FicSim: An Ethically Constructed Dataset for Long-Context Semantic Similarity Comparison within Fiction

Natasha Johnson Amanda Bertsch Emma Strubell
Language Technologies Institute
Carnegie Mellon University
nmj@alumni.stanford.edu abertsch@cs.cmu.edu strubell@cmu.edu

Abstract

As language models continue to advance in their ability to process long and complex texts, there has been growing interest in their application within computational literary studies (CLS). With the increasing development of CLS tools, many researchers have turned to public domain eBook collections, such as Project Gutenberg, to test their models. However, issues of large-scale web scraping and model contamination challenge the reliability of such evaluations and call for novel methods of data collection. In response to this, we assemble a dataset of literature that has been excluded from large-scale scraping, alongside similarity scores in a variety of literary categories. This dataset, *FicSim*, can be used to evaluate models for long-context semantic textual similarity comparison within fiction. Throughout our data-collection process, we prioritize author agency and rely on continual, informed author consent. We thus demonstrate how high-quality literary datasets can be constructed without undermining authors' rights to their work.

1 Introduction

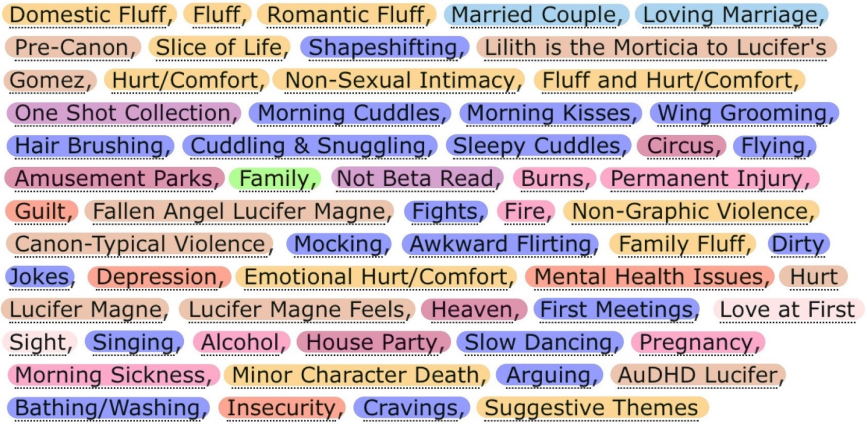
As language models become capable of processing increasingly long and complex texts [1, 2, 3, 4], interest in using such models for literary studies has also grown. An increasing number of computational approaches have been proposed to analyze elements of fictional texts that are of interest to literary scholars, such as character mobility [5], emotional arc [6], and narrative pacing [7].

To evaluate these methods, many computational literary studies (CLS) researchers [8, 9, 10] have turned to digital texts made accessible through public repositories like Project Gutenberg. However, these texts—alongside related data and analyses from other commonly-scraped sites like Wikipedia—are included in the pretraining data of most models [11], which could cause contamination and skew evaluation results [12]. Thus, despite the availability of benchmark datasets built upon public domain literature [13, 14, 8, 15], it has been difficult to verify the efficacy of recent CLS approaches.

In response to this gap, we present *FicSim*, an evaluation dataset for long-context semantic textual similarity (STS) constructed of long-form human-written narratives unlikely to appear in pretraining data, which are accompanied by author-labeled metadata and collected 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 different facets of fictional texts. Our dataset allows the fair evaluation of both existing CLS models and large language models on literary tasks. Our data collection process, which relies on communication with and consent from the authors of the texts, demonstrates how creative works can be collected without circumvention of the creators' rights and wishes.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Dataset



Domestic Fluff, Fluff, Romantic Fluff, Married Couple, Loving Marriage, Pre-Canon, Slice of Life, Shapeshifting, Lilith is the Morticia to Lucifer's Gomez, Hurt/Comfort, Non-Sexual Intimacy, Fluff and Hurt/Comfort, One Shot Collection, Morning Cuddles, Morning Kisses, Wing Grooming, Hair Brushing, Cuddling & Snuggling, Sleepy Cuddles, Circus, Flying, Amusement Parks, Family, Not Beta Read, Burns, Permanent Injury, Guilt, Fallen Angel Lucifer Magne, Fights, Fire, Non-Graphic Violence, Canon-Typical Violence, Mocking, Awkward Flirting, Family Fluff, Dirty Jokes, Depression, Emotional Hurt/Comfort, Mental Health Issues, Hurt Lucifer Magne, Lucifer Magne Feels, Heaven, First Meetings, Love at First Sight, Singing, Alcohol, House Party, Slow Dancing, Pregnancy, Morning Sickness, Minor Character Death, Arguing, AuDHD Lucifer, Bathing/Washing, Insecurity, Cravings, Suggestive Themes

Figure 1: Example of detailed tagging from AO3, with highlight colors corresponding to tag categories within Tables 1 and 2

We construct our dataset using fanfiction stories from Archive of Our Own (AO3), a digital repository hosting over 13M works. AO3 has made efforts to discourage web scraping, including disallowing Common Crawl scraping in 2022 [16] and implementing aggressive rate limiting[1]. Due to the site's construction and norms, many AO3 stories feature highly detailed tagging (e.g. Figure 1), which we leverage to create ground-truth similarity labels on various axes. We consider only stories with detailed tagging, written after December 2022, in English, with at least 30k words.

Author agency Though Archive of Our Own explicitly permits data collection done by academic researchers [16], we additionally seek each author's permission to include their work(s) in our dataset using an IRB-approved process. After identifying the authors whose works we would like to include, we reach out to each author individually through AO3 to explain our project and share guarantees about author privacy, story usage, and withdrawal of consent.[2] Then, we invite authors to sign a consent form using their pseudoynm, to protect privacy. Alongside this consent form, we provide authors with mechanisms to ask questions or withdraw consent at any point, for any reason.

Data In constructing our dataset, we only look at fanfics that have an original publication date after December 2022 and which have been marked as completed. Though novels are often expected to contain 40k words at minumum [17, 18], we set our lower inclusion threshold to 30k, to ensure that there are some texts in the dataset which are suitable for evaluating models with context windows of 32k tokens [3, 4]. Our final dataset includes 100 fanfics, which is the same number of texts upon which other annotated CLS datasets have been built [13, 14]. These 100 texts make up 4950 test pairs, which is greater than the number of tasks in LongBench [19]. In the future, we might expand our dataset to include stories below the 30k word minumum or to include additional manual annotation.

Simlarity score construction We sort the authors' tags into categories as seen in Tables 1 and 2. Some categories, such as "Narrative Action," capture general qualities of fictional literature and align with projects that compare textual similarity on the basis of actions, subjects, and themes [20, 21]. Other categories, such as "Relationship Mechanism," reflect fanfiction-specific tropes and serve to identify whether models can identify similarity based on genre-specific conventions. We standardize tag formatting, consolidate synonymous tags, and group similar tags. We then measure the overlap between these standardized tags and groups—weighing tag overlap more heavily than group overlap—to determine overall and category-specific ground-truth similarity scores for each fanfic pairing.

[1] While this does not *guarantee* that no fanfics posted after this date are used in pretraining corpora, it makes these texts relatively unlikely to be contaminated from pretraining.

[2] See Appendix A for the full text.

Table 1: Tag categories inspired by literary research

Category	Examples
Internal Character States	Depression Mental Instability Insecurity
Narrative Actions	Drinking Loss of Eyes Taking Care of Sick People
Entities & Physical Subjects	Cars Sick Character Alcohol
Story Themes	Cultural Differences Found Family Sibling Love
Time	Post-War 1970s Modern Era
Locations	Hospitals Pubs They All Live in the Same Apartment Building
Style	POV Third Person Omniscient Character Study References to Jane Austen
Uncategorized	No Dino No Gangs I Write This Instead of Sleeping Not Much Tho

Table 2: Tag categories inspired by fanfiction conventions

Category	Examples
Relationship Type	Established Relationship Enemies to Lovers Slow Burn
Relationship Mechanism	Sharing a Bed Fake/Pretend Relationship Arranged Marriage
AO3 Tone & Content Tags	Tooth-Rotting Fluff Mild Hurt/Comfort Angst
Fandom-Specific	Track Star Nobara POV Keith (Voltron) Autistic Sherlock

3 Evaluation and Intended Uses

We intend to benchmark the performance of CLS and general NLP models on our dataset to evaluate their suitability for long-context STS tasks, particularly within fiction.

To produce the similarity measurements for comparison against our truth labels, we will take the cosine similarity between pairs of stories using the embeddings generated by the models. For models which are not strictly encoder models, we will take the final hidden layer from the encoding process.

For models which have limited input capacities, we will split the story texts into smaller chunks to be encoded and then concatenate the resulting embeddings.

We will benchmark both STS-specific and general-purpose models. Of particular interest are two embedding approaches which were each designed for specific long-context STS tasks: *lib2vec* is an embedding method which captures the overall content of a book, as well as representations of individual facets, such as plot and setting; *lib2vec* has only been evaluated on public domain literature [8]. SDR is a model designed for ranking long-form documents according to their similarity, which has only been tested on Wikipedia articles [2]. Alongside these methods, we will benchmark the following models: SBERT [22], Mistral 7b v0.2 [3] , Unlimiformer [1], LongEmbed [23], and Gemma [24]. High performance across our entire test set would be suggestive of general suitability for STS tasks on long fiction texts. High performance in only one category would suggest that a model is particularly adept at capturing this type of similarity.

Furthermore, to see how model performance on this test set corresponds to performance in other NLP tasks, we will also evaluate the aforementioned models on the following datasets: the STS benchmark [25], LongBench [19], and the Wikipedia datasets developed to evaluate SDR [26, 27].

4 Conclusion

We present *FicSim*, a dataset of stories and similarity labels for benchmarking model performance on long-context STS tasks within fictional texts. Furthermore, we describe our method of data collection which equips authors with guarantees about how their work will be used, the choice to give or withhold consent, and the ability to withdraw consent at any time. We hope our dataset and evaluations will help CLS researchers make informed decisions about model selection for tasks relating to story similarity, and that this dataset's collection will also serve as an example of positive and honest communication between machine learning researchers and creative professionals.

References

[1] Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. Unlimiformer: Long-Range Transformers with Unlimited Length Input. *37th Conference on Neural Information Processing Systems*, 5 2023.

[2] Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. Self-Supervised Document similarity ranking via contextualized language models and hierarchical inference. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098, 8 2021.

[3] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego De Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 10 2023.

[4] Together. LLaMA-2-7B-32K, 5 2023.

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

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

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

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

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

[10] 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. The transformation of Gender in English-Language fiction. *Cultural Analytics*, 2 2018.

[11] 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. What's in my big data? *International Conference on Learning Representations*, 2024.

[12] Medha Palavalli, Amanda Bertsch, and Matthew R. Gormley. A taxonomy for data contamination in large language models, 7 2024.

[13] David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1 2019.

[14] David Bamman, Olivia Lewke, Anya Mansoor, and European Language Resources Association (ELRA). A dataset of literary coreference. *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 44–54, 5 2020.

[15] Matthew Sims, Jong Ho Park, and David Bamman. Literary Event Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1 2019.

[16] Organization For Transformative Works. AI and Data Scraping on the Archive, 2023.

[17] Word count guide: How long is a book, short story, or novella?

[18] family=Lyons given i=J, given=Jordan. How many words are in a novel? word count by genre more.

[19] 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. LongBench: a bilingual, multitask benchmark for long context understanding, 8 2023.

[20] Mark Algee-Hewitt and Erik Fredner. Typicality, 2023.

[21] Rabea Kleymann, Andreas Niekler, and Manuel Burghardt. Conceptual Forays: A corpus-based study of "Theory" in Digital Humanities Journals. *Journal of cultural analytics*, 7(4), 12 2022.

[22] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv (Cornell University)*, 8 2019.

[23] Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. LongEmbed: Extending embedding models for long context retrieval, 4 2024.

[24] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-Hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on Gemini research and technology, 3 2024.

[25] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo López-Gazpio, and Lucia Specia. SEMEVAL-2017 Task 1: Semantic Textual Similarity Multilingual and cross-lingual focused evaluation. *HAL (Le Centre pour la Communication Scientifique Directe)*, 8 2017.

[26] Dvir Ginzburg. Long document similarity dataset, Wikipedia excerptions for video games collections, 5 2021.

[27] Dvir Ginzburg. Long document similarity dataset, Wikipedia excerptions for wine collections, 5 2021.

A Author Consent Process

We reach out to authors by first leaving a comment on the stories we would like to use. This comment then directs them to our main Reddit post explaining the study and its terms.

A.1 Sample comment on fanfic

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]

A.2 Post on Reddit with information on how to consent

Hello! We are Natasha Johnson (`https://www.linkedin.com/in/natasha-m-johnson/`), 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: `https://forms.gle/JxHRsTbwMb78fCS77`

If you would like to give us permission to use your fanfic(s) in this way, please let us know via this consent form: `https://forms.gle/GrU86ZWEWm1kvY2C6`

Feel free to post any questions here in the comments, or you can reach out anonymously via this Google form: `https://forms.gle/94ejTXtwxZyVriqh8`