Characterizing Narrative Content in Web-scale LLM Pretraining Data

Teagan Johnson[♣] **Elliott Ash**[♠] **Andrew Piper**[♡] **Maria Antoniak**[♣]

[♣]University of Colorado Boulder [♠]ETH Zürich [♡]McGill University

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 (agency, setting, and events) 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 measurable at scale across extremely heterogeneous data, (ii) we uncover a continuous, multidimensional narrative structure 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.

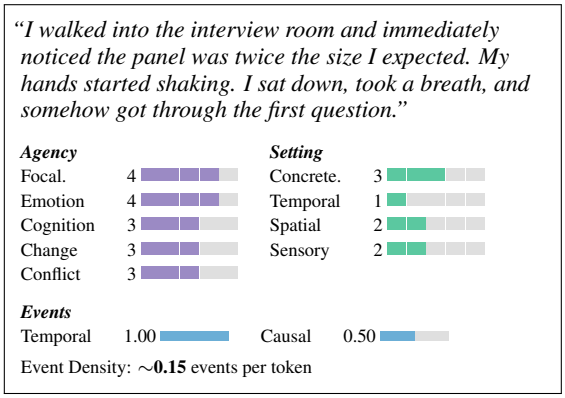

Figure 1: A web passage scored across our 12 narrative dimensions. Agency and setting dimensions are rated on a 1–5 Likert scale, temporal sequencing and causal density are passage-level proportions (0–1), and event density is the rate of event triggers per token. This passage scores high on agency and event features but low on setting, a "*narrative profile*" commonly seen across first-person web narratives.

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).

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 overrep-

https://huggingface.co/collections/teagrjohnson/narratives-in-llm-pretraining-data

https://github.com/johnsont4/narratives_in_pretraining_data_release

resented in ways that skew models toward particular event structures, perspectives, narrative forms, or safety violations. The well-documented creativity 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 structure** (Ochs et al., 2009), not as a binary classification (Pianzola, 2018; Antoniak et al., 2024). Our framework annotates passages along 11 narrative dimensions organized around three core elements from narrative theory: **events** (temporal ordering and causality), **agency** (focalization, emotion, cognition, change of state, and conflict), and **setting** (concreteness, temporal grounding, spatial grounding, and sensory detail). From a validated event detector we additionally derive **event density** (events per token), giving the 12 features shown in Fig. 1.

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 3M labeled passages sampled from the DOLMA corpus using a principled stratified pipeline, and
- a large-scale analysis revealing that narrativity in pretraining data forms a continuous, multidimensional structure 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; Wettig et al., 2025). 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. D.3.

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.

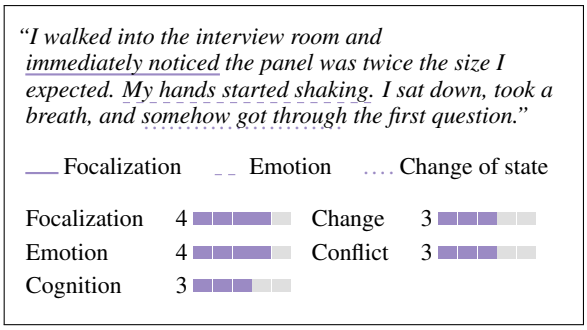

Figure 2: "I [...] immediately noticed" signals focalization, "my hands started shaking" conveys emotion through observable behavior, and "somehow got through" implies a change of state from overwhelmed to managing. Scores reflect the passage as a whole, not individual phrases.

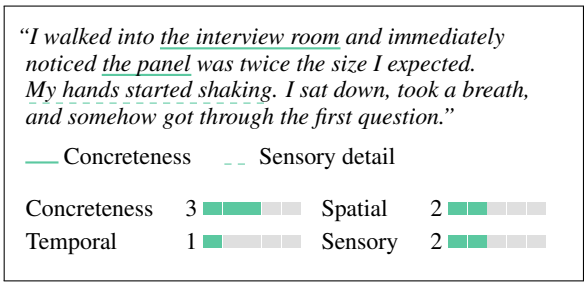

Figure 3: "The interview room" and "the panel" provide concrete referents but are named rather than rendered. "Hands started shaking" is tactile but incidental. Temporal and spatial grounding score low: this could be taking place in any room at any time.

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-

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), and is grounded in psycholinguistic evidence linking concreteness to imageability and lexical processing (Richardson, 1975).

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).

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, negated, and future events do not qualify.

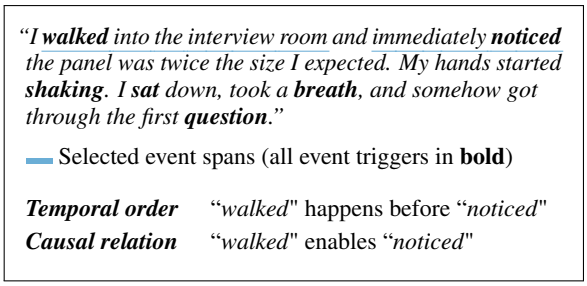

Figure 4: Two randomly selected adjacent event triggers are identified ("walked" and "noticed"). The event "walked into the room" precedes "noticed the panel". The events have a causal relationship: entering the room *enables* the character to notice the panel. Event density counts all triggers in the passage, while temporal order and causal relation describe the selected pair.

To measure 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 validate the model's performance on our web-scale corpus before proceeding (achieves an F1 of 0.85, see App. C.4).

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. Beyond the two annotated relation measures, a validated event detector (Sims et al. (2019)) also yields **event density**, the number of event triggers per token in the passage.

4 Sampling from Dolma

Dolma (Soldaini et al., 2024) is an open 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 A7).

Raw Dolma documents present several challenges for narrative annotation. Full webpage text often includes navigation elements, lists of unre-

lated 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[3] shards, allocating proportionally across sources according to the weights in Table A7. 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 A16 and A17), we assign a topic label using the WEBORGANIZER topic classifier (Wettig et al., 2025), which distinguishes 24 topics. 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 A7. 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 spanning ∼785K unique web documents. It follows the same stratified sampling procedure, with the unfiltered proportion shifted from 15% to 25% to increase coverage of non-narrative text. The raw DOLMA source distribution is shown in Fig. A16 and the distribution of NARRADOLMA categories (with topics applied to Common Crawl sources) is shown in Fig. A17.

[3] 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. D.2). 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 $N = 251$ 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.1.

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. A21, A22, and A23. 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.1.

5.3 NARRADOLMA

To scale narrative annotation beyond what LLM inference budgets allow, we distill the GEMMA labels into NARRABERT, a pair of ROBERTA 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 identified by the Sims et al. (2019) model, each pair with binary labels for temporal ordering and causal relation from GEMMA. We frame the Likert dimensions as regression 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 regression heads for the agency and setting dimensions and the other a dedicated encoder for event relations. Hyperparameters can be found in App. C.3.

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 (∼95% of span pairs labeled as temporally related, ∼75% labeled as not causally related, see Figure A1). Full per-dimension breakdowns for each feature are in App. A.1.

Scaling up annotation. We apply NARRABERT to the full NARRADOLMA corpus of ∼3M passages spanning ∼785K 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 Sims et al. (2019) 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 event detector itself also yields event density, the number of events per token. This produces a 12-dimensional narrative feature vector for each of the ∼3M passages, which together constitute the NARRADOLMA dataset. For all analyses in §6, we aggregate these passage-level vectors to the document level by averaging across each document's passages, yielding one narrative vector per document (∼785K documents). We perform a validity check by measuring Pearson correlations between automatic surface features and NARRABERT narrative predictions (App. Fig. A3).

6 The Narrative Landscape of Pretraining Data

6.1 Narrative Profiles Across Sources and Topics

We analyze narrative structure at the level of the ***category***: for Common Crawl documents this is the WEBORGANIZER topic, and for the four non–Common Crawl sources (Reddit, Gutenberg, Wikipedia, MegaWika) it is the source itself. For each category, we first characterize their narrative profiles by computing the mean z-scored value of every narrative dimension across its documents and visualize the result as a heatmap (Fig. 5).

Categories occupy distinct and coherent regions of the narrative space. For example, we can see a high-interiority cluster (Reddit, Literature, Gutenberg, Adult) which scores highly on focalization, emotion, and cognition while remaining low on setting. Another example is the cluster focused on rendered entities (Food & Dining, Fashion & Beauty, Travel, Home & Hobbies, Art & Design) which scores highly on concreteness and sensory detail. As a check that these profiles carry meaningful signal, we train classifiers to predict a document's category from its narrative features alone. They exceed chance and their misclassifications concentrate within the narrative clusters above (App. B.2).

Narrative variation *within* categories is substantial. Averaged across the 12 features, the within-category standard deviation of corpus-standardized features is 0.87 (where 1.0 matches the variability

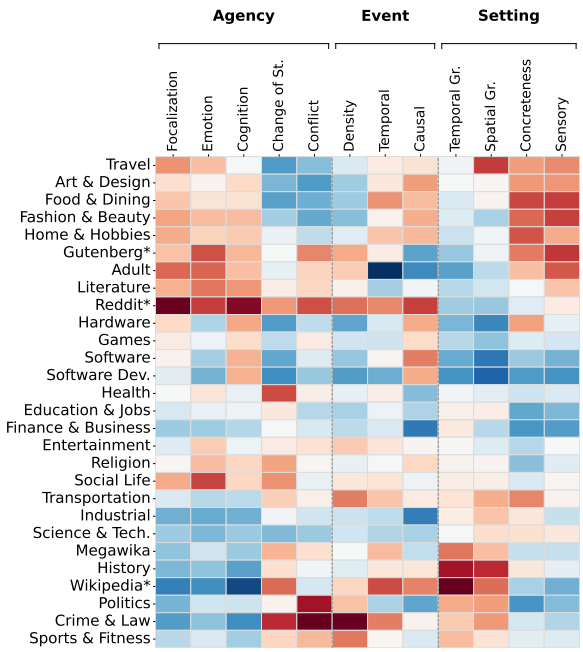

Figure 5: Mean z-scored narrative features by category, ordered by hierarchical clustering on profile similarity. ($*$) denotes original DOLMA sources, all other rows are Common Crawl topics assigned by WEBORGANIZER.

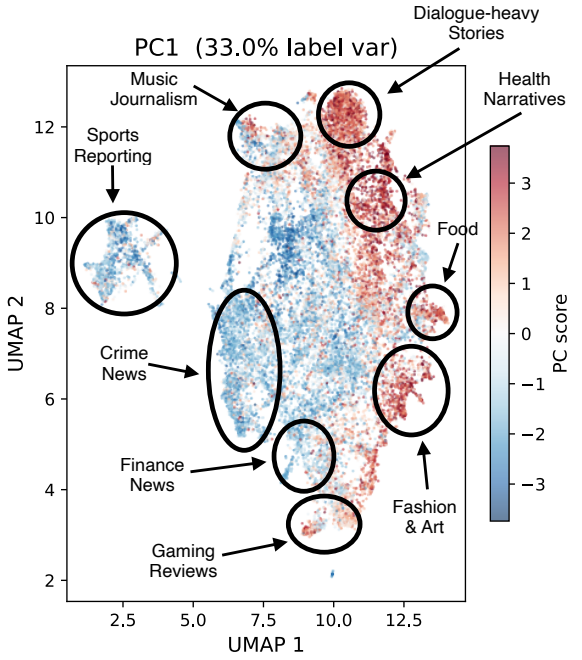

Figure 6: UMAP reduction of SBERT embeddings for 20,000 randomly sampled NARRADOLMA documents, colored by PC1 score (**interiority**). Labels are based on manual examination. Overlays for all three PCs appear in Fig. A5.

of the full corpus), so a category label removes only a small fraction of narrative variance. The homogeneity that does exist is concentrated in the factual and technical categories (Wikipedia (0.68) and Software Dev. (0.71)) while the narrative-rich categories are the most internally diverse (Reddit (1.07) and Literature (1.00)). The sources one would upweight to add narrative content are thus precisely those with the widest internal spread. This reinforces that **upweighting "high-narrative" sources does not uniformly increase narrative qualities**.

6.2 Dominant Modes of Narrative Variation

The narrative dimensions are strongly intercorrelated (Fig. A4). Specifically the interiority dimensions move together, as do concreteness and sensory detail, and change of state, conflict, and event density. To summarize this covariance compactly, we apply PCA to the correlation matrix of the ten always-defined features.[4] The first three components account for ∼72% of total variance (Fig. A6) and correspond to interpretable axes (Tab. A5).

PC1 captures *narrative interiority* with strong loadings for focalization, emotion, and cognition. **PC2** captures *grounded eventfulness* with strong loadings for change of state, conflict, and event density along with temporal and spatial grounding. **PC3** captures *storyworld texture* with strong loadings for concreteness, sensory detail, and spatial grounding.

To ask whether this variation aligns with lexical content, we overlay PC1 scores onto a UMAP reduction of SBERT embeddings for 20K randomly sampled NARRADOLMA documents (Fig. 6). PC1 shows clear spatial structure, with high-interiority documents concentrated in one region and topic clusters occupying distinct areas. This indicates that narrative interiority partly tracks lexical content. The remaining PC's overlays are shown in Fig. A5.

6.3 Where Narrativity Concentrates in Pretraining Data

Each document receives a **principal component score** (PC score) on each PC, computed as the dot product of its narrative feature vector with the PC's loading vector (Tab. A5). The distribution of PC scores per component is in Fig. A7. To characterize which categories concentrate at the extremes of each axis, we compute the proportion of each

[4]Temporal sequencing and causal density are undefined for the ∼34% of documents without event pairs and are near-orthogonal to the remaining dimensions. We therefore exclude them here. Excluding them raises the variance captured by the first three components from ∼60% to ∼72% and leaves the loadings essentially unchanged.

category's documents falling in the top quartile of each PC score (Fig. 7). Because NARRADOLMA over-represents narrative content (§4), these proportions describe relative concentration *within our sample* rather than prevalence in raw DOLMA. If narrative qualities were distributed uniformly, every category would place 25% of its documents in the top quartile. Categories above this baseline are over-represented at that axis's extreme.

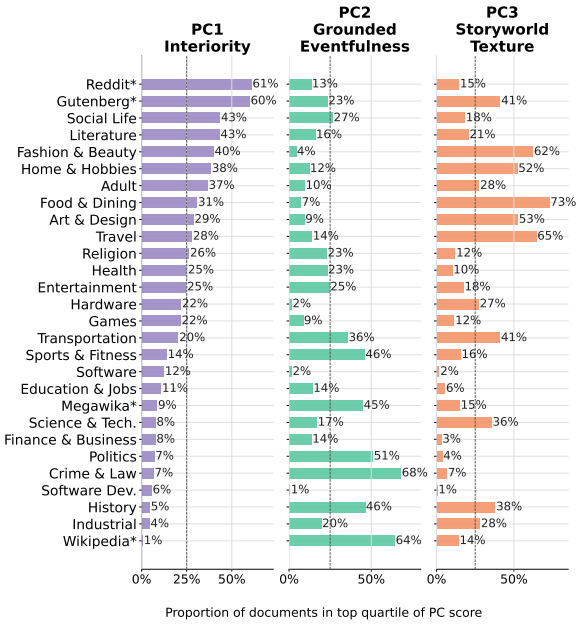

Figure 7: Proportion of each category's documents in the top quartile of each principal component's PC score. The dashed line marks the 25% uniform baseline. (∗) denotes original DOLMA sources, all other rows are Common Crawl topics assigned by WEBORGANIZER.

Interiority (PC1). Reddit and Gutenberg place more than half their documents in the top quartile, joined by Social Life and Literature. At the opposite extreme, Wikipedia, History, Politics, and Crime & Law are nearly absent. Categories dense with named entities and temporal markers contribute little interiority.

Grounded eventfulness (PC2). Crime & Law and Wikipedia lead, followed by Politics, History, and MegaWika, all narrating consequential change anchored in specific times and places. The texture-heavy categories (Food & Dining, Fashion & Beauty, Art & Design) and Software trail.

Storyworld texture (PC3). Food & Dining, Travel, and Fashion & Beauty lead, with Art & Design and Home & Hobbies close behind. Software Dev., Software, and Finance & Business are near the floor. Gutenberg scores high in storyworld texture while Reddit largely does not.

Cross-axis patterns. No single category dominates all three axes. Gutenberg is high on interiority and texture but only moderate on eventfulness. Reddit concentrates on interiority with little presence elsewhere. Wikipedia and Crime & Law lead eventfulness but are nearly invisible on interiority and texture. This indicates that narrative structure in pretraining data is a multidimensional configuration, not a single quantity that some categories simply have more of than others.

7 Conclusion

In this work we presented NARRADOLMA, a large-scale corpus characterizing narrative structure in pretraining data, together with NARRABERT, an efficient classifier for its 11 narrative dimensions. We developed an annotation framework grounded in narrative theory, validated through human annotation and LLM-assisted labeling, and applied it to produce narrative feature vectors for ∼3M passages across ∼785K documents.

Our analysis reveals three core empirical findings. First, the narrative dimensions covary into a small number of interpretable axes (interiority, grounded eventfulness, and storyworld texture). Second, these axes are unequally distributed across categories in ways that current curation practices neither measure nor account for. Third, narrative variation within categories 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.

Together these results position narrative structure as a measurable, multidimensional property of pretraining data rather than a single "narrativity" quantity. 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. Our framework, datasets, and models provide a foundation for 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, NARRADOLMA is a stratified subsample (∼3M passages) of a >3-trillion-token corpus, and it deliberately over-represents narrative content. The relative structural patterns we report are therefore informative but absolute prevalence figures do not transfer to raw DOLMA.

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. Primary annotation that was used for model validation was conducted by one author. Some agreement scores between annotators remain modest, particularly temporal ordering ($\kappa = 0.60$), suggesting some of these dimensions are intrinsically difficult regardless of annotator expertise. It is worth noting this lower κ score could be due to the severe class imbalance of temporal relations between events.

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.

Fourth, our analysis is limited to English text and to a single theoretical operationalization of narrative. Other languages, narrative traditions, and frameworks may surface different structure.

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).[5] 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. 🤗 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.

Acknowledgments

We thank our annotators for their careful work annotating the gold dataset, often across difficult and ambiguous passages, and for the many discussions that helped refine the annotation guidelines. We are grateful to our colleagues for feedback on earlier drafts of this work, and to members of the NLP Group at the University of Colorado Boulder for valuable discussion throughout the project. We also

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

🤗 https://huggingface.co/collections/teagrjohnson/narratives-in-llm-pretraining-data

https://github.com/johnsont4/narratives_in_pretraining_data_release

thank the anonymous reviewers for their constructive comments. Lastly we thank Rohan Das, Matt Pauk, Advait Deshmukh, and Uma Gunturi for their help with developing the annotation framework.

Computational resources were provided by the University of Colorado Boulder Research Computing group, including the Blanca condo cluster.

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.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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

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

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*.

John TE Richardson. 1975. Imagery, concreteness, and lexical complexity. *The Quarterly Journal of Experimental Psychology*, 27(2):211–223.

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*.

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.

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?

A Datasets & Validation

This appendix reports the evidence behind NARRABERT: gold-set label distributions, inter-annotator and model agreement at each validation stage, and a construct-validity check correlating NARRABERT's predictions against surface lexical features.

Figure A1 shows the distributions of gold annotated labels across the 11 manually-annotated features. Figure A2 shows the distributions of the full NARRADOLMA corpus labels.

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					
Is span event	502	—	—	0.813	0.926
Temporal	112	—	—	0.601	0.890
Causality	73	—	—	0.778	0.902

Table A1: Human inter-annotator agreement. Agency and setting report MAE and Krippendorff's α; event relations report Cohen's κ and F1. For event annotation, annotators are asked to determine whether pre-highlighted event spans (detailed in App. C.4) are events. If both pre-highlighted spans are events, annotators label temporal and causal relations.

A.1 Agreement Scores

Tables A1, A2, and A3 report full per-dimension agreement scores for each validation stage. Table A1 reports human inter-annotator agreement on the 100-passage overlap set for agency and setting and on the full event annotation set. Agency and setting dimensions report MAE and Krippendorff's α, while event relations report Cohen's κ and macro F1. Table A2 reports agreement between each of the three LLMs and gold split A ($N = 200$ for agency and setting, full event set for event relations), with the best-performing model per dimension bolded. Table A3 reports NARRABERT agreement against the held-out gold split B ($N = 200$ for agency and setting, full event set for event relations), using annotations not seen during LLM validation.

A.2 Automatic Lexical Features

To assess the construct validity of the NARRABERT labels, we computed Pearson correlations between each label and a set of 18 automatic lexical features extracted by a custom pipeline (see Figure A3). Features were computed using spaCy (`en_core_web_lg`), the Brysbaert et al. Brysbaert et al. (2014) concreteness lexicon, and the `cardiffnlp/roberta-base-sentiment` clas-

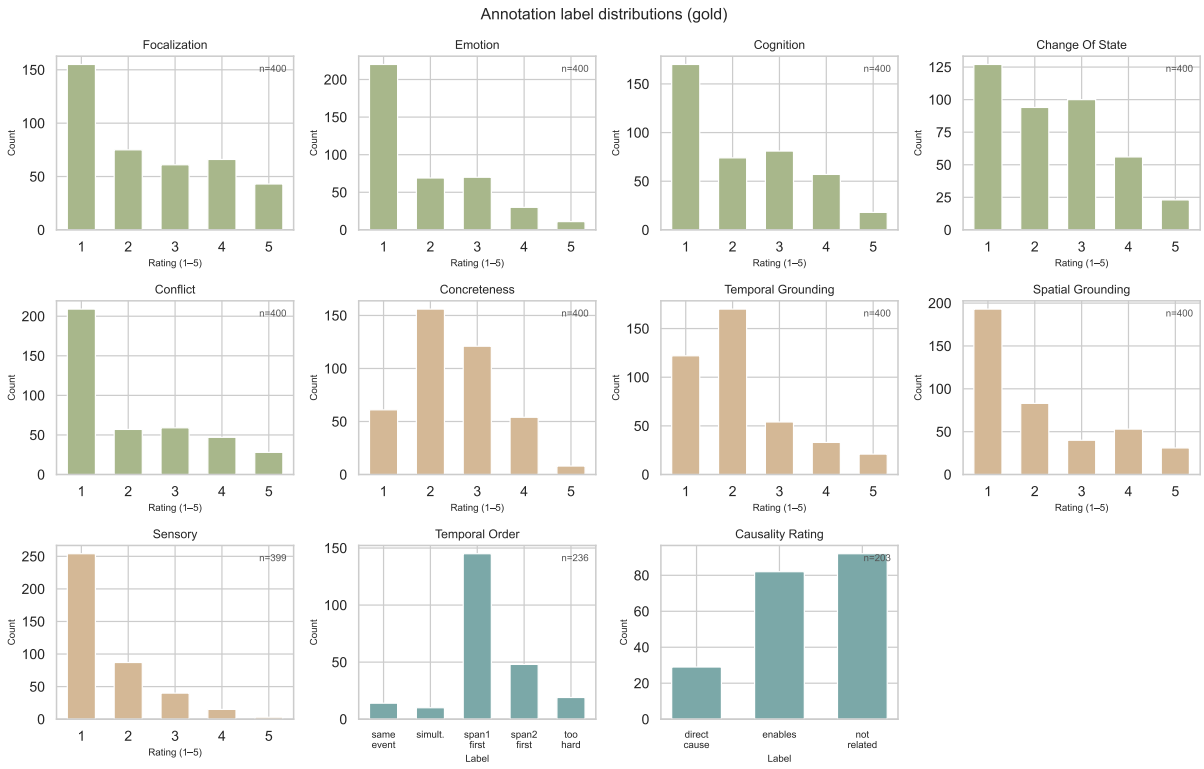

Figure A1: Distribution of gold annotated labels across the 11 manually-annotated features.

sifier (Barbieri et al., 2020). Table A4 lists each feature, its category, and its operational definition.

spaCy features. All token-level and entity-level features were extracted in a single pass using `en_core_web_lg` with the parser disabled and a sentence boundary component (`senter` or `sentencizer`) enabled.

Concreteness. Mean concreteness was computed over the subset of tokens present in the lexicon. Coverage (proportion of tokens matched) was also recorded but not used in the correlation analysis.

Sentiment. Texts were passed to `cardiffnlp/roberta-base-sentiment` with truncation to 512 tokens. All three class probabilities (positive, neutral, negative) were retained as separate features.

B Analysis Details

This appendix expands the corpus analysis: the feature correlations that motivate the PCA, the full PCA (loadings, scree, and score distributions), full UMAP overlays, within-category dispersion of PCA scores, representative passages, and the classification tests probing how narrative features separate categories.

Figure A4 shows the pearson correlation between each feature in NARRADOLMA.

Figure A5 shows each of the first 3 PCs overlayed upon a UMAP reduction of 20,000 randomly sampled passages. Figure A7 shows the distribution of PC scores over the full NARRADOLMA corpus. Figures A8 and A9 show the standard deviations of each PC for each source and category, respectively.

B.1 PCA Extreme Examples

Table A6 shows passages that have high PC scores for the first three PCs.

B.2 Classification Tests

As a check that the narrative features carry separable signal, we train logistic regression (LR) and gradient-boosted trees (GBT) to predict each document's category (28 classes) and a binary narrativity label, using the 12 features alone. Both beat chance by a wide margin: category macro-F1 is 0.32 (GBT) and 0.25 (LR) against 0.03 chance, and binary narrativity reaches 0.76/0.74 against a baseline of 0.50. The two tasks rely on different features. Category is driven by setting (sensory, concreteness, spatial grounding) and conflict (Figs. A10, A11), binary narrativity by

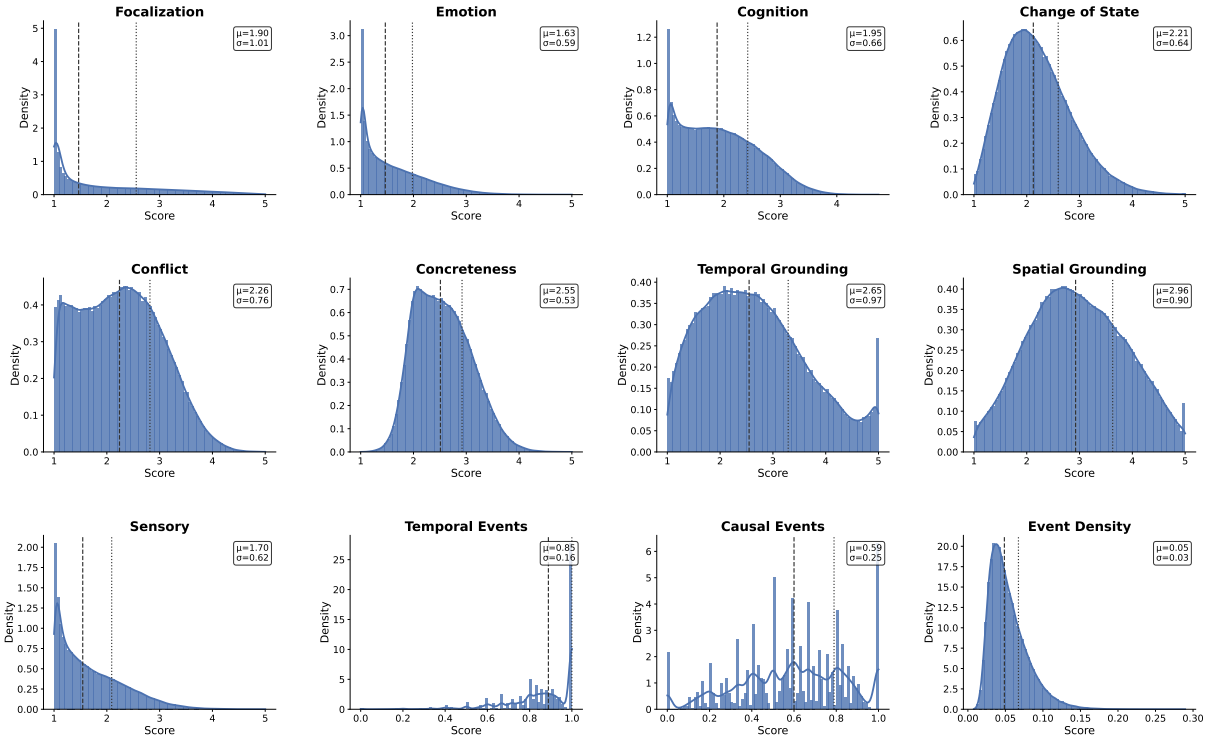

Figure A2: Distributions of each narrative feature in the full NARRADOLMA corpus.

event density, focalization, and change of state (Figs. A12, A13), with the event-relation rates weakest for both. This supports that narrative composition is distinct from topic. We attribute with the sign-free GBT permutation importances, since the collinear interiority features give misleading negative LR coefficients. The category confusion matrices (Figs. A14, A15), ordered by the Fig. 5 dendrogram, are block-diagonal, so errors fall between narratively similar categories.

C Sampling & Pipeline

This appendix details how passages were drawn from DOLMA: the source and topic distributions of the sample, the sources we excluded and why, the full sampling configuration, and the validation of the event-span detector that underlies our event features.

C.1 Dolma Source Distributions

Figures A16 and A17 show the distributions of Dolma sources and expanded categories, respectively. Table A7 shows the distribution of the gold set and NARRADOLMA corpus.

C.2 Excluded Sources from Dolma

Table A8 lists Dolma sources that were explicitly excluded (sampling weight set to zero) and the rationale for each.

C.3 Hyperparameters

Table A9 lists all configuration parameters used across the pipeline steps.

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.

		Sonnet			Qwen3			Gemma4		
Dimension	**N**	**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 A2: 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 A3: 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.

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 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 Qwen3-235B-A22B 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.

C.4 Event Span Detection

Event trigger spans are the lexical anchors around which event meaning is organized (Sims et al., 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[8]. On average, each passage contains 6.9 verb spans.

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

Feature	Category	Definition
First Person Rate	Point-of-view	Proportion of tokens that are first-person pronouns (*I, me, my, mine, myself, we, us, our, ours, ourselves*).
Second Person Rate	Point-of-view	Proportion of tokens that are second-person pronouns (*you, your, yours, yourself, yourselves*).
Third Person Rate	Point-of-view	Proportion of tokens that are third-person pronouns (*he, him, his, himself, she, her, hers, herself, they, them, their, theirs, themselves, it, its, itself*).
Masculine Pronoun Rate	Point-of-view	Masculine pronouns (*he, him, his, himself*) as a fraction of all gendered (masculine + feminine) pronouns.
Feminine Pronoun Rate	Point-of-view	Feminine pronouns (*she, her, hers, herself*) as a fraction of all gendered (masculine + feminine) pronouns.
Temporal Mention Rate†	Named entity	Named entities tagged `DATE` or `TIME` per 100 tokens.
Location Mention Rate†	Named entity	Named entities tagged `GPE`, `LOC`, or `FAC` per 100 tokens.
Named Entity Density†	Named entity	All named entities per 100 tokens.
Lexical Density	Lexical / syntactic	Unique content-word lemmas (nouns, verbs, adjectives, adverbs; non-stop-word, alphabetic) divided by total token count.
Past-Tense Verb Rate	Lexical / syntactic	Verbs carrying a past-tense morphological feature as a proportion of all verbs.
Average Sentence Length	Lexical / syntactic	Mean number of non-space tokens per sentence, using spaCy's sentence boundary detector.
Type-Token Ratio	Lexical / syntactic	Number of unique lowercase word forms divided by total token count.
Negation Density†	Lexical / syntactic	Occurrences of a fixed negation set (*not, n't, no, never, neither, nor, nobody, nothing, nowhere, none, without*) per 100 tokens.
Brysbaert Concreteness	Lexical / syntactic	Mean concreteness rating across all tokens found in the Brysbaert et al. 2014 lexicon.
Causal Connective Density†	Lexical / syntactic	Single-word causal connectives (*because, therefore, thus, hence, consequently, so, since, thereby, accordingly, wherefore*) plus multi-word phrases (*due to, owing to, for this reason, as a consequence, as a result*) per 100 tokens.
Sentiment Positive	Sentiment	Positive softmax probability from `cardiffnlp/roberta-base-sentiment`.
Sentiment Neutral	Sentiment	Neutral softmax probability from `cardiffnlp/roberta-base-sentiment`.
Sentiment Negative	Sentiment	Negative softmax probability from `cardiffnlp/roberta-base-sentiment`.

Table A4: Automatic lexical features used in the Pearson correlation analysis. All rate features are proportions of non-space tokens unless otherwise noted. Features marked with † are expressed per 100 tokens.

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.

D Annotation Framework

This appendix documents the full annotation framework: the rating scale for each of the 11 dimensions, the interface annotators used, the document summaries shown to orient them, and the prompts used to elicit the LLM annotations.

D.1 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."*

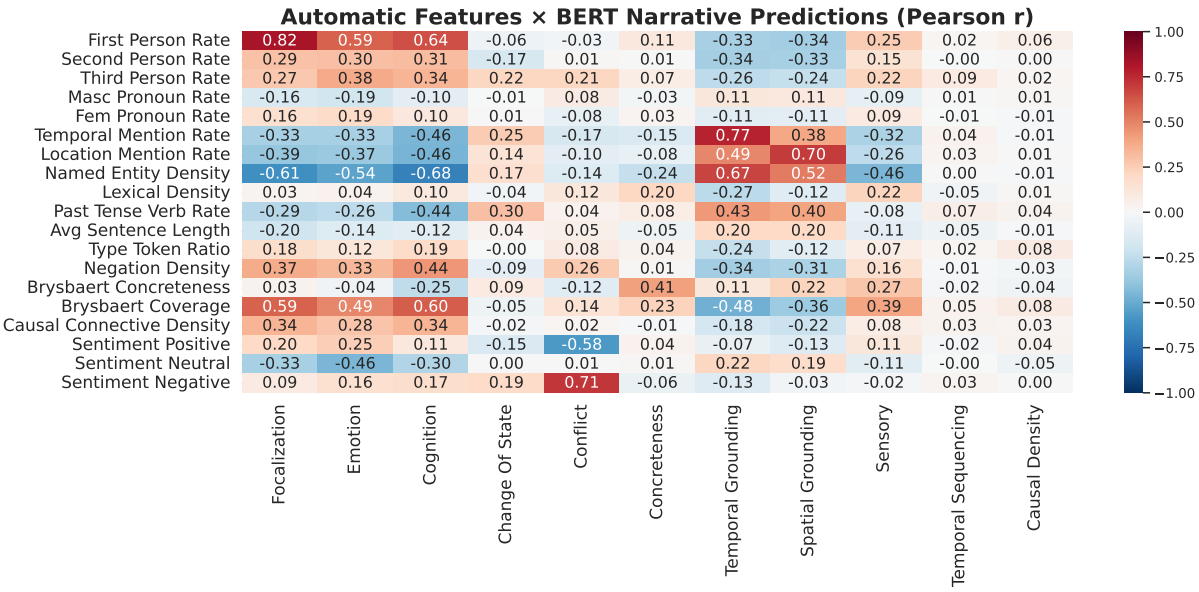

Figure A3: Pearson correlation between NARRADOLMA labels and a set of lexical-based features.

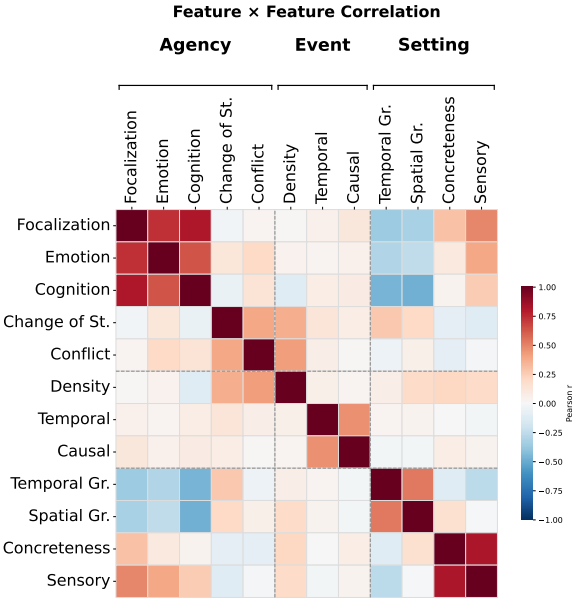

Figure A4: Pearson correlation of each feature across the full NARRADOLMA corpus.

	PC1	PC2	PC3	
Focalization	0.51	-0.09	-0.01	Agency
Emotion	0.47	-0.00	-0.13	
Cognition	0.47	-0.16	-0.21	
Change Of State	0.15	0.46	-0.27	
Conflict	0.23	0.34	-0.32	
Event Density	0.16	0.48	0.01	Event
Temporal Grounding	-0.17	0.44	-0.00	Setting
Spatial Grounding	-0.13	0.46	0.20	
Concreteness	0.20	0.11	0.64	
Sensory	0.33	0.04	0.55	

Table A5: Loadings for the first three principal components, explaining ∼72% of variance (see Fig. A6).

D.2 Potato Interface

Figures A18 and A19 show the UI that annotators use to label the sampled texts for agency and setting, respectively.

D.3 Annotator Feature scales

Tables A10-A18 show the feature scales given to annotators for each of the 11 fine-grained narrative features.

D.4 LLM Annotation Prompt: Agency and Setting

Fig. A21 displays the prompt used to elicit agency annotations and Fig. A22 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.

D.5 LLM Annotation Prompt: Event Relations

Fig. A23 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.

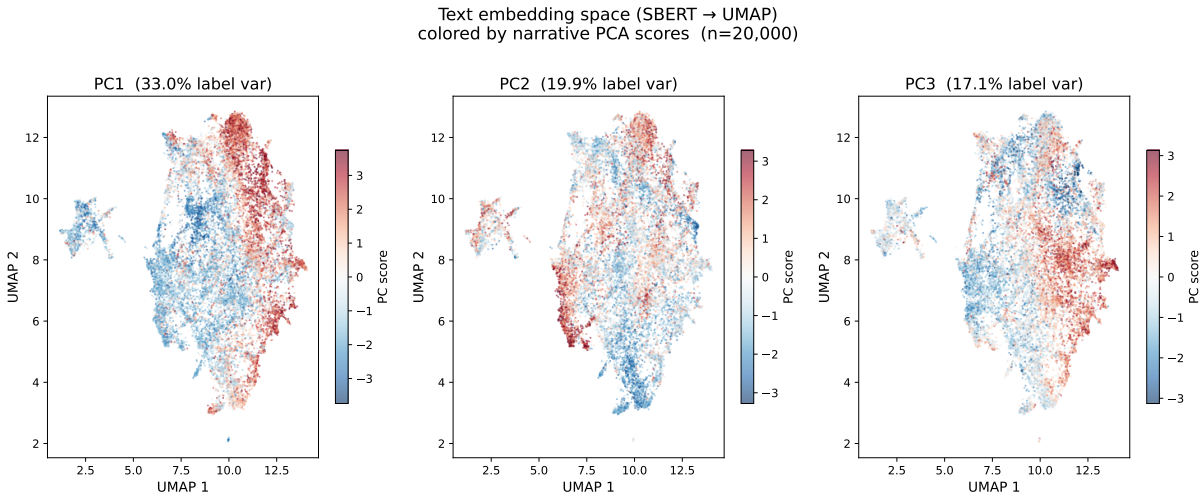

Figure A5: UMAP reduction of SBERT embeddings for 20,000 sampled documents, colored by narrative PCA scores. **PC1 (interiority)** shows clear spatial structure in embedding space. **PC2 (Grounded Eventfulness)** and **PC 3 (Storyworld Texture)** also show spatial structures that are each different.

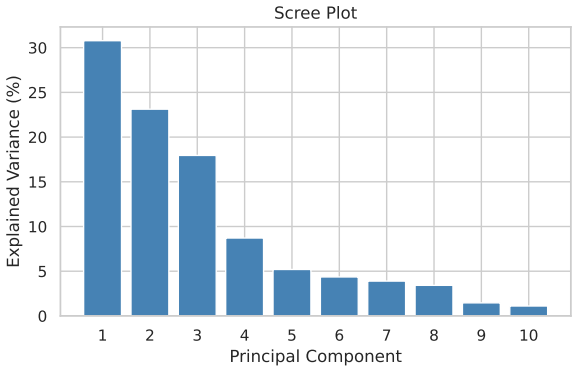

Figure A6: Scree plot of the 10 principal components across the 10 narrative features. We see a significant drop after PC3 to below 10%.

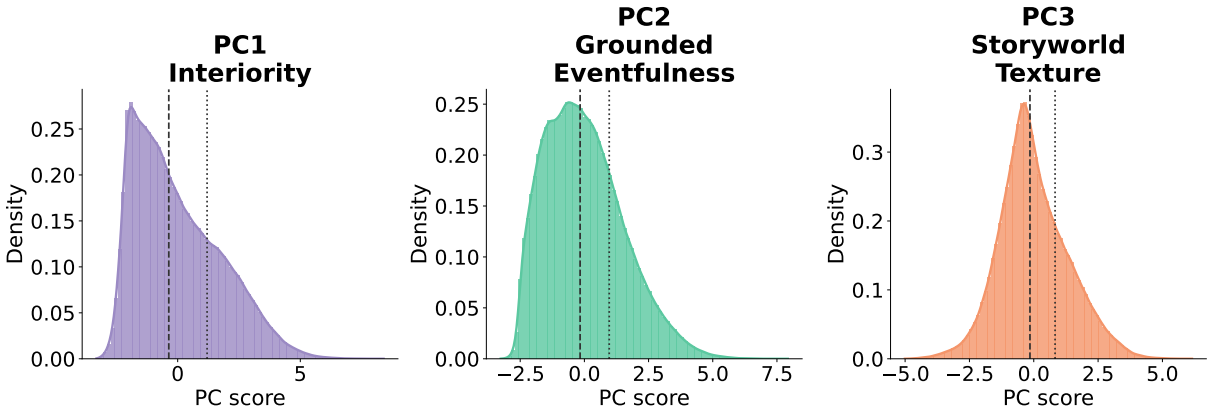

Figure A7: Distributions of the PC scores for each PC. Lines indicating the median and the top quartile. This is the distribution of all ∼785K documents from NARRADOLMA.

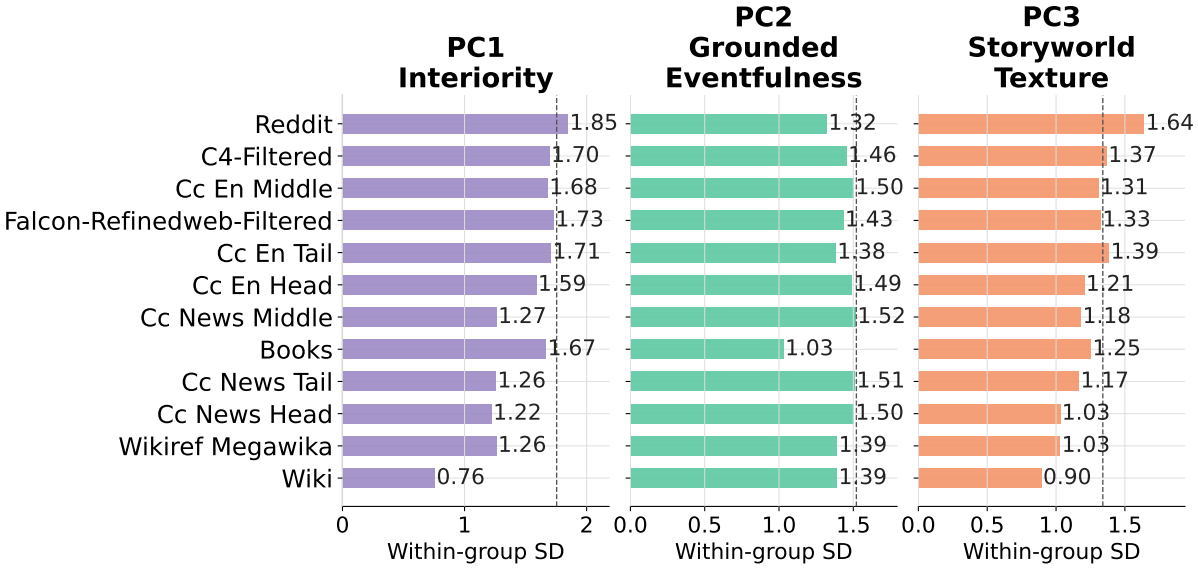

Figure A8: Standard deviation of principal components by original DOLMA source. These are the PC score SDs of all ∼785K documents from NARRADOLMA.

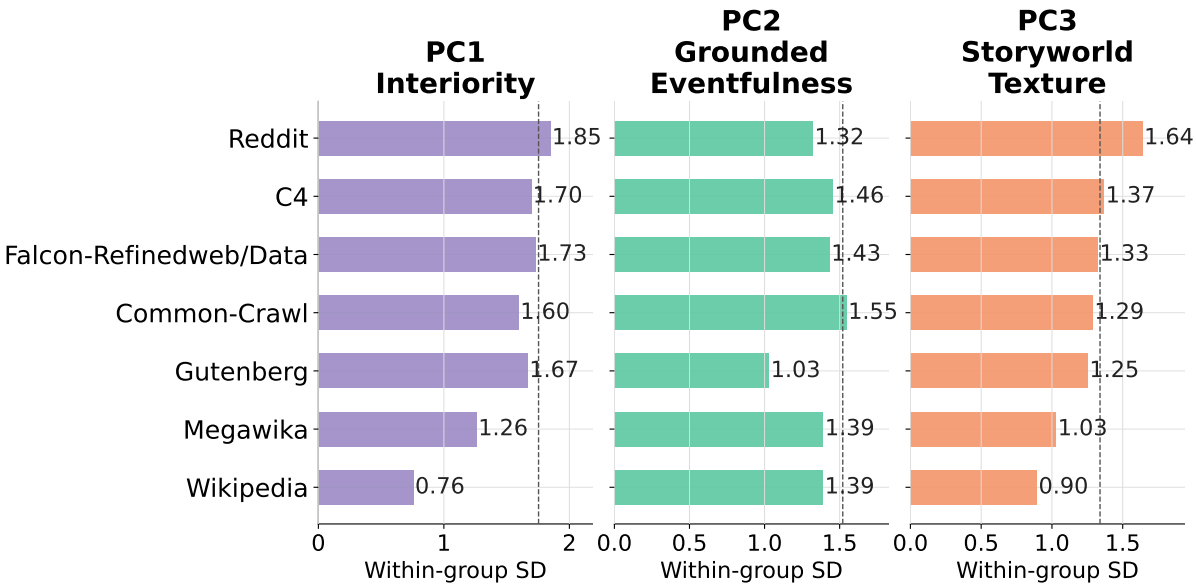

Figure A9: Standard deviation of principal component scores by category. These are the PC score SDs of all ∼785K documents from NARRADOLMA.

PC	Score	Topic/Source	Passage
PC1 Interiority	+8.22	Literature	*In my head everything was a washed-out shade of grey, the color leached away by horror and utter shock. I was still there—in the café, listening to him scream and scream, staring at myself as I stood frozen, the empty cup still clutched in my hand. Wasn't my fault.*
	−3.31	History	*University of Glasgow Research Conference, Glasgow, UK, 16-17 June 2021. Struan, A. and Alexander, M. (2010) Expressions of Civilisation and Colonisation in the Historical Thesaurus of the Oxford English Dictionary. From the Grand Tour to Mass Tourism: The Modern History of the British Abroad, Newcastle, UK, 01-02 Apr 2010.*
PC2 Grounded Eventfulness	+7.92	Crime & Law	*But at Norwich Magistrates Court on Thursday (August 25) she was found guilty after a doorman who witnessed the attack described seeing her punch the other woman in the face to leave her bloodied. Graham Howton said he had seen her demanding to know who had her handbag before throwing the punch.*
	−3.20	Reddit	*I have a few artists in mind, but I am not sure what the subject should be. I would like to have something meaningful that symbolizes growth; i don't really know of any nonreligious symbols/art/designs that would be good for that.*
PC3 Storyworld Texture	+6.14	Literature	*The sun peeks out long enough to highlight the changing hues, but grayish clouds are skittering across the sky, pushed by a blustery wind that is quickly relieving all the ash trees of their gorgeous golden leaves. As I walk to the post office, my hair blows in my face and the leaves swirl and dance into piles by the sidewalk.*
	−5.00	Reddit	*There was a lot of abuse and violence towards me during my early years and I've had to deal with the ramifications, but in order for me to heal I needed to forgive her for what she did, which ultimately means letting the mother figure I craved in my life die.*

Table A6: Representative passages at the positive and negative extremes of each principal component. High-scoring passages exemplify the narrative quality captured by each axis while low-scoring passages illustrate its absence.

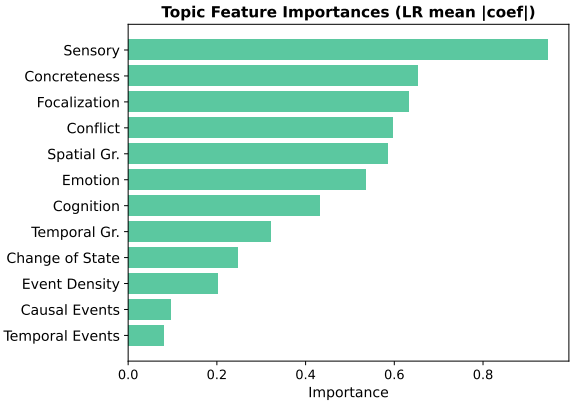

Figure A10: Importance of each feature for classifying category for the multinomial logistic regression model.

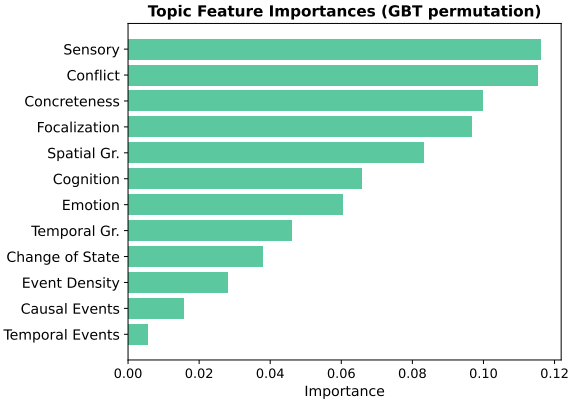

Figure A11: Importance of each feature for classifying category for the gradient boosted tree.

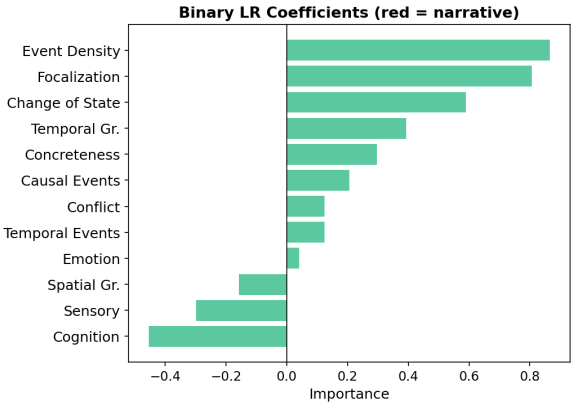

Figure A12: Importance of each feature for classifying binary narrativity for the multinomial logistic regression model.

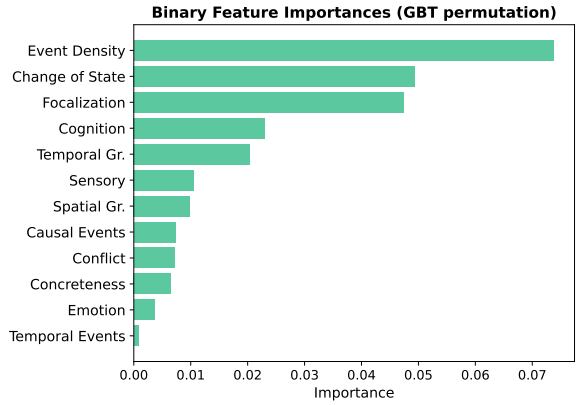

Figure A13: Importance of each feature for classifying binary narrativity for the gradient boosted tree.

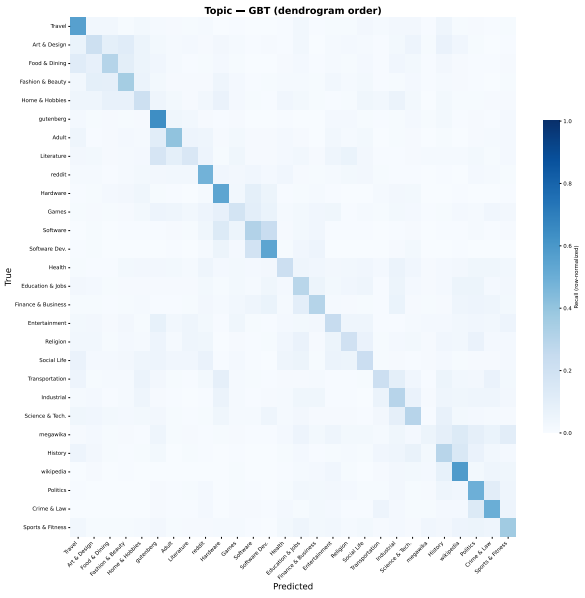

Figure A14: Confusion matrix of the gradient boosted tree category classifications.

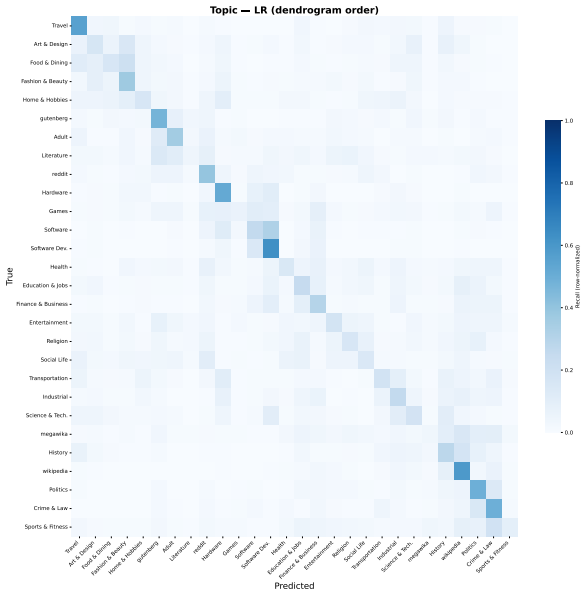

Figure A15: Confusion matrix of the multinomial logistic regression category classifications.

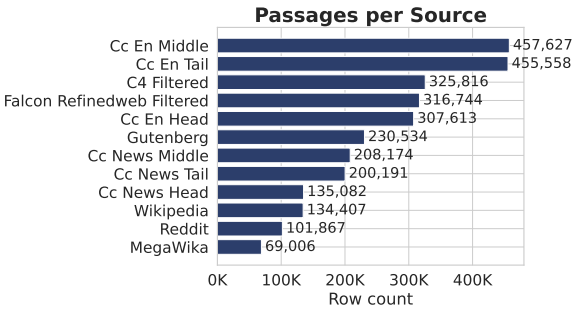

Figure A16: Distribution of Dolma sources that we sample for NARRADOLMA.

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 A7: *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.

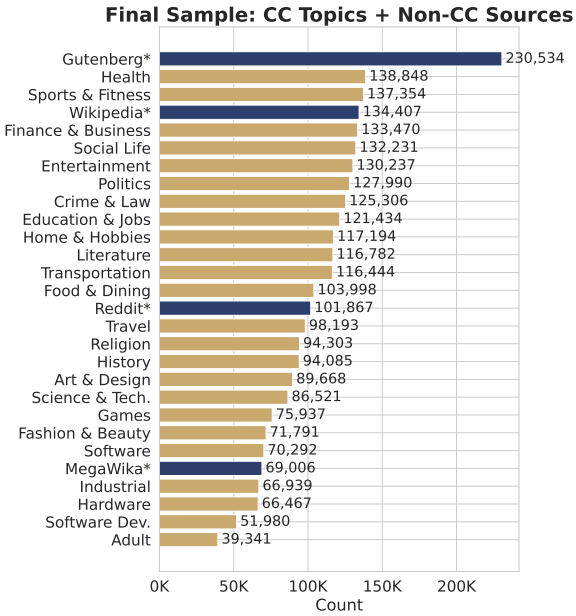

Figure A17: Distribution of Common Crawl topics and non-CC Dolma sources. Tan colors are the aggregated Common Crawl-related Dolma sources partitioned by topic from the WebOrganizer topic classifications.

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 A8: Dolma sources excluded from sampling.

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 A9: Full configuration for the Dolma sampling pipeline.

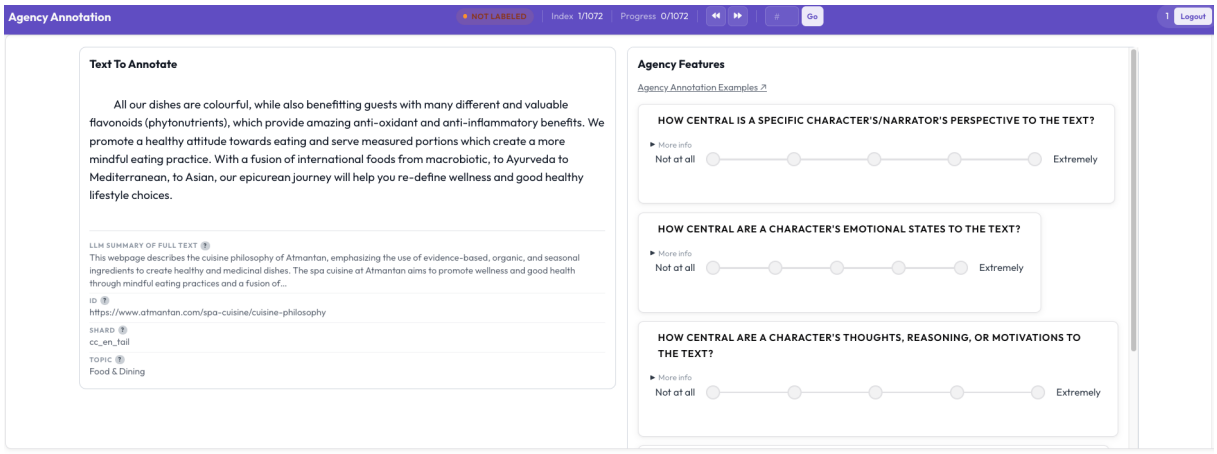

Figure A18: Screenshot of the agency annotation UI built upon the Potato software.

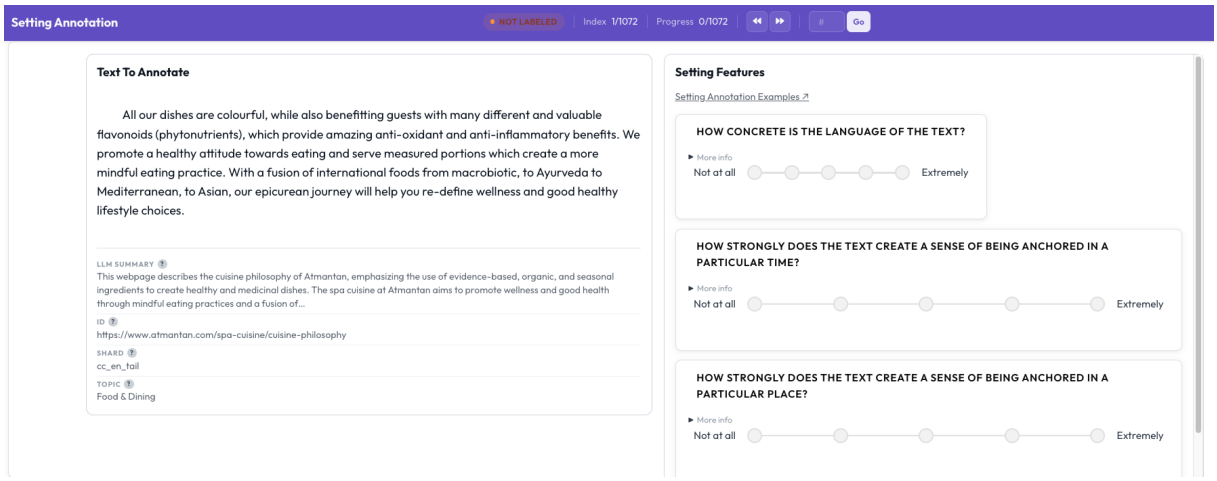

Figure A19: Screenshot of the setting annotation UI built upon the Potato software.

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 A10: 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 A11: Internal emotion rating scale.

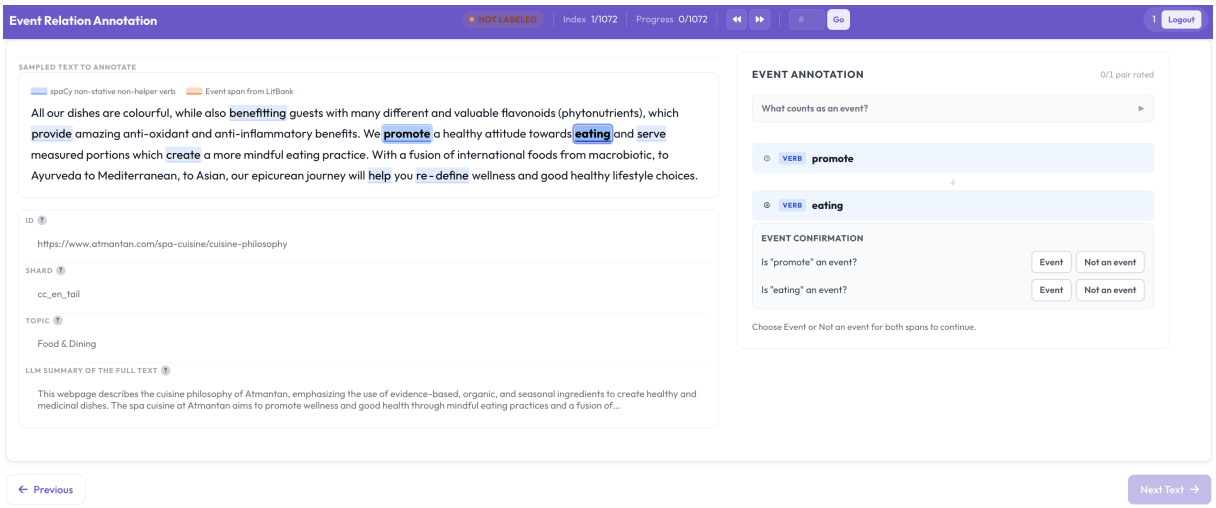

Figure A20: Screenshot of the event relation annotation UI built upon the Potato software.

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 A12: 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 A13: 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 A14: Conflict rating scale.

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 A21: 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 A22: Annotation prompt for setting.

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 A15: 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 A16: 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 A17: 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 A18: 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 A23: Annotation prompt for event relations.