Journal of Data Mining and Digital Humanities

Components of Character: Exploring the Computational Similarity of Austen's Characters

Carolyn Jane Anderson[1]

[1]Wellesley College, USA

Corresponding author: Carolyn Jane Anderson , carolyn.anderson@wellesley.edu

Abstract

Understanding characters is a key topic of interest in computational explorations of literary fiction. There are a growing number of tools available for extracting information about characters from text. Recent work proposes AustenAlike, a three-part benchmark for Jane Austen character similarity, to evaluate computational representations of character. I extend this exploration of feature-based representations of character with new experiments, including representations build from multiple features. I also explore a way of obtaining vector representations of characters using a large language model by generating short textual descriptions of characters using GPT-4, and then taking their embeddings. My results show surprisingly little advantage from combining multiple kinds of features, suggesting that the different features capture similar information to each other. I find that the LLM-generated textual descriptions perform best on expert benchmark, outperforming direct elicitation of character similarity judgments from the same model. However, the expert benchmark remains challenging across all representations explored, showing that there is much room for improvement in capturing the more nuanced aspects of character similarity that human readers identify.

Keywords

literary fiction; characters; vector similarity

I INTRODUCTION

There is a growing interest in using computational methods within literary analysis. Computational methods have become more accessible due to the increased availability of easy-to-run, inexpensive methods for analyzing text data. This includes both general purpose large language models, which can be used to annotate textual data [Gilardi et al., 2023, Alizadeh et al., 2025], and open-source systems designed specifically to extract information from literary fiction [Bamman et al., 2014, Yoder et al., 2021].

A key area of interest is in understanding characters from literary fiction. Pipelines for analyzing literary text can be used to identify character mentions and extract information about their attributes, event roles, and speech [Bamman et al., 2014, Yoder et al., 2021]. These features can then be used to analyze characters in different ways. Previous work has sought to extract and classify character types [Chambers and Jurafsky, 2009, Bamman et al., 2014, Jahan and Finlayson, 2019, Valls-Vargas et al., 2021, Stammbach et al., 2022], build networks of 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], and study character traits [Flekova and Gurevych, 2015, Pizzolli and Strapparava, 2019, Kim and Klinger, 2019].

James Morland from *Northanger Abbey*
Sibling to heroine and single 20-year-old clergyman 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

Yang and Anderson [2024] propose a benchmark that tests how well such systems extract information about characters using the works of Jane Austen as a test case. In their view, a system that produces useful character features is one whose features can be used to compare the similarity of characters. Since character similarity is a multi-faceted notion, their AustenAlike benchmark contains three parts: the first explores similarity by narratological role, the second by social characteristics, and the third takes a wisdom-of-the-crowd approach: it is based on character comparisons extracted from four decades of *Persuasions*, a journal published by the Jane Austen Society of North America. Figure 1 shows an example of how these three views of character similarity lead to different comparisons.

This paper extends Yang and Anderson [2024]'s in new directions. While Yang and Anderson [2024] compare features extracted by two systems, BookNLP [Bamman et al., 2014] and FanfictionNLP [Yoder et al., 2021], they construct character representations from only single features at a time. This paper explores character representations that integrate multiple kinds of character features. I also provide a clearer comparison with large language model methods by building vector representations from textual descriptions of characters produced by GPT-4. Finally, I explore the basis for GPT-4's judgments of character similarity in a qualitative coding analysis of GPT-4's reasoning.

My results show that though computational representations capture some broad social and narratological similarities, there remains a wide gap between the similarities they capture and the more nuanced similarities highlighted in the expert benchmark. The best feature-based representations exhibit only medium correlations with expert rankings of character similarity, and combining multiple kinds of features does not always improve character similarity measurements.

I find that the vector representations based on GPT-4-generated textual descriptions of characters align most closely with expert judgments. However, while these representations place the expert-identified most similar character in the top ten most similar characters 80% of the time, this is a very lenient measure of success, and the representations do not correlate with the full set of expert rankings any more strongly than feature-based representations. AustenAlike remains a challenging benchmark for evaluating computational representations of character: even for one of the best well-studied authors in English literature, much work remains to be done to achieve nuanced computational representations of literary characters.

II ANALYZING CHARACTERS

There is a growing interest in applying computational methods to analyze literary fiction, both in so called *distant reading* [Moretti, 2013] approaches that explore large collections of literature [Grayson et al., 2016, Jayannavar et al., 2015, Milli and Bamman, 2016] and in studies of individual authors and works [Agarwal et al., 2013, Wang and Iyyer, 2019, Liebl and Burghardt, 2020]. In this paper, I explore how well computational representations of character capture character similarity judgments. There has been much work exploring computational methods to understand literary characters. In this section, I discuss previous research applying computational methods to literary characters and contextualize this work.

2.1 What Makes a Character?

Literary scholars have proposed many definitions of what it means to be a character over the years. Some theorists treat characters similarly to real-life people, in that they have psychological interiority and existence before and after the events of the novel [Bradley, 1965]. Others argue that characters are inextricable from the text: from the events they participate in or even the words themselves [Weinsheimer, 1979]. Structuralist approaches seek to classify characters by their event participation: for instance, Propp defined seven character types in Russian folktales based on their "spheres of action" [Propp et al., 1975], and Frye distinguished between characters whose actions advance or thwart the protagonist's goals [Frye, 1957]. Others argue that characters consist of their stable personal traits [Chatman, 1978].

Beyond specific character taxonomies, theorists have also debated the appropriateness of classifying character complexity. Forster distinguished "flat" characters, who are static, from "round" characters, who experience change or growth during the work [Forster, 1927]. Ewen proposed multiple dimensions of character complexity, including development and "penetration into inner life" [Ewen, 1971, Rimmon-Kenan, 2002].

Because of the complexity of defining characters, the AustenAlike benchmark explores three different aspects of character similarity. The first two explore the kinds of stable traits that Chatman [1978] argues define characters, while the last benchmark includes all aspects of characters that experts may choose to discuss. Unlike some previous computational work, however, the benchmark does not explore how characters change or develop over the course of a novel [Iyyer et al., 2016, Chaturvedi et al., 2016]. Because the vectors representations are built from averaging all features extracted for a given character, regardless of where in the text they occur, they provide a whole-novel portrait of the character.

2.2 Computational character analysis

The rise of computational methods in literary analysis has led to a wealth of research on different aspects of characters.

Classifying character types. Structuralist treatments of literature explore common character types that recur throughout genres of fiction. Valls-Vargas et al. [2021] build directly on Propp's classification of character types from Russian folktales. Bamman et al. [2014]'s early work with Bayesian mixed effects models of characters' events, possessions, and modifiers, on the other hand, uses a more expansive notion of character *persona*. Other work exploring character classification includes Klenner et al. [2021]'s work on negative characters; Stammbach et al. [2022]'s work on heroes, victims, and villains; and Jahan and Finlayson [2019]'s narratologically-grounded framework for character identification.

Mapping character relationships. There is also an extensive body of work exploring the relationships between characters [Lee and Yeung, 2012, Jayannavar et al., 2015, Agarwal et al., 2013, Wohlgenannt et al., 2016]. Elson et al. [2010]'s early and influential approach used quotation features to identify character relationships. Labatut and Bost [2019] provide a more recent survey of different approaches. Some work also explores how these relationships change over the course of a narrative [Chaturvedi et al., 2016, Iyyer et al., 2016].

Understanding character depictions. Some existing work seeks to understand decisions in how characters are depicted in literature. Bullard and Ovesdotter Alm [2014] explore authorial decisions in representing characters. Besnier [2020] applies social network analysis to understand how character depictions evolve over retellings.

2.3 Extracting character information

Existing computational work on literary fiction span a variety of methodologies and focus areas. However, most existing work shares 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, I focus specifically on features related to literary characters.

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

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 [Yoder et al., 2021].

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 I 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.

2.4 Character representations

Once character features are extracted, they can be used to build computational representations of characters. My work explores vector representations of characters built from extracted features. Other researchers have attempted to learn character representations directly. Grayson et al. [2016]'s early work on word embeddings showed the simple vector representations learned from 19th-century works of fiction provide insight into characters. More recently, Holgate and Erk [2021] proposed learning vector representations using masked entity prediction as a training objective.

This paper extends Yang and Anderson [2024], which proposes AustenAlike, a benchmark for evaluating computational representations of characters written by Jane Austen. Inoue et al.

[2022] also propose a benchmark for evaluating character representations; their work is complementary in that it takes a broad multi-author, multi-task perspective, while AustenAlike explores three different aspects of character similarity for a single author's characters.

III THE AUSTENALIKE BENCHMARK

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. Yang and Anderson [2024]'s AustenAlike benchmark of character similarity contains three parts, each of which explores a different aspect of literary characterhood.[1] The first part of the benchmark explores similarity in social characteristics; the second, similarity in narratological role, and the third focuses the more nuanced and varied aspects of similarity that expert readers are sensitive to.

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*. It includes all named characters who speak more than once, except those who die in the first chapter.[2]

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 world of the novel.

The first part of AustenAlike explores similarity in social characteristics. It focuses on 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.[3] A summary of the social categories and the size of each group is shown in Table 1.

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.[4]

To achieve a more even balance across groups, AustenAlike partitions untitled gentry into two groups: New Gentle, characters whose fathers were not gentlemen, and Gentle, representing more established gentry. Professional characters are consolidated 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.

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

[2]Named characters excluded by these criteria include the senior Mr. Dashwood and Lady Elliot, who both die in the first chapter of their novels, and Mrs. Bates and Sarah (Sally) Morland, who are named but never speak.

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

[4]Female peers retain their titles if they marry someone of inferior rank, but this circumstance does not arise in Austen.

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; Yang and Anderson [2024] base their incomes on estimates from Heldman [1990] and Toran [2015]. Wealth for women is typically reported in Austen's novels as a total sum, while men's fortunes are stated in terms of yearly income. AustenAlike uses yearly incomes for all characters, converting lump sums using the 5% yearly dividend standard in Austen's time [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, AustenAlike estimates from the ages of their family members.

Marital status. Marital status is a key social characteristic of Austen characters. 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. AustenAlike contains four groupings: Single, Married, Widowed, and Transitional, a group comprising the handful of characters whose marital status changes before the last chapter of the novel.

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 1: Social Characteristics benchmark summary

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. AustenAlike explores seven narrative roles that commonly occur across Austen's novels. These groupings are shown in Table 2.

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; AustenAlike treats both as protagonists/heroines. Heroines are particularly easy to distinguish from other narrative roles since they are the main viewpoint characters in Austen's novels.

Heroes. AustenAlike uses 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

Heroines:	Emma Woodhouse, Elizabeth Bennet, Elinor Dashwood, Marianne Dashwood, Fanny Price, Catherine Morland, Anne Elliot
Heroes:	George Knightley, Colonel Brandon, Edward Ferrars, Edmund Bertram, Henry Tilney, Frederick Wentworth, Fitzwilliam Darcy
Deceivers:	John Thorpe, George Wickham, John Willoughby, William Elliott, Henry Crawford, Frank Churchill
Rivals:	Caroline Bingley, Lucy Steele, Louisa Musgrove, Mary Crawford, Harriet Smith, Jane Fairfax
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, John Dashwood, Margaret Dashwood
Parents:	Mr. Bennet, Sir Walter Elliot, Lieutenant Price, Mr. Woodhouse, Mrs. Bennet, Mrs. Dashwood, Mrs. Price, Mrs. Morland, Lady Betram, Mrs. Norris, Sir Thomas Bertram, Lady Russell

Table 2: Narrative Roles benchmark summary

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 all but one novel (*Northanger Abbey*), there is a female character who serves as an alternate love interest for the hero (*rivals*). Similarly, in every novel but *Sense & Sensibility*, there is at least one male character who unsuccessfully courts the heroine (*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. AustenAlike looks at two groups: parents and siblings. The parents group includes some non-parent characters who serve a guardian role; for instance, in *Mansfield Park*, the heroine is raised in her uncle's family, so her aunts and uncle are included in addition to her birth parents. Similarly, AustenAlike includes Lady Russell from *Persuasion*, who mentors Anne Elliot after her mother's death, and the Allens in *Northanger Abbey*, who serve as Catherine Morland's temporary guardians during her stay in Bath.

3.3 Wisdom-of-the-Experts Character Pairs

The final part of AustenAlike explores characters who have been identified as similar by literary scholars. It uses a wisdom-of-the-crowds approach, but with an expert crowd: authors of articles published in *Persuasions*, the Jane Austen Society of North America's peer-reviewed journal.

Pairwise character comparisons were manually extracted from 43 volumes of *Persuasions*. All instances of a similarity or shared property of two characters discussed in an article were extracted. When an article mentioned a similarity between more than two characters, all pairings from the set were added. The resulting dataset contains 5740 character comparison pairs.

The identified comparisons are diverse, encompassing traits from the 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.

IV BUILDING COMPUTATIONAL REPRESENTATIONS OF CHARACTER

Yang and Anderson [2024] build computational representations of character from the features extracted using two literary pipelines: BookNLP and FanfictionNLP. They construct vector representations for each feature type: for BookNLP, events, quotes, and modifiers; for FanfictionNLP, quotes and assertions. My extended experiments also consider combinations of features. I explore representations based on all FanfictionNLP features; all BookNLP features, and the combined features of both pipelines. I also explore representations based on textual descriptions of characters generated by GPT-4, for a new LLM-based baseline.

4.1 Character Mentions

The first step in extracting character features is to identify and disambiguate character mentions. I use each pipeline to identify character mentions, perform coreference resolution, and aggregate character mentions. I then merge and filter character clusters using a handwritten alias map for Austen character names.

This step is necessary because both pipelines mainly cluster characters by proper names, but Austen characters are often referred to differently by various characters. For instance, Elizabeth Bennet from *Pride and Prejudice* is called *Lizzy* by her family members, *Miss Eliza Bennet* by Caroline Bingley, and either *Miss Bennet* or *Miss Elizabeth Bennet* by acquaintances, depending on whether her elder sister is present.

4.2 Feature Extraction

Next, the two pipelines are used to extract the features associated with each character. The output is a JSON file with lists of each kind of feature associated with a character ID. BookNLP identifies the events that characters participate in and their role (agent or patient); modifiers that are used to describe the character; and quoted speech of the character. FanfictionNLP extracts quotations and assertions. Assertions are descriptions of characters, like modifiers, but they can consist of spans of text rather than single words.

The quantity of each feature extracted varies according to the prominence of each character. For a main character like Elinor Dashwood from *Sense and Sensibility*, BookNLP identifies 125 modifiers, 2051 events, and 435 quotes, and FanfictionNLP identifies 1076 assertions and 501 quotes. For a minor character like her youngest sister, Margaret Dashwood, BookNLP identifies only 3 modifiers, 11 quotes, and 35 events, and FanfictionNLP identifies only 14 assertions and 2 quotes.

4.3 Feature Embeddings

Once the features have been extracted, I use them to construct vector representations of the characters. Vector representations are commonly used for semantic representation because they can capture multiple dimensions of meaning: each element of a vector can correspond to a different aspect of meaning (or in this case, character). Moreover, similarity can be measured as distance in the vector space, allowing for easy similarity comparisons.

To construct these vector representations, I retrieve contextualized embeddings for each kind of feature from a neural network model. Embeddings are vector representations that come from the activations of the neural network layers, typically, from the last layer before the prediction layer. These embeddings are vector representations of a word's meaning within the context of the piece of text that is used as input to the model.

For events and modifiers, which are single words, I retrieve a contextualized embedding of the word in its context (the sentence surrounding it) using T5 (11B) [Raffel et al., 2020], a large language model that performs well on embedding tasks. For quotes and assertions, I retrieve sentence embeddings using NV-Embed (7.85B) [Lee et al., 2024], which is better-suited for embedding longer spans of text.

Next, each kind of feature embedding is centered by subtracting the mean of all embeddings for the feature. This reduces noise and makes the comparisons more numerically stable.

4.4 Constructing Representations

For each feature and character, I construct a character representation by averaging the embeddings of the character's features. For events, I average the character's agent events and patient events separately and concatenate the vectors.

Yang and Anderson [2024] use this technique to produce 5 representations per character: an assertion vector, a modifier vector, an event vector, and two quote vectors (one per pipeline).

My extended experiments explore combinations of features. I construct an all-features representation by concatenating all 5 vectors produced by Yang and Anderson [2024]. For characters where the pipeline failed to identify any instances of a particular feature, I use a dummy vector of all zeros. This is an appropriate neutral choice since all the vectors have been centered around the mean.

I also explore vectors composed of all BookNLP features; all Fanfiction features; all quotes; and modifiers plus assertions. These last two allow us to quantify the overlap between the features extracted from each pipeline. If the combined quote vectors do not outperform the single pipeline quote vectors, this would tell us that each pipeline extracts the same information from quotations.

4.5 GPT-4 comparisons

Yang and Anderson [2024] 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 can assume that GPT-4's training data contains all six novels, as well as many web pages discussing them.

Yang and Anderson [2024] extracted character similarities using three prompting approaches. They asked GPT-4 to select the most similar character from a list of all benchmark characters; to select the most similar character and explain its choice; and to choose the ten most similar characters from a list of all benchmark characters.

I add an additional large language model baseline by using GPT-4 to generate textual descriptions of each character, and then using NV-Embed (7.85B) to produce embeddings from each description. An example interaction is shown in Appendix A. This method produces vector representations of characters that are more directly comparable to the feature-based representations. I also present an analysis of the kinds of information that GPT-4 uses in its reasoning-based ranking versus its textual descriptions of characters (Section IX).

V EVALUATING CHARACTER SIMILARITY

The AustenAlike benchmark contains three subsets 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 the characters that are most similar to a target character are the same as those with whom experts pair them.

Vector similarity. The representations capture different aspects of the characters as different dimensions in a vector space. Similar character vectors will be closer to each other in this space. I adopt a commonly used distance metric to assess the similarity of the character representations: cosine similarity.

Cosine similarity is the cosine of the angle between two vectors. It is computed by taking the dot product of two vectors and normalizing by the product of their lengths. Consequently, cosine similarity does not capture differences in magnitude.

1. **Cosine Similarity**: S_c(A,B) = $\frac{A \cdot B}{||A||||B||}$

Cosine similarity ranges from -1, which indicates orthagonal vectors, to 1, for completely similar vectors.

5.1 Grouping evaluation

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

In-group Cosine Similarity. I explore whether characters are more similar to characters within their group than those outside of their group. I 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.

I call this *in/out-group cosine similarity difference*.

2. **In-group Similarity**: $mean_{c \in RG} \frac{(t \cdot c)}{(||t||||c||)}$ where t is the target character and c is a character in their narrative role group RG.
3. **Out-group Similarity**: $mean_{c \notin RG} \frac{(t \cdot c)}{(||t||||c||)}$ where t is the target character and c is a character outside of their narrative role group RG.

Most Similar Character. I also ask whether very similar characters come from the same groups. I 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, I measure the extent to which the cosine similarities of each kind of representation align with the expert-identified pairs using three metrics:

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

Ranking similarity. Literary experts may be more interested in identifying highly similar characters than in quantifying degrees of dissimilarity. I 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

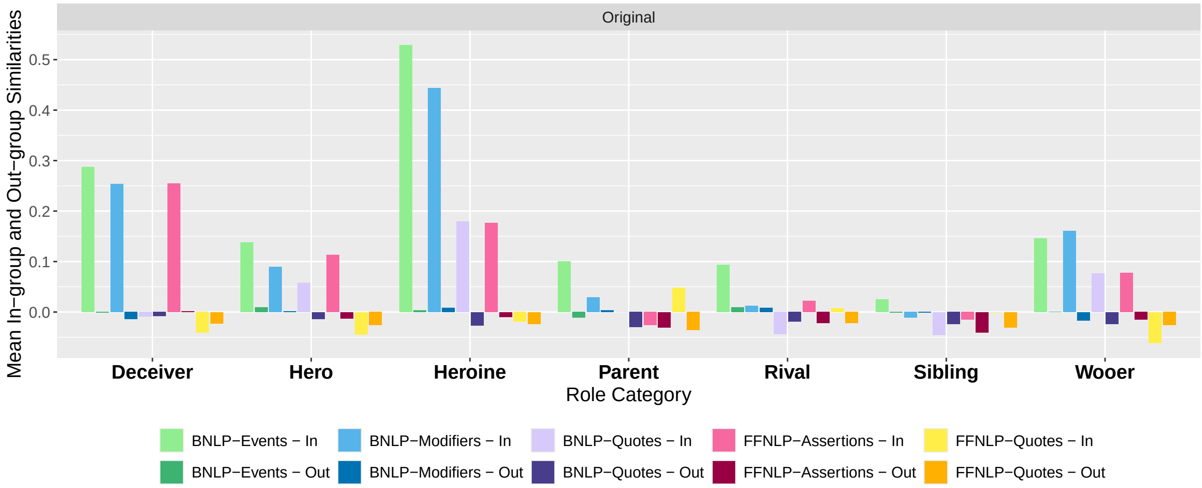

Figure 2: Narrative Role Benchmark: Mean cosine similarities between same-group characters and other characters, Yang and Anderson [2024] representations.

union. If the two lists are completely different, their Jaccard similarity is 0; if they mostly agree, it is close to 1. This is a more appropriate metric than alternatives for ranked data, like the Kendall correlation coefficient, because the rankings may contain different characters.

Top character in ten-most similar. Finally, I focus on the top expert-identified pairings. I 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. Given that there are 109 characters in AustenAlike, there is a 4.7% chance for the model to succeed by random guessing.

VI 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.

My extended experiments also explore how much information is captured about narrative role similarity by each pipeline; by all features extracted by pipelines; and by GPT-4-generated descriptions of the characters.

6.1 Are same-role characters more similar?

One way of exploring the narrative roles benchmark is to ask whether characters are more similar to characters that share their narrative role than they are to characters who do not play this role.

To measure this, I compare the average cosine similarity of representations within a narrative role group with their average cosine similarity to non-group members. I compute the in-group and out-group scores for each character in a target role group and average them.

6.1.1 Previous findings

Figure 2 plots the cosine similarity for characters within the same narrative role group compared to characters outside of the group for Yang and Anderson [2024]'s representations.

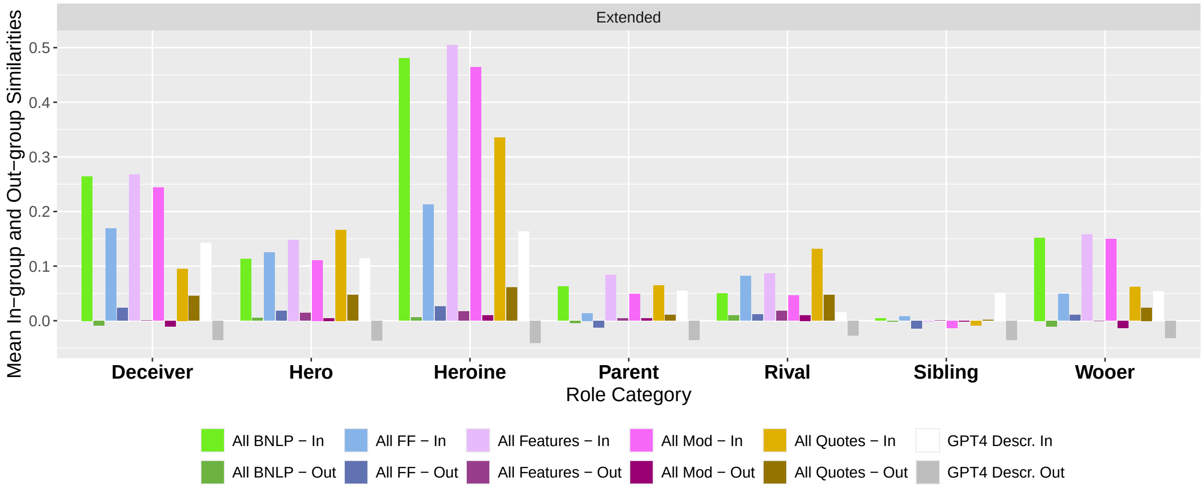

Figure 3: Narrative Role Benchmark: Mean cosine similarities between same-group characters and other characters, extended representations.

Yang and Anderson [2024] previously found 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.

6.1.2 Extended results

Figure 2 plots the cosine similarity for characters within the same narrative role group compared to characters outside of the group for the new representations, including the GPT-4-generated textual descriptions.

For most categories, the best-performing combined representation is the one derived from all features, following by the combined BookNLP representation. However, neither of these representations capture narratological role similarity more strongly than the BookNLP event representations, suggesting that the other features do not add much more information.

There is little difference between the combined modifier representations and the BookNLP modifier-based representations, suggesting that the FanfictionNLP assertions do not add much information. On the other hand, the combined quote representations do better than either original quote-based representation at capturing in-group similarity for Heroines, Rivals, and Heroes, suggesting that each pipeline captures different quotes with relevant narratological information.

The GPT-4 textual description-based representations do not capture in-group similarity more than the feature-based representations. However, they do capture out-group dissimilarity more strongly than many of the feature-based representations.

6.2 Is the most similar character from the same group?

In addition to exploring overall differences between in-group and out-group members, I also look at whether the single most similar character to a target character belongs to the same narrative role group. For each character in a role group, I count how often the character with highest cosine similarity belongs to their 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

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
GPT-4 Textual Description	0.21	0.5	0.33	0	0.17	**0.42**	**0.38**
All BookNLP	0.21	**1**	0.42	**0.25**	**0.5**	0.27	0
All FanfictionNLP	0.14	0.43	0.17	0.08	0.08	0.09	0.13
All Features	0.21	**1**	0.42	**0.25**	**0.5**	0.27	0.08
All Modifiers	0.07	**1**	0.33	**0.25**	0.08	0.18	0.17
All Quotes	0	0.71	0.08	0.17	0	0.18	0.08

Table 3: Narrative Role Benchmark: Average occurrence of most similar character in same narrative role group, all characters.

the novel. Qualitatively, I observe that GPT-4 is also somewhat biased towards drawing comparisons with characters in the same novel. I therefore present results both with (Table 3) and without (Table 4) characters from the same novel.

6.2.1 *Previous findings*

The strength of narratological similarity varies greatly by role: heroines are more often similar to heroines for all of Yang and Anderson [2024]'s single feature representations, while other groups have lower rates of same-group membership. This is true regardless of whether same-novel characters are excluded.

Among the single feature representations, 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.

Excluding same-novel characters tends to help feature-based representations more than the GPT-4-derived rankings. We see 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 not reliably better than the feature-based representations: it never selects wooers and parents as most similar to other wooers and parents.

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

6.2.2 *Extended findings*

When we consider combined representations, we see that the all-feature and all BookNLP representations capture the most similarity by narrative role, with the highest scores in two of the

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	0.67	**1**	0.75	0	0	0.43	**0.5**
GPT-4 Reasoning	**1**	**1**	**0.83**	0.17	0.5	0.45	0
GPT-4 Textual Description	0.43	0.71	0.67	0	0.33	**0.5**	0.46
All BookNLP	0.22	**1**	0.45	0.22	**0.55**	0.32	0
All FanfictionNLP	0	0.57	0.33	0.17	0.17	0.09	0.08
All Features	0.25	**1**	0.45	0.22	**0.55**	0.25	0.10
All Modifiers	0.1	**1**	0.36	0.22	0.11	0.19	0.17
All Quotes	0	0.89	0.17	**0.33**	0	0.18	0.08

Table 4: Narrative Role Benchmark: Average occurrence of most similar character in same narrative role group, characters from same novel excluded.

seven categories.

The combined quote representations again perform better on this metric than either the FanfictionNLP quote representations or the BookNLP quote representations, suggesting that each pipeline captures different quote information. The combined modifier representations also do better than the individual assertion and modifier representations in most categories.

Surprisingly, however, the all-feature representations do not reliably capture stronger narratological similarity than the combined BookNLP representations; this suggests that the FanfictionNLP features do not contribute much new information.

The GPT-4 generated descriptions do not seem to align as strongly with narratological similarity as GPT-4's reasoning-prompted rankings, though they have high scores in the Parent and Sibling categories, where feature-based representations do not seem sensitive to narratological similarity.

VII 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, event- and quote-based representations may also reflect similarities based on these characteristics.

7.1 How similar are characters with shared social characteristics?

I first explore whether characters within the same group in each of the social categories are most similar to each other by comparing in-group and out-group cosine similarities.

7.1.1 Previous findings

Figure 4 plots the average cosine similarity for characters within the same social group compared with non-group members for Yang and Anderson [2024]'s original representations.

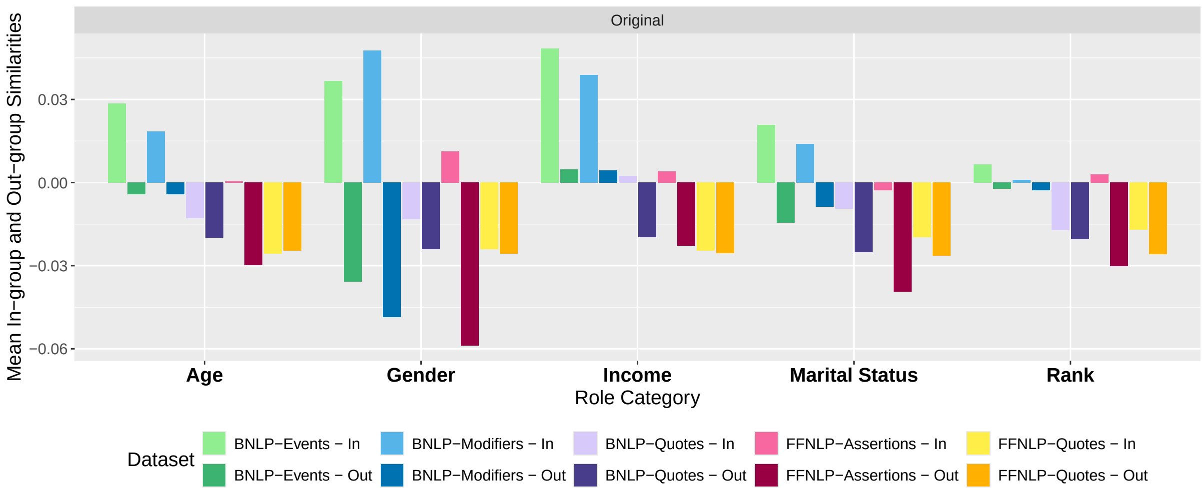

Figure 4: Social Benchmark: average differences in cosine similarity between same-group characters and other characters by social role group, Yang and Anderson [2024] representations.

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.

7.1.2 Extended results

Figure 5 plots the average cosine similarity for characters within the same social group compared with non-group members for the extended representations. The combined BookNLP and all-features representations capture in-group similarity strongly in most categories.

The combined assertion and modifier representations capture similarity well in the Gender and Income categories, with particularly strong differences between same-gender and different-gender characters. Combining these features is very effective because the Fanfiction assertions show strong out-group dissimilarities, while the BookNLP modifiers capture strong in-group similarities.

The GPT-4 textual description representations also capture social similarity fairly well. They show some of the strongest dissimilarities for non-group members, but their in-group similarities are weaker than the combined feature-based representations in most categories (though stronger than many of Yang and Anderson [2024]'s single feature representations).

7.2 Is the most similar character from the same group?

I also focus more narrowly on the top-most similar character. Table 5 shows how often the character with the highest cosine similarity to the target character occurs in the same social group for all characters, while Table 6 shows the same results with same-novel characters excluded.

7.2.1 Previous findings

Yang and Anderson [2024] find that 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.

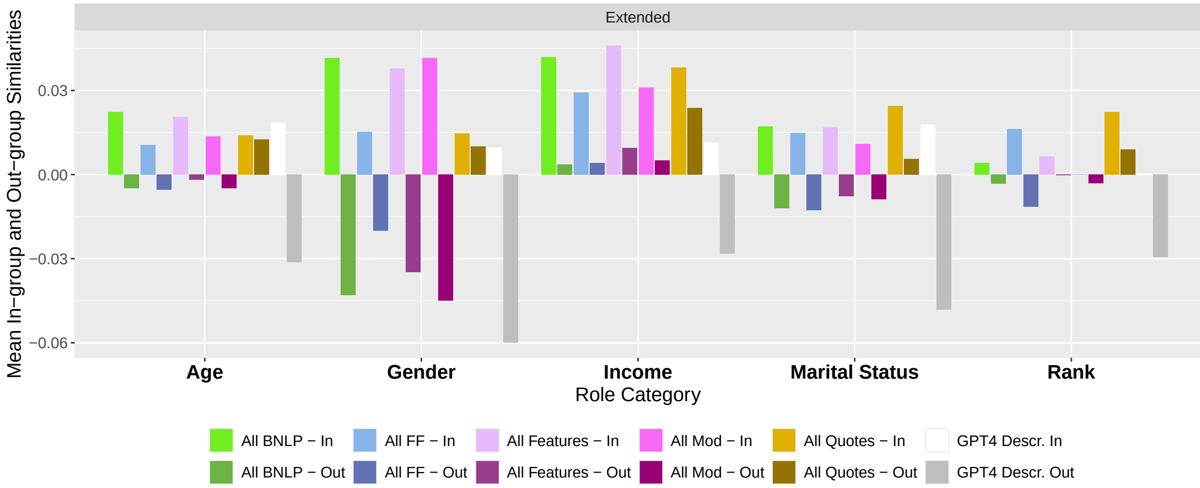

Figure 5: Social Benchmark: average differences in cosine similarity between same-group characters and other characters by social role group, extended representations.

They also find that GPT-4's similarity judgments align with social characteristics more strongly than any of their single-feature representations. Quote-based representations do not seem to capture similarity by social characteristics as well as the other feature-based representations in most categories.

7.2.2 Extended results

The combined representations show little advantage over single-feature representations. The all-feature and all BookNLP representations seem to capture only as much about social similarity as the BookNLP modifier and event representations.

On the other hand, the embeddings based on GPT-4's textual descriptions of characters capture a lot of information about social similarity. They show comparable scores to the GPT-4 reasoning comparisons in all but one category (Age) when same-novel characters are omitted, and higher similarity by Income.

VIII EXPERT BENCHMARK

The last portion of the AustenAlike benchmark explores more nuanced aspects of character similarity using an expert wisdom-of-the-crowd approach. The expert benchmark contains counts of character similarity pairings that occur in *Persuasions*. I 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.

If computational representations actually capture how similar two characters are, we expect their cosine similarity to correlate strongly with how often experts discuss the two characters as similar.

8.1 Does cosine similarity correlate with expert judgments?

I examine how well computational character representations align with expert judgments by measuring the correlation between expert character pairings and cosine similarity. High quality computational representations should produce higher cosine similarity between the characters that are more frequently deemed similar by experts. Table 7 shows the correlation between

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
GPT-4 Textual Description	0.25	0.69	0.14	**0.65**	**0.54**
All BookNLP	0.23	0.79	0.08	0.47	0.22
All FanfictionNLP	0.13	0.63	0.13	0.46	0.37
All Features	0.22	0.79	0.12	0.47	0.21
All Modifiers	0.20	0.74	0.06	0.44	0.20
All Quotes	0.13	0.48	0.09	0.51	0.30

Table 5: Social Benchmark: average occurrence of most similar characters in the same social group by character representation. Characters from same novel are included.

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	0.25	0.91	0.12	0.56	0.31
GPT-4 Reasoning	**0.34**	0.98	0.03	0.60	**0.38**
GPT-4 Textual Description	0.25	**0.99**	0.07	**0.62**	**0.38**
All BookNLP	0.25	0.80	0.07	0.48	0.22
All FanfictionNLP	0.12	0.73	0.07	0.41	0.32
All Features	0.22	0.79	0.07	0.47	0.19
All Modifiers	0.21	0.77	0.03	0.45	0.19
All Quotes	0.15	0.47	0.04	0.49	0.18

Table 6: Social Benchmark: average occurrence of most similar characters in the same social group by character representation. Characters from same novel are excluded.

expert pairing counts and cosine similarity for each of the computational representations, including both Yang and Anderson [2024]'s single-feature representations and my combined representations.

8.1.1 Previous findings

Yang and Anderson [2024] observe moderate positive correlations between the cosine similarity of character representations and the number of expert similarity pairings. Of their single-feature representations, the BookNLP events correlate most strongly with expert pairings, while the FanfictionNLP quotes correlate less strongly. This converges with their social and narratological similarity findings. However, none of the feature-based representations are strongly correlated with expert judgments.

Dataset	Pearson's ρ	Jaccard Similarity	Top in Top 10
FanfictionNLP Assertions	0.29	**0.03**	0.69
BookNLP Events	**0.40**	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	-
GPT-4 Textual Description	0.36	**0.03**	**0.80**
All BookNLP	0.37	0.02	0.40
All FanfictionNLP	0.29	**0.03**	0.65
All Features	**0.40**	0.02	0.39
All Modifiers	0.31	0.02	0.28
All Quotes	0.32	**0.03**	0.52

Table 7: Expert Benchmark: measures of alignment between expert pairing counts and computational similarity.

8.1.2 Extended results

Yang and Anderson [2024] find that none of the single feature representations correlate strongly with the expert pairing counts. I find that combining multiple kinds of features does not improve over single feature representations. This is disappointing, as each kind of feature might in principle capture different information about character similarity. However, we find that the all-feature representations have the same medium correlation with expert judgments as the BookNLP events alone (ρ=0.4). Figure 6 shows an example of how two representations correlate with the twelve characters that experts compare most frequently to Elizabeth Bennet from *Pride and Prejudice*.

The combined BookNLP and combined FanfictionNLP features do not outperform the best single feature representations from their respective pipelines. This suggests that adding less informative features does not improve the representations, and can even hurt them: the combined BookNLP features have a slightly weaker correlation than the BookNLP events alone. Although combining multiple feature types does not seem to help, the combined quote representations and combined modifier representations do perform better than their single-feature components. This shows that combining information extracted by different tools can be helpful.

Yang and Anderson [2024]'s previous language model-based results could not be directly correlated with the expert pairing counts, since they took the form of top character selections or lists, rather than similarity scores for each character pair. The GPT-4 textual description-based vector representations, on the other hand, can be directly correlated like the feature-based representations, providing a more straightforward comparison. I find that the GPT-4 textual descriptions do not correlate any more strongly with expert judgments than the feature-based representations.

8.2 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. Consequently, I look at two additional measures of agreement for the most similar characters.

For each character, I retrieve the ten characters with the highest cosine similarity, and the ten characters with whom they are most frequently paired by experts. I then measure agreement by computing the Jaccard similarity of the two sets; Figure 7 shows an example of a Jaccard similarity comparison between two top ten lists.

We can also look at how often the top expert-selected character occurs in a target character's top ten list. This is the most lenient measure of alignment; the representation essentially gets ten chances to guess the top most similar character.

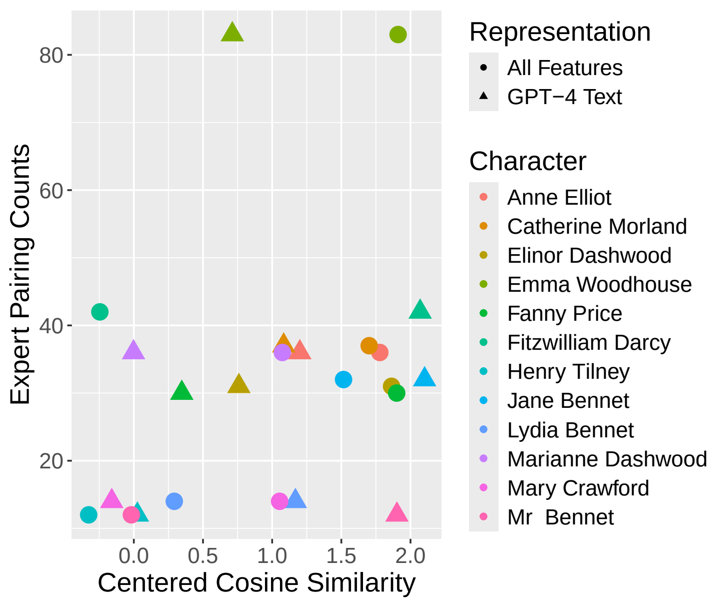

Figure 6: Correlation between expert pairings and the cosine similarity of select representations for Elizabeth Bennet.

8.2.1 Previous findings

Table 7 shows the average Jaccard similarity the top ten sets for Yang and Anderson [2024]'s single-feature representations. If the highest ten cosine similarities are for the same characters that experts pair most frequently, then the Jaccard scores should be high. However, we observe uniformly low Jaccard scores, indicating that cosine similarity tends not to identify the same set of highly similar characters as experts. 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.

Yang and Anderson [2024] also explore 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. They find that the expert benchmark is challenging even by this very lenient metric. 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.

8.2.2 Extended results

Neither the combined feature representations nor the GPT-4 generated textual descriptions achieve high Jaccard similarity scores when compared to the expert rankings. This shows that computational representations and experts do not tend to agree on the top ten most similar characters to a target character.

There is no observable advantage from combining feature-based representations; however, we also do not see any improvement for the GPT-4 embeddings over feature-based representations. This illustrates the large gap that remains between similarity by computational representations of character, pretrained LLM understanding of character similarity, and expert evaluations.

When we explore the most lenient measure, whether the top expert-paired character occurs in the list of ten closest characters, the GPT-4 textual descriptions have a strong advantage over other computational representations.

When asked to produce textual descriptions of each character, the resulting GPT-4 representations have the top character in their top ten most similar representations 80% of the time.

This is a marked improvement over asking GPT-4 to compare characters directly, and it is much closer to expert judgments than the combined feature-based representations. However, it is important to note that this is an extremely lenient measure of success. Overall, the findings show that there is still much room for improving the correlation between expert judgments of character similarity and computational representations of characters.

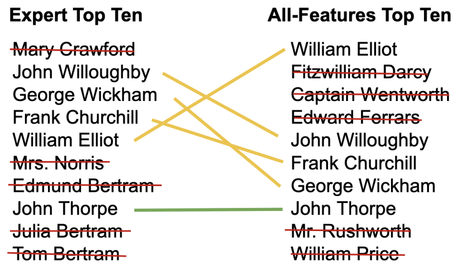

Figure 7: Alignment between top matches for Henry Crawford from *Mansfield Park*. Yellow lines indicate reranking; green indicates matches; red lines indicate characters that appear on just one list.

IX HOW DOES GPT-4 REASON ABOUT CHARACTERS?

In this section, I explore the reasons that GPT-4 gives for its character comparisons. Above we have seen that GPT-4's judgments of Austen character similarity correlate only moderately well with expert judgments. There could be several reasons for the lack of alignment, including reliance on different aspects of character similarity and incorrect or incomplete knowledge of Jane Austen's characters.

I explore the quality and nature of the evidence that GPT-4 uses for its similarity judgments by manually coding the reasoning chains that GPT-4 produces when prompted to provide reasoning. The qualitative coding considers the five aspects of social similarity explored in the Social benchmark; two aspects of narratological similarity, narrative role and growth over the course of the novel; and multiple aspects of similarity that experts might identify, including personality traits, actions, values, appearance, and goals. This produced twelve codes for aspects of character similarity.

In addition, I identify three kinds of errors that the model makes: misidentification of the target character; comparison to a character not included in the AustenAlike set; and factually incorrect information about character properties. This resulted in a total of 15 codes.

Figure 8 shows the frequency of each coding category in GPT-4's reasoning responses. Most reasoning chains contain some comparison of the personality traits of the two characters. GPT-4 also frequently references characters' values. For instance, it describes Elizabeth Bennet and Anne Elliot as similar because each "values personal integrity," while Mrs. Dashwood and Isabella Knightley "place a high value on family." GPT-4 also draws comparisons of characters' goals, particularly for characters who are focused on social advancement. Action comparisons mainly focus on characters who are very social, such as Mr. Weston and Sir John Middleton, who "enjoy hosting social gatherings."

Comparisons of social characteristics are relatively less frequent. The most commonly mentioned are rank and income, but these are also one of the most frequent sources of factual errors. GPT-4 describes Miss Bates as a widow, even though she is unmarried in the novel *Emma*; Dr. Grant is described as a medical man rather than a clergyman; and Catherine Morland is described as similar to Harriet Smith in social standing, when in fact Harriet Smith is illegitimate and has no known fortune, while Catherine Morland's family is comfortably affluent.

GPT-4 makes mistakes in character attributes about 10% of the time. While social attributes are the most common kinds of mistakes, GPT-4 also makes factual errors about the narrative roles of characters, such as claiming that Colonel Fitzwlliam plays a "significant role" in the life of the heroine in *Pride & Prejudice*.

More rarely, GPT-4 makes mistakes about the identity of the characters. Fanny Price's father is conflated with her brother in all five attempts, and Elinor Dashwood is conflated with Mrs. John Dashwood once. GPT-4 also struggles with the Bateses from *Emma*, alternating between references to Miss and Mrs. Bates within a single paragraph.

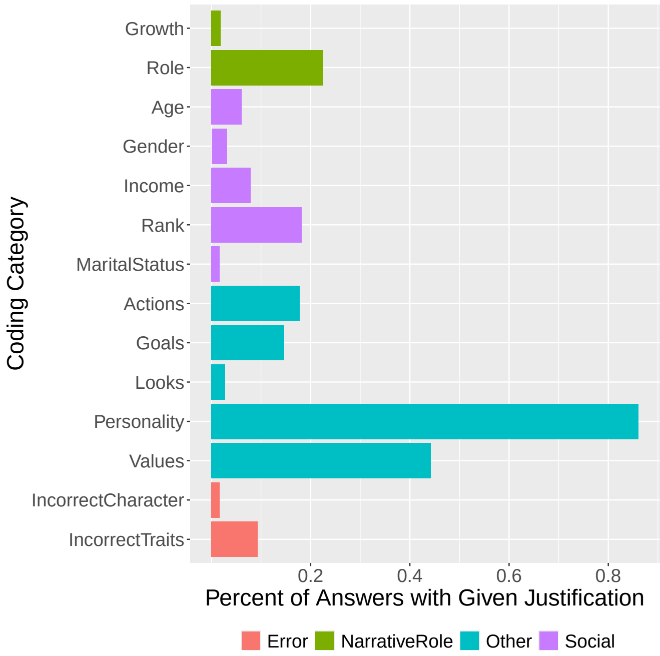

Figure 8: Qualitative Coding of GPT-4 Reasoning

Overall, the qualitative exploration of GPT-4's reasoning shows that GPT-4 largely relies on personality traits and values to assess character similarity. Given that these are nuanced aspects that experts may also discuss, the lack of alignment between expert judgments and GPT-4's similarity judgments is surprising. However, GPT-4's 10% error rate in assessing character traits likely affects the quality of its comparisons.

X CONCLUSION

I extend Yang and Anderson [2024]'s evaluation of feature-based computational representations of characters from the works of Jane Austen. I revisit Yang and Anderson [2024]'s three-part AustenAlike benchmark to evaluate vector representations of characters built from multiple kinds of features. My results show little consistent advantage from combining features, suggesting that each kind of feature captures similar information about characters. On the other hand, there are cases where combining the same kind of feature extracted by multiple pipelines improves performance.

I provide a clearer comparison with large language model capabilities. Rather than prompting the model to rank character similarity directly, I generate textual descriptions of each character using GPT-4 and embed them using the same performant textual embedding model used for feature-based representations. This allows for a more direct comparison with feature-based representations. My results reveal that GPT-4's textual representations outperform other methods in the challenging expert benchmark: they place the expert-identified top character in the top ten most similar characters 80% of the time. However, even these representations do not display a strong correlation with expert judgments overall, showing that there is considerable room for improvement in computational representations of character.

LIMITATIONS

My approach has a number of limitations:

Noisy Character Data. Pipelines like FanfictionNLP and BookNLP produce character clusters with some amount of inconsistency and error. In some cases, pipelines fail to resolve multiple ways of referring to the same character (*Miss Tilney*, *Eleanor Tilney*). I post-process the output with a manually-created Austen-specific alias map. To extend this work to other works of literature, this post-processing step would need to be recreated.

Missing Characters. Both pipelines failed to extract features for some characters included in the AustenAlike benchmark. BookNLP failed to identify twelve characters and FanfictionNLP failed to identify four. For combined feature-based representations, I use a dummy vector of all zeros when a certain kind of feature is missing for a character. This is appropriate because I center all embeddings on the mean. I exclude characters from analysis for whom no representation could be built. This was most impactful in the siblings and parents subsets of the narrative roles benchmark.

Generalizability. Although I find that GPT-4 textual description-based representations perform best on the challenging expert benchmark, it is important to note that their performance likely depends on the fact that Austen's work is extremely well-represented online. We can expect much weaker performance of large language model methods for less prominent authors; for more obscure works of fiction, feature-based representations might perform best. This is an important direction for future work to explore.

References

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

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 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, October 2013. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I13-1171.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *J. Comput. Soc. Sci.*, 8(1):17, 2025.

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 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, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1005. URL https://aclanthology.org/E14-1005.

David Bamman, Ted Underwood, and Noah A. Smith. 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, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1035. URL https://aclanthology.org/P14-1035.

David Bamman, Sejal Popat, and Sheng Shen. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1220. URL https://aclanthology.org/N19-1220.

David Bamman, Olivia Lewke, and Anya Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.6.

Clément Besnier. 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, December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.latechclfl-1.1.

A.C. Bradley. *Shakespearean Tragedy*. Macmillan, 1965.

Julian Brooke, Adam Hammond, and Timothy Baldwin. Bootstrapped text-level named entity recognition for literature. In *Annual Meeting of the Association for Computational Linguistics*, 2016. URL https://api.semanticscholar.org/CorpusID:14818594.

Joseph Bullard and Cecilia Ovesdotter Alm. 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, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0902. URL https://aclanthology.org/W14-0902.

Nathanael Chambers and Dan Jurafsky. 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, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1068.

Seymour Chatman. *Story and Discourse*. Cornell University Press, 1978.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. Modeling evolving relationships between characters in literary novels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016. doi: 10.1609/aaai.v30i1.10358. URL https://ojs.aaai.org/index.php/AAAI/article/view/10358.

Liviu P. Dinu and Ana Sabina Uban. 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, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2210. URL https://aclanthology.org/W17-2210.

Anton Ehrmanntraut, Leonard Konle, and Fotis Jannidis. LLpro: A Literary Language Processing Pipeline for German Narrative Texts. In *Conference on Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:267770726.

David Elson, Nicholas Dames, and Kathleen McKeown. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1015.

Joseph Ewen. The theory of character in narrative fiction. *Hasifrut*, 3:1–30, 1971.

Lucie Flekova and Iryna Gurevych. 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1208. URL https://aclanthology.org/D15-1208.

E.M. Forster. *Aspects of the Novel*. Harcourt, 1927.

Northrop Frye. *Anatomy of Criticism: Four Essays*. Princeton University Press, 1957.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2305016120.

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. Novel2vec: Characterising 19th century fiction via word embeddings. In *Irish Conference on Artificial Intelligence and Cognitive Science*, 09 2016.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 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, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1129.

James Heldman. How wealthy is Mr. Darcy – Really? Pounds and Dollars in the World of *Pride and Prejudice*. *Persuasions*, 12:38–49, 1990. ISSN 1.

Eric Holgate and Katrin Erk. "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), June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.iwcs-1.19.

Naoya Inoue, Charuta Pethe, Allen Kim, and Steven Skiena. Learning and evaluating character representations in novels. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1008–1019, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.81. URL https://aclanthology.org/2022.findings-acl.81.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1180. URL https://aclanthology.org/N16-1180.

Labiba Jahan and Mark Finlayson. Character identification refined: A proposal. In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2402. URL https://aclanthology.org/W19-2402.

Prashant Jayannavar, Apoorv Agarwal, Melody Ju, and Owen Rambow. 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, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0704. URL https://aclanthology.org/W15-0704.

Evgeny Kim and Roman Klinger. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1067. URL https://aclanthology.org/N19-1067.

Manfred Klenner, Anne Göhring, and Sophia Conrad. Getting hold of villains and other rogues. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 435–439, Reykjavik, Iceland (Online), May 31–2 June 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.49.

Vincent Labatut and Xavier Bost. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5), sep 2019. ISSN 0360-0300. doi: 10.1145/3344548. URL https://doi.org/10.1145/3344548.

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

John Lee and Chak Yan Yeung. 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, November 2012. Faculty of Computer Science, Universitas Indonesia. URL https://aclanthology.org/Y12-1022.

Bernhard Liebl and Manuel Burghardt. "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, December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.latechclfl-1.7.

Smitha Milli and David Bamman. 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, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1218. URL https://aclanthology.org/D16-1218.

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

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 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, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1044.

Daniele Pizzolli and Carlo Strapparava. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3411. URL https://aclanthology.org/W19-3411.

V.I.A. Propp, V. Propp, L. Scott, and L.A. Wagner. *Morphology of the Folktale: Second Edition*. American Folklore Society Bibliographical and Special Series. University of Texas Press, 1975. ISBN 9780292783911. URL https://books.google.com/books?id=cyc7AQAAIAAJ.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Shlomith Rimmon-Kenan. *Narrative Fiction: Contemporary Poetics*. Routledge, 2nd edition, 2002.

Ina Roesiger and Simone Teufel. 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, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-3006. URL https://aclanthology.org/E14-3006.

Matthew Sims, Jong Ho Park, and David Bamman. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1353. URL https://aclanthology.org/P19-1353.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wnu-1.6. URL https://aclanthology.org/2022.wnu-1.6.

Katherine Toran. The Economics of Jane Austen's World. *Persuasions On-Line*, 36:1817–1853, 2015. ISSN 1.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1088. URL https://aclanthology.org/D15-1088.

Josep Valls-Vargas, Santiago Ontañón, and Jichen Zhu. Toward character role assignment for natural language stories. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2021. URL https://api.semanticscholar.org/CorpusID:3656347.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2504. URL https://aclanthology.org/W19-2504.

Shufan Wang and Mohit Iyyer. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1130. URL https://aclanthology.org/N19-1130.

Joel C. Weinsheimer. Theory of character: Emma. *Poetics Today*, 1:185, 1979.

Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 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, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-4004.

Funing Yang and Carolyn Jane Anderson. Evaluating computational representations of character: An Austen character similarity benchmark. In Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni, editors, *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 17–30, Miami, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4dh-1.3. URL https://aclanthology.org/2024.nlp4dh-1.3.

Michael Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn Rosé. FanfictionNLP: A text processing pipeline for fanfiction. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 13–23, Virtual, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nuse-1.2. URL https://aclanthology.org/2021.nuse-1.2.

A FURTHER DETAILS OF GPT-4 EXPERIMENTS

I 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. I 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

System message: *Consider the following list of Jane Austen characters:*

1. Anna Weston
2. Augusta Elton
...
108. Sir John Middleton
109. Thomas Palmer

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

Example Assistant Response: *15*

Top Character with Reasoning Prompt

System message: *Consider the following list of Jane Austen characters:*

1. Anna Weston
2. Augusta Elton
...
108. Sir John Middleton
109. Thomas Palmer

User: *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.*

Example Assistant Response: *Augusta Elton is a character from Jane Austen's "Emma." She is known for her social climbing, pretentiousness, and somewhat overbearing nature. Given these traits, the character most similar to Augusta Elton from the list, excluding herself, would be Caroline Bingley from "Pride and Prejudice."*

Caroline Bingley is also known for her social ambitions, her desire to marry well, and her condescending attitude towards those she considers beneath her. Both characters exhibit a sense of superiority and a keen interest in securing advantageous social connections.

Therefore, the character most similar to Augusta Elton is Caroline Bingley.

The number of the character is 69.

Top Ten Characters Prompt

System message: *Consider the following list of Jane Austen characters:*

1. Anna Weston
2. Augusta Elton
...
108. Sir John Middleton
109. Thomas Palmer

User: *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.*

Example Assistant Response: *3, 4, 5, 8, 15, 18, 20, 24, 25, 26*

Textual Description Prompt

System message: *Consider the works of Jane Austen.*

User: *Write a couple of paragraphs describing the character of Catherine Morland from Northanger Abbey.*

Example Assistant Response: *Catherine Morland, the protagonist of Jane Austen's "Northanger Abbey," is a charming and relatable character whose innocence and imagination shape much of her journey throughout the novel. She's initially introduced as an unremarkable young girl with a love for gothic novels, which fuels her vibrant imagination and romantic ideals. Catherine's upbringing in a rural, unassuming environment contributes to her naivety and lack of worldliness. However, what she lacks in sophistication, she makes up for with her sincerity, amiability, and strong moral character.*

As she ventures from her quiet life at Fullerton to the bustling city of Bath, Catherine's experiences broaden her perspective on the world and human nature. Her friendship with the vivacious Isabella Thorpe exposes her to both the delights and deceits of society. Meanwhile, her infatuation with the enigmatic Henry Tilney provides a more genuine and morally sound contrast to Isabella's manipulative tendencies. Catherine's stay at Northanger Abbey, with its dark, mysterious allure, serves as the pinnacle of her imaginative fantasies. Yet, it is also where she confronts the limitations of her overactive imagination and grows in emotional maturity. By the end of the novel, Catherine emerges more perceptive and judicious, having navigated the complexities of social relationships and personal growth. Through her earnestness and eventual self-awareness, Austen crafts Catherine Morland into a character of endearing authenticity and depth.