Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark

Funing Yang and **Carolyn Jane Anderson**
Wellesley College
Wellesley, MA
carolyn.anderson@wellesley.edu

Abstract

Several systems have been developed to extract information about characters to aid computational analysis of English literature. We propose character similarity grouping as a holistic evaluation task for these pipelines. We present AustenAlike, a benchmark suite of character similarities in Jane Austen's novels. Our benchmark draws on three notions of character similarity: a structurally defined notion of similarity; a socially defined notion of similarity; and an expert defined set extracted from literary criticism.

We use AustenAlike to evaluate character features extracted using two pipelines, BookNLP and FanfictionNLP. We build character representations from four kinds of features and compare them to the three AustenAlike benchmarks and to GPT-4 similarity rankings. We find that though computational representations capture some broad similarities based on shared social and narrative roles, the expert pairings in our third benchmark are challenging for all systems, highlighting the subtler aspects of similarity noted by human readers.

1 Introduction

There is growing interest in using computational techniques to analyze works of literary fiction. Several systems have been developed to automatically extract information about characters from English literary text (Bamman et al., 2014; Yoder et al., 2021). In this paper, we explore character similarity as a holistic evaluation task for literary pipelines. We use character similarity to explore the information about characters that is captured by the different kinds of features these pipelines extract: their events, utterances, and attributes.

Because characters can be similar along multiple axes, we construct a multi-part benchmark, AustenAlike, that uses three different notions of character similarity to group characters in Jane Austen's novels. The first is a structurally defined notion of similarity to group Austen's characters: characters are similar if they fill similar narrative roles. The second is a socially defined notion of similarity: characters are similar if they share demographic features. The final benchmark takes a wisdom-of-the-crowd approach, but with an expert crowd: we extract comparisons of characters from four decades of *Persuasions*, a journal dedicated to the analysis of Austen's work. Figure 1 shows an example of how these three views of character similarity can lead to different comparisons.

James Morland from *Northanger Abbey*
Sibling to heroine and single 20-year-old male clergy with income of £400/year
Social Pairings: Charles Hayter, Edward Ferrars, Robert Martin
Narrative Role Pairings: Isabella Knightley, John Dashwood, Margaret Dashwood, Susan Price, William Price, Elizabeth Elliot, Mary Musgrove, Jane Bennet, Mary Bennet, Kitty Bennet, Lydia Bennet
Expert Pairings: Edmund Bertram, Edward Ferrars, Henry Tilney, Philip Elton

Figure 1: Example character from AustenAlike

We use AustenAlike to explore how much information about characters is captured by the different kinds of features that literary pipelines extract. We extract character events, quotes, modifiers, and assertions using the BookNLP (Bamman et al., 2014; Sims et al., 2019) and FanfictionNLP Yoder et al. (2021) pipelines. We build character representations using contextualized embeddings of these features, and compare how well these representations align with the three sets of character groupings in the AustenAlike benchmarks. We also compare a non-feature-based approach by extracting similarity judgments from ChatGPT.

Our results show that event- and assertion-based representations capture more information about

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 17–30
November 16, 2024. ©2024 Association for Computational Linguistics

character similarity than quote-based representations. Overall, however, we show that though computational representations capture some broad social and narratological similarities, there is a wide gap between the similarities they capture and the more nuanced similarities highlighted in our wisdom-of-the-expert-crowd benchmark. The best feature-based representations exhibit only medium correlations with expert rankings of character similarity, and GPT-4 lists the expert-identified most similar character in a top ten similarity list only half of the time. AustenAlike illustrates how much work remains to achieve nuanced computational representations of literary characters.

2 Related Work

There is a growing interest in applying computational methods to analyze literary fiction, both in analyses of large collections (*distant reading* (Moretti, 2013)) (Grayson et al., 2016; Jayannavar et al., 2015; Milli and Bamman, 2016) and of individual authors and works (Agarwal et al., 2013; Wang and Iyyer, 2019; Liebl and Burghardt, 2020). Though these projects range in scope, they share a foundation of feature extraction: literary evidence must be identified before it can be interpreted.

To facilitate computational analysis, a number of pipelines for extracting features from literary text have been developed (Bamman et al., 2014; Sims et al., 2019; Yoder et al., 2021; Ehrmanntraut et al., 2023). In this paper, we focus specifically on features related to literary characters.

Character mentions The first step in computational studies of character is to identify character mentions using named entity recognition and coreference resolution. There is a large body of existing work on these tasks (Vala et al., 2015; Brooke et al., 2016; Roesiger and Teufel, 2014) given their complexity in a literary setting and their importance for downstream tasks.

Some pipelines further disambiguate character references in a *character clustering* step. BookNLP is a pipeline trained on data from LitBank, which provides annotated training data drawn from 19th- and early 20th-century English fiction, including annotations for named entity recognition (Bamman et al., 2019) and coreference resolution (Bamman et al., 2020). FanfictionNLP is a similar pipeline that is trained on and tailored to fanfiction.

Character features Once character mentions have been identified, the surrounding text can be used to extract information related to characters.

Some previous work focuses on character personality traits and emotions (Flekova and Gurevych, 2015). Kim and Klinger (2019) analyzes how emotions are expressed nonverbally in a corpus of fan fiction short stories, while Pizzolli and Strapparava (2019) train classifiers to identify personality traits in Shakespeare characters. The pipelines we study target more general descriptions: for FanfictionNLP, *assertions*, descriptions of physical and mental attributes; for BookNLP, modifiers and possessions.

What characters do and say is also of interest. Although quote attribution remains a challenging task with a number of approaches (He et al., 2013; Almeida et al., 2014; Muzny et al., 2017), it is useful for analyzing both the content and style of characters' speech (Dinu and Uban, 2017; Vishnubhotla et al., 2019). BookNLP extracts both events and quotes, while FanfictionNLP extracts only quotes.

There is also much work on mapping and analyzing relationships between characters (Elson et al., 2010; Lee and Yeung, 2012; Jayannavar et al., 2015; Agarwal et al., 2013; Wohlgenannt et al., 2016; Labatut and Bost, 2019). For instance, Chaturvedi et al. (2016) and Iyyer et al. (2016) automatically identify how relationships between characters change over the course of narratives.

Character models Once character features are extracted, they can be used to build computational representations of characters. Some work seeks to classify characters into types (Chambers and Jurafsky, 2009; Valls-Vargas et al., 2021; Stammbach et al., 2022; Bamman et al., 2014). For instance, Jahan and Finlayson (2019) propose a narratologically-grounded framework for character identification and a simple rule-based system for extracting characters and their roles.

Others explore authorial decisions in representing characters (Bullard and Ovesdotter Alm, 2014) or how they evolve over retellings (Besnier, 2020).

Some approaches learn character representations directly. Grayson et al. (2016) show that word embeddings learned from 19th-century works of fiction provide insight into characters.Holgate and Erk (2021) learn vector representations using masked entity prediction as a training objective. Most similar to our work, Inoue et al. (2022) propose a benchmark for evaluating character representations.

Their work takes a broad multi-author, multi-task perspective, while ours dives more deeply into characters by a single author, exploring character similarity from three different angles.

3 A Three-Part Benchmark for Evaluating Character Similarity

Character similarity is a multi-faceted concept. Two characters may play the same role in a narrative or follow the same plot trajectory. They may have similar personality traits or fill similar social roles. AustenAlike uses a multi-faceted approach to character similarity that explores three aspects of literary characterhood: shared narrative roles, shared social characteristics, and pairwise comparisons from expert analysis.[1] The AustenAlike benchmark focuses on characters from the six Jane Austen novels published within or immediately after her lifetime: *Sense and Sensibility*, *Pride and Prejudice*, *Mansfield Park*, *Emma*, *Persuasion*, and *Northanger Abbey*. We include all named characters who speak more than once, except those who die in the first chapter.

3.1 Social Characteristics

Jane Austen's novels highlight how her character's choices are impacted by their position in society. Although her characters struggle to varying degrees to reconcile their desires with constraints imposed by gender, rank, and wealth, these social characteristics play a large part in determining the options available to them within the novel.

We consider five demographic dimensions that define social relationships within Austen's writing: marital status, gender, rank, age, and wealth. There are other social characteristics that demarcated opportunities within Austen's historical context, such as race and nationality; however, the characters under consideration are homogeneously White and English.[2] A summary of the social categories and the size of each group is in Appendix A.

Rank Although almost all of Jane Austen's characters belong to the upper middle or lower upper classes, their relative social rank is nonetheless important to their prospects. Most characters are gentry: independently wealthy, often landowners. Lower-ranked characters belong to professions. Following social conventions of the time, an unmarried woman has her father's rank and a married woman her husband's.

Wealth Austen novels center on questions of wealth, particularly as they relate to marital prospects. As a result, the wealth of unmarried characters is typically stated. The wealth of married characters is not always stated. We draw on estimates from Heldman (1990) and Toran (2015).

Gender The genders of all Austen characters are overt and stable. All characters are Male or Female.

Age Character ages are reasonably stable as almost all plot events take place within a year. If a character's age is not mentioned, we estimate from the ages of their family members.

Marital status Marital status is a key social characteristic of Austen characters. We divide characters into four groups: Single, Married, Widowed, and Transitional, a group comprising the handful of characters whose marital status changes before the climax of the novel.

3.2 Narrative Roles

Another way in which characters can resemble each other is in the role they play in the narrative structure of the work. We define seven narrative roles:

- Heroine: each novel has at least one protagonist who is an unmarried woman seeking a marriage partner.
- Hero: the character that each protagonist marries at the novel's end.
- Deceiver: each novel features a character who sets key events in motion by lying about himself or the heroine.
- Rival: an alternate love interest for the hero.
- Wooer: an alternate love interest for the heroine.
- Parents: the parents of the heroine.
- Siblings: the siblings of the heroine.

.

These groupings are shown in Table 1.

3.3 Wisdom-of-the-Experts Character Pairs

In our most fine-grained benchmark, we look at characters who have been identified as similar by literary scholars. We use a wisdom-of-the-crowds approach, but with an expert crowd: authors of

[1] The dataset and support code are available at https://github.com/Wellesley-EASEL-lab/AustenAlike.

[2] Given the exclusion of Austen's unfinished *Sanditon*.

Heroines:	Emma Woodhouse, Elizabeth Bennet, Elinor Dashwood, Marianne Dashwood, Fanny Price, Catherine Morland, Anne Elliot
Heroes:	George Knightley, Fitzwilliam Darcy, Edward Ferrars, Edmund Bertram, Henry Tilney, Frederick Wentworth, Colonel Brandon
Deceivers:	John Thorpe, George Wickham, John Willoughby, William Elliott, Henry Crawford, Frank Churchill
Rivals:	Caroline Bingley, Lucy Steele, Louisa Musgrove, Mary Crawford, Harriet Smith
Wooers:	Henry Crawford, William Elliot, Philip Elton, Charles Musgrove, William Collins, John Thorpe
Siblings:	Marianne Dashwood, Jane Bennet, Lydia Bennet, Mary Bennet, Kitty Bennet, Susan Price, Mary Musgrove, Elizabeth Elliot, Isabella Knightley, James Morland, William Price
Parents:	Mr. Bennet, Sir Walter Elliot, Lieutenant Price, Mr. Woodhouse, Mrs. Bennet, Mrs. Dashwood, Mrs. Price, Mrs. Morland

Table 1: Narrative Roles benchmark summary

articles published in *Persuasions*, the Jane Austen Society of North America's peer-reviewed journal.

We manually reviewed 43 volumes of *Persuasions* to create a set of character pairings. We extract all instances of a similarity or shared property discussed in an article. When an article mentions a similarity between more than two characters, we add all pairings from the set. The resulting dataset contains 5740 character comparison pairs.

The identified comparisons are diverse, encompassing traits from our other benchmarks, such as rank, age, and narrative role, as well as more nuanced commonalities. For instance, *Persuasions* authors describe Edward Ferrars and Frank Churchill as similar because both are secretly engaged; Emma Woodhouse and Lady Catherine de Bourgh because they oversee charitable work; and Isabella Thorpe and Lydia Bennet because of their flirtatiousness. These expert-identified pairings provide a comprehensive view of character similarity.

4 Building Computational Representations of Character

We build computational representations of character from the output of two literary pipelines. We construct representations out of the features they extract: for BookNLP, events, quotes, and modifiers; for FanfictionNLP, quotes and assertions.

4.1 Character Mentions

We use each pipeline to identify all character mentions, perform coreference resolution, and aggregate character mentions. We then merge and filter character clusters using a handwritten alias map for Austen character names.

4.2 Feature Embeddings

We retrieve contextualized embeddings for each kind of feature. For events and modifiers, which are single words, we retrieve a contextualized embedding of the word in its context using T5 (11B) (Raffel et al., 2020). For quotes and assertions, we retrieve sentence embeddings using NV-Embed (7.85B) (Lee et al., 2024). We center each kind of feature embedding by subtracting the mean of all embeddings for the feature.

For each feature and character, we construct a character representation by averaging the embeddings of the character's features. For events, we average the character's agent events and patient events separately and concatenate the vectors. This process produces 5 representations per character: an assertion vector, a modifier vector, an event vector, and two quote vectors (one per pipeline).

Having produced these 5 representations for each character, we are interested in exploring the effectiveness of each kind of feature-based representation in capturing character similarity. Thus, we compute each result presented in Section 6 for each of the 5 representations.

4.3 GPT-4 comparison

We provide a non-featured-based comparison by querying a pretrained large language model, GPT-4 (Achiam et al., 2023), for character similarity rankings. Given the popularity of Austen's work, we assume that GPT-4's training data contains all six novels and many web pages discussing them.

We extract character similarities using three approaches: asking GPT-4 to select the most similar character from a list of all benchmark characters; asking GPT-4 to select the most similar character and explain its choice; and asking GPT-4 to choose the ten most similar characters from a list of all benchmark characters. We repeat each experiment 5 times (further details in Appendix B).

5 Evaluating character similarity

We have proposed three benchmarks that capture different aspects of character similarity. For the social and narrative roles benchmarks, we are interested in the similarity between characters in the same groupings. For the expert benchmark, we are interested in whether characters are most similar to those they are paired with by experts.

5.1 Grouping evaluation

The Social and Narrative benchmarks define groupings of characters. We explore how strongly these groupings are captured by computational character representations using two evaluation metrics.

In-group Cosine Similarity We explore whether characters are more similar to characters within their group than those outside of their group. We compute the average cosine similarity between a grouped character and all other group members, and compare it to the average cosine similarity between the character and non-group characters. We call this *in/out-group cosine similarity difference*.

Most Similar Character We also ask whether very similar characters come from the same groups. We count how often the single character with highest cosine similarity to the target character belongs to the same group.

5.2 Pairing evaluation

For the Expert benchmark, we measure the extent to which the cosine similarities of each kind of representation align with the expert-identified pairs using three metrics:

Correlation We look at the correlation between cosine similarity of two character representations and the number of times experts describe the two characters as similar. We calculate Pearson's ρ to measure the strength of the correlation.

Ranking similarity Literary experts may be more interested in identifying highly similar characters than in quantifying degrees of dissimilarity. We identify the ten most similar characters according experts and to cosine similarity, and compute the alignment between the lists using Jaccard similarity. Jaccard similarity measures the intersection of the groups divided by their union. If the two lists are completely different, their Jaccard similarity is 0; if they mostly agree, it is close to 1.

Top character in ten-most similar Finally, we focus on the top expert-identified pairings. We count how often the character who experts pair most with a target character has one of the ten highest cosine similarities to the target character.

6 Results

We explore how well computational representations of character capture aspects of character similarity using the three-part AustenAlike benchmark.

6.1 Narrative Roles Benchmark

The narrative roles benchmark explores similarity between characters who play similar roles in the plot of a novel. Are heroines similar to other heroines? Are parents similar to other parents? If parents are described similarly to other parents, assertion- and modifier-based representations should capture their similarity; if they say and do similar things as other parents, their quote- and event-based representations should be similar.

6.1.1 Are same-role characters more similar?

We test whether characters who share the same narrative role are more similar than characters who do not. We compare the average cosine similarity of representations within a narrative role group to their similarity to non-group members. We compute the in-group and out-group scores for each character in a target role group and average them.

Figure 2 plots the cosine similarity for characters within the same narrative role group compared to characters outside of the group. We observe that event- and assertion-based representations are the best at showing dissimilarity for characters outside of the role group. The FanfictionNLP quote-based representations show the weakest differences between in-group and out-group members.

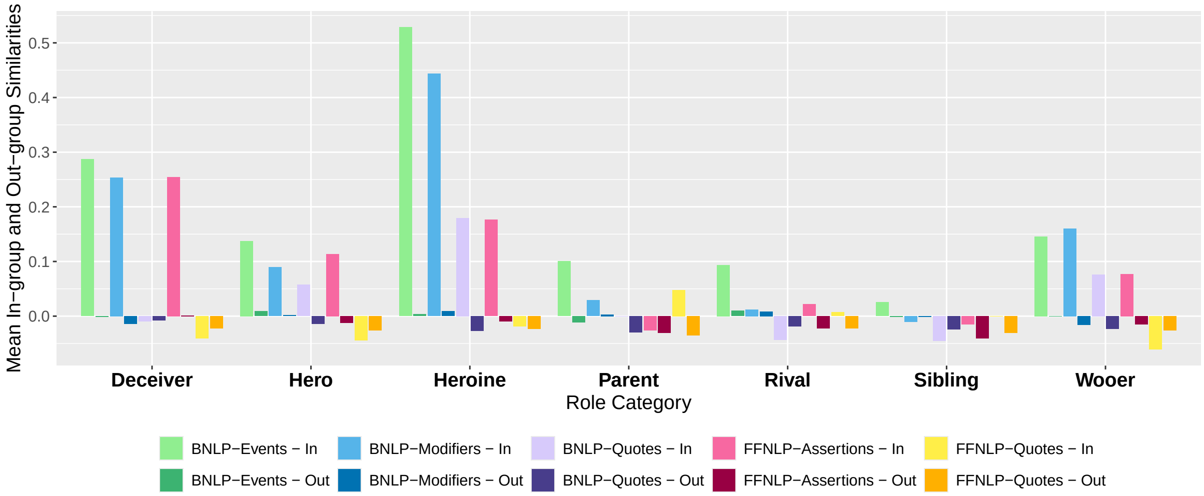

Figure 2: Narrative Role Benchmark: Mean cosine similarities between same-group characters and other characters by representation type.

System	Hero	Heroine	Deceiver	Rival	Wooer	Parent	Sibling
FanfictionNLP Assertions	0.29	0.43	0.33	0	0	0.18	**0.29**
BookNLP Events	0	**1**	0.36	0.09	0.18	**0.35**	0
BookNLP Modifiers	0	0.86	0.33	0.2	0	0.27	0.18
BookNLP Quotes	0.13	0.78	0.57	**0.33**	0.43	0.08	0
FanfictionNLP Quotes	0	0.43	0	0.14	0	0.18	0.08
GPT-4	0.43	0.43	0.5	0	0	0.33	0.25
GPT-4 Reasoning	**0.86**	**1**	**0.83**	0.17	**0.5**	0.42	0.08

Table 2: Narrative Role Benchmark: Average occurrence of most similar character in same narrative role group by character representation. Characters from same novel are excluded.

6.1.2 Is the most similar character from the same group?

We also explore whether a target character's most similar character belongs to the same narrative role group. For each character, we count how often the character with highest cosine similarity belongs to the same role group. Feature-based representations can be skewed towards same-novel similarity: for instance, characters in *Northanger Abbey* are more likely to engage in reading events since this is a theme of the novel. We therefore explore results with and without characters from the same novel.

Table 2 reports how often the most similar character occurs in the same role group, with same-novel characters excluded (inclusive version in Appendix C). We see marked differences between categories. Heroines are frequently similar to heroines for all representations, while other groups have lower rates of same-group membership.

The BookNLP quote representations capture narrative role similarity better than the FanfictionNLP quote representations, perhaps because BookNLP is trained on literary fiction. However, FanfictionNLP assertions perform competitively in two of the most challenging categories for feature-based representations, Hero and Sibling.

We observe that GPT-4, when asked to justify its decision, is more sensitive to narrative role than the feature-based representations in about half of the categories. However, without reasoning-prompting, it is no better than the feature-based representations, identifying selecting a heroine as the most similar to heroines only 43% of the time.

Qualitatively, a challenging aspect of this benchmark seems to stem from young single characters with different narrative roles. Like heroes and heroines, deceivers, wooers, and rivals tend to be unmarried and of a similar age. We observe that heroes tend to be similar to deceivers (10/69 out-group cases) and vice versa (12/50 out-group cases), and rivals to heroines (26/64) and vice versa (6/31 out-group cases), aligning with the social characteristics of each set. The error patterns for the remaining categories seem less clear, perhaps reflecting the

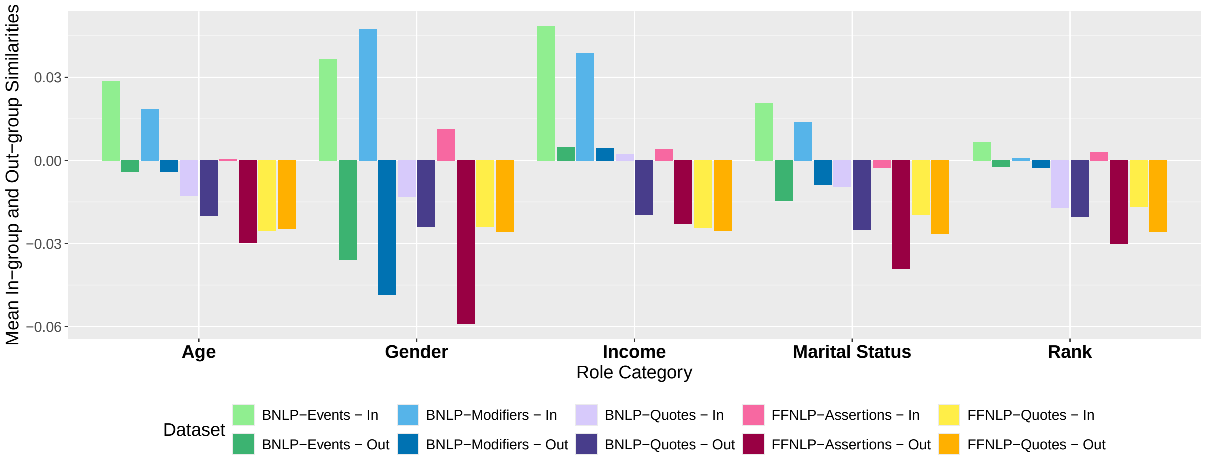

Figure 3: Social Benchmark: average differences in cosine similarity between same-group characters and other characters by character representation and social role group.

System	Age	Gender	Income	Marital Status	Rank
FanfictionNLP Assertions	0.16	0.9	0.02	0.5	0.34
BookNLP Events	0.23	0.76	0.07	0.51	0.29
BookNLP Modifiers	0.22	0.80	0.05	0.46	0.19
BookNLP Quotes	0.06	0.63	0.15	0.42	0.26
FanfictionNLP Quotes	0.13	0.54	0.02	0.3	0.25
GPT-4	0.26	0.80	**0.21**	0.52	**0.42**
GPT-4 Reasoning	**0.32**	**0.98**	0.07	**0.58**	0.39

Table 3: Social Benchmark: average occurrence of most similar characters in the same social group by character representation. Characters from same novel are excluded.

limited mentions of parent characters and the more heterogeneous characteristics of siblings.

6.2 Social Benchmark

The second AustenAlike benchmark evaluates character similarity on the basis of social characteristics. It groups characters based on five demographic features: rank, wealth, gender, age, and marital status. Modifiers and assertions may directly describe these characters. However, given that a character's social status delimits the set of actions and utterances available to them, we also expect event- and quote-based representations to echo back similarities based on these characteristics.

6.2.1 How similar are characters with shared social characteristics?

We explore whether characters within the same group in each of the social categories are most similar to each other. Figure 3 plots the average cosine similarity for characters within the same social group compared with non-group members.

We observe that the event-based representations are the most reliable for distinguishing social similarity. Gender shows the sharpest in-group/out-group differences for all three categories, followed by income. Quote-based representations struggle to capture similarity by social group: the FanfictionNLP quote-based representations do not capture differences for any of the criteria, while the BookNLP quote-based representations show only a (weak) in-group/out-group difference for income.

6.2.2 Is the most similar character from the same group?

We also focus more narrowly on the top-most similar character. Table 3 shows how often the character with the highest cosine similarity to the target character occurs in the same social group. Top character representations most commonly share gender and then marital status. This makes sense, since Austen's plots center around courtship: these key aspects of identity should be reflected in how they are described and the events they participate in.

GPT-4's similarity judgments align with social characteristics more strongly than any of the

Dataset	Pearson's ρ	Jaccard Similarity	Top in Top 10
FanfictionNLP Assertions	0.29	**0.03**	**0.69**
BookNLP Events	**0.4**	0.02	0.34
BookNLP Modifiers	0.28	0.01	0.29
BookNLP Quotes	0.27	**0.03**	0.56
FanfictionNLP Quotes	0.15	0.02	0.49
GPT-4	-	-	0.52
GPT-4 Reasoning	-	-	0.56
GPT-4 Top Ten List	-	0.02	-

Table 4: Expert Benchmark: measures of alignment between expert pairing counts and computational similarity.

feature-based representations. Quote-based representations do not seem to capture similarity by social characteristics as well as the other feature-based representations in most categories.

6.3 Expert Benchmark

Our last benchmark takes an expert wisdom-of-the-crowd approach. The expert benchmark contains counts of character similarity pairings. We compare these pairing counts to the cosine similarity between the computational representations of the two characters to evaluate how well computational representations aligns with expert judgments of character similarity.

6.3.1 Does cosine similarity correlate with expert judgments?

We examine how well computational character representations align with expert judgments by measuring the correlation between expert character pairings and cosine similarity. We posit that high quality computational representations should produce higher cosine similarity between the characters that are more frequently deemed similar by experts.

Table 4 shows the correlation between expert pairing counts and cosine similarity for each of the computational representations.

Overall, we observe moderate positive correlations between the cosine similarity of character representations and the number of expert similarity pairings. The BookNLP event representations correlate most strongly with expert pairings, while the FanfictionNLP quote-based representations correlate less strongly than other feature-based representations. This converges with our social and narratological similarity findings.

Although the expert benchmark is useful in differentiating among feature-based representations, it is also important to note that none of the feature-based representations are strongly correlated with expert judgments. This shows that there are many aspects of character similarity that are apparent to human readers that remain uncaptured in the computational character representations we explore.

6.3.2 Is there agreement on the most similar characters?

Correlations between cosine similarity and expert pairing counts may be skewed by very dissimilar characters, whose expert pairings are few. We also look at two measures of agreement for the most similar characters.

For each character, we retrieve the ten characters with the highest cosine similarity, and the ten characters with whom they are most frequently paired by experts. We then measure agreement by computing the Jaccard similarity of the two sets.

Table 4 shows the average Jaccard similarity these top ten sets. The Jaccard scores are uniformly low, indicating that cosine similarity tends not to identify the same set of highly similar characters as experts. Interestingly, GPT-4 does not appear any more successful at identifying expert-aligned similar characters than the feature-based approaches, despite its success in identifying socially and narratologically similar characters.

We also examine how often the single character that experts compare most to a target character occurs within the target's top ten closest representations by cosine similarity. Table 4 shows the average success on this lenient measure.

Even with this easier measure, the expert benchmark is quite challenging. GPT-4 includes the expert top character in its top ten list only half of the time. The best feature-based representation, FanfictionNLP assertions, include it 69% of the time. Since this is a very lenient measure of success, this illustrates the large gaps that remain between similarity by computational representations of character, pretrained LLM understanding of character similarity, and expert evaluations.

7 Conclusion

We present AustenAlike, a three-part Jane Austen benchmark for evaluating multiple aspects of character similarity: narrative role similarity, social similarity, and expert judgments of character similarity drawn from prior scholarly analysis. We use AustenAlike to evaluate five computational representations of character built atop features extracted by pipelines for analyzing English literature.

We find that event- and assertion-based representations tend to capture character similarity better than quote-based representations. Overall, however, our results show how much work still remains to be done to improve computational representations of character: feature-based representations and GPT-4 alike struggle to place the expert-identified most similar character in their top ten lists of character similarity. We hope that by providing a multi-faceted benchmark with expert judgments, AustenAlike can guide future work on computational representations of character.

Limitations

We have evaluated five kinds of feature-based character representations across two systems. However, our approach has a number of limitations.

Noisy Character Data Both pipelines produce character clusters with some amount of inconsistency and error. In some cases, the pipelines failed to resolve multiple ways of referring to the same character (*Miss Tilney*, *Eleanor Tilney*). We post-process the output with an Austen-specific alias map; to extend our work to other works of literature, this post-processing step would need to be manually extended.

Missing Characters Both pipelines failed to extract features for some characters included in our benchmark. BookNLP failed to identify twelve characters and FanfictionNLP failed to identify four. This was most impactful in the siblings and parents subsets of the narrative roles benchmark.

Generalizability Our benchmark focuses on characters from the work of Jane Austen. As a result, it may favor methods of deriving computational representations that are trained on similar literary text. This may affect our comparison of FanfictionNLP and BookNLP quotes, as noted above.

Combining Character Data In this paper, we compare 5 different kinds of feature-based representations: events, assertions, modifiers, and quotations extracted from two pipelines. However, it would also be possible to combine these different sources of information about a character, and use them together. Future work could explore this kind of merged representation.

Ethics Statement

Our work does not involve any human data. The literary works we analyze are in the public domain. The computational resources involved in our experiments are also modest: all contextualized embeddings were extracted using less than 12 hours on a single Nvidia RTX A6000 GPU.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1202–1208, Nagoya, Japan. Asian Federation of Natural Language Processing.

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.

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, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *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)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Clément Besnier. 2020. History to myths: Social network analysis for comparison of stories over time. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9, Online. International Committee on Computational Linguistics.

Julian Brooke, Adam Hammond, and Timothy Baldwin. 2016. Bootstrapped text-level named entity recognition for literature. In *Annual Meeting of the Association for Computational Linguistics*.

Joseph Bullard and Cecilia Ovesdotter Alm. 2014. Computational analysis to explore authors' depiction of characters. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 11–16, Gothenburg, Sweden. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Liviu P. Dinu and Ana Sabina Uban. 2017. Finding a character's voice: Stylome classification on literary characters. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 78–82, Vancouver, Canada. Association for Computational Linguistics.

Anton Ehrmanntraut, Leonard Konle, and Fotis Jannidis. 2023. LLpro: A Literary Language Processing Pipeline for German Narrative Texts. In *Conference on Natural Language Processing*.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal. Association for Computational Linguistics.

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. 2016. Novel2vec: Characterising 19th century fiction via word embeddings. In *Irish Conference on Artificial Intelligence and Cognitive Science*.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

James Heldman. 1990. How wealthy is Mr. Darcy – Really? Pounds and Dollars in the World of *Pride and Prejudice*. *Persuasions*, 12:38–49.

Eric Holgate and Katrin Erk. 2021. "politeness, you simpleton!" retorted [MASK]: Masked prediction of literary characters. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 202–211, Groningen, The Netherlands (online). Association for Computational Linguistics.

Naoya Inoue, Charuta Pethe, Allen Kim, and Steven Skiena. 2022. Learning and evaluating character representations in novels. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1008–1019, Dublin, Ireland. Association for Computational Linguistics.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Labiba Jahan and Mark Finlayson. 2019. Character identification refined: A proposal. In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Prashant Jayannavar, Apoorv Agarwal, Melody Ju, and Owen Rambow. 2015. Validating literary theories using automatic social network extraction. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 32–41, Denver, Colorado, USA. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *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)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5).

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models.

John Lee and Chak Yan Yeung. 2012. Extracting networks of people and places from literary texts. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 209–218, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Bernhard Liebl and Manuel Burghardt. 2020. "shakespeare in the vectorian age" – an evaluation of different word embeddings and NLP parameters for the detection of shakespeare quotes. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 58–68, Online. International Committee on Computational Linguistics.

Smitha Milli and David Bamman. 2016. Beyond canonical texts: A computational analysis of fanfiction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2048–2053, Austin, Texas. Association for Computational Linguistics.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Daniele Pizzolli and Carlo Strapparava. 2019. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ina Roesiger and Simone Teufel. 2014. Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Katherine Toran. 2015. The economics of jane austen's world. *Persuasions On-Line*, 36:1817–1853.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Josep Valls-Vargas, Santiago Ontañón, and Jichen Zhu. 2021. Toward character role assignment for natural language stories. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2019. Are fictional voices distinguishable? classifying character voices in modern drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–34, Minneapolis, USA. Association for Computational Linguistics.

Shufan Wang and Mohit Iyyer. 2019. Casting Light on Invisible Cities: Computationally Engaging with Literary Criticism. In *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)*, pages 1291–1297, Minneapolis, Minnesota. Association for Computational Linguistics.

Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 2016. Extracting social networks from literary text with word embedding tools. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 18–25, Osaka, Japan. The COLING 2016 Organizing Committee.

Michael Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn Rosé. 2021. FanfictionNLP: A text processing pipeline for fanfiction. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 13–23, Virtual. Association for Computational Linguistics.

Category	Group	N
Rank	Nobility	2
	Titled Gentry	15
	Gentle	48
	New Gentle	5
	Clergy	12
	Military	13
	Profession	14
Wealth	£50	8
	£51-£250	7
	£251-£500	9
	£501-£1000	8
	£1001-£3000	6
	£3001+	5
Gender	Male	50
	Female	59
Age	< 18	8
	18-20	13
	21-24	16
	25-27	18
	28-30	12
	31-40	13
	41-50	19
	51+	10
Marital Status	Single	48
	Transitional	6
	Married	42
	Widowed	13

Table 5: Social Characteristics benchmark summary

A Further Details of Benchmark Construction

A.1 Social Benchmark

Rank To achieve a more even balance across groups, we partition untitled gentry into two groups: New Gentle, characters whose fathers were not gentlemen, and Gentle, representing more established gentry. We consolidate professional characters into three groups: a military group encompassing the army and navy; a professional group encompassing business, law, and farming; and a clergy group. This totals six categories: New Gentle, Gentle, Gentry, Military, Profession, Clergy, and Nobility.

Wealth Wealth for women is generally reported as a total sum, while men's fortunes are typically stated in terms of yearly income. We convert all figures to yearly incomes assuming the 5% yearly dividend standard during Austen's time (Toran, 2015).

Marital Status Marital status tends to remain stable until the end of each novel: although many single characters marry, most marriages take place in the last chapter.

A.2 Narrative Roles Benchmark

Heroines All Jane Austen novels involve young people finding marriage partners. Each novel has at least one protagonist who is an unmarried woman seeking a marriage partner. *Sense and Sensibility* focuses on a pair of sisters who both marry by the end of the novel; we treat both as protagonists/heroines. Heroines should be particularly easy to distinguish from other narrative roles since they are the main viewpoint characters in Austen's novels.

Heroes We use the term *hero* for the character that each protagonist marries at the novel's end.

Deceiver Each of Austen's novels features at least one character who lies in a way that sets key events in motion. Frequently, this character misrepresents himself to the heroine in a key way (Wickham in *Pride and Prejudice*; Willoughby in *Sense and Sensibility*); in other cases, the character lies to conceal an ulterior motive (William Elliot in *Persuasion*; Frank Churchill in *Emma*). In one case, this character spreads lies about the heroine herself (John Thorpe in *Northanger Abbey*).

Rivals and Wooers In each of the six novels, there is at least one character who serves as a rival, an alternate love interest for the hero. In all but one novel (*Sense & Sensibility*), there is a character who unsuccessfully courts the heroine; we refer to these characters as *wooers*.

Family roles Austen's novels are concerned with domestic settings and interactions within a relatively confined society. As a result, there are numerous family members. We look at two groups: parents and siblings. In the case of *Mansfield Park*, in which the heroine is raised in her uncle's family, we considered including her guardians but excluded them to be consistent with other mentors (Lady Russell in *Persuasion*) and temporary guardians (the Allens in *Northanger Abbey*).

B Further Details of GPT-4 Experiments

We run three experiments to extract character similarities from GPT-4: a top character experiment, a top character experiment with reasoning, and a top ten characters experiment. We run each experiment five times at temperature=0.2.

The prompts are shown below (full list of characters omitted for readability). *c* represents the name of the target character, and *cIndex* is that character's number in the list.

Top Character Prompt
Consider the following list of Jane Austen characters:
1. Anna Weston
2. Augusta Elton
...
108. Sir John Middleton
109. Thomas Palmer

Which character is *c* most similar to (other than *c*)? Respond with only a number. Do not choose *cIndex*.

Top Character with Reasoning Prompt
Consider the following list of Jane Austen characters:
1. Anna Weston
2. Augusta Elton
...
108. Sir John Middleton
109. Thomas Palmer

Which character is *c* most similar to (other than *c*)? Describe your reasoning and then reply with the number of the character. Do not choose *cIndex*.

Top Ten Characters Prompt
Consider the following list of Jane Austen characters:
1. Anna Weston
2. Augusta Elton
...
108. Sir John Middleton
109. Thomas Palmer

List the 10 characters that are most similar to *c* (other than *c*). Consider characters from all Austen novels. Reply with just their numbers. Do not choose *cIndex*.

C Further Results

C.1 Narrative Role Benchmark

Table 6 shows how often the most similar character is within the same narrative role set as the target character, with all books included. Table 2 excludes characters from the same book.

C.2 Social Benchmark

Table 7 shows how often the most similar character is within the same social role set as the target character, with all books included. Table 3 excludes characters from the same book.

C.3 Expert Benchmark

Tables 8 and 9 shows Pearson's ρ correlations between cosine similarity and expert pairing counts by novel, with characters from the same novel included and excluded respectively.

System	Hero	Heroine	Deceiver	Rival	Wooer	Parent	Sibling
FanfictionNLP Assertions	0.14	0.36	0.17	0	0	0.18	0.25
BookNLP Events	0.07	1	0.33	0.08	0.17	0.36	0
BookNLP Modifiers	0	0.86	0.33	0.25	0	0.27	0.18
BookNLP Quotes	0.07	0.64	0.33	0.17	0.25	0.09	0
FanfictionNLP Quotes	0.14	0.21	0	0.08	0	0.14	0.08
GPT-4	0.43	0.43	0.5	0	0	0.33	0.25
GPT-4 Reasoning	0.86	1	0.83	0.17	0.5	0.42	0.08

Table 6: Narrative Role Benchmark: Average occurrence of most similar character in same narrative role group by character representation. Characters from same novel are included.

System	Age	Gender	Income	Marital Status	Rank
FanfictionNLP Assertions	0.18	0.75	0.13	0.52	0.41
BookNLP Events	0.23	0.77	0.13	0.51	0.30
BookNLP Modifiers	0.21	0.78	0.07	0.46	0.19
BookNLP Quotes	0.09	0.58	0.15	0.40	0.34
FanfictionNLP Quotes	0.10	0.49	0.05	0.37	0.34
GPT-4	0.26	0.80	**0.21**	0.52	**0.42**
GPT-4 Reasoning	**0.32**	**0.98**	0.07	**0.58**	0.39

Table 7: Social Benchmark: average occurrence of most similar characters in the same social group by character representation. Characters from same novel are included.

Novel	*Emma*	*MP*	*NA*	*Pers.*	*P&P*	*S&S*	All
FanfictionNLP Assertions	0.30	0.38	0.29	0.27	0.23	0.28	0.29
BookNLP Events	0.44	0.44	0.43	0.31	0.37	0.43	**0.4**
BookNLP Modifiers	0.31	0.31	0.25	0.26	0.26	0.29	0.28
BookNLP Quotes	0.26	0.28	0.28	0.20	0.24	0.35	0.27
FanfictionNLP Quotes	0.21	0.20	0.15	0.11	0.10	0.11	0.15

Table 8: Expert Benchmark: Pearson's ρ correlation between cosine similarity and expert pairing count by character representation. Character pairs with no expert mentions are excluded.

Novel	*Emma*	*MP*	*NA*	*Pers.*	*P&P*	*S&S*	All
FanfictionNLP Assertions	0.3	0.38	0.34	0.23	0.33	0.27	
BookNLP Events	0.47	0.48	0.50	0.47	0.45	0.48	
BookNLP Modifiers	0.34	0.37	0.33	0.38	0.34	0.35	
BookNLP Quotes	0.16	0.11	0.27	0.13	0.28	0.27	
FanfictionNLP Quotes	-0.01	0.04	0.15	0.04	0.04	-0.07	

Table 9: Expert Benchmark: Pearson's ρ correlation between cosine similarity and expert pairing count by character representation. Characters from the same novel are excluded. Character pairs with no expert mentions are excluded.