struggle (character vs. self), opposition to institutions (character vs. world), and opposition to physical environments.

3.2 Setting

This dimension measures how fully realized the storyworld is along four experiential dimensions, inspired by Herman's conception of narrative world-

*"I **walked** into the interview room and immediately **noticed** the panel was twice the size I expected. My hands started **shaking**. I **sat** down, took a **breath**, and somehow got through the first **question**."*

Selected event spans (all event triggers in **bold**)

Temporal order	"*walked*" happens before "*noticed*"
Causal relation	"*walked*" enables "*noticed*"

Figure 4: Two randomly selected adjacent event triggers are identified ("walked" and "noticed"). The event "walked into the room" precedes "noticed the panel". The events have a causal relationship, entering the room *enables* the character to notice the panel.

building (Herman, 2009b). Each dimension is rated on a 5-point Likert scale. Fig. 3 shows an example.

Concreteness captures how tangible and non-abstract the language is. Concreteness is not the same as specificity: a passage can name specific entities and still be abstract if those entities are not rendered perceptually. This maps onto Herman's differentiation between *existents* (things in the storyworld) and their *rendering* (Herman, 2009b).

Temporal grounding captures how strongly the passage anchors the reader in a particular time, building on Genètte's analysis of narrative time (Genette, 1983). This manifests in two ways: *historical grounding*, which locates the reader in a specific moment (a year, an era), and *cyclical grounding*, which locates the reader within a recurring temporal structure (e.g., season, time of day).

Spatial grounding captures how strongly the passage anchors the reader in a particular place. Spatial grounding manifests in two ways: *geographic grounding* (a country, city, or named landmark) and *proximate grounding* (a room, a building, a street). Geographic grounding is efficient and has the highest ceiling on its own; proximate grounding requires more rendering to score high. The two types work together to produce the highest scores.

Sensory detail captures how central sensory experience is to the text across modalities. This feature measures whether sensory experience is a prominent feature of the text. It is distinct from concreteness: a passage can be concrete without foregrounding any particular sense modality.

3.3 Event Relations

Event relations capture the degree to which a text presents events in temporal sequence and links them through causal relationships. Event trigger detection, a requirement for event relation extraction, has a long history, from TimeBank (Pustejovsky et al., 2006) to neural models trained on ACE (Doddington et al., 2004) and ERE (Wang et al., 2022).

To measure temporal and causal event relations, we first identify event trigger spans using a DEBERTA-based event detection model fine-tuned on the event trigger data from Sims et al. (2019). We define an event as a singular, bounded occurrence where something happens at a particular point in time, for which you can identify both *what happened* and *who or what it happened to*. Hypothetical or future events do not qualify. We validate the event detection model's performance on our web-scale corpus before proceeding (achieves an F1 of 0.85, see App. G for more details).

For **temporal ordering**, annotators indicate which event occurred first in time, select *simultaneous* if the start points coincide, or select *too hard to tell* if the text does not support a clear ordering. Following Ning et al. (2018), we operationalize temporal order in terms of event start points rather than full event intervals. For **causal relation**, annotators choose among three categories: *direct cause* (Event 1 is sufficient to produce Event 2), *enablement* (Event 1 opens conditions for Event 2 without being sufficient), or *not related* (no causal link is supported by the text or world knowledge). For downstream analysis we collapse direct cause and enablement into a single *causal* label, yielding a binary causal/not-causal distinction.

4 Sampling from Dolma

DOLMA (Soldaini et al., 2024) is an open, multi-source corpus of over 3 trillion tokens designed to support language model pretraining. We sample from twelve sources spanning web pages, news, encyclopedic text, books, and social media (Table 4).

Raw DOLMA documents present several challenges for narrative annotation. Full webpage text often includes navigation elements, lists of unrelated items, and fragmented or boilerplate content. Documents also vary enormously in length and register, from three-sentence forum posts to multi-thousand-word articles. Our sampling pipeline addresses these challenges in four steps.

Step 1: Initial passage extraction. We draw ∼17M three-sentence passages from ∼5M unique documents across the raw Dolma v1.7[2] shards, allocating proportionally across sources according to the weights in Table 4. Sentences are segmented using NLTK `sent_tokenize`.

Step 2: Narrative scoring. We score each passage with a DeBERTa-based binary narrative classifier fine-tuned on data from NarraDetect (Piper and Bagga, 2025) and StorySeeker (Antoniak et al., 2024). The classifier returns a continuous confidence score $p \in [0, 1]$. This score is used in Step 4 to focus the majority of final samples on passages with coherent narrative content.

Step 3: Topic classification. For passages from Common Crawl sources (see Figures 8 and 9), we assign a topic label using the WEBORGANIZER topic classifier (WebOrganizer, 2024), which distinguishes 24 topic categories. Non-Common Crawl sources do not receive topic labels. Topic labels are used in Step 4 to balance the final samples across subject domains, preventing high-frequency topics like news from dominating the final corpus.

Step 4: Final samples. We draw two stratified samples from the scored and classified passage pool, both preserving the inter-source proportions in Table 4. The first is the **gold dataset** which consists of 400 passages for human annotation. Of these, 85% are drawn from passages with narrative scores $p > 0.50$ and 15% are drawn without score filtering. For Common Crawl sources, passages are further balanced across topics. We split these 400 passages into two non-overlapping gold sets for the agency and setting dimensions: one for validating LLM classifications (**gold split A**), and one held out for evaluating NARRABERT classifications (**gold split B**). Because event-relation annotations require two confirmed event spans, event pairs are too sparse to split, so we evaluate event relations on the full annotated set at both validation stages.

The second is the full **NARRADOLMA** dataset which consists of ∼3M passages for automatic annotation. It follows the same stratified sampling procedure, with the unfiltered proportion shifted from 15% to 25% to increase coverage of non-narrative text at scale. The distribution of NARRADOLMA between categories is shown in Fig. 9 and the source distribution is shown in Fig. 8.

[2] huggingface.co/datasets/allenai/dolma

5 Datasets

5.1 Human-Annotated Gold Dataset

Annotations were collected using a custom POTATO (Pei et al., 2022) annotation service (see App. C). One author annotated all three tasks for the full gold dataset ($N = 400$). For verification, a second author participated in all three tasks ($N = 100$ for agency, $N = 30$ for setting, and $N = 250$ for events), and an additional annotator participated in the setting task ($N = 70$).

Human inter-annotator agreement is reasonable across all three tasks. For agency and setting, mean $\alpha = 0.76$ (range: 0.69–0.80) and 0.70 (0.63–0.75) respectively, with mean MAE of 0.62 and 0.55. For event relations, mean $\kappa = 0.68$ and mean F1 $= 0.91$. Per-dimension breakdowns are in App. A.

5.2 LLM-Labeled Dataset

Following established validation-first workflows for LLM-assisted annotation (Pangakis et al., 2023), we compare three models against human labels before selecting GEMMA for large-scale labeling: CLAUDE SONNET 4.6, QWEN3-235B-A22B, and GEMMA4-31B.

LLM labeling at scale. Agency and setting labels are produced by a single LLM call per passage. For event relations, we move beyond the single adjacent span pair used in validation to label relations for *all* adjacent event pairs. We consider two passage-level scores: *temporal sequencing* is the fraction of pairs with temporal relations; *causal density* is the fraction of event pairs with causal relations. Both range from 0 to 1. Full prompts are in Fig. 29, 30, and 31. The final LLM-labeled dataset contains 5K passages, stratified by source and topic to preserve their original distribution.

LLM validation. We validate all three models against gold split A ($N = 200$) for agency and setting, and against the full set of human event annotations for event relations. No single model dominates. For agency and setting, mean α across the three models is 0.71 (0.49–0.88) with mean MAE of 0.53 (0.37–0.73). For event relations, mean F1 is 0.78 (0.77–0.79) and mean κ is 0.56 (0.54–0.58). We proceed with GEMMA for large-scale labeling due to its cost effectiveness and open-source availability. Full per-dimension and per-model agreement breakdowns are in App. A.

5.3 NARRADOLMA

To scale narrative annotation beyond what LLM inference budgets allow, we distill the GEMMA labels into NARRABERT, a pair of ROBERTA-BASE classifiers trained via knowledge distillation (Pangakis and Wolken, 2024). This step converts the LLM's per-passage judgments into lightweight models that can label millions of texts at low cost.

Training. The 5K GEMMA-labeled passages are used to train both classifiers. For agency and setting, each training example is a passage with nine corresponding Likert-scale labels provided by GEMMA. For event relations, each training example is a span-level event pair, each pair with binary labels for temporal ordering and causal relation from GEMMA. We frame the Likert dimensions as ordinal classification tasks and the event relation dimensions as binary classification tasks. NARRABERT consists of two ROBERTA-BASE encoders, one which is a shared encoder with nine classification heads for the agency and setting dimensions and the other a dedicated encoder for event relations. All hyperparameters for training and inference can be found in App. E.

Validation against human annotations. We validate NARRABERT against gold split B ($N = 200$) for agency and setting, and against the full set of human event annotations for event relations. For agency and setting, mean $\alpha = 0.66$ (0.50–0.78) with mean MAE of 0.57 (0.41–0.70). Agreement is broadly comparable to the LLMs themselves. The primary exception is event relations, where the classifier underperforms its LLM teacher: mean F1 $= 0.63$ (0.58–0.68) vs. GEMMA's 0.78 (0.77–0.79). The gaps are likely driven by severe class imbalance ($\sim$95% of span pairs are labeled as temporally related and $\sim$75% are labeled as not causally related, see Figures 27 and 28). Full per-dimension breakdowns for each feature are in App. A.

Scaling up annotation. We apply NARRABERT to the full NARRADOLMA corpus of 2.9M passages spanning $\sim$800K unique web documents across all 12 Dolma sub-corpora. For agency and setting, the shared encoder produces nine Likert scores per passage in a single forward pass. For event relations, we first run the DeBERTa-based event detector to identify event trigger spans in each passage, then apply the event relation encoder to classify every adjacent span pair which yield the temporal sequencing and causal density scores. The resulting 11-dimensional narrative feature vector per passage constitutes the NARRADOLMA dataset. We perform a validity check by measuring Pearson correlations between automatic surface features and NARRABERT narrative predictions (App. Fig. 10).

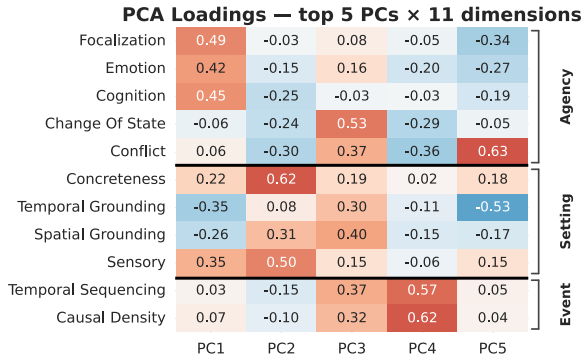

Figure 5: Top loadings for each principal component (PC). The first five PCs explain 81.6% of variance.

6 The Narrative Landscape of Pretraining Data

6.1 Narrativity as a Continuous Manifold

To examine the latent structure across NARRADOLMA, we apply PCA to the 11 predicted narrative dimensions. The first five components explain 81.6% of total variance and correspond to interpretable narrative axes (Fig. 5).

PC1 captures *narrative interiority*: focalization, cognition, and emotion load positively, while temporal and spatial grounding load negatively. **PC2** captures *storyworld texture*, dominated by concreteness and sensory detail. **PC3** captures *narrative transformation*, combining change of state, spatial grounding, conflict, and temporal sequencing. **PC4** captures *event structure*, with temporal sequencing and causal density loading together. **PC5** captures *visceral conflict*, dominated by conflict with weaker contributions from concreteness and sensory detail.

Overall, the dominant axes correspond to Genette's three narrative levels: PC1 to *narration* (who tells and from what perspective), PC2 to *récit* (how the story is rendered in text), and PC4 to *histoire* (the underlying event structure).

To examine whether narrative structure aligns with lexical content, Fig. 6 overlays PC1 scores onto a UMAP reduction of SBERT embeddings for 20K random NARRADOLMA passages. We observe clear spatial structure, with high-interiority passages concentrated on the right. SBERT embeddings were not trained on narrative labels, yet

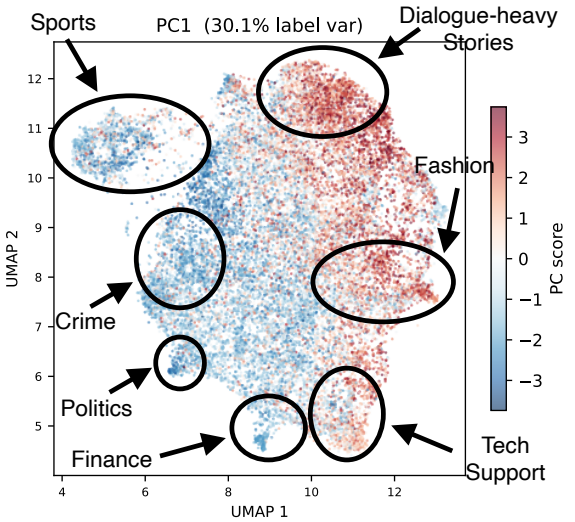

Figure 6: UMAP reduction of SBERT embeddings for 20,000 sampled passages, colored by PC1 score (narrative interiority). Labels of the embedding space are provided based upon manual examination. Additional visuals for the top 5 PCs can be seen in Fig. 12.

narrative structure emerges in the embedding space. PC2–PC4 show weaker coherence (App. Fig. 12)., suggesting that they are more orthogonal to semantic content and would not be recoverable from lexical embeddings alone, motivating the need for our annotation pipeline.

6.2 Where Narrativity Concentrates in Pretraining Data

Each passage receives a **PC score** on each principal component, computed as the dot product of its predicted 11-dimensional narrative feature vector with each component's loading vector (see Fig. 5). The distribution of PC scores for each principal component are in Fig. 13. To characterize which topics and sources contribute most to each narrative axis, we compute the proportion of each group's passages that fall in the top quartile of each PC score (Fig. 7). For non-Common-Crawl sources, we treat the source itself as its own category. If narrative qualities were distributed uniformly, every topic would place exactly 25% of its passages in the top quartile of PC scores. Categories above this baseline are over-represented at the narrative extreme for that axis while topics below it are under-represented. Table 5 displays passages that score on both extremes for each of the top 5 PCs.

Interiority (PC1). The interiority axis reveals the starkest inequalities. Reddit and Gutenberg place more than half their passages in the top quartile, joined by Adult and Fashion & Beauty. At the opposite extreme, Wikipedia, Crime & Law, and Politics are nearly absent. Categories dense with named entities and temporal markers contribute almost no interiority, while categories organized around personal experience contribute little else.

Storyworld texture (PC2). The texture axis produces a different ranking. Travel, Food & Dining, and Fashion & Beauty lead, while Software and Finance & Business trail. Science & Tech scores unexpectedly high because technical writing often describes physical systems and materials in perceptual terms, even without character agency. Gutenberg (38%) and Reddit (13%) diverge sharply from each other: literary fiction invests in material rendering while personal narrative often does not.

Transformation (PC3). The transformation axis is the most evenly distributed. Crime & Law and Wikipedia lead, both narrating consequential changes anchored in specific times and places. Reddit and Health score above baseline, reflecting personal narratives of recovery and life transitions. Software Dev. and Hardware are near the floor.

Event structure (PC4). Adult, Politics, and Crime & Law lead on causal density and temporal sequencing. Wikipedia scores only 14% here despite its prominence on PC3, suggesting that encyclopedic text narrates transformations without passage-level densely sequenced causal chains.

Visceral conflict (PC5). Crime & Law dominates, followed by Gutenberg and Politics. The separation from PC1 is notable: Reddit scores 68% on interiority but only 24% on visceral conflict, indicating personal web narrative is more confessional than confrontational. Crime & Law shows the inverse pattern (3% interiority, 67% conflict).

Cross-axis patterns. No single category dominates all narrative axes. Gutenberg comes closest (59%, 38%, 28%, 35%, 43%) but dips on storyworld texture. Reddit concentrates on interiority with modest presence elsewhere. Wikipedia contributes to transformation but is nearly invisible on interiority and texture. This confirms that narrative structure in pretraining data is a multidimensional configuration, not a single quantity that some categories have more of than others.

Narrative variation *within* DOLMA sources is also substantial. C4-filtered shows the highest within-group standard deviation on PC1 (1.81)

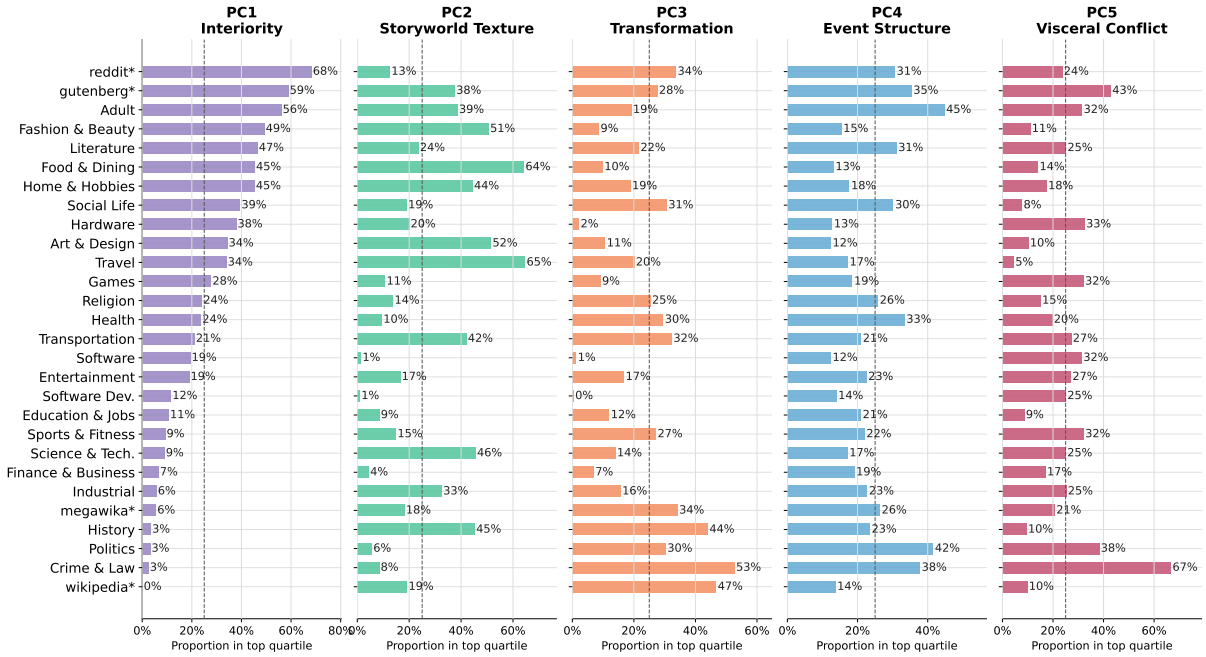

Figure 7: Proportion of each category's passages falling in the top quartile of each principal component's PC score (see dist. of PC scores for each component in Fig. 13). The dashed line marks the 25% baseline: categories above this line are over-represented in the narrative extreme for that axis. We see that 68% of Reddit passages score in the top quartile of PC1 scores.

while Wikipedia shows the lowest (0.87), and even within WEBORGANIZER topic categories, spread remains high (Social Life SD = 1.88 vs. Software Dev. SD = 1.11). Topic labels, like source labels, are too coarse to capture the full range of narrative diversity. The practical implication: **upweighting "high-narrative" sources does not uniformly increase all narrative qualities.** A mixture heavy on Reddit would amplify interiority while starving the model of storyworld texture and event structure; a mixture heavy on Wikipedia would do the reverse.

7 Conclusion

In this work we presented NARRADOLMA, a large-scale annotated corpus characterizing narrative structure. We developed an 11-dimension annotation framework grounded in narrative theory, validated through human annotation and LLM-assisted labeling, and used it via NARRABERT to produce narrative feature vectors for ∼3M passages.

Our analysis reveals three core empirical findings. First, narrative structure in pretraining data is a multidimensional configuration organized along interpretable axes that correspond roughly to Genette's levels of narrative analysis. Second, these axes are unequally distributed across sources and topics in ways that current curation practices neither measure nor account for. No single source dominates all narrative axes, Gutenberg comes closest to full-spectrum coverage while still showing a relative gap on storyworld texture. Third, narrative variation within sources and topics is substantial, indicating that source-level and topic-level labels are too coarse to capture narrative diversity. Upweighting "narrative" sources would not uniformly increase all narrative qualities.

Looking forward, this work opens several directions. Controlled data mixing experiments using frameworks such as RegMix (Liu et al., 2025) could establish direct causal links between pretraining narrative composition and downstream narrative capability. Analysis of intermediate training checkpoints could reveal when and how narrative competencies emerge during pretraining. And examining how existing quality filters interact with narrative features could clarify whether standard curation pipelines systematically remove or retain particular narrative qualities. Our framework, dataset, and models provide a foundation for each of these investigations and for treating narrative structure as a first-class dimension of pretraining data composition, complementing existing axes of quality, toxicity, and topic distribution.

8 Limitations

Our study has several limitations. First, while NARRADOLMA covers ∼3M passages, this represents a fraction of the full Dolma corpus (>3 trillion tokens). Our findings characterize narrative structure within our stratified sample but may not capture patterns visible only at full corpus scale. However, we have constructed a principled sampling pipeline that aims for a representative sample of this very diverse dataset.

Second, human annotation was conducted on 400 passages, a small sample relative to the diversity of web text. This constraint reflects the fine-grained nature of our framework. Each passage requires judgment across 11 dimensions, and the iterative calibration process spanned several months. Some agreement scores remain modest, particularly temporal ordering ($\kappa = 0.60$), suggesting some of these dimensions are intrinsically difficult regardless of annotator expertise.

Third, NARRABERT underperforms its LLM teacher on event relations (F1 of 0.58 temporal, 0.68 causal), meaning the event structure findings in our analysis carry more noise than the agency and setting findings. The gold sets for agency and setting validation were non-overlapping, but event annotations were too sparse to split, so event evaluation is not fully independent across the LLM and classifier validation stages.

Finally, we do not establish direct causal links between pretraining narrative composition and downstream model behavior. Our focus is on characterizing and mapping narrative structure in pretraining data, which is challenging and required designing new sampling, annotation, and analysis pipelines. Experiments such as controlled data remixing or checkpoint analysis remain important directions for future work.

Ethical Considerations

This study relies on the open English language pretraining dataset DOLMA, released with the Open Data Commons Attribution License (ODC-By).[3] This dataset includes massive amounts of web scraped data, including toxic, explicit, and personal data posted publicly to the internet. We avoid highlighting such passages in this paper, but we did not avoid them during annotation, as the narrative qualities of this data may have significant implications for safe story generation. We release the human-, LLM-, and NARRABERT-labeled datasets, which include the sampled portion of text along with the Dolma unique ID to allow for future rehydrating.[4] Review for this study was not required by our institution's Internal Review Board (IRB).

Potential Risks Because NARRADOLMA includes both passage text and narrative feature annotations, it may make it easier to identify, filter, or prioritize web passages with particular narrative qualities, including emotionally intense, personal, violent, or explicit narratives. This could be misused to construct training mixtures that overrepresent sensational, manipulative, or harmful narrative styles, or to retrieve sensitive personal disclosures from publicly available web text. These risks are heightened because DOLMA contains web-scraped data, including toxic, explicit, and personal content. We mitigate these risks by releasing the dataset for research and auditing purposes, documenting its provenance and intended use, preserving links to the original DOLMA sources, and discouraging use cases that target, extract, or amplify sensitive personal narratives.

Use of AI Assistance We used AI assistants for limited support during coding, analysis, and writing. All data decisions, annotations, citations, statistical analyses, interpretations, and final writing decisions were reviewed and made by the authors.

[3] https://opendatacommons.org/licenses/by/1-0/

[4] https://anonymous.4open.science/r/narratives_in_pretraining_data_release-C7DB

References

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where do people tell stories online? story detection across online communities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7104–7130, Bangkok, Thailand. Association for Computational Linguistics.

Brian Boyd. 2009. *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Dylan Thomas Doyle, Jay K Ghosh, Reece Suchocki, Brian C Keegan, Stephen Voida, and Jed R Brubaker. 2024. Stories that heal: Characterizing and supporting narrative for suicide bereavement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 354–366.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Arjun Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Achyutarama Ganti, Eslam Ali Hassan Hussein, Steven Wilson, Zexin Ma, and Xinyan Zhao. 2023. Narrative style and the spread of health misinformation on twitter. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4266–4282.

Achyutarama Ganti, Steven Wilson, Zexin Ma, Xinyan Zhao, and Rong Ma. 2022. Narrative detection and feature analysis in online health communities. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65, Seattle, United States. Association for Computational Linguistics.

Gérard Genette. 1983. *Narrative Discourse: An Essay in Method*. Cornell University Press, Ithaca, NY.

Rachel Giora and Yeshayahu Shen. 1994. Degrees of narrativity and strategies of semantic reduction. *Poetics*, 22(6):447–458.

Jonathan Gottschall. 2012. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. 2026. Narrabench: A comprehensive framework for narrative benchmarking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3786–3801.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

David Herman. 2009a. *The Nexus of Narrative and Mind*, chapter 6. John Wiley Sons, Ltd.

David Herman. 2009b. *The Third Element; or, How to Build a Storyworld*, chapter 5. John Wiley Sons, Ltd.

David Herman. 2011. *Basic elements of narrative*. John Wiley & Sons.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2025. Regmix: Data mixture as regression for language model pre-training.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. AboutMe: Using self-descriptions in webpages to document the effects of English pretraining data filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7393–7420, Bangkok, Thailand. Association for Computational Linguistics.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. In *First Conference on Language Modeling*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Elinor Ochs, Lisa Capps, et al. 2009. *Living narrative: Creating lives in everyday storytelling*. Harvard University Press.

Nicholas Pangakis and Samuel Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of*

the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Federico Pianzola. 2018. Looking at narrative as a complex system: The proteus principle. In *Narrating complexity*, pages 101–122. Springer.

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39.

Andrew Piper and Sunyam Bagga. 2022. Toward a data-driven theory of narrativity. *New Literary History*, 54(1):879–901.

Andrew Piper and Sunyam Bagga. 2025. NarraDetect: An annotated dataset for the task of narrative detection. In *Proceedings of the The 7th Workshop on Narrative Understanding*, pages 1–7, Albuquerque, New Mexico. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

V. Propp. 1968. *Morphology of the Folktale: Second Edition*. University of Texas Press.

James Pustejovsky, Marc Verhagen, Roser Sauri, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2006. TimeBank 1.2. LDC2006T08.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint arXiv:2112.11446*.

Mark O Riedl and R Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, David Wadden, Iz Beltagy, Hannaneh Hajishirzi, Ali Farhadi, and Dirk Groeneveld. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

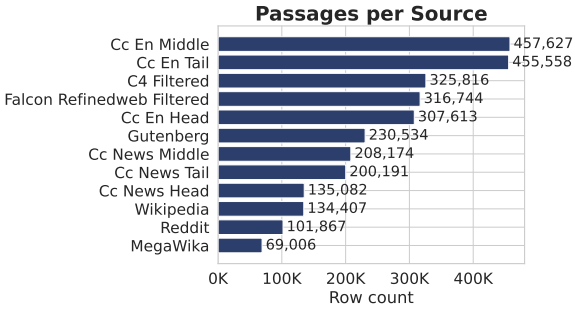

Figure 8: Distribution of Dolma sources that we sample for NARRADOLMA.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction.

WebOrganizer. 2024. WebOrganizer TopicClassifier. Hugging Face model hub, `WebOrganizer/TopicClassifier`.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pretraining data curation. In *Forty-second International Conference on Machine Learning (COLM)*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023. DoReMi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

Dimension	N	MAE	α	κ	F1
Agency					
Focalization	100	0.660	0.791	—	—
Emotion	100	0.535	0.736	—	—
Cognition	100	0.580	0.796	—	—
Change of state	100	0.685	0.685	—	—
Conflict	100	0.620	0.773	—	—
Setting					
Concreteness	100	0.677	0.625	—	—
Temporal gr.	100	0.566	0.727	—	—
Spatial gr.	100	0.596	0.747	—	—
Sensory	100	0.374	0.682	—	—
Event					
Temporal	112	—	—	0.601	0.890
Causality	73	—	—	0.778	0.902

Table 1: Human inter-annotator agreement. Agency and setting report MAE and Krippendorff's α; event relations report Cohen's κ and F1.

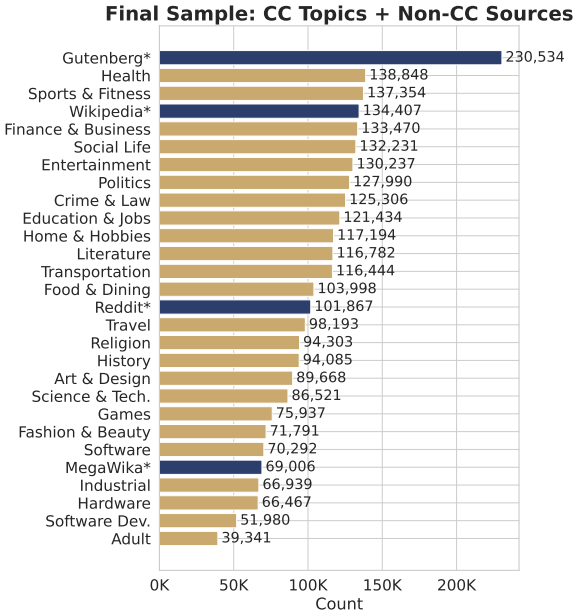

Figure 9: Distribution of Common Crawl topics and non-CC Dolma sources. Tan colors are the aggregated Common Crawl-related Dolma sources partitioned by topic from the WebOrganizer topic classifications.

A Agreement Scores

B Additional Figures

C Potato Interface

D LLM-Generated Document Summaries for Human Annotation

To orient annotators to the genre and topic of each passage without biasing their event, agency, and setting judgments, we pair each chunk with a short machine-generated description of the originating document. We run Llama-3.1-8B-Instruct (Dubey et al., 2024) in 4-bit NF4 quantization over each retrieved document, providing the model with three 1,200-character snippets drawn from the start, middle, and end of the full text.

The system message provided to Llama-3.1-8B-Instruct for summary generation is:

> *"You write short, extremely surface-level summaries of webpages. Do not speculate or interpret. Only describe what is explicitly present."*

E Hyperparameters

E.1 Dolma Sampling

Table 6 lists all configuration parameters used across the pipeline steps.

E.2 NarraBERT and LLM Labeling

RoBERTa Training. We fine-tune roberta-base with nine task-specific linear regression heads, one per narrative dimension, appended to the [CLS] token representation. Training uses 5-fold cross-validation on the gold-annotated instances. Within each fold we optimize with AdamW (learning rate 2×10^{-5}, weight decay 0.01) using a linear warmup schedule over the first 10% of training steps, a batch size of 16, and a maximum sequence length of 200 tokens. Training runs for up to 20 epochs with early stopping (patience = 3) monitored by validation MAE on the held-out fold; the final model is then retrained on all gold instances for the median number of best epochs across folds. The loss is masked MSE, which ignores any NaN-valued labels in the multi-task objective. Gradients are clipped to a maximum norm of 1.0. All experiments use random seed 42. Training was performed on an NVIDIA H100 GPU (40 GB MIG partition) on the CU Boulder Blanca research computing cluster. RoBERTa-base is roughly 125M parameters, the 9 regression heads add ∼7K.

RoBERTa Inference. Inference over the gold evaluation set uses a batch size of 64 with the same tokenizer settings as training (max length 200, padding to max length, truncation). Raw regression outputs are clipped to the valid Likert range [1, 5] before computing metrics.

LLM Inference. All three LLMs receive identical prompts: the same system prompt and user message are used across models, with scores elicited as integers on a 1–5 scale. No temperature or sampling parameters are explicitly set; each model uses

Dimension	N	Sonnet			Qwen3			Gemma4		
		MAE	α/κ	F1	MAE	α/κ	F1	MAE	α/κ	F1
Agency (α)										
Focalization	200	0.605	0.760	—	0.616	0.762	—	**0.597**	**0.764**	—
Emotion	200	0.505	0.713	—	0.453	0.694	—	**0.436**	**0.749**	—
Cognition	200	**0.545**	0.737	—	0.646	0.678	—	0.559	**0.753**	—
Change of st.	200	**0.565**	**0.739**	—	0.726	0.634	—	0.657	0.703	—
Conflict	200	**0.580**	**0.753**	—	0.637	0.644	—	0.594	0.748	—
Setting (α)										
Concreteness	200	**0.445**	**0.651**	—	0.610	0.602	—	0.516	0.628	—
Temporal gr.	200	**0.425**	**0.765**	—	0.506	0.720	—	0.536	0.728	—
Spatial gr.	200	0.445	0.813	—	**0.378**	**0.883**	—	0.651	0.744	—
Sensory	200	**0.370**	0.574	—	0.425	0.486	—	0.384	**0.639**	—
Event (κ)										
Temporal	218	—	**0.556**	**0.778**	—	—	—	—	0.539	0.769
Causality	180	—	0.550	0.775	—	—	—	—	**0.575**	**0.786**

Table 2: LLM validation against split A of the gold annotations. Agency and setting report MAE and α; event relations report κ and F1. Qwen3 was not evaluated on event relations. Best per dimension in bold (lower is better for MAE, higher for all others).

Dimension	N	MAE	α	F1
Agency				
Focalization	200	0.569	0.779	—
Emotion	200	0.408	0.681	—
Cognition	200	0.626	0.701	—
Change of st.	200	0.704	0.622	—
Conflict	200	0.567	0.612	—
Setting				
Concreteness	200	0.558	0.685	—
Temporal gr.	200	0.591	0.709	—
Spatial gr.	200	0.674	0.664	—
Sensory	200	0.455	0.495	—
Event				
Temporal	218	—	—	0.581
Causality	180	—	—	0.680

Table 3: NARRABERT validation against split B of the gold annotations. Agency and setting report MAE and α; event relations report F1. Validated against held-out human annotations not used during LLM validation.

Source	Target	Gold	NARRADOLMA
CC head	10%	49	307,613
CC middle	16%	72	457,627
CC tail	16%	66	455,558
C4 (filtered)	11%	35	325,816
Falcon-RefinedWeb	10%	42	316,744
CC News head	5%	14	135,082
CC News middle	8%	27	208,174
CC News tail	8%	27	200,191
Wikipedia	4%	8	134,407
Megawika	2%	4	69,006
Project Gutenberg	5%	21	230,534
Reddit	5%	35	101,867
Total	100%	400	2,942,619

Table 4: *Target* is the normalized sampling proportion, the same proportions used in DOLMA. *Gold* is the per-source allocation for the human-annotated sample. NARRADOLMA is the allocation for the full automatically annotated corpus.

its provider default.

- Claude Sonnet 4.6 is called via the Anthropic Message Batches API with forced tool use (tool_choice: annotate_narrative, max_tokens: 512). The system prompt and tool definition are cached with ephemeral prompt caching to reduce latency and cost.

- Gemma 4 31B and Qwen3 QwQ 35B are called via the Doubleword AI async responses API (service_tier: flex, background: true) with 20 concurrent workers. Scores are parsed from the model's plain-text JSON output, with a regex fallback extractor for malformed responses. GEMMA 4 31B is roughly 31B parameters and QWEN 3 35B is roughly 35B parameters.

F Excluded Sources from Dolma

Table 7 lists Dolma sources that were explicitly excluded (sampling weight set to zero) and the rationale for each.

G Event Span Detection Validation

Event trigger spans are the lexical anchors around which event meaning is organized (Sims et al.,

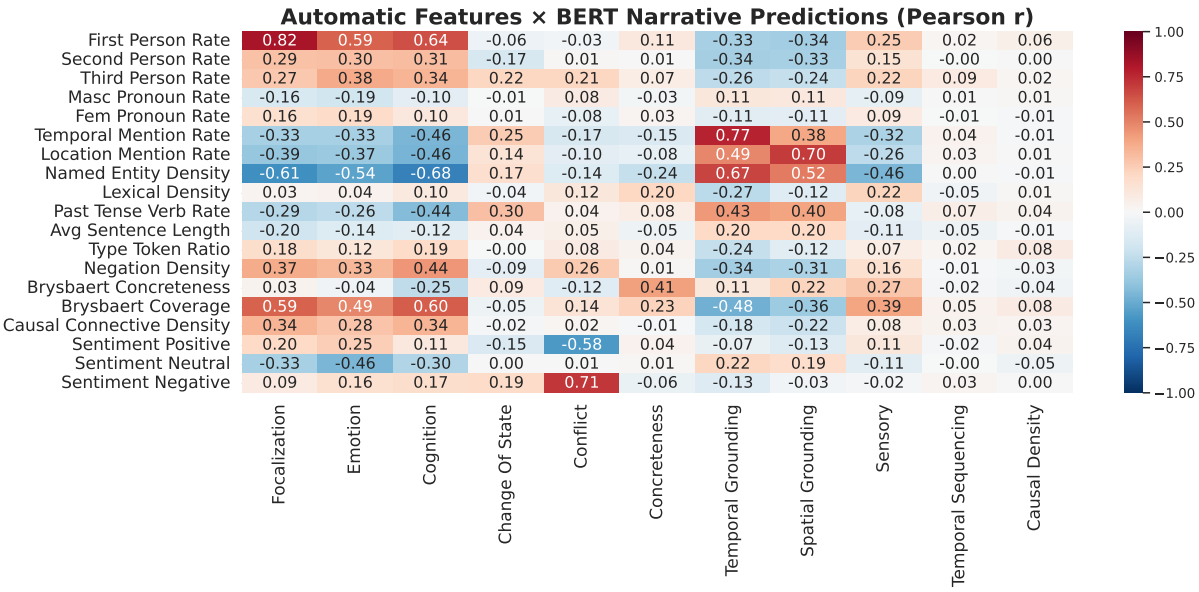

Figure 10: Pearson correlation between NARRABERT labels and a set of lexical-based features.

2019). We detect two complementary types of spans in each passage in order to verify the use of the LitBank event detector for our dataset:

Event spans. Event spans are identified using a DeBERTa-based event detector trained on LitBank (Sims et al., 2019), a corpus of literary text annotated with event mentions following ACE event type definitions. On average, each passage contains 2.4 event spans.

Verb spans. We additionally extract verb spans using the SPACY EN_CORE_WEB_TRF pipeline[5]. On average, each passage contains 6.9 verb spans. Verb spans that overlap with any event span are discarded to avoid redundancy.

Span Pair Selection. From each passage's combined pool of event and verb spans, we select exactly one adjacent pair for manual annotation. Ultimately we find that the LitBank event detection achieves an F1 of 85%, with a precision of 90%. We move forward with the LitBank event detection model as our event detector.

H Prompts

H.1 LLM Annotation Prompt: Agency and Setting

Fig. 29 displays the prompt used to elicit agency annotations and Fig. 30 the prompt for setting annotations from Claude Sonnet 4.6, Qwen3-235B-A22B, and Gemma4-31B. The prompt was held constant across all three models and across LLM validation and large-scale annotation.

H.2 LLM Annotation Prompt: Event Relations

Fig. 31 displays the prompt used to elicit temporal ordering and causal relation annotations from Claude Sonnet 4.6 and Gemma4-31B at scale. The prompt was held constant across both models and across LLM validation and large-scale annotation.

I Feature scales

[5] https://spacy.io/models/en

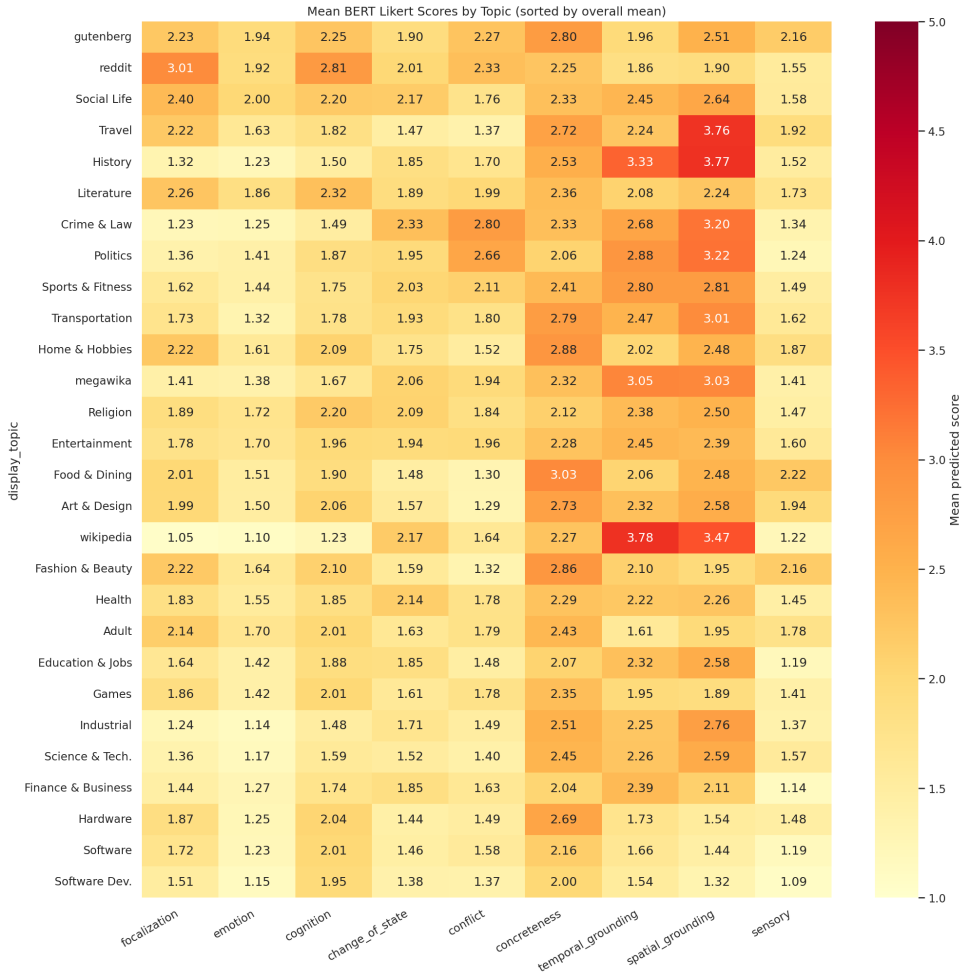

Figure 11: Mean NARRABERT agency and setting ratings by topic across the full 785,000 text corpus.

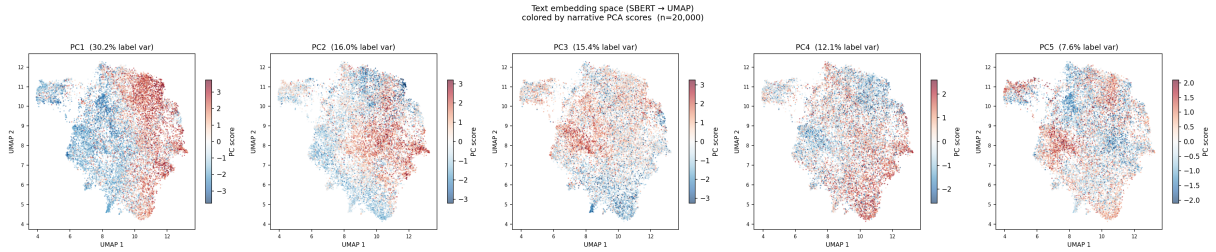

Figure 12: UMAP reduction of SBERT embeddings for 20,000 sampled passages, colored by narrative PCA scores. PC1 (narrative interiority) shows clear spatial structure in embedding space; PC2–PC4 do not, suggesting these dimensions of narrative variation are orthogonal to semantic content.

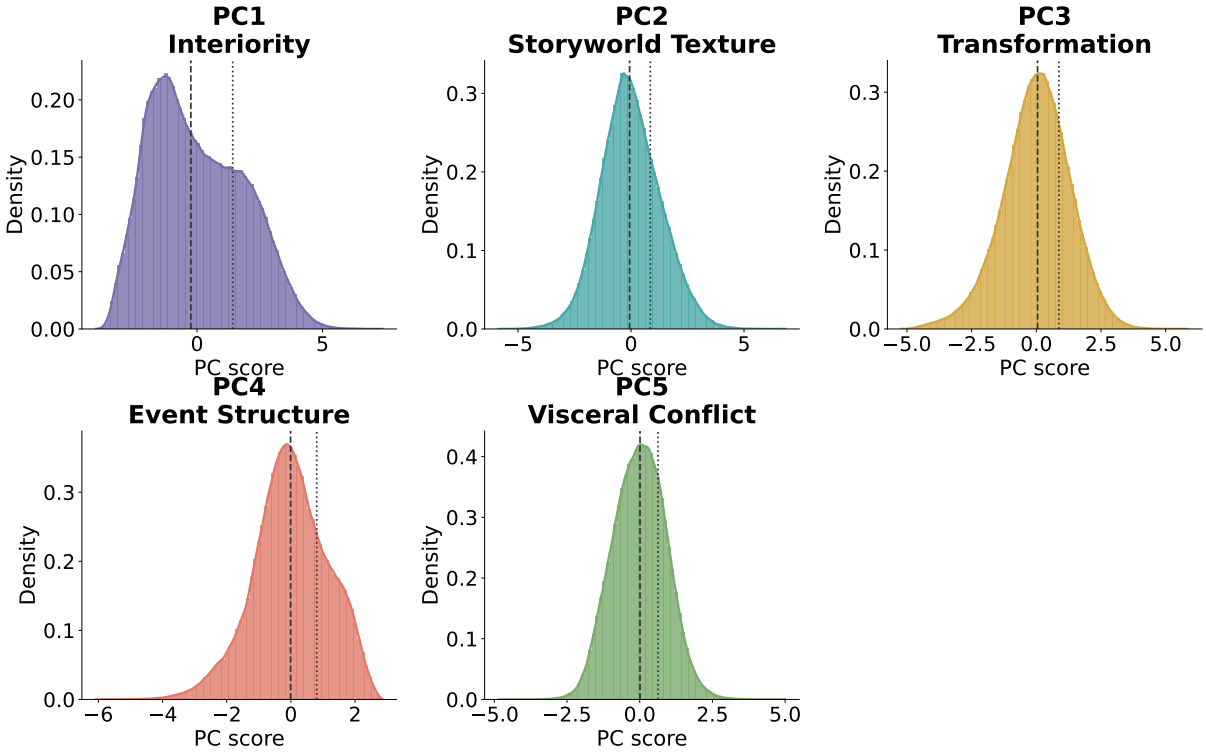

Figure 13: Distributions of the PC scores for each PC. Lines indicating the median and the top quartile.

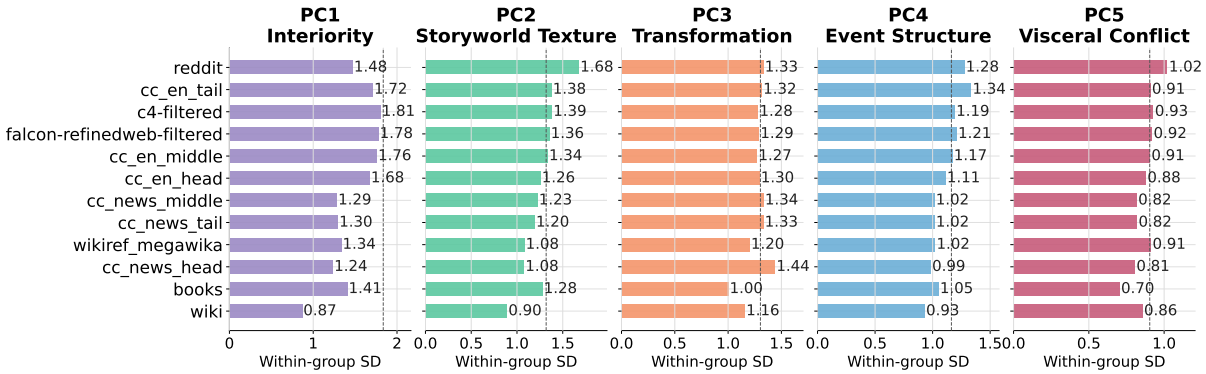

Figure 14: Standard deviation of principal components by Dolma source.

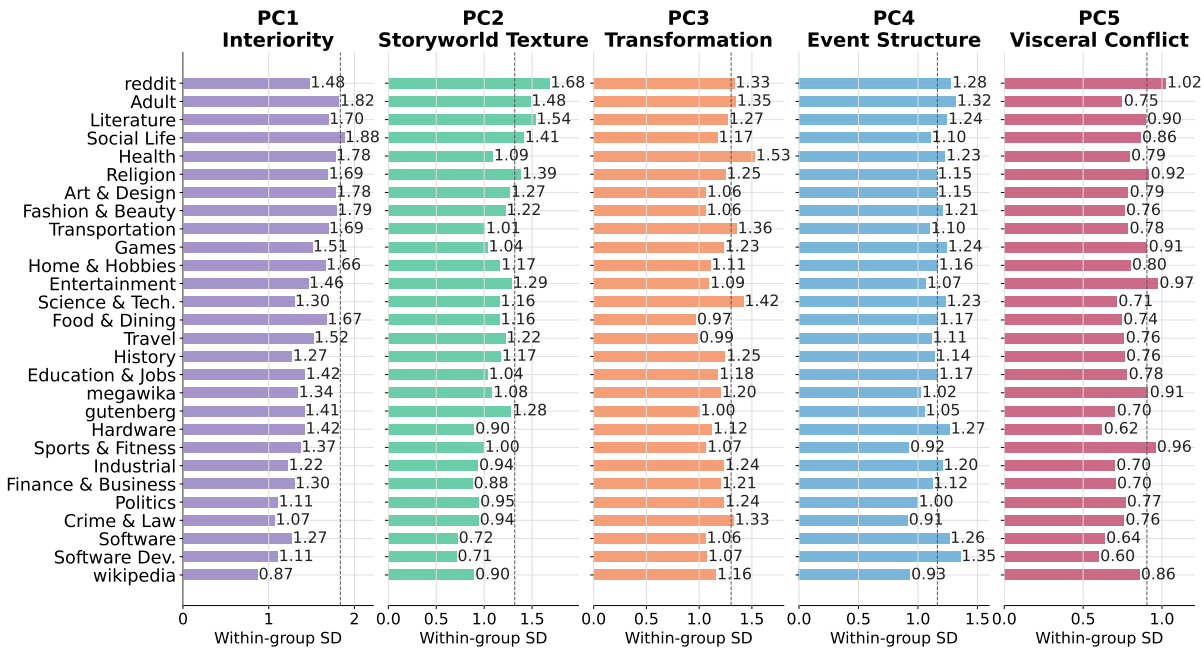

Figure 15: Standard deviation of principal component scores by topic (and Reddit, Gutenberg, and Wiki/megawiki).

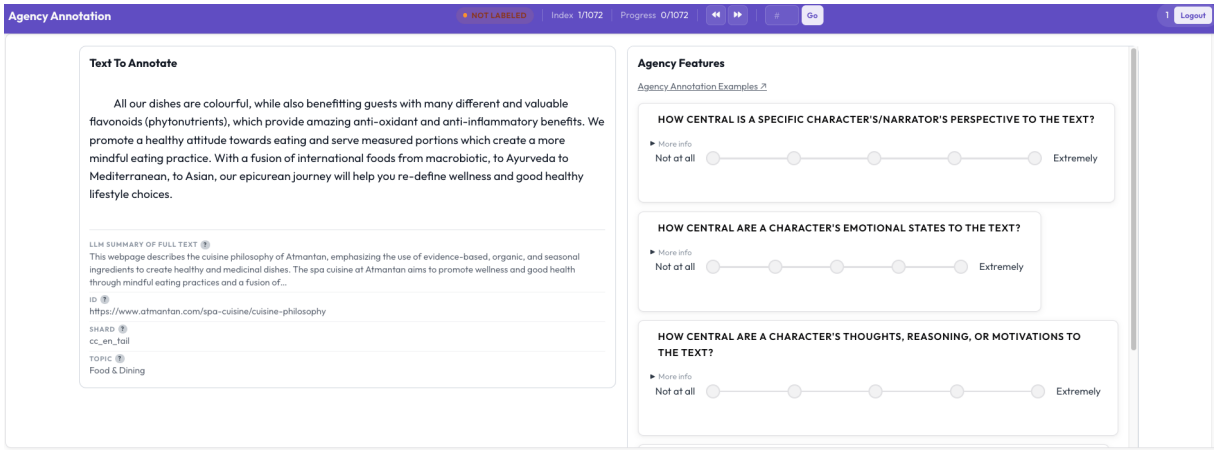

Figure 16: Screenshot of the agency annotation UI built upon the Potato software.

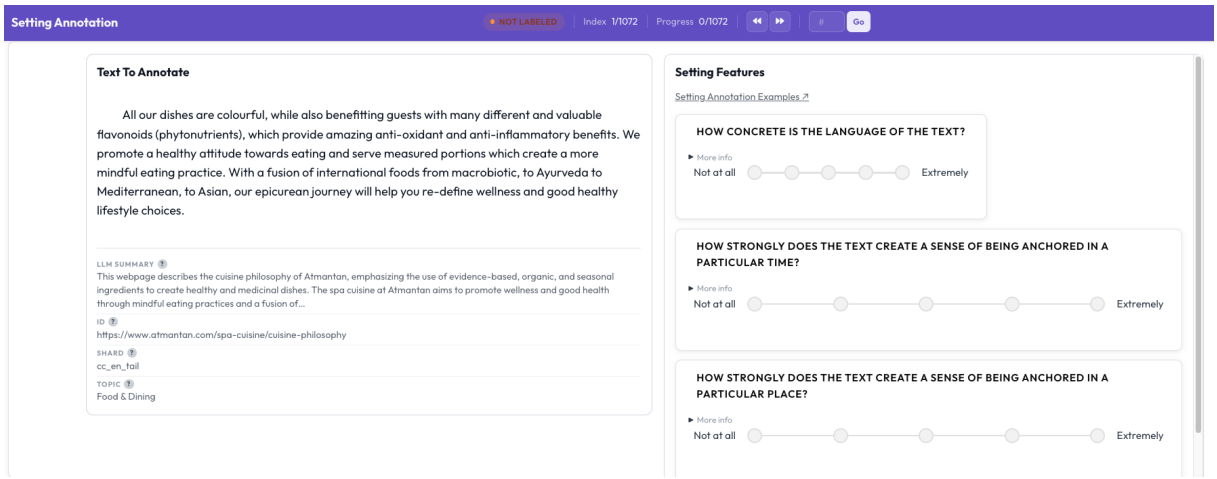

Figure 17: Screenshot of the setting annotation UI built upon the Potato software.

PC	Score	Topic/Source	Passage (truncated)
PC1 Interiority	+7.24	Reddit	*I knew I was eating a dead animal in the back of my head but at that moment it hit different and I realised there was a corpse in my mouth and I felt a sense of disgust that never left me. Another time I was tripping and I spent what felt like hours staring at my forearms, just thinking that they're made of meat...*
	−4.00	History	*Brooks, the oldest World War II veteran in the U.S. — and believed to be the oldest man in the country — died on Wednesday, Jan. 5, 2022, at the age of 112.*
PC2 Storyworld Texture	+6.52	Literature	*The sun was up, white-yellow through clouds, wafting salmon pink mist across the city like a blanket. Sprawling chimneys and gangling tower blocks dangled through the cloud. Layers of light broke away towards the hills on the horizon, shading each rise and fall...*
	−5.85	Reddit	*What really cemented the decision for me was the realization that I would rather live with a small chance potentially regretting not having children, than the soul crushing resentment and regret of having a child I didn't want.*
PC3 Transform.	+5.69	Health	*In 2008 Wesley was rushed to the hospital after having pain in his groin area due to an incident that happened in his home. "I woke up the next day, and my scrotum was the size of a honeydew melon," "My personal hell was just beginning."*
	−5.20	Literature	*Emile Durkheim is known as a functionalist states that everything serves a function in society and his main concern to discover what that function was. On the other hand Karl Marx, a conflict theorist, stresses that society is a complex system characterized by inequality and conflict...*
PC4 Event Structure	+2.83	Megawika	*In the case of zymase, an enzyme in yeast, he isolated its coenzyme, cozymase, and determined its chemical structure. Euler-Chelpin helped determine the chemical structure of several vitamins, which form portions of coenzymes.*
	−6.07	Literature	*born in the woods, raised among the jackals / epic battles were to come; forlorned he stood as the stars were falling one by one / he raised his hands and cursed eternal frost in soul to blame...*
PC5 Visceral Conflict	+4.45	Crime & Law	*The railroaders and citizens waited upon Johnson and were going to lynch him, but upon his solemn promise that his brutal acts should cease he was released, but was again taken in hand, and after being soundly flogged was ducked in the railway tank.*
	−4.26	Travel	*I was just talking about my first trip to NYC in 1996 (I was 26 at the time). I remember being amazed by all the activity, the beautiful and city-evolved people, the scale and the pace, and the deep sense of entanglement with everyone and anyone nearby.*

Table 5: Representative passages at the positive and negative extremes of each principal component. High-scoring passages exemplify the narrative quality captured by each axis; low-scoring passages illustrate its absence. Passages are lightly edited for length.

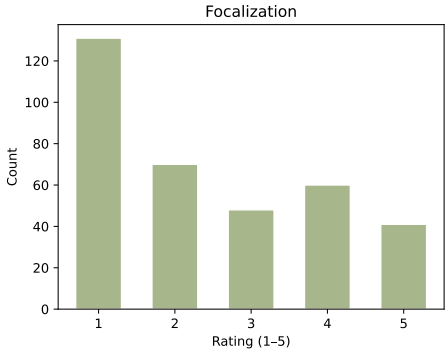

Figure 18: Distribution of gold annotated focalization labels.

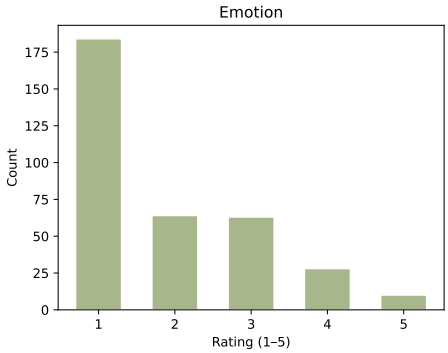

Figure 19: Distribution of gold annotated emotion labels.

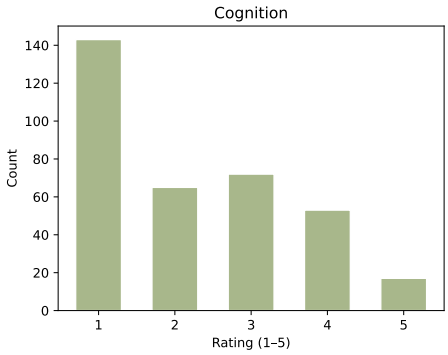

Figure 20: Distribution of gold annotated cognition labels.

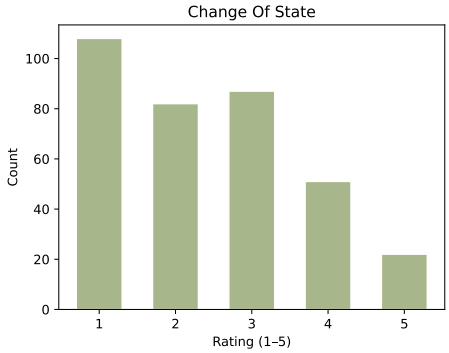

Figure 21: Distribution of gold annotated change of state labels.

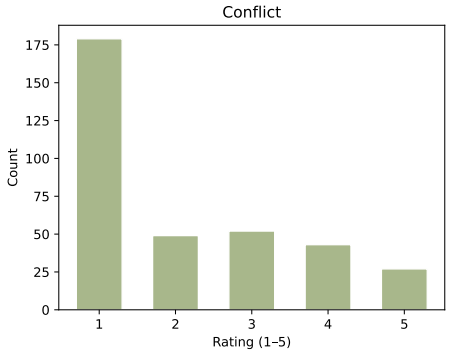

Figure 22: Distribution of gold annotated conflict labels.

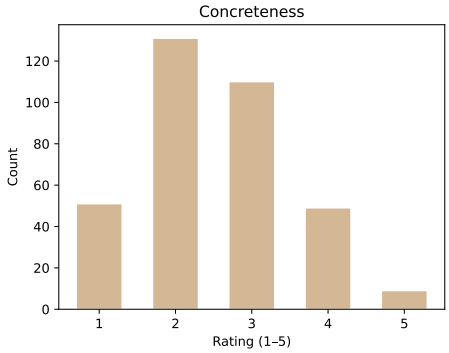

Figure 23: Distribution of gold annotated concreteness labels.

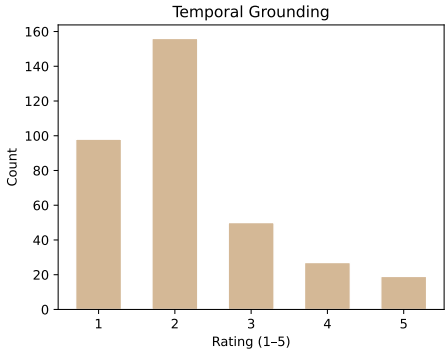

Figure 24: Distribution of gold annotated temporal grounding labels.

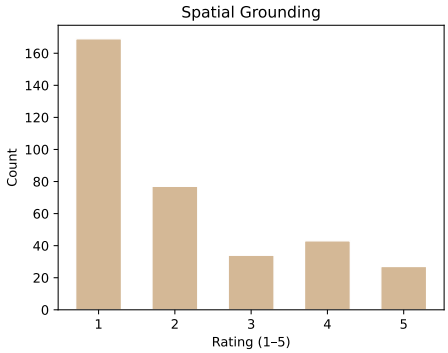

Figure 25: Distribution of gold annotated spatial grounding labels.

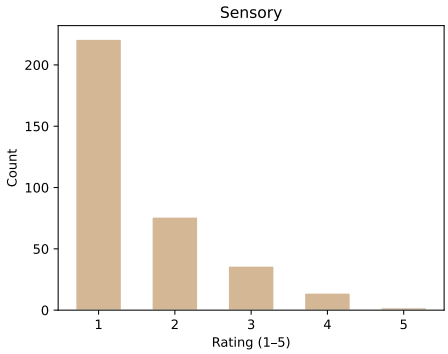

Figure 26: Distribution of gold annotated sensory labels.

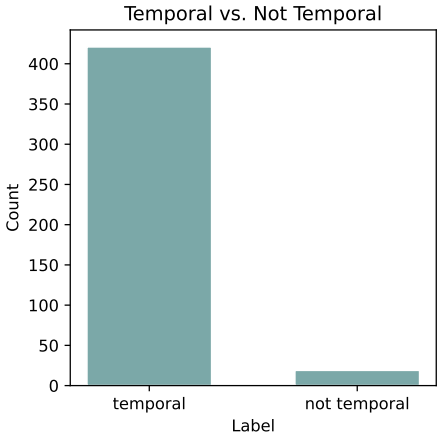

Figure 27: Distribution of gold annotated temporal event relation labels.

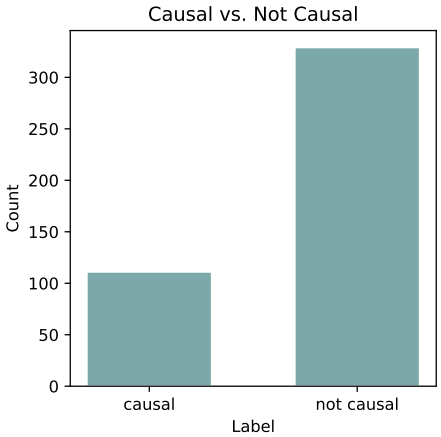

Figure 28: Distribution of gold annotated causal event relation labels.

Parameter	**Value**
Initial pool size	17,267,212 passages
Unique documents	4,757,629 documents
Chunk length	3 sentences
Narrative classifier	DeBERTa
batch size	64
max input length	512 tokens
Narrative confidence threshold	$p \geq 0.50$
Topic classifier	WebOrganizer
batch size	32
max input length	512 tokens
LLM model for annotators	Llama-3.1-8B-Instruct
quantization	4-bit NF4 (bfloat16)
snippet length	1,200 chars $\times$
max new tokens	120
temperature	0.2
top-p	0.9
summary character cap	300 chars

Table 6: Full configuration for the Dolma sampling pipeline.

You are an expert at identifying narrative qualities in web text. Your task is to read a passage and rate it on nine dimensions across two categories (Agency and Setting) each on a scale from 1 (not at all) to 5 (extremely).

AGENCY DIMENSIONS

These five dimensions capture how characters are represented as agents in the text. Importantly, these dimensions rate the presence and centrality of the feature in the text as written, not what can be inferred about the content.

1. Focalization
How central is a specific character's/narrator's perspective to the text?
A higher score reflects text in which the details and events are primarily presented through a specific character's perspective: their perceptions, thoughts, and feelings shape how the story is told. A lower score reflects text that reports details and events from the outside, describing what is observable without granting access to any character's internal experience.
Focalization is about whether the reader experiences events through a specific perceiving consciousness's inner life, regardless of the formal mechanism. That can be achieved through: direct quotation (if the content renders inner experience), free indirect discourse (*"she couldn't shake the feeling..."*), first person narration with genuine interiority, or close third person narration. When "you" appears in instructional or legal text, it should score low on focalization. However a specific second-person perspective that immerses the reader can score high.

- Score 5: *"she scanned the faces in the crowd, certain that someone was watching her"*
- Score 3: *"He wasn't sure the meeting had gone well. He gathered his things and headed for the door."*
- Score 1: *"she walked through the crowd."*

2. Emotion
How central are a character's emotional states to the text?
A higher score reflects text in which a character's emotional experience is a prominent feature. A lower score reflects text with little or no reference to how a character feels.

- Score 5: *"Completely paralyzed with fear, he was all of a sudden flooded with relief when he heard her voice"*
- Score 3: *"She was nervous about the presentation. She went into the interview room."*
- Score 1: *"he answered the phone."*

3. Cognition
How central are a character's thoughts, reasoning, or motivations to the text?
A higher score reflects text in which a character's thoughts, beliefs, reasoning, goals, or desires are a prominent feature. A lower score reflects text with little or no reference to what a character thinks, intends, or wants.

- Score 5: *"she kept turning the problem over in her mind, convinced there was something she had missed. Was it because they were so quiet in the meeting? Or their body language? She couldn't tell."*
- Score 3: *"She read the email twice. It wasn't entirely clear what they were asking for, but she thought she understood the gist. She drafted a reply."*
- Score 1: *"she sat at her desk"*

4. Change of State
How central is a change in a character's condition or state to the text?
A higher score reflects text in which a change in a character's condition/state is a prominent or organizing feature. A lower score reflects text where characters remain essentially unchanged. The change does not need to be completed within the passage; an in-progress or partially implied change still counts. If a passage describes a world event (a company going bankrupt, an earthquake, a death) that necessarily entails a change in a character's condition, that counts toward the score even if the character's internal response is not elaborated. Change of state should be rated on presence and centrality of the changes themselves, independent of how vividly or dramatically they are rendered.

- Score 5: *"by the time she reached the door, something in her had shifted. She wasn't the same person who had walked in"*
- Score 3: *"By the end of the conversation he felt somewhat better about the situation. It wasn't resolved, but it seemed more manageable than it had before."*
- Score 1: *"she sat at her desk."*

5. Conflict
How central is conflict involving characters to the text?
A higher score reflects text in which conflict is a dominant or organizing feature. A lower score reflects text in which conflict is absent or only incidentally present. Conflict can take many forms: tension between characters, internal psychological struggle, or opposition to institutions, environments, inanimate objects like technology, social forces, or physical events.

- Score 5: *"they had been arguing for hours, neither willing to give ground. The people around them started to stare. They were so caught up in the fight that they didn't even notice."*
- Score 3: *"They disagreed about the route. They took the highway and didn't talk much for the first hour."*
- Score 1: *"they sat across from each other at the table."*

Scoring Scale: 1 = Not at all 2 = Slightly 3 = Moderately 4 = Considerably 5 = Extremely

Figure 29: Annotation prompt for agency.

SETTING DIMENSIONS

These four dimensions capture how the text constructs a sense of place, time, and physical presence.

6. Concreteness
How concrete is the language of the text?
A higher score reflects text in which most content could be explained by pointing, demonstrating, or showing (objects, physical states, bodily actions, perceptual qualities). A lower score reflects text in which most content can only be explained using other words (categories, principles, relationships, institutions, ideas). The key distinction is not just whether physical things are present, but how much the text renders them. When a passage enumerates many specific, tangible, pointable things, the cumulative effect creates strong concrete presence.

- Score 5: *"the chipped blue mug sat on the edge of the sink, handle cracked from years of use"*
- Score 3: *"She had been commuting to the same office for six years. The building was on a corner downtown, glass and steel, indistinguishable from the ones on either side of it."*
- Score 1: *"people often form attachments to everyday objects"*

7. Temporal Grounding
How strongly does the text create a sense of being anchored in a particular time?
A higher score reflects text that creates a vivid sense of temporal location. A lower score reflects text that feels as if events could be taking place at any time. Two types contribute to the score: *historical grounding* locates the reader in a specific, unrepeatable moment (a year, an era, a named event); *cyclical grounding* locates the reader within a recurring temporal structure (a season, a time of day). Historical grounding can reach the upper end of the scale even with sparse language; cyclical grounding alone typically reaches a 3.

- Score 5: *"It was the summer of 2015 at 9:02am, before the world changed forever."*
- Score 3: *"This unique summer school gives teenagers the opportunity to work with tutors and musicians."*
- Score 1: *"he walked outside."*

8. Spatial Grounding
How strongly does the text create a sense of being anchored in a particular place?
A higher score reflects text that creates a vivid sense of spatial location. A lower score reflects text that feels as if events could be taking place anywhere. Two types contribute to the score: *geographic grounding* locates the reader on the globe (a country, city, named landmark); *proximate grounding* locates the reader in an immediate physical environment (a room, a building, a street). Geographic grounding can reach a 5 without proximate rendering; vividly rendered proximate grounding without geographic grounding caps at a 4.

- Score 5: *"In the narrow streets of Naples, the apartment's tiled floors kept cool even in August heat."*
- Score 3: *"She worked at an office in London." or "The kitchen was small, cluttered with dishes."*
- Score 1: *"She decided to quit her job."*

9. Sensory
How central are sensory details to the text?
A higher score reflects text in which one or more senses are central to how the passage is organized, where the sensory experience drives, anchors, or sustains the content. A lower score reflects text where events and ideas are conveyed without meaningful appeal to the senses. What matters is both whether a sense modality is central to the passage, and whether it is vividly described.

- Score 5: *"the bread was still warm, its crust crackling under her fingers, the whole kitchen thick with the smell of it"*
- Score 3: *"The bread was still warm when she pulled it from the oven. She set it on the rack and went to wash her hands."*
- Score 1: *"she made breakfast which consisted of bread and eggs."*

Scoring Scale: 1 = Not at all 2 = Slightly 3 = Moderately 4 = Considerably 5 = Extremely

Figure 30: Annotation prompt for setting.

Source	Reason for exclusion
Proof Pile 2 (Algebraic Stack)	Formal mathematics
Proof Pile 2 (OpenWebMath)	Formal mathematics
StarCoder	Source code
RedPajama-arXiv	Academic abstracts
RedPajama-Stack Exchange	Q&A, code snippets
PeS2o	Scientific papers
FLAN-TULU	Instruction templates

Table 7: Dolma sources excluded from sampling.

Score	Description
1	Not central. The passage describes events and details from a purely external vantage point with no access to any character's inner life.
2	Minimally central. A character's perspective is faintly implied but does not shape how events are presented. The passage remains largely external.
3	Moderately central. A character's perspective is present and shapes some of the passage, but external description and internal access are roughly balanced.
4	Considerably central. Events are predominantly filtered through a character's perceptions and inner life, with only occasional external description.
5	Extremely central. The passage is fully organized around a specific character's or narrator's perspective. Every detail is filtered through their perceptions, feelings, and inner experience.

Table 8: Focalization rating scale.

Score	Description
1	Not central. The passage contains no reference to how a character feels, either through observable behavior or internal access.
2	Minimally central. A character's emotional state is faintly implied through observable behavior (e.g., a gesture or expression), but feelings are not elaborated or sustained.
3	Moderately central. A character's emotional states are present and noticeable, but emotion is not the dominant or organizing feature of the passage.
4	Considerably central. Emotional experience is prominent and shapes how the passage is organized. The reader has meaningful access to how a character feels, either through behavior or internal description.
5	Extremely central. Emotional experience dominates the passage. The reader is given sustained and direct access to a character's feelings, which are foregrounded throughout.

Table 9: Internal emotion rating scale.

Score	Description
1	Not central. The passage contains no reference to what a character thinks, believes, reasons, or wants.
2	Minimally central. A character's mental state is faintly implied but not elaborated. Cognition is incidental to the passage.
3	Moderately central. A character's thoughts or reasoning are present and contribute to the passage, but are not its dominant or organizing feature.
4	Considerably central. A character's beliefs, deliberations, goals, or reasoning are prominent and shape how the passage is organized.
5	Extremely central. The passage is dominated by a character's inner cognitive life including their thoughts, reasoning, beliefs, and desires.

Table 10: Internal cognition rating scale.

Score	Description
1	Not central. Characters remain essentially unchanged across the passage. No change in physical, psychological, relational, or existential condition is present or implied.
2	Minimally central. A minor or incidental change is present or implied, but it is not an organizing feature of the passage.
3	Moderately central. A change in a character's condition is present and contributes meaningfully to the passage, but is not its dominant feature.
4	Considerably central. A change in a character's condition – physical, psychological, relational, or existential – is prominent and shapes how the passage is organized.
5	Extremely central. Change is the dominant and organizing feature of the passage. The transformation (whether completed, in progress, or strongly implied) is what the passage is fundamentally about.

Table 11: Change of state rating scale.

Score	Description
1	Not central. No conflict of any kind is present or implied. The passage describes actions or states without opposition or tension.
2	Minimally central. A hint of tension or opposition is present but remains entirely in the background and does not shape the passage.
3	Moderately central. Conflict is present and noticeable – between characters, within a character, or against external forces – but is not the dominant feature of the passage.
4	Considerably central. Conflict is prominent and shapes how the passage is organized, whether interpersonal, internal, or against external forces or circumstances.
5	Extremely central. Conflict dominates the passage. The opposition or tension is what the passage is fundamentally about, and it is sustained throughout.

Table 12: Conflict rating scale.

Score	Description
1	No physical referents. The passage is composed entirely of abstract ideas, generalizations, categories, or institutional descriptions. Nothing could be explained by pointing.
2	Physical objects or entities are named but not rendered. The text identifies things that exist in the world but provides no perceptual detail. No shape, color, texture, or material quality.
3	Physical referents are present with some rendering. Objects and actions are described with enough specificity to suggest a physical scene, but perceptual detail is sparse or thin.
4	Physical referents are rendered with clear perceptual detail. The passage provides enough specific, tangible description (through surface qualities, physical components, or rendered actions) that the reader can picture the scene.
5	Physical referents are richly and fully rendered. The passage is dense with perceptual particulars. Objects, environments, and actions are described with enough sensory and material detail that the world feels concretely present.

Table 13: Concreteness rating scale.

Score	Description
1	No temporal content. The passage has no markers of time whatsoever; events could be happening at any point in history or none at all.
2	Temporal language present but only sequencing events. Relative markers (e.g., *before*, *after*, *recently*, *a few years ago*) order events without locating the reader in any particular time.
3	Modest temporal location. The passage provides cyclical grounding – a season, a time of day, a day of the week – that gives the reader a felt sense of when without historical specificity. Most cyclical-only passages cap here.
4	Strong temporal location. The passage provides historical grounding – a specific year, era, named event, or cultural period – that anchors the reader in an identifiable moment. In rare cases, rich and sustained cyclical grounding can reach this level.
5	Vivid and sustained temporal immersion. Both historical and cyclical grounding work together, creating a strong and atmospheric sense of a specific moment in time.

Table 14: Temporal grounding rating scale.

Score	Description
1	No spatial content. The passage has no markers of place; events could be happening anywhere or nowhere.
2	Spatial language present but minimal. Either a bare proximate location with no rendering (e.g., *she sat in the kitchen*) or a vague geographic reference (e.g., *somewhere in Europe*). Places the reader no further than knowing a space exists.
3	Modest spatial location. Either a named geographic location that places the reader on the globe without evoking it (e.g., a city name), or a proximate location with some modest rendering that begins to create a felt sense of place.
4	Strong spatial location. Either geographic grounding with atmospheric texture, or a proximate environment rendered vividly enough that the reader feels present in the space.
5	Vivid and sustained spatial immersion. Both geographic and proximate grounding work together with rich rendering, creating a strong and atmospheric sense of a specific place.

Table 15: Spatial grounding rating scale.

Score	Description
1	No sensory content. Events and ideas are conveyed without any appeal to the senses; no modality is invoked.
2	Sensory content present but incidental. One or more senses are mentioned or implied, but the sensory experience is peripheral to the passage's meaning. Removing the sensory language would not substantially change the passage.
3	Sensory content present and noticeable. At least one sense modality recurs or contributes meaningfully to the passage, but sensory experience is not the organizing feature of the text.
4	Sensory experience is central. One or more senses are prominent and clearly organize how the passage is structured, even if the sensory content is not rendered in rich perceptual detail.
5	Sensory experience is dominant and richly rendered. The passage is organized around sensory experience across one or more modalities, with sustained and detailed evocation of what it feels like to perceive the described world.

Table 16: Sensory detail rating scale.

TEMPORAL ORDERING

Which event started first in the narrative timeline?

- **span1_first** — SPAN1 started before SPAN2
- **span2_first** — SPAN2 started before SPAN1
- **simultaneous** — Both events genuinely share a start point and are distinct events
- **same_event** — The two spans refer to the same event from different angles
- **too_hard_to_tell** — The start times could feasibly be in either order, or the events occur in completely unrelated temporal frames

If either span is not an event, use **not_applicable**.

Inferring order. Use explicit connectives ("before," "after," "then," "when," "while," "as"), backward-looking cues ("preceded," "followed," "prior to"), and logical necessity. A cause precedes its effect, a question is asked before it is answered, a proposal precedes an agreement. When one event logically presupposes the other, infer the ordering.

Reporting and retrospective verbs. Story-world events precede the speech acts or memory acts that describe them, even when the reporting verb appears first in the text. *"The coach expressed postgame that losing the player had impacted the offense"* → span2_first. Memory verbs (*recalls, remembers, reflects*) similarly occur after the events they describe.

Simultaneous vs. sequential. Use simultaneous only when neither event initiates the other and the text marks coincident starts ("just as X, Y"). When "as" or "while" appears, ask whether both events genuinely begin at the same instant (→ simultaneous) or whether one was already underway when the other began (→ span1_first or span2_first). Process/result pairs and physical chains are sequenced even when compressed.
Example: *"While touring in Germany, he released a record."* → span1_first. Touring was already underway when the release happened; "while" marks an ongoing background activity, not a coincident start.

Same event. Use same_event when both spans describe the same occurrence and neither adds new information the other omits. Continuation verbs (*adding, continuing, noting*) extend a prior speech act → use span1_first.

Too hard to tell. Use when: (a) two events describe different aspects of the same episode with no logical ordering between them; (b) the events involve different unrelated actors with no shared context; or (c) the ordering requires world knowledge inference the text does not support. If you find yourself inferring order from narrative position or plausible sequence rather than explicit textual evidence, use too_hard_to_tell. Uncertainty is not a failure.

CAUSAL RELATION

Given a confirmed temporal ordering, how are the two events causally related?

- **direct_cause** — E1 is sufficient to produce E2; given E1, E2 was bound to happen without any intervening decision or action.
  - Physical chain: *"She dropped her phone. The screen cracked."*
  - Involuntary reaction: *"He read the message. His stomach dropped."*
- **enables** — E1 creates a necessary precondition for E2: if E1 had not happened, E2 could not have happened in the same way. The dependency must be traceable in the text; mere co-occurrence is not enough. Enablement can run in either direction.
- **not_related** — No traceable causal dependency. Events that are merely co-present in the narrative without a direct conditional link.

Story-world vs. discourse-world. Story-world events (actions, occurrences, mental states) and discourse events (*said, reported, wrote, announced*) operate at different levels and cannot causally relate to each other. A fire does not cause a spokesperson's statement; the statement is an independent communicative act. When one span is a reporting verb and the other is a story-world event, use **not_related**. Two discourse events can causally relate to each other.

Distinguishing tip. Was E2 bound to happen given E1 alone? → direct_cause. Did E1 create a clear necessary precondition? → enables. Otherwise → not_related.

Figure 31: Annotation prompt for event relations.