FICSIM: A Dataset for Multi-Faceted Semantic Similarity in Long-Form Fiction

Natasha Johnson[1] **Amanda Bertsch**[1] **Maria-Emil Deal**[2] **Emma Strubell**[1]
[1] Language Technologies Institute, Carnegie Mellon University
[2] School of Library and Information Studies, University of Oklahoma
nmj@alumni.stanford.edu abertsch@cs.cmu.edu

Abstract

As language models become capable of processing increasingly long and complex texts, there has been growing interest in their application within computational literary studies. However, evaluating the usefulness of these models for such tasks remains challenging due to the cost of fine-grained annotation for long-form texts and the data contamination concerns inherent in using public-domain literature. Current embedding similarity datasets are not suitable for evaluating literary-domain tasks because of a focus on coarse-grained similarity and primarily on very short text. We assemble and release FICSIM, a dataset, of long-form, recently written fiction, including scores along 12 axes of similarity informed by author-produced metadata and validated by digital humanities scholars. We evaluate a suite of embedding models on this task, demonstrating a tendency across models to focus on surface-level features over semantic categories that would be useful for computational literary studies tasks. Throughout our data-collection process, we prioritize author agency and rely on continual, informed author consent.[1]

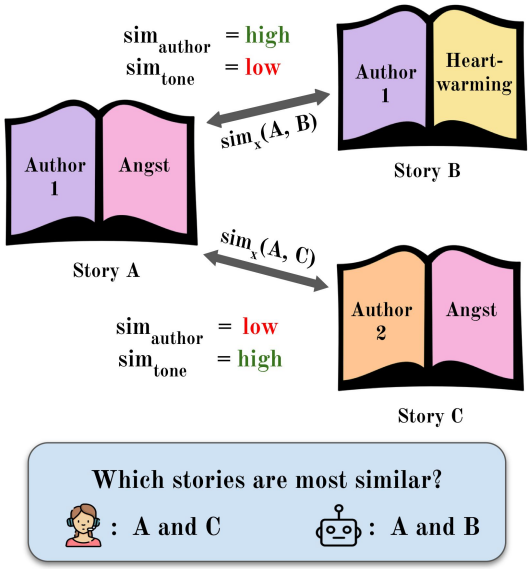

Figure 1: Similarity between literary texts can be defined along many dimensions. Computational literary studies scholars generally seek to measure specific, semantic types of similarity such as similarity in tone, but embedding models over-index on more obvious features such as the author's writing style.

1 Introduction

The last few years have been a time of immense progress in long-context processing in NLP. Several language models now support context lengths in excess of a million tokens. Embedding models for 32k context inputs abound. While challenges remain in long context modeling, successive approaches have made strong progress on benchmarks and have found applications downstream (Kapoor et al., 2024; Godbole et al., 2024; Nie et al., 2024).

In parallel, there is interest in applying NLP methods within digital humanities (DH), particularly within computational literary studies. Many DH scholars have incorporated NLP methods such as topic modeling, sentiment analysis, and semantic textual similarity (STS) tasks such as clustering and measuring similarity into their research (Kleymann et al., 2022; Algee-Hewitt and Fredner, 2023).

Yet evaluations of these approaches have been limited, particularly with regards to STS tasks. Many NLP models which have been (or have the potential to be) applied in DH research are not evaluated on literary applications. When they are, it is often on digital texts made accessible through public repositories such as Project Gutenberg (Kohlmeyer et al., 2021; Kryściński

[1]Dataset can be accessed at https://huggingface.co/datasets/ficsim/ficsim. Additional documentation can be found at https://github.com/natashamariejohnson330/FicSim

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25228–25246
November 4-9, 2025

Source	Long	Publicly available	Not publicly analyzed	Author labeled	Multiple axes of similarity
Recent novels	✓	✗	✓	✗	✗
Project Gutenberg	✓	✓	✗	✗	✓
MTEB STS tasks	✗	✓	✓	✗	✗
AO3 Fanfiction	✓	✓	✓	✓	✓

Table 1: Comparison of data sources for semantic textual similarity tasks. Fanfiction represents our approach.

et al., 2019; Underwood et al., 2018; Bamman et al., 2024; Kočiský et al., 2018; Xu et al., 2022). However, these texts—alongside related data and analyses from other commonly-scraped sites like Wikipedia—are included in the pretraining data of most models (Elazar et al., 2024), which could cause direct or indirect contamination and skew evaluation results (Palavalli et al., 2024; Zhang et al., 2024b).

Furthermore, while existing literary datasets evaluate model suitability for tasks such as summarization (Kryściński et al., 2019), question answering (Kočiský et al., 2018; Xu et al., 2022), and identifying literary co-reference (Bamman et al., 2020), DH scholars have expressed the need for embedding methods that capture semantic textual similarity within novel-length texts along several axes such as plot, tone, and setting (Sobchuk and Šeļa, 2024).

In response to this gap, we present FICSIM, an evaluation dataset for fine-grained semantic textual similarity (STS), constructed of long-form human-written narratives that are unlikely to appear in pretraining data, are accompanied by author-labeled metadata, and are included in this dataset *with author consent*. We describe our processes for selecting text not included in CommonCrawl scrapes, for obtaining and maintaining author consent for the use of their works, and for constructing pairwise similarity measurements corresponding to 12 different facets of fictional texts, in consultation with both literary scholars and authors (§3). We then describe the resulting dataset and its use (§3.3). We evaluate existing approaches on this multi-faceted STS task (§4). Models struggle to capture salient characteristics of long-form texts, only weakly disambiguate between categories, and over-index on surface features of the text (§5). We conclude by discussing the relevance of our results for both literary studies scholars and NLP researchers (§7). We hope that FICSIM enables more focus on narrowing the gap between models' general capabilities and their applicability to literary domain tasks.

2 Sourcing Data

2.1 Desiderata

To effectively measure similarity in long-form fiction, we need a corpus of stories with several characteristics, summarized in Table 1: The stories must be (1) *long*, coherent narratives (2) *publicly available* for evaluation (3) *not publicly analyzed online*, in order to prevent potential contamination (4) *well-annotated*, ideally by an expert, capturing *multiple axes of similarity*, beyond superficial similarities that are easy to detect but limited in the value for literary scholars.

Recently published novels offer one compelling solution (Karpinska et al., 2024; Duarte et al., 2024) but limit public release of the dataset.[2] Furthermore, the suitability of such novels for evaluation purposes decays over time, as summaries and analyses of the texts become increasingly likely to have been incorporated into model training. Public domain literature, such as the texts made available through Project Gutenberg, satisfies the length and public availability requirements, but is deeply present both in pretraining corpora and in public culture—with many analyses online of the themes, plot, and character arcs of each story, it is unclear whether a model identifying these characteristics is doing so through memorization (Palavalli et al., 2024).

2.2 Our Approach

In response, we turn towards *fanfiction*–fictional texts inspired by existing media, often sharing the characters or setting of the source work. Many fanfiction texts are complex long-form narratives, reflecting both the fandom subculture and major cultural movements of the time. In recent decades, fan studies has become an active subfield within literary, media, and cultural studies.

Significantly for our purposes, popular fanfiction websites allow users to assign tags to their fanfiction (e.g. Figure 2) to help with fanfic discoverability and categorization. Tags range from purely descriptive to analytical to conversational and identify important elements of a fanfic from the author's perspective. They are intended to help readers find the content that is interesting to them among millions of stories. We use such tags as the

[2]Copyright laws do permit physical copies of books to be purchased, scanned, and shared with certain provisions. However, digitizing physical texts can be a labor and cost intensive.

basis for our gold-standard similarity scores.

2.3 Story Selection

Archive of Our Own. We source our fanfiction from Archive of Our Own (AO3), a digital repository hosting over 15M works. We selected this venue for two main reasons: First, AO3 has made significant efforts to discourage web scraping, including blocking Common Crawl scraping in 2022 (Works, 2023) and implementing aggressive rate limiting policies.[3] Second, due to the site's construction and norms, many AO3 stories feature highly detailed tagging (e.g. Figure 2), which we leverage to compute similarity along various axes.

Requirements for stories. We only consider stories that are written in English, exceed 10,000 words, and feature detailed tagging. Furthermore, because AO3 restricted web crawling in December 2022, we only consider stories that were started and completed after this date. Focusing on texts over 10,000 words ensures that the stories in our dataset are similar in length to fictional texts commonly studied by literary scholars, including short stories, novelettes, novellas, and novels (Gioia and Gwynn, 2006; Harmon, 2003). We do not screen for or remove explicit content, although the majority of the stories in the dataset are not marked as explicit.[4] Additionally, when identifying stories that we hoped to include in our dataset, we looked for a variety of tropes, settings, and writing styles.

2.4 Author Consent

Many artistic communities have strong negative feelings towards the machine learning community. Writers have described the AI community as exhibiting "complete lack of respect" for artistic work (Gero et al., 2025). One survey found that 96% of authors were against the use of their work for AI training without their explicit consent (The Authors Guild, 2023). Fanfiction authors are no exception: fan communities have expressed dismay over learning that some fanfiction has been included in CommonCrawl datasets, and Archive of Our Own lawyers went before the U.S. Copyright Office to argue that fan authors should be able to opt-out their work from model pretraining corpora (Archive of Our Own, 2023).

Therefore, although Archive of Our Own explicitly permits metadata collection done by academic researchers (Works, 2023), we additionally received each author's permission to include their work(s) in our dataset using an IRB-approved process. We reached out to each author individually through AO3 to explain our project and share guarantees about author privacy, story usage, and withdrawal of consent. Then, we invited authors to sign a consent form and provided mechanisms to ask questions or withdraw consent at any time. Furthermore, in consultation with several members of fan communities, we decided to commit to not using FICSIM to train models; to access the dataset, we will require other researchers to agree to the same terms of usage. For more information on the consent process and the full text of the outreach documents, see Appendix C.

2.5 Fanfiction versus Traditional Literature

Over the past decade in particular, the line between fanfiction and published literature has been blurred in terms of both story content and writing style. Many stories originally written as fanfiction have been published as standalone books (with minimal edits, such as changing the characters' names) and have seen commercial success (Arzbaecher, 2023). And many tropes created or popularized by fan communities have been adopted outside of fan spaces by authors and consumers of contemporary genre fiction (Majnaric, 2024; Jerasa and Boffone, 2021). Furthermore, when performing stylistic comparison between fanfiction stories and their inspiration texts (e.g. the Harry Potter novels), researchers have found that fanfiction stories do not stylistically diverge from their source material in statistically significant ways (Jacobsen et al., 2024). Thus, we believe fanfiction texts are suitable for evaluating model performance on literature, particularly contemporary fiction.

2.6 Use of Fanfiction Tags

We derive gold-standard annotations for document similarity from the tags each author has ascribed to their story, which may be either canonical [5] or user-authored.

Our use of user-generated tags as the basis for our similarity calculations was informed by the

[3] While these do not *guarantee* that no fanfics posted after this date are used in pretraining corpora, these restrictions, along with AO3's lack of an official API, make these texts relatively unlikely to appear in pretraining web scrapes.

[4] Explicit texts (which may be tagged as such because of violent or sexual content) are an active area of research in media studies (e.g. Popova (2018); Fazekas (2022)).

[5] That is, standardized tags that are internally linked to synonymous and related tags by AO3's tag-wrangling team.

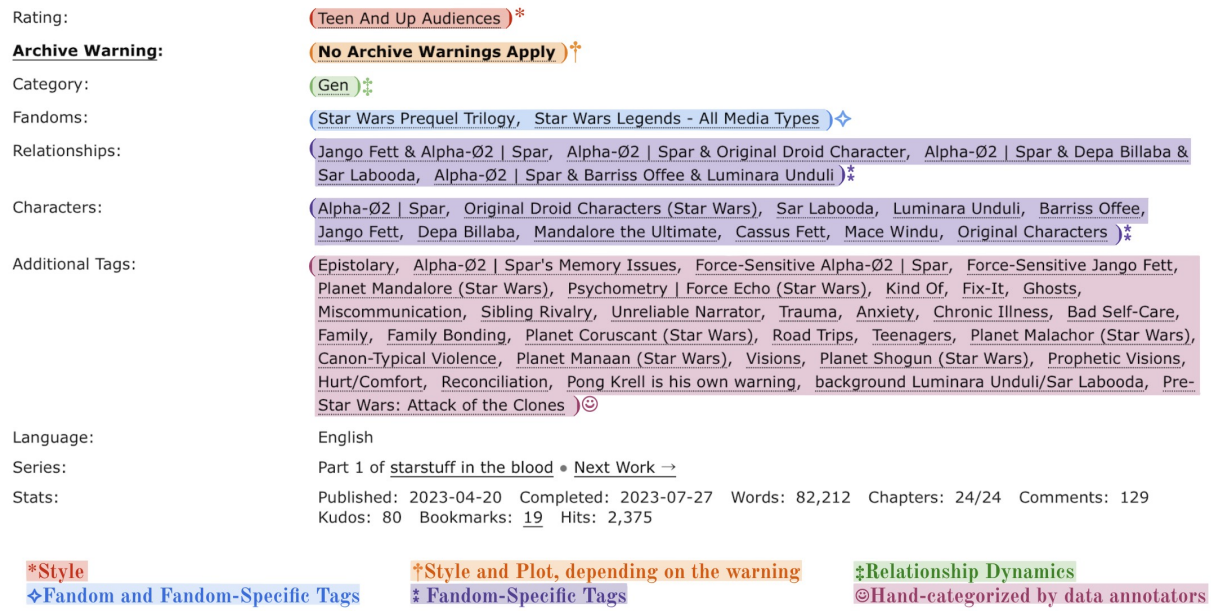

Figure 2: Example of fic tagging and metadata on AO3. Colored annotations mark data that inform similarity scores.

work of Lyons and Tappeiner (2008), who explored using user-generated tags to expand access to library resources. They discuss how user-based tagging offers a form of natural language keyword categorization that can help capture particular narrative features which are not standard subject headings for narrative works. In this way, user-generated tags do a better job of helping guests find materials that relate to their interests. This is essentially the purpose that tags serve on AO3—its tagging system, handling of user-generated tags, and user norms have allowed readers to find the content that is interesting to them among millions of other stories.

Though AO3 tags might seem like a limited source from which to construct similarity scores, within DH work, genre labels are often used as ground-truth labels for evaluating literary embedding, topic modeling, and clustering methods (Sobchuk and Šeļa, 2024; Schöch, 2017; Allison et al., 2012). Within fan spaces, fanfiction tags serve many of the same purposes and are held in the same common understanding (Hellekson and Busse, 2014) as genre labels. In fact, many tags created or popularized by fan communities such as "Enemies to Lovers" and "Slow burn" have notably been adopted outside of fan spaces by authors and consumers of contemporary genre fiction (Majnaric, 2024). These tags are commonly used in advertisements, recommendations, and reviews of books (Jerasa and Boffone, 2021). Goodreads has even added "Enemies to Lovers" as a book genre on their site (goodreads).

Thus, our process calculates story similarity according to a well-established framework of what is important in a story, as developed and refined by fanfiction authors and readers, and as adopted by many traditional authors, publishers, and readers.

3 Constructing FicSim

3.1 Tag Categorization

We place tags into 12 different categories, corresponding with various types of similarity we hope to measure. Some categories, such as "Plot," "Theme," and "Time" capture general qualities of fictional literature and align with projects that discuss narrative similarity on the basis of actions, subjects, themes, and temporal setting (Algee-Hewitt and Fredner, 2023; Kleymann et al., 2022; Sobchuk and Šeļa, 2024; Piper, 2022). Others, such as "Fanfiction Tone and Content Tags" capture fanfiction-specific qualities and serve to identify whether models can identify similarity based on genre-specific conventions. To determine the categories, we relied on the aforementioned DH scholarship, as well as the annotators' analysis of the tag set. When applicable, we place tags into multiple categories. Table 2 describes the categories with tag examples from each.

Tag cleaning. In order to increase tag interpretability and allow for clearer tag comparison, we clean and standardize the tags prior to constructing

Category	Description	Example Tags
Plot	Narrative actions and concrete subjects; describes what happens in a story and what is *in* a story.	Letters Blood and Torture Artificial Intelligence
Character States	The emotions, attributes, roles, and physical characteristics of characters in the text.	Trans Woman Character Is Bad at Feelings Dissociative Identity Disorder
Relationship Dynamics	Key characteristics of both platonic and romantic relationships in the text.	Established Relationship Possibly slowest ever burn F/F
Theme	Abstract ideas explored throughout the story.	Racism Found Family What Is The Impact Of A Mother
Time	Temporal setting.	Alternate Universe - 19th Century Modern Retelling Post-Apocalypse
Style	Features of the writing style or narrative technique.	POV Third Person Omniscient Epistolary Dialogue Heavy
Fanfiction Tone and Content Tags	Fan-community language for the type of story; often relates to both tone and plot (e.g. "fluff" generally involves lighthearted and domestic scenes).	Angst Fluff and Hurt/Comfort Tooth-Rotting Fluff
Fandom-specific	Tags that involve settings, characters, or events from a canon text, and thus reveal information about the fandom the story belongs to.	Phantom of the Opera AU Yule Ball (Harry Potter) Capitano/Mavuika
Overall (Fandom-Agnostic)	An aggregate grouping of tag categories 1-7; captures similarity between all tags that do not reveal the fandom or author identity.	[any of the above tags]
Overall (Fandom-Specific)	An aggregate grouping of tag categories 1-8; captures similarity between all tags	[any of the above tags]
Fandom	Captures whether two texts are inspired by the same piece(s) of media (e.g. books, movies, television shows).	Genshin Impact Grey's Anatomy Star Wars
Author	Captures whether two texts are written by the same fan author.	[author IDs]

Table 2: Types of similarity and example tags that align with each category. Some of the above tags were placed into multiple categories within our dataset, but they nonetheless serve as useful references for the particular category they represent above.

similarity scores. The cleaning process involves removing unnecessary punctuation, standardizing capitalization, and correcting obvious spelling errors. Our tag standardization involves replacing tags with their canonical counterparts and rephrasing or restructuring tags to increase their semantic interpretability.

Fandom-agnostic tags. In order to allow our dataset to evaluate model suitability for applications outside of fan studies, we want all tag categories except Fandom, Fandom Specific, and Overall (fandom-specific) to contain only fandom-agnostic terms which could be used to compare story similarity across fandoms. For the nearly 50% of the tags that contained fandom-specific references, we create a fandom-agnostic version of the tag by removing or replacing fandom-specific references (e.g. "Protective Cristina Yang" becomes "protective character"). We then place this fandom-agnostic tag into the appropriate fandom-agnostic category(/ies) while also placing the original tag into the Fandom and Fandom Specific category.

Annotation process. Our tag categorization and rewriting was performed by two authors who are experts in fan studies, one with a background in computational literary studies and the other with a background in library and information science. Following the qualitative methods tradition of interpretative analytical process (Lincoln et al., 2011; Strauss and Corbin, 1990), the annotating authors arrived at annotations through a series of consensus-building discussions. In a few instances where tags contained fandom-specific terminology for a particular fandom that neither annotator had specific

expertise in, the annotators reached out to fanfiction authors within that fandom to confirm their interpretation of the tag prior to categorization. In these cases, the authors were acquaintances that each annotator knew through their own engagement with fan communities.

3.2 Similarity Score Calculation

After tags are cleaned and categorized, we calculate category-specific similarity scores between stories. We embed each tag using Gemini Embedding (Lee et al., 2025). Then, given category-specific tag lists from two fanfics, we calculate the similarity between tag lists A and B as the average of the pairwise tag cosine similarities:

$$\text{sim}(A, B) = \text{avg}\,\text{sim}_{a \in A, b \in B}(a, b)$$

Gemini embedding. We selected Gemini Embedding as our embedding model for the following reasons: (1) It has the overall highest score on the MTEB leaderboard (Muennighoff et al., 2023); (2) Among the best-performing models on the MTEB English-language STS tasks[6], Gemini Embedding is the highest ranking model that is not from the same model family (or built upon the same model family) as the models we evaluate below. Furthermore, because Gemini Embedding has an input limit of 2048 tokens, selecting it as our tag embedding model does not then preclude us from including it in our evaluation section.

Gold score validation. We validate our tag handling and similarity score construction in two steps. Prior to tag cleaning, one author annotated a set 330 story triplets (thirty 3-way story comparisons in each of the 11 non-author categories), identifying whether the first story in a triplet was more similar to the second or third story from that triplet in a given category.[7] Tag cleaning, standardization, and embedding processes were then adjusted to align the resulting gold similarity scores with the author's annotations. Then, to further validate the scores, two other authors each annotated identical sets of 220 story triplets (20 stories x 11 categories) on the same task. Since the two annotators were not both experts in fan studies or literature, they were allowed to skip comparisons for which they did not think they could identify the more similar story. Of the 158 triplets that neither annotator skipped we measured annotator agreement of 82% (Cohen's $\kappa = 0.65$). Our gold truth labels were then evaluated against the 129 triplets for which both authors provided the same rating, demonstrating 80% alignment with the annotations.

3.3 FICSIM Statistics

Our final dataset includes 90 stories[8] and gold similarity scores along 12 axes, for a total of 33,790 pairwise comparisons. The stories range from 10,001 to 488,772 words and span 46 fan communities (fandoms). The cleaned tagset has 9448 total tags (2133 unique) across 12 categories. For more detailed dataset statistics and license information, see Appendices A and D.

3.4 Similarity scores

For every story pair, FICSIM presents up to 12 similarity scores, described in detail in Table 2. We divide these scores into three evaluation groups:

Fine-grained similarity: Plot, character states, relationship dynamics, theme, time, style, fanfiction tone & content tags.

Broader notions of similarity: Overall (fandom-agnostic and fandom-specific). The overall (fandom-specific) category captures a mix of both superficial and more integrated narrative elements, while the overall (fandom-agnostic) category primarily targets the latter.

Superficial similarity: Fandom-specific tags, fandom, and author. Because these are always provided (even when a story is not otherwise tagged well) and generally obvious to deduce from the text, we do not consider these axes of similarity to evaluate embedding quality, only to measure how embeddings capture these in contrast to the more fine-grained features above.

4 Evaluation

For each model, we compute cosine similarity between embeddings for all story pairs. Following Muennighoff et al. (2023), we measure Spearman's ρ between the model-induced ranking of story similarities and our tagset-derived gold ranking. We report ρ out of 100 instead of 1.0 for readability, and highlight values that are significant ($p < 0.05$).

[6] As of April 2025.

[7] Story triplets were drawn from a randomly-generated set of 100 story triplets in each category. The author identified the first 30 non-ambiguous comparisons in each category, skipping triplets where the author thought an argument could be made for either ranking.

[8] Literary datasets with expert annotations are often comparable in magnitude (Sims et al., 2019; Bamman et al., 2020).

Model	Plot	Character States	Relationship Dynamics	Theme	Time	Style	Fanfiction Tone & Content	All (Fandom Agnostic)	All (Fandom-Specific)	Fandom Specific	Fandom	Author
Linq-Embed	**18.65**	8.63	5.52	0.14	28.85	3.33	9.95	16.59	34.61	34.99	30.47	**40.91**
+ SW	10.17	5.69	9.12	6.87	21.51	7.03	6.09	7.37	15.17	18.02	29.25	40.73
GTE-Qwen2	15.84	4.36	9.62	**13.92**	31.23	13.43	**20.87**	**24.38**	40.50	45.48	45.87	40.82
+ SW	13.00	-1.01	10.74	7.52	24.84	12.96	1.74	11.02	20.49	20.65	26.73	40.73
SFR-Embed	17.06	0.98	5.83	4.96	24.34	-5.06	8.45	18.07	40.65	36.62	34.35	38.76
+ SW	10.42	7.14	10.73	7.98	22.66	9.28	8.05	10.47	17.88	19.29	31.49	40.64
GTE-ModernBERT	13.49	**11.87**	15.39	0.17	-0.30	8.20	13.28	16.77	42.70	48.50	**50.08**	36.33
+ SW	18.07	4.68	**23.82**	6.71	**40.24**	16.43	5.96	22.13	41.37	42.03	46.57	37.60
m2-BERT-32k	5.95	8.06	17.08	2.66	-0.44	-1.66	14.30	9.99	14.11	13.49	14.40	17.47
+ SW	8.96	7.77	13.54	6.35	0.90	5.00	19.65	12.81	20.68	20.02	22.87	20.18
Voyage-3-large	9.87	4.60	9.93	10.81	27.62	**22.42**	13.71	20.01	**43.66**	42.21	31.72	37.92
+ SW	10.80	4.25	8.29	11.01	22.13	19.32	9.22	19.81	43.15	42.06	35.99	37.72
Claude+SW	-1.19	-0.74	5.94	11.64	27.54	17.05	-0.43	8.99	39.27	**49.41**	44.95	40.19

Table 3: Spearman correlation of embedding cosine similarity to our tagset similarity measures, for several representative open-source and API-based embedding models. All models struggle at category similarity and overindex on authorial style. +SW denotes the use of a sliding window; Statistical significance highlighted.

Models. We select the 3 best-performing open-weights models on the Hugging Face (Wolf et al., 2020) MTEB leaderboard with 32k context lengths: Linq-Embed-Mistral (Choi et al., 2024), GTE-Qwen2-7B-instruct (Li et al., 2023), and SFR-Embed-Mistral (Meng et al., 2024). These are all 7B models; however, computational literary studies scholars often have access to limited computational resources. Thus, we consider a much smaller model– GTE-ModernBERT-base (Zhang et al., 2024a)– and two API-based solutions: m2-BERT-80M-32k-retrieval (Fu et al., 2023) through the TogetherAI API and using Voyage-3-large (Voyage AI, 2025) through the Voyage.ai API. Finally, we consider whether large language models could perform this task. We use Claude-3.7-Sonnet (Anthropic, 2025) to summarize each story, then embed the much shorter summary documents with Voyage.[9] We follow each model's default strategy for constructing embeddings; for additional details on models and pooling methods, see Appendix B.

Prompt. Linq-Embed, SFR-Embedding, and both GTE models support providing an instruction in a special format at the beginning of a text to be embedded. We experiment with using this to produce category-specific embeddings, by providing instructions to focus on each type of similarity in turn. We provide the prompt (see Appendix B) to each embedding model at the start of the document.[10] For Claude, we use a modified form of category-specific prompt for summarizing and omit the prompt for the embedding stage.

Sliding window. There are stories in FICSIM that exceed the context length of every model evaluated. When stories exceed the maximum context length, we consider two options. As a baseline, we naively truncate each story to the maximum length.We then consider a sliding-window approach where we chunk the text to windows of the maximum context length and then pool embeddings across windows (Wang et al., 2019). For the sliding window approach, we overlap windows by 2048 characters (depending on the model and story, approximately 500 words).

5 Results: Can existing models perform fine-grained literary STS?

We present results using a single general-purpose embedding from each model in Table 3 and results on category-specific embeddings for each model in Table 5. All models struggle to perform fine-grained STS; unexpectedly, the larger models are not consistently better than the small embedding models. We discuss overall trends below.

Truncation and sliding windows show marginal differences. Surprisingly, naively truncating to the first segment of the story and taking the mean of sliding window embeddings are similarly success-

[9] Anthropic does not have its own embedding model; we use Voyage on the Claude outputs because this is the embedding model Anthropic recommends in their documentation.

[10] For the models that do not support a custom instruction, we simply prepend the prompt without special formatting.

Gold score category	Overall (Fandom Specific)	Fandom Specific	Fandom	Author
Plot	31.85	7.56	8.43	12.14
Character States	4.52	0.09	15.10	13.55
Relationship Dynamics	36.11	10.58	17.25	17.53
Theme	10.30	1.95	3.98	5.10
Time	12.53	14.36	**34.65**	**21.28**
Style	24.18	8.90	11.21	13.05
Fanfic Tone & Content	00.00	-3.61	-1.05	5.19
Overall (Fandom-Agnostic)	**48.01**	**14.04**	18.90	19.20

Table 4: Gold category-specific scores range in correlation with the overall (fandom-specific), fandom-specific, fandom, and author categories. The comparatively low scores in the latter three categories indicate that the gold scores weigh surface-level features far less heavily than our embedding models. Statistical significance highlighted

ful across the board. Though the mean Spearman correlation for truncation is slightly higher than the mean sliding window correlation (11.70 vs 11.17), when comparing individual correlations for a given model and category, truncation beats sliding window 51% of the time. [11] Despite allowing a model to consider more of the text, sliding window embeddings may fail to increase model performance because models are not generally trained for this approach; because the pooling strategy needs to be adjusted when pooling across more data; or simply because some of the features that capture story similarity can be extracted from the start of the text alone. That being said, while sliding windows do not improve correlation scores in any of the fandom-agnostic categories, they do decrease correlations in the confounding categories (mean ρ of 30.47 as opposed to 35.69), suggesting that their embeddings weigh surface-level features less heavily.

Models overindex on surface features. Across all models, Spearman's ρ is higher for the four "confounder" categories than any of the fine-grained similarity categories. Notably, the author category (which is computed solely from exact match of author IDs) has the highest score more frequently than any other category. This indicates that embeddings are much more sensitive to author-specific stylistic factors than to the fine-grained semantic factors captured in the remaining categories.

Some sensitivity to author and fandom is expected—some authors will focus on different types of stories, and some conventions or styles will be more common in one fandom than another. The overall (fandom-specific) category takes this into account by considering similarity based on both fandom-specific and non-fandom tags. However, 77% of the models in Table 3 score higher in the fandom-specific, fandom, or author category than in the overall (fandom-specific). This indicates that they are not only capturing the fingerprints of certain authors and fandoms, but that they are furthermore failing to capture other narrative elements. In contrast, Table 4 shows how the gold scores correlate with fandom and author-based categories. Many of the gold scores are positively correlated with the overall (fandom-specific) category, reflecting the fact that this category incorporates all other narrative features into its ranking. However, the majority are not as strongly correlated with the fandom-specific, fandom, and author scores, indicating that the information they capture extends beyond these categories.

While capturing author and fandom information is not inherently harmful, the outsized impact of these (trivially computable from metadata) features on embedding-based similarity scores limits their applicability to analysis looking for more subtle phenomena like theme or trope similarity.

Category-specific embeddings show minimal impact. Table 5 shows the performance with category-specific instructions across each model and category. When comparing correlations for a given model and context-handling method, category-specific embeddings outperform non-specific embeddings exactly 50% of the time.

6 Related Work

Long-context and embedding evaluation. A number of datasets for long-context evaluation have included literary texts. BookSum (Kryściński et al., 2019) involves summarization over public domain books. LongBench, LongBenchv2, and HELMET (Bai et al., 2023, 2024; Yen et al., 2025) include question answering over NarrativeQA (Kočiský et al., 2018); (Zhang et al., 2024c) introduces summarization and QA tasks over a set of novels with entity names replaced to reduce the impact of potential contamination. Embedding-focused datasets include STS tasks but focus primarily on very short

[11] Note that we do not compare between sliding window and truncation for the Claude results because only four stories exceed Claude's 200k context window, so there is not enough data to make a meaningful comparison between the methods.

Model	Plot	Character States	Relationship Dynamics	Theme	Time	Style	Tone & Content	Overall (Fandom-Agnostic)
Linq-Embed	8.36	6.00	4.94	4.38	24.96	6.63	9.90	16.59
+ SW	10.30	5.90	9.36	7.03	21.67	7.30	6.26	7.37
GTE-Qwen2	15.44	5.36	9.43	14.43	33.19	14.01	**20.35**	**24.38**
+ SW	13.06	-1.30	10.41	7.27	24.75	12.78	1.86	11.02
SFR-Embed	**18.82**	-0.18	5.79	2.59	24.30	-5.28	7.52	18.07
+ SW	10.46	7.33	10.91	8.12	22.50	9.62	8.22	10.47
GTE-ModernBERT	11.64	**12.78**	12.98	0.74	5.07	8.82	16.20	16.77
+ SW	18.32	6.24	**23.59**	7.72	**38.21**	15.65	6.48	22.13
m2-BERT-32k	5.82	7.88	16.62	2.88	-2.14	-2.54	12.98	9.99
+ SW	8.82	7.11	13.07	6.58	2.28	4.61	18.17	12.81
Voyage-3-large	14.96	7.66	16.93	11.53	21.84	**21.91**	15.79	20.01
+ SW	13.99	7.04	12.51	**15.24**	23.54	19.04	11.61	19.81
Claude+SW	6.23	8.16	7.33	13.07	21.77	5.71	9.94	8.99

Table 5: When using category-specific instructions, rank-correlation does not show notable improvement and is still quite poor on average. Statistically significant results are highlighted.

inputs (Muennighoff et al., 2023); LongEmbed (Zhu et al., 2024), which evaluates long-context embedding but not on STS, instead using QA tasks over NarrativeQA and SummScreen screenplays (Chen et al., 2022). None of these benchmarks measure performance on long-context STS tasks, which are of particular interest to digital humanities and literary scholars (Sobchuk and Šeļa, 2024).

NLP tools for literary studies. A number of works have studied the applicability of NLP methods to digital humanities tasks on public-domain literature. Bamman et al. (2024) compare LLMs to traditional supervised methods on a wide range of tasks within literary studies. Other works propose novel computational approaches to analyze elements of fictional texts that are of interest to literary scholars, such as character mobility (Wilkens et al., 2024), emotional arc (Öhman et al., 2022), and narrative pacing (Bamman et al., 2014). Kohlmeyer et al. (2021) propose *lib2vec*, a method for representing facets of fictional texts using multiple embeddings; because our similarity categories differ, direct application of their method to FICSIM is challenging. The (in)applicability of NLP systems to downstream uses has also been studied in other domains, including law (Kapoor et al., 2024) and materials science (Gururaja et al., 2025).

7 Conclusion

We present *FicSim*, a dataset of stories and similarity labels for benchmarking model performance on long-context STS tasks within fictional texts. Using FICSIM, we show that there is no single model that performs well across all types of similarity—and there are types of similarity for which no model performs well. In corpora with strong superficial similarities, like author or fandom overlap, embeddings may capture this information at the expense of other types of similarity. For this specific type of task, bigger (or more expensive) models are not uniformly better than their smaller, cheaper alternatives. Our evaluation of sliding window attention and category-specific embeddings also demonstrates that sensible modifications to the model to adapt to long-form or literary texts have a minimal impact on performance. We call for the careful evaluation of models on the particular task they are applied to, with annotation or validation by subject-matter experts.

The poor performance of otherwise strong models on FICSIM highlights the substantial gap that exists between current models and their utility for literary applications. Our models fail to capture fine-grained literary similarity and overindex on superficial features of the text in their embeddings. We expect that clever system design or additional domain-specific training could improve performance within this generation of embedding models, and we encourage the evaluation on literary tasks for future embedding model releases.

We hope FICSIM will help digital humanities researchers make informed decisions about model selection for tasks relating to story similarity, encourage more evaluation of embedding models on DH tasks, and serve as an example of how creative works can be used for academic research without circumventing creators' rights and wishes.

Limitations

Data Collection While we were originally informed by an AO3 support member that leaving comments on fanfics would be an appropriate method for soliciting consent, our account was later temporarily suspended on the basis that we were leaving spam messages. (We had used the same introductory message to reach out to each author, in alignment with IRB protocols.) Our attempt to appeal the suspension was unsuccessful, despite our explanation that we were following instructions we had been given by another AO3 team member. When the temporary suspension was lifted, we decided not to attempt further data collection, because we ultimately did not want to be using AO3 in a way that further increased tension between machine learning researchers and fanfiction writers.

Thus, while we were able to assemble a dataset using the methods outlined in this project, an exact replication of our process would not be appropriate.

Embedding methods It is not possible to consider every possible means of constructing embeddings; while we aimed to capture a set of models and methods that were representative of those applied in digital humanities works with embeddings, it is possible that there exist other methods that would outperform those presented as baselines here. In particular, computing similarities using multiple-embedding strategies is likely to improve performance. We leave devising better embedding strategies for literary domain text to future work.

Language We consider only stories written in English because of our need to reach out to each author individually. While we believe the fanfictions within FICSIM represent an interesting selection of works across these similarity dimensions, differences exist between literary corpora. It is possible that models that excel at similarity on FICSIM would nevertheless struggle on STS tasks for 19th century English literature, short-form satirical poetry from social media, or any other number of specialized literary domains. We see FICSIM as an initial step towards improved literary-domain evaluation.

Acknowledgments

We extend our deepest gratitude towards the fanfiction authors who have given us permission to include their works in our dataset and without whom this project would not have been possible.

We would also like to thank David Mimno, Maarten Sap, Chap Morack, Suguru Ishizaki, and our reviewers for their helpful feedback on our work.

AB was supported by a grant from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE2140739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

Mark Algee-Hewitt and Erik Fredner. 2023. Typicality.

Sarah Allison, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. 2012. Quantitative formalism: An experiment. *N+1*, 13:81–108.

Anthropic. 2025. Claude 3.7 sonnet and claude code. Accessed: 2025-05-19.

Archive of Our Own. 2023. A statement on ai and fanworks. Accessed: 2025-05-20.

Lauren Arzbaecher. 2023. 10 books you didn't know started out as fan fiction, from 'Twilight' to 'Star Wars'-inspired stories.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. LongBench: a bilingual, multitask benchmark for long context understanding.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. *arXiv (Cornell University)*.

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On Classification with Large Language Models in Cultural Analytics. *arXiv (Cornell University)*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian Mixed Effects Model of Literary Character. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy yong Sohn. 2024. Linq-embed-mistral technical report. *Preprint*, arXiv:2412.03223.

André Duarte, V, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. 2024. DE-COP: Detecting copyrighted content in language models training data. *arXiv (Cornell University)*.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? *International Conference on Learning Representations*.

Angela Fazekas. 2022. *Creative Becomings: Explicit Fanfiction, Reinventing Adolescence, and Queer Relationality*. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2024-02-16.

Daniel Y Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. 2023. Monarch mixer: A simple sub-quadratic gemm-based architecture. In *Advances in Neural Information Processing Systems*.

Katy Ilonka Gero, Meera Desai, Carly Schnitzler, Nayun Eom, Jack Cushman, and Elena L. Glassman. 2025. Creative writers' attitudes on writing as training data for large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–16. ACM.

Dana Gioia and R. S. Gwynn. 2006. *The art of the short story*. Longman Publishing Group.

Aditi Godbole, Jabin Geevarghese George, and Smita Shandilya. 2024. Leveraging long-context large language models for multi-document understanding and summarization in enterprise applications. *Preprint*, arXiv:2409.18454.

goodreads. Enemies to Lovers Books.

Sireesh Gururaja, Nupoor Gandhi, Jeremiah Milbauer, and Emma Strubell. 2025. Beyond text: Characterizing domain expert needs in document research. *Preprint*, arXiv:2504.12495.

William Harmon. 2003. *A handbook to literature*, 9 edition. Prentice Hall.

Karen Hellekson and Kristina Busse. 2014. *The Fan Fiction Studies Reader*. University of Iowa Press.

Mia Jacobsen, Ross Deans Kristensen-McLachlan, Denmark Center for Humanities Computing, Aarhus University, Cognitive Science Department of Linguistics, and Denmark Semiotics, Aarhus University. 2024. Admiration and Frustration: A Multidimensional Analysis of Fanfiction. *CHR 2024: Computational Humanities Research Conference*.

Sarah Jerasa and Trevor Boffone. 2021. BookTok 101: TikTok, Digital Literacies, and Out-of-School Reading Practices. *Journal of Adolescent Adult Literacy*, 65(3):219–226.

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. 2024. Promises and pitfalls of artificial intelligence for legal applications. *Preprint*, arXiv:2402.01656.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

Rabea Kleymann, Andreas Niekler, and Manuel Burghardt. 2022. Conceptual Forays: A corpus-based study of "Theory" in Digital Humanities Journals. *Journal of cultural analytics*, 7(4).

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Lasse Kohlmeyer, Tim Repke, and Ralf Krestel. 2021. Novel views on Novels: embedding multiple facets of long texts. *WI-IAT '21: Web Intelligence and Intelligent Agent Technology*, page 8.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2019. BookSum: A Collection of Datasets for Long-form Narrative Summarization. *arXiv.org*.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. 2025. Gemini Embedding: Generalizable Embeddings from Gemini. *arXiv.org*.

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

Yvonna S Lincoln, Susan A Lynham, and Egon G Guba. 2011. Paradigmatic controversies, contradictions, and emerging confluences, revisited. *The Sage handbook of qualitative research*, 4:97–128.

Catherine Lyons and Elisabeth Tappeiner. 2008. Cataloging 2.0: Metadata research and initiatives at a community college library. *Journal of Library Metadata*, 8(2):155–157.

Lucija Majnaric. 2024. Novel in the Time of the Internet: A Closer Look at the Fanfiction Phenomenon.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

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

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *Preprint*, arXiv:2406.11903.

Medha Palavalli, Amanda Bertsch, and Matthew R. Gormley. 2024. A taxonomy for data contamination in large language models.

Andrew Piper. 2022. The CONLIT Dataset of Contemporary Literature. *Journal of Open Humanities Data*, 8.

Milena Popova. 2018. *"Slight dub-con but they both wanted it hardcore": Erotic fanfiction as a form of cultural activism around sexual consent*. Ph.D. thesis.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Christof Schöch. 2017. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, 11(2).

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Oleg Sobchuk and Artjoms Šeļa. 2024. Computational thematics: comparing algorithms for clustering the genres of literary fiction. *Humanities and Social Sciences Communications*, 11(1).

Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research*. Sage publications.

The Authors Guild. 2023. New authors guild ai survey reveals that authors overwhelmingly want consent and compensation for use of their works. Accessed: 2025-05-20.

Ted Underwood, David Bamman, Sabrina Lee, Boris Capitanu, Hoyt Long, Richard Jean So, Teddy Roland, Laura Mandell, Allen Riddell, Andrew Piper, and Stephen J. Downie. 2018. The transformation of Gender in English-Language fiction. *Cultural Analytics*.

Voyage AI. 2025. voyage-3-large: the new state-of-the-art general-purpose embedding model. Accessed: 2025-05-19.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China. Association for Computational Linguistics.

Matthew Wilkens, Elizabeth F. Evans, Soni Sandeep, David Bamman, and Andrew Piper. 2024. Small Worlds. Measuring the Mobility of Characters in English-Language Fiction. *Journal of Computational Literary Studies [preprint]*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Organization For Transformative Works. 2023. AI and Data Scraping on the Archive.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. Helmet: How to evaluate long-context language models effectively and thoroughly. In *International Conference on Learning Representations (ICLR)*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024a. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Xinhao Zhang, Olga Seminck, Pascal Amsili, and Lattice. 2024b. Remember to Forget: A study on verbatim memorization of literature in large language models. In *CHR 2024: Computational Humanities Research Conference*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024c. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

Emily Öhman, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2022. EmotionArcs: Emotion Arcs for 9,000 Literary Texts. *Proceedings of LaTeCH-CLfL*.

Comparison axis	Pair count
Plot	4005
Character States	3240
Relationship Dynamics	2278
Theme	1431
Time	105
Style	1431
Fanfiction Tone & Content Tags	1275
Overall (All fandom-agnostic tags)	4005
Overall (not fandom-agnostic tags)	4005
Fandom-Specific Tags	4005
Fandoms	4005
Author	4005
Total	33,790

Table 6: The number of pairwise comparisons in FICSIM by category. We exclude stories from pairwise comparisons in categories where they lack tagging.

A Additional dataset documentation

Tag standardization Much of our tag categorization and standardization process was inspired by AO3's own practices. AO3 has a team dedicated to *tag wrangling*, which is the task of maintaining a database of canonical tags, sorting and organizing those tags, and linking tags to their canonicalized form. Thanks to this high standard of organization, many tags can be mapped back to canonical tags. Non-canonical tags often come in the forms of meta commentary, merging of multiple canonical tags, or a modification of canonical tags to include fandom-specific references. These tags still contain valuable information about their stories, and looking at them in conjunction with similar canonical tags sometimes helped us determine appropriate categorizations.

Comparisons by category. Not all stories have a similarity score along every axis. Table 6 lists the number of comparisons possible in each category.

Length Figure 3 shows the length distribution of texts in FICSIM.

Additional metadata. In addition to the similarity scores and full texts, FICSIM contains many other metadata fields about each story, including chapter splits, author IDs, author-written summaries (where available), and a number of AO3-imposed classifications (e.g. the genders of the characters in the primary relationship in the story). While we do not explicitly clean data for additional non-tag categories, we make these data available in the hope that they will be useful to other researchers working on literary applications.

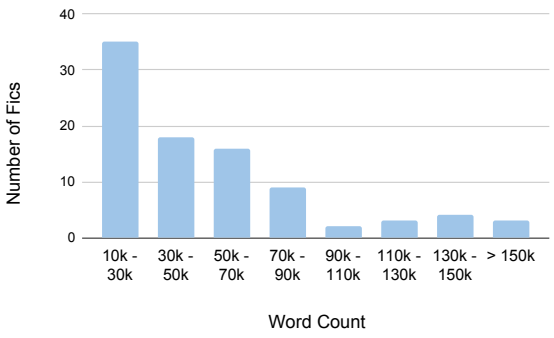

Figure 3: Story lengths in FicSim range from 10k to >400k words.

B Additional documentation of models

This section contains additional details for reproducing the embedding methods.

Models. We evaluate on 7 models, described in detail in Table 7.

Pooling. We follow each model's default strategy for constructing embeddings: Linq-Embed and SFR-Embedding use last-token pooling, GTE-Qwen2 uses mean pooling of all tokens, and GTE-ModernBERT uses a CLS token. In cases where we obtain multiple embeddings (i.e., when using a sliding window), we average all embeddings to produce a single embedding for each document. When mean-pooling multiple embeddings, we average *after* pooling all token embeddings instead of averaging a single pooled embedding from each window. In mean-pooled embeddings, we include prompt tokens but take only one embedding of each overlapped token in sliding windows with overlap.

Software. For single-window (truncation) approaches, we use `sentence_transformers` (Reimers and Gurevych, 2019). For sliding window approaches, we use Hugging Face `transformers` (Wolf et al., 2020). We call Voyage and Claude through their respective APIs, which (at the time of writing) do not retain user data for model training. A limited amount of language model assistance was used for writing simple data processing scripts; all code was verified by the authors.

Computational resources. All local models were run on a mixture of L40S and H100 GPUs; we estimate that the total compute time in development and running the final embedding methods did not exceed 200 GPU-hours. The total cost of development and running the API-based methods was approximately $80, of which $78 was the cost of running Claude.

Prompts. We use the same prompt for all models except Claude; the category-specific prompts are in Table 8. For Claude, we use the system prompt "Below is a long-form fanfiction written in English. You will be asked to summarize this story." We provide the full text of the story as a user message, then provide an additional user message with instructions. The instruction message always begins "Please write a detailed summary of this story, using up to 5,000 words." It optionally also has a category-specific instruction; these instructions are listed in Table 9.

In the rare (3) cases where a story exceeds Claude's context window, we summarize as much of the story as possible in a first API call and provide the summary plus the remainder of the story in a second API call. In this second call, the system prompt is changed to "Below is a long-form fanfiction written in English. The first section is a summary of the first portion of the story, and then the remainder of the story follows. You will be asked to summarize this story." The two sections of the input are labeled "Summary:" and "Remainder of the story:" and separated by a line of dashes. The instruction is changed to "Please write a detailed summary of the full story, using up to 5,000 words. You may copy the summary of the first portion of the story exactly, or modify it as you wish." along with any category-specific instruction.

Abbreviated name	HF or API name	Max context	Pooling strategy	Param count
Linq-Embed	Linq-Embed-Mistral	32,768	Last-token	7B
SFR-Embedding	SFR-Embedding-Mistral	32,768	Last-token	7B
GTE-Qwen2	gte-Qwen2-7B-instruct	32,768	All tokens	7B
GTE-ModernBERT	gte-modernbert-base	8,192	CLS	149M
m2-BERT-32k	m2-bert-80M-32k-retrieval	32,768	CLS	80M
Voyage-3-large	voyage-3-large	32,768	Unknown	Unknown
Claude-3.7-Sonnet	claude-3-7-sonnet-20250219	200,000	n/a	Unknown

Table 7: Additional details on the models ran.

Category	Prompt
Plot	Identify the main plot arc of the fanfiction based on the text.
Character state	Identify the main character states of the fanfiction based on the text.
Relationship dynamic	Identify the main relationship dynamics of the fanfiction based on the text.
Theme	Identify the main themes of the fanfiction based on the text.
Time	Identify the main time period of the fanfiction based on the text.
Style	Identify the main literary style of the fanfiction based on the text.
Tone & content	Identify the main fanfiction-specific tone and content descriptors of the fanfiction based on the text.
Overall	[no prompt]

Table 8: Prompt for all embedding models. The prompt (with any applicable model-specific formatting) is prepended to the beginning of the text and the start of every sliding window.

Category	Prompt
Plot	In your summary, pay particular attention to the plot of the text.
Character state	In your summary, pay particular attention to the attributes of the characters in the text.
Relationship dynamic	In your summary, pay particular attention to the relationship dynamics of the characters in the text.
Theme	In your summary, pay particular attention to the themes of the text.
Time	In your summary, pay particular attention to the temporal setting of the text.
Style	In your summary, pay particular attention to the literary style of the text.
Tone & content	In your summary, pay particular attention to any fanfiction-specific tone or tropes exhibited in the text.
Overall	[no additional prompt]

Table 9: Prompt for Claude summarization. This is appended as part of the last user message, after the system message and a user message containing the full text of the story.

C Author Consent Process

Archive of Our Own does not have a private messaging feature, and authors do not generally post contact information (or real names) on their fanfictions. After consulting with the AO3 policy team, we agreed to reach out to authors by leaving a comment on the stories we would like to use. This comment then directs them to our main Reddit post, which links to the project's webpage, explains the study and its terms, and offers a locale for authors to ask questions directly of the authors. This process was approved as Carnegie Mellon University IRB Study 00000260.

Revoking consent. We maintain a Google Form for requesting removal of a story at any time, with no questions asked. We commit to monitoring this form in perpetuity and removing fanfiction promptly if authors choose to revoke consent. For replication of results on the dataset, we will clearly label the dataset on Hugging Face and the repository with a version number, and ask that anyone using the dataset report the evaluation version.

C.1 Outreach process documents

We provide the exact text of the comments to reach out to authors (Figure 4) and the text of the Reddit post (Figure 5). [12]

[12] Our original post cites 30k as our desired lower word limit for fanfiction contributions, but after seeing the volume and quality of fanfic contributions we received below this threshold, we decided to include texts above 10k words in our dataset, provided they had adequately detailed tagging.

Hi! My name is Natasha Johnson :) I'm a recent graduate from Carnegie Mellon University's English Department. I'm working alongside Emma Strubell and Amanda Bertsch at CMU on a project involving fanfiction , and we're hoping to include your fanfic(s) in our research. If you would like to learn more about our project and consent for us to include your work, please take a look at the post we made about it here: `[link to post]`

Figure 4: Sample comment on fanfiction

Hello! We are Natasha Johnson (https://natashamariejohnson330.github.io/), Emma Strubell (https://strubell.github.io/), and Amanda Bertsch (https://www.cs.cmu.edu/~abertsch/). We're interested in exploring the capabilities and limitations of digital tools in the context of humanities research. We are currently conducting a research project that looks at quantifying fanfiction similarity, focusing on fics over 30k words.
Because of the detailed tagging you use on your work, we're asking for your consent to use your fanfiction for this project.
If you consent to us using your fanfic(s), here are our promises:

1. We might make observations about fanfic content, but we will not critique fanfics in any way.
2. We will actively avoid seeking any personal information about you.
3. We will not use your fanfics to train AI models.
4. We will use your fanfics to test how well AI models capture similarity in literary contexts, to see if these models could be useful for literary scholars. During testing, the models do not retain any history or memory of input text, and the models are not trained on the inputs.
5. If we publish our research, we will release our dataset alongside it. This dataset will include the fanfic texts, the fanfiction tags, and a numerical author identifier in place of your AO3 pseudonym.
6. In order to access the dataset, we will ask viewers to agree not to use the data for AI training purposes.
7. At your request, we will remove your fanfic(s) from the dataset at any time, for any reason. Here's the form you can use to submit a removal request: `[removal form link]`

If you would like to give us permission to use your fanfic(s) in this way, please let us know via this consent form: `[consent form link]`
Feel free to post any questions here in the comments, or you can reach out anonymously via this Google form: `[google form link]`

Figure 5: Post on Reddit with information on how to consent

D License