Personality Understanding of Fictional Characters during Book Reading

Mo Yu[1*] **Jiangnan Li**[2*] **Shunyu Yao**[3] **Wenjie Pang**[1]
Xiaochen Zhou[4] **Xiao Zhou**[1] **Fandong Meng**[1] **Jie Zhou**[1]
[1]Pattern Recognition Center, WeChat AI [2]Institute of Information Engineering, Chinese Academy of Sciences
[3]Princeton University [4]Syracuse University
moyumyu@global.tencent.com lijiangnan@iie.ac.cn

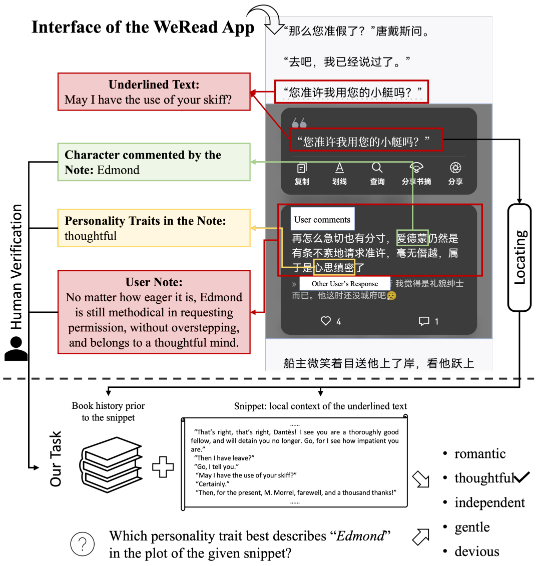
Abstract

Comprehending characters' personalities is a crucial aspect of story reading. As readers engage with a story, their understanding of a character evolves based on new events and information; and multiple fine-grained aspects of personalities can be perceived. This leads to a natural problem of **situated and fine-grained** personality understanding. The problem has not been studied in the NLP field, primarily due to the lack of appropriate datasets mimicking the process of book reading. We present the first labeled dataset PERSONET for this problem. Our novel annotation strategy involves annotating user notes from online reading apps as a proxy for the original books. Experiments and human studies indicate that our dataset construction is both efficient and accurate; and our task heavily relies on long-term context to achieve accurate predictions for both machines and humans.[1]

1 Introduction

Lively fictional characters with distinct personalities are the first drive of the plotline developments. The authors shape the characters with various personality types, which distinguish a character from others and explain the motivations and behaviors of the characters. As a reverse process, the readers grasp the characters' personalities during reading a story, which helps to understand the logics of a plot and predict its future developments.

The NLP community has also recognized the values of personality understanding; and conducted studies (Bamman et al., 2013; Flekova and Gurevych, 2015; Sang et al., 2022b) along this direction. In the problem definition of the existing tasks, the input is an entire book. By construction, they ask for the general impression of character personalities. Also for this reason, they only focus on coarse-grained personality types, *e.g.*, the four coarse MBTI types (Myers and McCaulley, 1988).


Interface of the WeRead App
Underlined Text:
Character commented by the Note: Edmond
Personality Traits in the Note: thoughtful
User Note:
Human Verification
复制
划线
查询
分享书摘
分享
User comments
Other User's Response
4
1
Locating
Book history prior to the snippet
Snippet: local context of the underlined text
Our Task
romantic
thoughtful✓
independent
gentle
devious

May I have the use of your skiff?

No matter how eager it is, Edmond is still methodical in requesting permission, without overstepping, and belongs to a thoughtful mind.

"那么您准假了？"唐戴斯问。

"去吧，我已经说过了。"

"您准许我用您的小艇吗？"

"您准许我用您的小艇吗？"

再怎么急切也有分寸，爱德蒙仍然是有条不紊地请求准许，毫无僭越，属于是心思缜密了

» 我觉得是礼貌绅士而已。他这时还没城府吧😥

船主微笑着目送他上了岸，看他跃上

......
"That's right, that's right, Dantès! I see you are a thoroughly good fellow, and will detain you no longer. Go, for I see how impatient you are."
"Then I have leave?"
"Go, I tell you."
"May I have the use of your skiff?"
"Certainly."
"Then, for the present, M. Morrel, farewell, and a thousand thanks!"
......

Which personality trait best describes "*Edmond*" in the plot of the given snippet?

Figure 1: Illustration of the WeRead App interface enabling our dataset construction (user IDs masked for privacy); and an example of our task, *situated personality prediction*.

To make a personality prediction task more practical and useful, we consider two important aspects of character understanding in real life that have not been studied in the context of machine reading. First, we aim at predicting **fine-grained** personality types, with an exhaustive vocabulary of personality traits as the targets. Second and more importantly, we study the continuous-process nature of story reading — As people read, they form dynamic impressions of the characters and plots. We name this process **situated** comprehension. Specific to personality understanding, a character may have multi-faced personalities. In a certain point of the story, the character's behaviors can reflect one of them when faced the situation and events at the time. Human readers have the ability to use their knowledge of what has happened so far (*i.e.,* the **history**) to understand the character in the current situation. We hence propose to study **situated personality prediction**, which differs from the static prediction problem studied before.

While the aforementioned two problems are practical and common in real life, they create new

[*]Authors contributed equally to this paper.
[1]Available at https://github.com/Gorov/personet_acl23.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14784–14802
July 9-14, 2023 ©2023 Association for Computational Linguistics

challenges in dataset creation, especically the latter. To accurately mimic the human reading process, annotators would need to read entire books, which is not practical due to the significant time required.

We overcome this annotation difficulty and create a large-scale dataset for personality prediction in the reading processes. To achieve this goal, we propose a novel annotation strategy that utilizes publicly available book notes. Recent online reading apps such as WeRead[2] (shown in Figure 1) allow users to take notes while reading a book. As users read, they can add notes at the current reading position. These notes are linked to specific text in the book, which is highlighted with a dotted underline, referred to as **underlined texts** throughout this paper. This mechanism ensures that the notes accurately reflect the thoughts the user had while reading the surrounding text of the underlined text.

Based on this resource, we propose our strategy of annotating *user notes as a delegate of the book reading process*. Specific to our task of personality prediction, this corresponds to (1) identifying if a user note discusses the personality trait of a character; and (2) associating the trait label to the context at the note location. We take user notes that contain at least a character name and a personality trait word, and ask human annotators to confirm if the trait is a modifier of the character in the note text (*i.e.*, the user note mentions that the character has the trait). The verified notes serve as nature labels of character personalities reflected by the surroundings of the underlined texts. By using this approach, we collect labeled data that only requires annotators to read short notes, without the need for knowledge about the books themselves.

With our new strategy, we create our situated personality prediction dataset, PERSONET, that contains ∼32K instances from 33 books in the classic literature domain. We prove that our annotation strategy is *efficient* as each worker only requires a median of 23.7s to finish one sample. The whole annotation process costed in total $2,400 and 471.8 hours (distributed to 20 working days by 11 annotators). It is also *accurate* evidenced by both the over 88% inter-annotator agreement. In addition, we make the dataset bilingual in both English and Chinese with automatic book sentence alignment and manual character alignment.

We conduct experiments with our dataset in two folds. First, we develop various improvement over the standard pre-trained models, including enabling the models to use different types of long contexts, equipping the models with oracle history trait information, and task-oriented unsupervised training. Second, we conduct extensive human studies with people who have read the books (*i.e.*, with the knowledge of the book history) and not. Our results show that (1) our task is challenging as humans with knowledge of book history can achieve more than 70% accuracy, compared to the best model accuracy of ∼45%; (2) our task heavily requires the long context modeling, as introducing characters' history information significantly improves the model accuracy; and humans without the book history can only perform on par with models.

We make the following contributions:

- A dataset, PERSONET, that is the first benchmark of *situated reading comprehension* and of *fine-grained personality prediction* on books. We prove that our dataset is a valid assessment to long context understanding for both machines and humans without significant shortcuts.
- A novel dataset creation approach for book comprehension problems based on user notes, which is efficient and accurate.
- Task-oriented unsupervised training and character history enhancement methods that improve on our task, providing insights to future work.

2 Related Work

Story book understanding has been recognized as a challenging and rewarding direction (Piper et al., 2021; Sang et al., 2022a). Many evaluation benchmarks on various narrative understanding tasks have been developed, such as plot structure analysis (Saldias and Roy, 2020; Papalampidi et al., 2019), question answering (Richardson et al., 2013; Kočiskỳ et al., 2018; Xu et al., 2022), summarization (Ladhak et al., 2020; Kryściński et al., 2021; Chen et al., 2022), character identification and relationship extraction (Elson et al., 2010; Elsner, 2012; Elangovan and Eisenstein, 2015; Iyyer et al., 2016; Chaturvedi et al., 2016; Kim and Klinger, 2019; Sang et al., 2022c).

All of the prior work takes the entire long story as input to a model for predictions. None of them considers the *situated* reading process like ours.

Existing strategies of dataset construction over **long stories** fall into the following categories: •A straightforward way is **to have labelers read the entire stories**. Because of the huge efforts, it only

[2] https://weread.qq.com/.

works for short stories for young children (Xu et al., 2022) or simpler tasks like coref (Bamman et al., 2019), NER (Bamman, 2020) and quotation attribution (Vishnubhotla et al., 2022). **•Using the book summaries as proxy** of the original stories, *e.g.*, the creation of book-level question answering task (Kočiskỳ et al., 2018). The created data usually only covers abstract and major events in the book, as shown in (Mou et al., 2021). Thus the types of comprehension skills that can be assessed with this strategy are limited. **•Exploiting Web resources created by fans or experts**. Flekova and Gurevych (2015) used fans' rated MBTI types to create a classification task for book characters; Ladhak et al. (2020); Kryściński et al. (2021) created a book chapter summarization task based on summaries on the English learning websites; and Thai et al. (2022) created a book retrieval task based on quotes in literature reviews. The drawback of this strategy is that the tasks can be supported are limited by the available resources. **•Automatically created cloze tests** is a traditional strategy. With specifically designed techniques, the clozes can be made resolvable only with global context, *e.g.*, (Rae et al., 2020; Sang et al., 2022c; Yu et al., 2022). The problem of this method is that the created datasets usually have unclear assessment goals.

The limitations of these strategies make them insufficient to create datasets for our task of situated personality understanding.

3 Problem Definition

Our PERSONET is the first task on situated prediction of characters' personality traits in book contexts. That is, we aim to predict the traits reflected by a local snippet of book, given all the previous book content as the background (Figure 1).

Formally, we consider a local book snippet $\mathcal{S}^{(i)} = \{s_{k_1^{(i)}}, s_{k_2^{(i)}}, ..., s_{k_J^{(i)}}\}$. Each $s_{k_j^{(i)}}$ is a sentence from the book, with $k_j^{(i)}$ the absolute position of the sentence in the book. Each $\mathcal{S}$ in our task depicts a character's personality. Therefore, it is associated with a pair of (c, t), where c is a character name or alias and t is the personality trait of c that reflected by $\mathcal{S}$. Note that different pairs may share a same snippet. Our task is then to predict:

$$P(y = t|c, \mathcal{S}^{(i)}, \mathcal{H}^{(i)} = s_{1:k_1^{(i)}}), \qquad (1)$$

where $s_{1:k_1^{(i)}}$ refers to all the sentences before $\mathcal{S}^{(i)}$ in the book. We split the books into training, dev and test sets, so that the evaluation characters are unseen during training. For evaluation, we adopt a multi-choice setting. For each instance, we sampled 4 negative candidates, two from the top-20 most frequent traits and the rest from the whole list. Combining the negative choices with t, we have a candidate set $\mathcal{T}$. Our data thus form a tuple $(\mathcal{S}, \mathcal{H} = s_{1:t_k^{(i)}}, c, t, \mathcal{T})$.

4 Our PERSONET Dataset

4.1 Data Source

List of Personality Traits Following previous work (Shuster et al., 2019), we use the list of 818 English personality traits from the MIT Ideonomy project.[3] We translate the traits into Chinese with Youdao dictionary,[4] then ask human annotators to select all the translated meanings that depict personality in Chinese. There are 499 English traits and 565 Chinese traits left that are bilingually aligned.

Books and Notes We collect 100 public books available in the Gutenberg project. For each book, we find all its Chinese-translated versions on the WeRead app and collect all their user notes. We kept notes that (1) contain any traits, (2) contain any person names[5] and (3) with lengths less than 100 words (relatively shorter notes can improve human annotation efficiency). We filtered out the books with less than 100 notes left, leaving 33 books and 194 of their Chinese translations. These books have 110,114 notes that contain 140,268 traits in total.

Note Clustering It is common for multiple users to comment on the same part of a book, discussing the same character. When these users express similar opinions about a character, it leads to duplication. To remove this duplication for data annotating, we group the notes according to their positions, defined as the center token offset of its associated snippet $\mathcal{S}^{(i)}$ (*i.e.*, its underlined text). Notes with distances smaller than 100 tokens are grouped, leading to 27,678 note clusters. We take the unique traits within each cluster for human labeling, which corresponds to 113,026 samples as defined in Section 3. The notes are anonymized for human annotation.

Extension of the Snippets The lengths of underlined texts can vary significantly, which means they may not always provide a representative context

[3] http://ideonomy.mit.edu/essays/traits.html.
[4] https://cidian.youdao.com/.
[5] We use Spacy (zh_core_web_lg) for NER.

for reflecting a character's personality, particularly when the texts are very short. We address this issue by extending each $\mathcal{S}^{(i)}$ from the underlined text to a window of 480 tokens. This window is generally large enough to encompass a scene and ensures that the context relevant to the user note is included. The reason for choosing this window size is that it is typically longer than one page displayed by the WeRead App (as shown in Figure 1) — users often write notes on the same page while reading the context, rather than flipping through previous or subsequent pages.[6]

4.2 Dataset Construction

Our dataset construction consists of two major steps: (1) human annotation of user notes; (2) projection of labeled data from Chinese to English. In addition, we show that (3) our data construction strategy enables to build an accurate note classifier for automatic weakly-supervised data labeling.

Step 1: Human Annotation This step requires the annotators to read each user note, and determine if it discusses the personality of a character. We present the annotators with notes that contain at least one trait word in our vocabulary in Section 4.1. The note is paired with the *underlined book content*, which is optional to read, if they think the note itself is ambiguous. The annotators are then asked to (1) judge if the note is indeed about a certain character's trait; then (2) marked the target character name with the trait from the note.

The first step takes most of the human efforts. We wrote concrete guidelines (Figure 4 in Appendix A) for the decision making process. The annotators are citizens in China who have received at least high school education (which, in the Chinese education system, covers most of the general knowledge about classic literature). Therefore it is more convenient for them to work in Chinese; and Figure 4 lists both the original guidelines in Chinese and their English translations.

Our annotation interface (with English translations) is shown in Appendix A. Once the annotators confirm that the given trait word describes some characters, they are required to annotate the character name by dragging from the note text. If not, the character name will be left empty.

Step 2: Bilingual Projection The human annotation step has created a personality prediction dataset in Chinese. Next we project the data to English. Since the same English book may have multiple translated books in Chinese, their labeled data scattered. By projecting the labeled data to English books, the book version is unified and the annotations become dense.

According to Section 3, to create an English version of our dataset, we only need to project the traits t, the characters c and the snippets (positions) $\mathcal{S}$. The trait t is already from a bilingual vocabulary, so we only need to focus on the latter two.

•**Book Alignment** The projection of $\mathcal{S}$ is equivalent to finding each labeled instance's sentence positions in the English book, which is essentially a sentence alignment problem. Specifically, we sentencize the books firstly with Spacy; then utilize the *vecalign* (Thompson and Koehn, 2019) toolkit to achieve sentence alignments among books. We represent each sentence with the default number (10) of its consecutive sentences, and employ the multilingual sentence embedding *LASER*[7] to embed the sentences. After that, *vecalign* performans sentence alignments with dynamic programming based on the embeddings.

With bilingual sentence alignment, the position of each labeled instance can be mapped to the corresponding position in the English book, *i.e.*, $\mathcal{S}_{\text{en}} = \{a(s)|\forall s \in \mathcal{S}\}$, where $a(s)$ refers to aligned position of the Chinese sentence s in the English book. For most of the $\mathcal{S}$ in our dataset, we can find consecutive $\mathcal{S}_{\text{en}}$ as the aligned results. There are a few instances mapped to empty. We excluded these cases in our English dataset. There are also a few instances mapped to inconsecutive English sentences, sometimes in a wide range. For this situation, we take the median position of the mapped English sentences and include the consecutive context in a window as the projection.

•**Character Name Projection** We manually project the list of 377 frequent (appear >10 times in our labeled data) character names to English. We askeds two annotators to find the English names of these characters; and resolved all the inconsistency after they complete their own annotation jobs.

Step 3: Weakly-Supervised Data Our method reduces the problem of annotation over books to annotation over notes. This makes it possible to build a note classifier for automatic data augmentation.

We collect another 65,521 notes from the same book collection that consists of at least one trait

[6] Therefore, if the context is not covered by the window, it suggests that the note should not be taken on that page.

[7] https://github.com/facebookresearch/LASER.

word and one person name. By pairing traits with names within the same notes, we create 154,030 examples. Then we train a binary *roberta-wwm-ext* (Cui et al., 2020) classifier over our human-labeled data to determine if the note discusses the character's trait, *i.e.*, the same task in human annotation but without the need of marking target characters. For each human annotated note, if the note is recognized as describing a trait of a character, it is used as a positive example. For those labeled as irrelevant to character traits, *i.e.*, no characters are annotated, we denote them as negative examples. Cross-validation on the human-labeled data shows that our classifier is accurate: 91.1% and 90.2% on the dev and test set. Applying our classifier to these unlabeled examples, we recognize 31,346 examples as describing characters' traits.

4.3 Quality of the Annotated Data

This section proves the accuracy of our data construction method via human study.

Correctness of Book Notes First of all, we need to prove that the user notes are indeed an accurate delegate of books. That is, when a note mentions a personality of a character, whether it is highly consistent to what the book content reflects.

This study requires annotators who have read the books to make the correct judgement. We selected four books with two annotators who have read and are familiar with them. Each annotator labeled two books. We sampled in total 431 notes from these books. The annotators are required to judge if the note is accurate about the character or not. We present the corresponding underlined content along with the note, so that the annotators can identify which part the note is commenting. The results in Table 1 show that 89.1% of the notes are accurate understanding of the books. There are 9.7% *ambiguous* examples, meaning the annotated traits are implied by the current place of the books, but might be falsified later, *e.g.*, the authors may intend to mislead the readers to create surprisal or tension. These ambiguous labels give valid data for our problem of *dynamic personality prediction*, according to our description at the beginning of Section 1 and Eq. (1).

Accuracy of Human Labels Next, we proved that our annotation process leads to accurate human labels. This accuracy is verified in two ways. First, we compute the inter-annotator agreement, with a duplicated set of 3,000 notes during annotation. 88.67% of the duplicated samples receive consistent labels. The Cohen's Kappa (Cohen, 1960) is 0.849, which indicates nearly perfect agreement (Viera et al., 2005). Second, as shown in the Step 3 in Section 4.2, a fairly accurate note classifier can be trained on our human-labeled data (91.1% and 90.2% accuracy on dev and test).

correct	89.1
ambiguous	9.7
incorrect	1.2

Table 1: Notes (%) that consistently reflect the character personalities in the stories.

perfect match	187
high overlap	7
low overlap	1
no match	5

Table 2: Human study: quality of bilingual alignment.

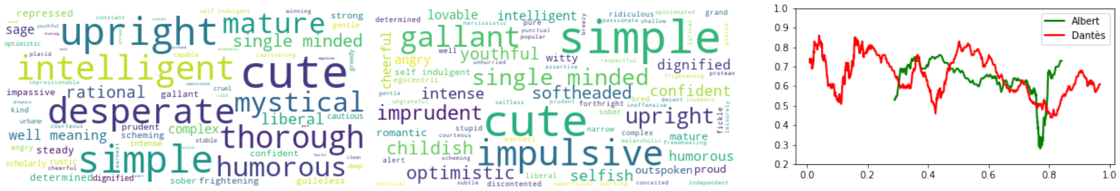
Set			**#Instance**	
	#Books	**#Chars**	**English**	**Chinese**
Train	17	148	18,190	18,273
Weakly Sup			26,244	26,331
Development	6	54	3,745	3,751
Test	10	72	3,624	3,647
Total	33	274	51,803	52,002

Table 3: Data statistics of our PersoNet dataset, including the number of unique books, characters and the numbers of instances in English and Chinese datasets. The weakly supervised data is used for training only.

Both tests confirm the accuracy of our annotation strategy. Considering the relevance of the book notes (Table 1), this gives an estimation of overall accuracy around 87.6∼89.1%. The two endpoints are computed with inter-annotator agreement and classifier accuracy accordingly. It confirms that our dataset is overall accurate.

Table 7 in Appendix B gives some difficult examples that created disagreements. There are two major sources of difficulties: (1) the trait word has multiple meanings in Chinese and the usage does not represent the sense of the trait; (2) a trait word is used to recall the general impression or history behavior of a character in an implicit way.

Accuracy of Cross-Lingual Alignment Finally, we evaluate the quality of the bilingual alignment. We randomly sampled 200 labeled instances for human study. We present to the annotators the snippet $\mathcal{S}$ of each instance in the Chinese book and their aligned sentences from the English books. The human annotators were asked to rate the alignments into four grades: *perfect/high overlap/low overlap/no match*, *i.e.*, all/>50%/<=50%/none of the Chinese sentences have their translations in the paired English sentences. Table 2 show that >97% of the cases fall into the *perfect* and *high overlap* categories. When taking texts from the median



(a) Dantès (*i.e.*, Count Monte Cristo) (b) Albert (*The Count of Monte Cristo*) (c) Plots of sentiments of traits along time

Figure 2: Word clouds and plots of sentiments of traits along time for the characters.

position of the sentences for model inputs, these categories both can make accurate projections of annotations to the English books.

4.4 Data Statistics and Visualization

Data Statistics Table 3 shows the statistics of our PERSONET. We give the full list of books in Appendix C. We can also see that our dataset contains a wide range of book characters. In the annotated training set, approximately 41% of the notes are about positive traits, 36% are about negative traits, and 23% are about neutral traits. This distribution reveals a slight bias, which can be attributed to the fact that users are more inclined to write notes when they have strong sentiments or opinions about a character.

Visualization of Our Dataset Figure 2 visualizes the major traits and the polarity of traits over time for two of the most popular characters. It can be found that the major traits match readers' common impressions; and the trends well align with the common feelings of readers during reading. This further confirms the quality of our data.

Detailed explanations of the figures and more examples can be found in Appendix D and Figure 7.

5 Models for Persona Prediction

We design models based on two different types of pre-trained models, BERT (Devlin et al., 2019) and Longformer (Beltagy et al., 2020). We use the latter model to investigate the strength of models that are pre-trained to handle long contexts.

5.1 Input to the Reader Models

Our data instance consists of a tuple $(\mathcal{S}, \mathcal{H}, c, t, \mathcal{T})$. Here $\mathcal{S}$ is a book snippet that expresses a personality trait t of a character c. $\mathcal{H}$ is the previous history of $\mathcal{S}$ in the book. $\mathcal{T}$ is a set of candidate traits with t as an element. The task is to rank t to the top within $\mathcal{T}$ given $(\mathcal{S}, \mathcal{H})$ and c. We represent the input $(\mathcal{S}, \mathcal{H}, c)$ with the following format options:

•**No history** represents the input as $x = [c \text{ [SEP] } \mathcal{S}]$, *i.e.*, does not use the history $\mathcal{H}$.

•**Extended history:** $x = [c \text{ [SEP] } \mathcal{S} \text{ [SEP] } \mathcal{H}_{\text{prev}}]$, where $\mathcal{H}_{\text{prev}} \subset \mathcal{H}$ includes sentences that are adjacent to $\mathcal{S}$, truncated by models' length limited.

•**Character history:** $x = [c \text{ [SEP] } \mathcal{S} \text{ [SEP] } \mathcal{S}_c]$. $\mathcal{S}_c \subset \mathcal{H}$ includes snippets to the left of $\mathcal{S}$ that contains the character c in our dataset.

5.2 Model Architectures

Our methods compute the score of an input x having a trait t, based on the siamese model.

Text Encoding Firstly, we use a pre-trained LM (PLM, either BERT or Longformer) to encode x and t to the embedding space. The encoded contextualized embeddings of input and output are denoted as $\mathbf{X} = \text{PLM}(x) \in \mathbb{R}^{l_x \times d}$, where l_x is the length of x and d is the size of hidden states; and $\mathbf{T} = \text{PLM}(t) \in \mathbb{R}^{l_t \times d}$, with l_t the length of t.

Baseline Siamese Model As our baseline models, we compute a weighted sum over $\mathbf{X}$ to get a vector representation of the input. Specifically, we use a linear model to compute the attention score over each token of x:

$$A = \text{Attention}(\mathbf{H}),\ \alpha = \text{Softmax}(A).$$

The attention α_x is then used to summarize the hidden states $\mathbf{X}$ a vector $\mathbf{x} = \mathbf{X}^T\alpha$.

The sequence of a trait t is usually short (*e.g.*, a single word's BPE tokenization). Therefore we simply take the average $\mathbf{t} = \text{mean}(\mathbf{T})$. The model makes prediction with $t = \arg\max_{t \in \mathcal{T}} <\mathbf{x}, \mathbf{t}>$.

Contextualization with History When the input x contains the extended or character history as defined in Section 5.1, we need to separate the information of the history from the current context. We maintain a mask $H \in \mathbb{R}^{l_x \times 1}$, such that $H[j] = 1$ if the j-th word belongs to the appended history and 0 otherwise. Two attention vectors are computed for the current snippet and the history:

$$\alpha_s = \text{Softmax}(A \odot (1 - H)),\ \alpha_h = \text{Softmax}(A \odot H).$$

The corresponding summaried vectors are $\mathbf{s} = \mathbf{X}^T \alpha_s$ and $\mathbf{h} = \mathbf{X}^T \alpha_h$. The prediction function is then modified with a gating function $\sigma(\mathbf{s})$ added:

$$t = \arg\max_{t \in \mathcal{T}} \sigma(\mathbf{s}) < \mathbf{s}, \mathbf{t} > +(1 - \sigma(\mathbf{s})) < \mathbf{h}, \mathbf{t} > . \quad (2)$$

5.3 Unsupervised Training

Finally, we propose an unsupervised training task to improve personality prediction. The unsupervised task is used to pre-train the classifiers, before they are fine-tuned on our labeled data. The task mimics the problem definition in Section 3 and constructs tuples of $(\mathcal{S}, t)$. We first extract sentences that contain trait words. If a sentence s_j contains a trait t, we keep a local window of it as the book snippet, with the sentence itself removed. That is, $\mathcal{S}^{(i)} = \{s_{j-w}, \cdots, s_{j-1}, s_{j+1}, \cdots, s_{j+w}\}$. Intuitively, since $\mathcal{S}$ provides the context of s_j, it is informative for inferring the appearance of the trait described in s_j. Therefore this unsupervised task helps to find narrative clues of traits thus can help to better pre-train the encoders.

The method has the limitation of not being character-specific, hence not compatible with our character-history-based models. We leave it to future work.

6 Experiments

6.1 Experimental Settings

We use *bert-base-uncased* and *longformer-base-4096* as backbones for English experiments; and *Roberta-wwm-ext* for the Chinese experiments.

Hyperparameters For our siamese models with and without history, the most important hyperparameter is the lengths of $\mathcal{S}$ and $\mathcal{H}$. We set the maximal length of $\mathcal{S}$ to 480 tokens for most of the models. For models with history we set the maximum of $|\mathcal{S}| + |\mathcal{H}|$=1,600. To show the better performance of our usage of history, we also compare with Longformer with a maximum $|\mathcal{S}|$=2K tokens (the best a single A100 GPU can handle).

The batch size is 40 for BERT-based models; and 8 for Longformer-based models with gradient accumulation every 5 batches. Each epoch of BERT and Longformer models takes ∼7 and ∼40 minutes respectively on a single A100 GPU. We set the learning rate to $2e^{-5}$. We conduct early-stopping on the dev set; and run 5 times to compute the average and stand derivation for all the methods.

Additional Baselines Besides the models in Section 5, we further compared with the follows: •**Models with Oracle Traits in History**, which uses the character's history traits in replace of the history texts. For each instance, we take its target character c's other instances prior to it, and concatenate their groundtruth traits as a sequence to replace $\mathcal{H}$ in the model of Eq. (2). •**Char-Majority**, which always predicts the most frequent trait for a character. This is used to show the diversity of traits for the same character (*i.e.*, necessity of situated prediction). •**GPT-davinci** (text-davinci-003), the few-shot instruct-GPT (Ouyang et al., 2022). •**ChatGPT**, which conduct zero-shot prediction on our task thus can take longer inputs. We test $|\mathcal{S}|$=480 and 1.6K as in our experiments with trained models. •**Humans:** we present the same format of our instances with maximal $|\mathcal{S}|$=480 to humans to get their performance.

Furthermore, we added LoRA (Hu et al., 2022) fine-tuning of the **LLaMA** (Touvron et al., 2023) and **WeLM** (Su et al., 2022) on our PERSONET as additional baselines. The fine-tuning of large language models and the usage of ChatGPT reflect the latest state-of-the-arts in concurrence with our work.

6.2 Overall Results

Our main results are shown in Table 4. First, all the three models without the usages of history achieve similar results. The Longformer with a 2K window does not give better performance, showing that simply increasing the length of input without including useful history information is not helpful for our task. Second, our model with character history achieves the best results. Replacing the character history with extended history slightly reduces the dev performance but lead to significant test performance drop (according to the standard derivation). Among all the supervised-only methods, this model is the only on that maintains consistent dev and test accuracy. Third, our unsupervised training significantly improve the accuracy for all the models.

Fourth, the oracle history traits improve the supervised accuracy with a large margin. Yet for Longformer, adding character history and unsupervised training makes the gap smaller. Finally, the best human performance with knowledge of story history greatly outperforms all the models with and without oracle information with 20∼23%, showing the challenges and great potential of our PERSONET. These results highlight the importance of incorporating history information in solving our

System	Len	Accuracy	
		Dev	Test
Random	–	20.00	20.00
Frequent Traits	–	14.10	12.75
BERT (no-hist)	480	45.01±0.64	42.96±1.07
+ unsup	480	46.18±0.49	44.93±1.01
Longformer (no-hist)	480	45.02±0.45	42.75±0.97
Longformer (no-hist)	2K	45.00±0.44	42.42±0.39
Char-Hist-Longformer	1.6K	45.50±0.54	45.33±1.11
+ unsup	1.6K	**46.39**±0.63	**45.85**±0.72
w/ extend-hist	1.6K	45.46±0.67	43.44±0.72
+ unsup	1.6K	45.93±0.52	44.54±1.49
w/ Oracle Information			
BERT + hist traits	480	50.15±1.01	50.02±1.03
Longformer + hist traits	2K	48.66±0.96	48.11±1.21
Char-Majority	–	16.10	17.25
GPT-davinci 5-shot*	480	34.88	31.51
ChatGPT 0-shot*	480	33.72	42.47
ChatGPT 0-shot*	1.6K	36.05	36.99
LLaMA + LoRA-sft*	1.6K	47.67	49.32
Human w/o history*	480	44.19	40.54
Human w/ history*	480	69.77	65.75

Table 4: Overall performance (%) on our PERSONET-en task. (*) Results were conducted on a subset of the dataset.

task; and reveal that characters exhibit dynamic personalities that evolve over time, thus solely relying on history traits (even oracle) is limited.

The two methods based on large language models, namely GPT-davinci and ChatGPT, performed worse than the models trained on our dataset. This indicates that our task is still a challenge for these general-purpose models. Moreover, although ChatGPT performed better than GPT-davinci, it was not better overall to use the longer context length of 1.6K as compared to using shorter contexts. This suggests that ChatGPT may not have been trained to effectively utilize long context in our situated reading setting.

Chinese Task Performance Table 5 shows results on the Chinese version of PERSONET. The results are in general higher than those in the English setting for two reasons: (1) during annotation we have the semantic space of traits in Chinese, so their English translations may not be the most commonly used words. (2) the user notes tend to reuse words in the books, so there is higher change that some traits explicitly appear in Chinese books.

System	Dev	Test
BERT Reader	49.72	48.70
Multi-Row BERT Reader	50.25	49.25
BERT w/ Trait-History	53.29	51.25
GPT-davinci 5-shot*	33.72	32.78
ChatGPT 0-shot*	34.88	41.89
WeLM + LoRA-sft*	51.16	54.05
Human w/ history*	73.26	68.92

Table 5: 5-choice accuracy (%) on our PERSONET-zh task.

Performance of Fine-Tuned LLMs To fine-tune the LLMs, we adopt the same setup in the ChatGPT experiments, where the same prompts serve as inputs and the ground truth answers are used as outputs. The optimization focuses on minimizing perplexity concerning the outputs. Regarding hyperparameter tuning, we specifically adjust the rank r, weight α, and number of training epochs. For model selection, we rely on the accuracy on the development subset utilized in our human study, which sets $r = 8$, $\alpha = 1$ and 10 training epochs.

The results in Table 4 and 5 show that the fine-tuned LLM achieves slightly better results compared to our proposed baselines. However, it still significantly lags behind human performance by a considerable margin. Interestingly, unlike the other models and humans, the fine-tuned LLM perform better on the testing subset compared to the development one. Our hypothesis is that the testing book *Notre-Dame de Paris* is more popular on the Internet, thus may be more sufficiently trained during the pre-training stages of LLaMA and WeLM.

The LLM fine-tuning results can be potentially improved by employing a contrastive training approach similar to our proposed models. We leave this to future study.

6.3 Human Study

We conduct human study to understand the challenges of our task. We sampled instances from the two books that have most instances from the development and testing sets; and asked human annotators (who are co-authors of the paper but have not seen the labeled data before) to complete our multi-choice task. There are two types of annotators: Type-I who have not read the books before (*human w/o history*); and Type-II who have read the books (*human w/ history*).

We have annotated in total 160 samples. Each sample is guaranteed to be annotated by two humans, one with history and one without history.

Ratio of Ambiguous Instances Sometimes an event in a book can depict multiple aspects of personality. When the sampled negative choices share similarity to these personality traits, it leads to ambiguous cases with more than one correct answers. To investigate these cases, we require the Type-II annotators to mark the instances that they believe have ambiguous labels.[8] There are 41 ambiguous samples recognized, *i.e.*, ∼25% of the cases have more than one correct answers. This indicates an **∼87.2% approximated upperbound** accuracy of our task, if we consider each ambiguous instance has two choices that are correct.

In the future, we can leverage our note clusters to mitigate this ambiguity by ensuring that negative candidates do not appear in the cluster from which the snippet originates.

Main Findings The knowledge of book history is not only important to models, but also to humans. Table 6 compares humans performance with and without history. There is an ∼25% performance gap. Furthermore, human performance without history is only comparable to the best model performance (selected according to dev accuracy, which performs 47.18% and 47.21% on the full dev and test data). These results confirm that our task raises the core challenge of long context understanding.

Detailed results show that the Type-I annotators labeled ∼35% of cases that they believe unsolvable because of their lacking of the book history. After verification by Type-II annotators, there are 37 cases left for close examination. It reveals that the history information is critical for these cases for two major reasons: (1) there are multiple possible answers given the snippets but with the knowledge of the characters' history behavior the incorrect traits can be resolved (17 of 37); (2) the plots in the snippets cannot be understood and linked to any personality without book history (11 of 37). There is a third difficult category (9 of 37), where reasoning is required to draw connections, *e.g.*, consequence or analogy between the current snippet and a character's previously demonstrated personality. Examples of these categories can be found in Table 10 in Appendix E.

6.4 Analysis

Learning Curve Figure 3 plots the learning curve of our PERSONET task. The curves shows that the size of our dataset is large enough as the curves become flat after the point of 30K. More importantly, the results justify the accuracy of our data construction strategy. As adding weak supervision (all) significantly outperforms training with only human-labeled data (dotted lines).

System	Data	
	All	Unamb
Best model	48.75	49.58
GPT-davinci 5-shot	33.33	38.46
ChatGPT zero-shot	37.74	41.03
LLaMA + LoRA-sft	48.43	52.63
Human w/o history	42.50	50.42
Human w/ history	67.92	73.50

Table 6: Comparison of performance among models and humans. The *Unamb*iguous subset is annotated by annotators who have read the books.

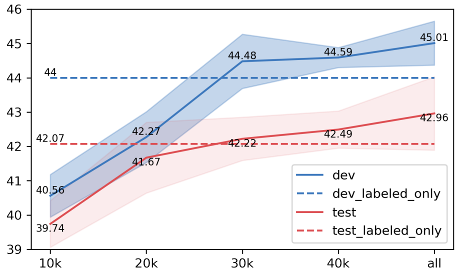


Figure 3: Learning curves with varying sizes of training data.

Difficult Trait Types We examine the traits that appear more than 20 times in the dev set. The most difficult types include *Confident* (0.00%), *Mature* (5.56%), *Liberal* (7.41%), *Humorous* (7.69%), *Impressionable* (8.82%), *Gentle* (9.09%), *Optimistic* (10.81%), *Rational* (11.36%), *Imprudent* (14.29%) and *Insincere* (16.00%). It can be found that most of the difficulty types are abstract, which are usually not explicit depicted in the books but require reasoning from characters' behaviors.

7 Conclusion

We propose a dataset PERSONET for the new problem of situated personality understanding of book characters. We overcome the difficulty in dataset construction with a new strategy of annotating the user notes as a proxy for the original books. Our dataset constuction method maintains both efficiency and accuracy. Experiments show that the task raised challenges of long-text understanding for both humans and machines.

Limitations

Our propose annotation strategy can be applied to labeling other MRC problems, no matter situated comprehension ones or not. However, when generalizing to other problems other than personality prediction we studied here, the accuracy of the user notes may vary with the difficulty of tasks. Additional human verification on the correctness of

[8] Because these people have memory of the books, they can accurately distinguish the ambiguous cases from those can be disambiguated by the history.

notes like in our Section 4.3 need to be conducted.

Our unsupervised training technique does not support the Longformer reader with character history (Char-Hist Longformer) yet. Therefore, the improvement from unsupervised training for our this model is smaller.

While Longformer is common in benchmarking for long story understanding tasks. There are other families of models (Rae et al., 2020; Izacard and Grave, 2021; Ainslie et al., 2020; Xiong et al., 2021; Pang et al., 2022) handling long text encoding. We leave the comparison with these models to future work.

Potential Risks Like the other work that based on the similar set of books (Bamman et al., 2019; Bamman, 2020; Vishnubhotla et al., 2022; Thai et al., 2022), the classic literature may be limited by the time of writing, thus raise fairness considerations. However, please note that our dataset construction strategy is not limited to these books, but can work with any books on WeRead to create a sampled book set without such biases. The main reason we stick with the current list of books is for reproducibility since they are publicly available.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284.

David Bamman. 2020. Litbank: Born-literary natural language processing. *Computational Humanites, Debates in Digital Humanities (2020, preprint)*.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Vinodh Krishnan Elangovan and Jacob Eisenstein. 2015. "you're mr. lebowski, i'm the dude": Inducing address term formality in signed social networks. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626.

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. pages 874–880.

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.

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

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain QA techniques: A comprehensive study. *Trans. Assoc. Comput. Linguistics*, 9:1032–1046.

Isabel Briggs Myers and Mary H McCaulley. 1988. *Myers-Briggs type indicator: MBTI*. Consulting Psychologists Press Palo Alto.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2022. Long document summarization with top-down and bottom-up inference. *arXiv preprint arXiv:2203.07586*.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Belén Saldias and Deb Roy. 2020. Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events, NUSE@ACL 2020, Online, July 9, 2020*, pages 78–86. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022a. A survey of machine narrative reading comprehension assessments. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5580–5587. ijcai.org.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022b. MBTI personality prediction for fictional characters using movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates*, pages 6715–6724. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022c. Tvshowguess: Character comprehension in stories as speaker guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526.

Hui Su, Xiao Zhou, Houjin Yu, Yuwen Chen, Zilin Zhu, Yang Yu, and Jie Zhou. 2022. Welm: A well-read pre-trained language model for chinese. *CoRR*, abs/2209.10372.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. Relic: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848. European Language Resources Association.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148.

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

Mo Yu, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Jing Li, Yue Yu, and Jie Zhou. 2022. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. *arXiv preprint arXiv:2211.04684*.

A Annotation Guidelines and Interface

We show our guidelines in Figure 4; and the annotation interface with translations in Figure 5.

B Notes that Are Difficult or Ambiguous to Label

Table 7 in Appendix B gives some difficult examples that created disagreements. There are two majors sources of difficulties. First, the trait word has multiple meanings in Chinese and the usage does not represent the sense of the trait. In the first example, "可怕的敌人 (frightening enemy)" in Chinese usually means "an very-strong enemy that is hard/impossible to beat", *i.e.*, a terrible enemy. The enemy, here refers to the protagonist Dantès, does not necessary has the *frightening* personality. Similarly, in the second example, the annotators have disagreement because some people believe in Chinese, "非凡 (extraordinary)" can be used as a personality trait only when a person possesses exceptional characteristics. While some annotators think the trait can also describe a person with exceptional abilities.

Second, a trait word is used to recall the general impression or history behavior of a character in an implicit way. In the third example, the user wanted to expresses that Elizabeth used to be clear-headed but becomes a fool at the dance party. This recall of the general impression *clear-headed* is not explicit, but can be inferred from the next sentence that this note is commenting on a snippet of the dance party. Therefore the user aims to comment the *foolish* trait on the snippet instead of *clear-headed*.

Target Trait: 可怕 (*frightening*) **Note Text (Chinese):** 反目的朋友才是最可怕的敌人，因为最了解你的是朋友，知道你短板最多的也是朋友，螳螂是唐格拉尔，黄雀是唐代斯 **Note Text (Translated):** *A friend who turns against you is the most terrible enemy, because the friend who knows you best is the friend, and the friend who knows your most shortcomings is also the friend. The praying mantis is Tanglar, and the oriole is Dantes.*
Target Trait: 非凡 (*extraordinary*) **Note Text (Chinese):** 维尔福的政治头脑在这一刻发挥了它最大的用处，任何时候都是极端的利己主义者，相对应也一定有非凡的时局判断力，才能在哪儿都保全的了自己 **Note Text (Translated):** *Villefort's political mind is at its best use at this moment, he is always an extreme egoist, correspondingly, he must have extraordinary judgment of the situation, in order to be able to protect himself everywhere.*
Target Trait: 清醒 (*clear-headed*) **Note Text (Chinese):** 对于别人的事情，伊丽莎白却又清醒得很，对于自己的事情却变成一个小傻瓜。这一次舞会在众人面前的出丑，或许成为了彬格莱先生后来的一次不了了之的决定性因素" **Note Text (Translated):** *For others' matters, Elizabeth is quite clear-headed, but for her own matters she becomes a fool. This time the embarrassment in front of everyone at the dance party, may have become a decisive factor for Mr. Bingley's later decision to forget about it.*

Table 7: Example of a human mistake.

System	**Slump**	**All**
BERT (no-hist)	35.98	40.33
+ unsup	56.25	44.58
Char-Hist-Longformer	46.53	45.75

Table 8: Accuracy on the slump of Figure 2(c) for the character *Albert* (144 instances) versus on all (424) of the *Albert* instances.

C Full Book List

Table 9 shows the detailed information of each book included in our PERSONET.

D Visualization

Trait Clouds Figure 6 include more word clouds for different characters.

Sentiment Plots Our trait vocabulary contains in total 818 traits with polarity annotations. Specifically, there are 234 positive traits, 292 neutral traits and 292 negative traits. Figure 7 visualizes readers' sentiments towards four popular characters through the lens of traits. We map the labeled traits to their sentiments, *i.e.*, positive or negative, and then plot the sentiment along time. Here the x-axis is the position of an note with trait label, normalized by the lengths of the books. The curves are smoothed within a window of 50 for *The Count of Monte Cristo* and 20 for *Notre-Dame de Paris*, depending on the sparsity of the samples.

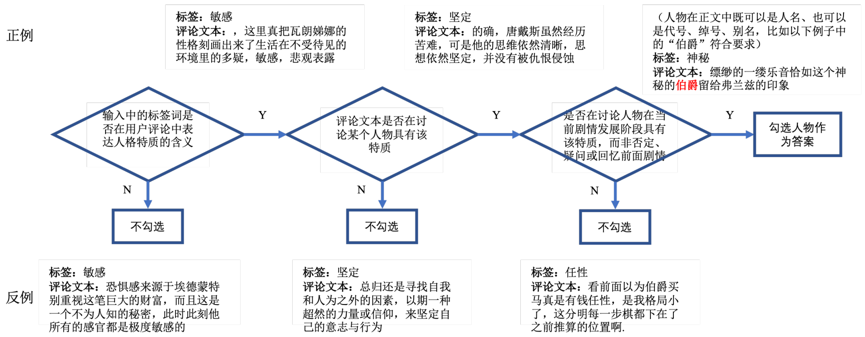
The trends well align with the common feelings of readers during the reading process. For example, the character *Albert* is in general a brave and decent person. Most readers liked his personality until he recklessly challenged *Dantès* for a duel. Then the character's reputation is saved after he found out that his father framed many people including *Dantès* and decided to give up the duel and live off his father's ill-gotten gains. One the other hand, *Claude Frollo* received monotone decreased rates, who appeared first as a pious and highly knowledgeable man then turned to be evil and morbid because of his obsessive for *Esmeralda*.

标注目标：

在阅读故事的过程中，读者在阅读到某一位置时所写的评论（任务输入），是否在评论当前故事情节下，某个人物（任务输出）的某种人格特质（任务输入）。

任务内容：

（1）对给出的人格表述词，找到“评论文本”中人格表述词的描述对象（被描述人物名字）。
（2）每个人格表述词可以选择多个符合要求的人物，但是每个人物只需要选择一次。
（3）如果在“评论文本”中没有找到符合要求的人物，或没有出现人物的姓名，可不做选择，直接提交。
（4）（可选阅读）任务同时提供了评论所在位置的“原书文本”作为参考。


正例
输入中的标签词是否在用户评论中表达人格特质的含义
Y
N
不勾选
评论文本是否在讨论某个人物具有该特质
Y
N
不勾选
是否在讨论人物在当前剧情发展阶段具有该特质，而非否定、疑问或回忆前面剧情
Y
N
不勾选
勾选人物作为答案
反例

标签：敏感
评论文本：，这里真把瓦朗娣娜的性格刻画出来了生活在不受待见的环境里的多疑，敏感，悲观表露

标签：坚定
评论文本：的确，唐戴斯虽然经历苦难，可是他的思维依然清晰，思想依然坚定，并没有被仇恨侵蚀

（人物在正文中既可以是人名、也可以是代号、绰号、别名，比如以下例子中的“伯爵”符合要求）
标签：神秘
评论文本：缥缈的一缕乐音恰如这个神秘的伯爵留给弗兰兹的印象

标签：敏感
评论文本：恐惧感来源于埃德蒙特别重视这笔巨大的财富，而且这是一个不为人知的秘密，此时此刻他所有的感官都是极度敏感的

标签：坚定
评论文本：总归还是寻找自我和人为之外的因素，以期一种超然的力量或信仰，来坚定自己的意志与行为

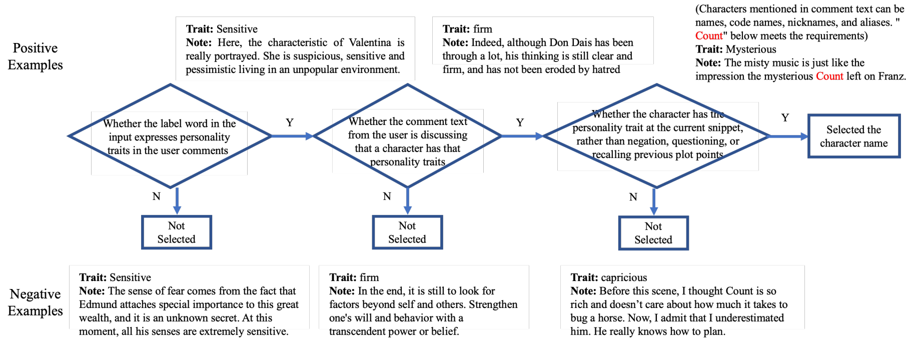
标签：任性
评论文本：看前面以为伯爵买马真是有钱任性，是我格局小了，这分明每一步棋都下在了之前推算的位置啊.

Annotation Goals:

If a reader writes a comment (task input) at a certain position of the book during the reading process, please judge whether the comment is commenting on a certain personality trait (task input) of a certain character (task output).

Tasks:

(1) For the given word expressing a personality, find the name of the character described by the personality-expressing word in the " Comment Text ".
(2) Each personality-expressing word can select multiple qualified characters, but each character only needs to select once.
(3) If the "Comment Text" does not contain any characters satisfying the requirement or even no character appears in it, you can pass the selection of character and submit this item directly.
(4) (Optional reading) The task provides the "original book snippet" where the comments is located as a reference.


Positive Examples
Whether the label word in the input expresses personality traits in the user comments
Y
N
Not Selected
Whether the comment text from the user is discussing that a character has that personality traits
Y
N
Not Selected
Whether the character has the personality trait at the current snippet, rather than negation, questioning, or recalling previous plot points
Y
N
Not Selected
Selected the character name
Negative Examples

Trait: Sensitive
Note: Here, the characteristic of Valentina is really portrayed. She is suspicious, sensitive and pessimistic living in an unpopular environment.

Trait: firm
Note: Indeed, although Don Dais has been through a lot, his thinking is still clear and firm, and has not been eroded by hatred

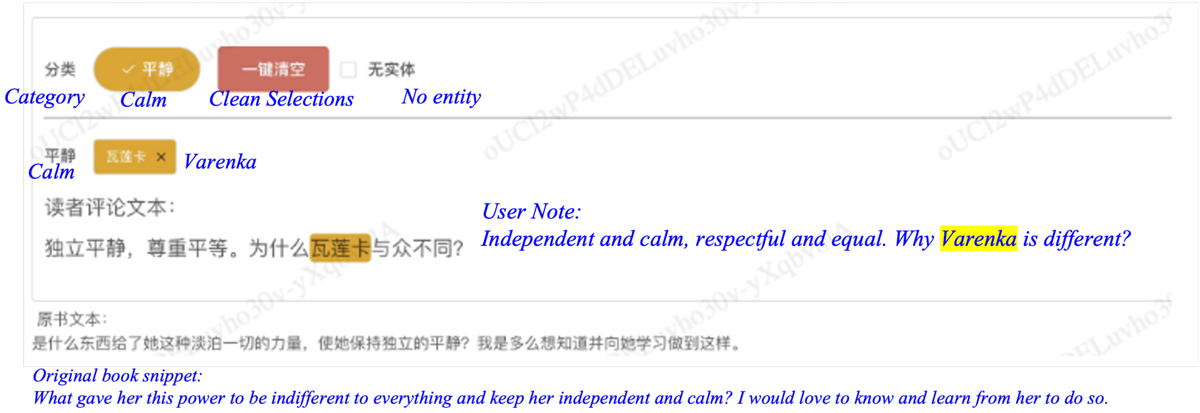
(Characters mentioned in comment text can be names, code names, nicknames, and aliases. " Count" below meets the requirements)
Trait: Mysterious
Note: The misty music is just like the impression the mysterious Count left on Franz.

Trait: Sensitive
Note: The sense of fear comes from the fact that Edmund attaches special importance to this great wealth, and it is an unknown secret. At this moment, all his senses are extremely sensitive.

Trait: firm
Note: In the end, it is still to look for factors beyond self and others. Strengthen one's will and behavior with a transcendent power or belief.

Trait: capricious
Note: Before this scene, I thought Count is so rich and doesn’t care about how much it takes to bug a horse. Now, I admit that I underestimated him. He really knows how to plan.

Figure 4: Our annotation guidelines. Top: the original Chinese guidelines. Bottom: the English translation.



读者评论文本：

独立平静，尊重平等。为什么瓦莲卡与众不同？

User Note:
Independent and calm, respectful and equal. Why Varenka is different?

原书文本：

是什么东西给了她这种淡泊一切的力量，使她保持独立的平静？我是多么想知道并向她学习做到这样。

Original book snippet:
What gave her this power to be indifferent to everything and keep her independent and calm? I would love to know and learn from her to do so.

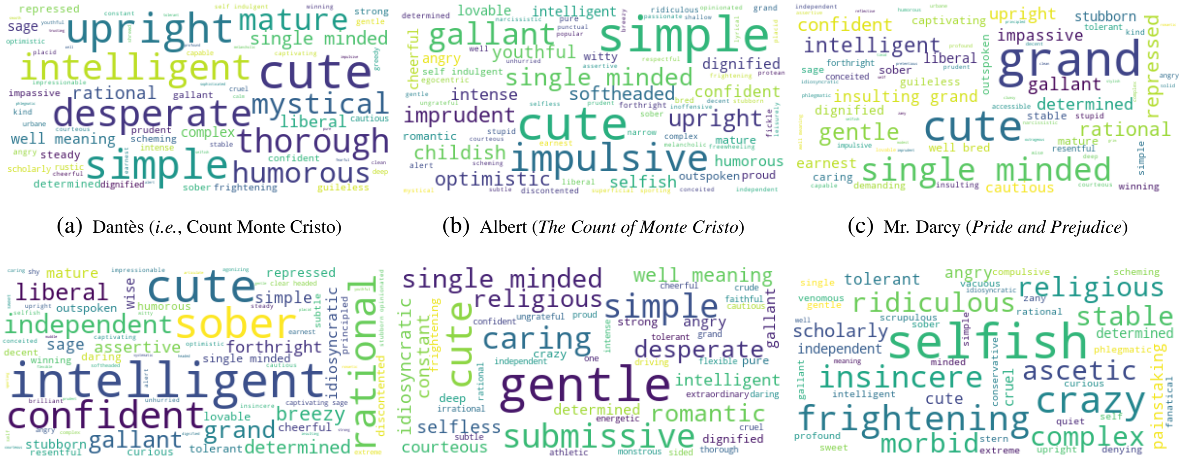
Figure 5: Our annotation interface (with English translations in *blue words*).

A Case Study We assessed the model performance on the points where people's ratings of *Albert* have dramatic fluctuations (around x=0.8). Specifically, we compared three models: the baseline BERT model without any history, the BERT model enhanced with our unsupervised objective, and the Char-Hist Longformer, which can leverage longer historical information. The results are shown in Table 8.

Our findings revealed that both the enhanced models—BERT with the unsupervised objective and Char-Hist Longformer—achieved a similar level of improvement over the BERT baseline when considering the entire evaluation set of *Albert*. These results align with our experimental observations from the comprehensive evaluation data. However, it is noteworthy that the model incorporating the unsupervised objective exhibited a significantly greater enhancement at the slump of the curve. As mentioned earlier in this section, the author explicitly portrayed Albert's reckless personality through his actions and dialogues in this particular case. Even without prior knowledge of the events leading up to this point, humans can intuitively grasp Albert's personality traits. Our unsupervised task aims to capture the correlation between personality and the external expressions manifested within the narrative. This is why it proves to be more effective in this specific case.

E Examples of Cases that Require History Information

The cases where history information is necessary to solve can be roughly categorized into three types according to our human study in Section 6.3. We include examples for each type in Table 10.



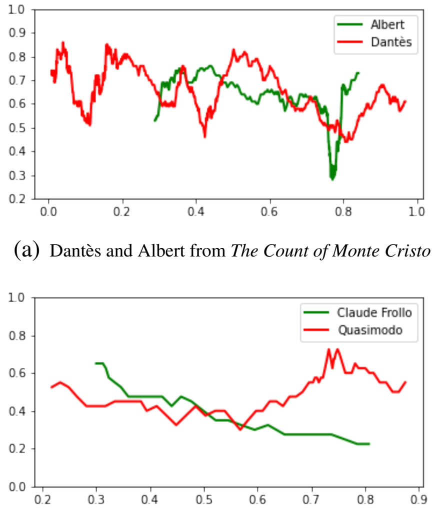
(a) Dantès (*i.e.*, Count Monte Cristo) (b) Albert (*The Count of Monte Cristo*) (c) Mr. Darcy (*Pride and Prejudice*)

(d) Elizabeth (*Pride and Prejudice*) (e) Quasimodo (*Notre-Dame de Paris*) (f) Claude Froll (*Notre-Dame de Paris*)

Figure 6: Word clouds for the characters.

Book Name	**#Characters**	**#Instances**	**#Sentences**
	Training Books		
Of Human Bondage	18	6539	16542
Pride and Prejudice	21	8632	5954
Madame Bovary	10	5440	6952
Anna Karenina	14	8204	20898
Anne Of Green Gables	8	3755	6834
Little Women	7	2726	9409
War and Peace	21	3448	31784
Don Quixote	3	767	9384
Wuthering Heights	13	2471	6753
Jane Eyre	10	1048	9692
Twenty Thousand Leagues under the Sea	4	471	6614
Jude the Obscure	3	174	9191
The Sorrows of Young Werther	1	74	2400
Father Goriot	5	198	6678
Uncle Tom's Cabin	5	138	10122
Vanity Fair	5	171	13125
Oliver Twist	5	178	9166
	Development Books		
The Red and the Black	6	721	11061
The Count of Monte Cristo	27	2488	26437
The Adventures of Tom Sawyer Complete	2	115	4913
David Copperfield	15	312	19195
The Gadfly	1	76	6875
A Tale of Two Cities	3	33	7757
	Testing Books		
Crime and Punishment	18	1498	14347
The Brothers Karamazov	12	638	24101
Les Miserables	14	557	35139
Eugenie Grandet	4	217	3797
Tess of the d'Urbervilles	3	162	8074
Notre-Dame de Paris	8	206	11278
The Call of the Wild	1	42	1696
The Idiot	9	215	16072
Moby Dick; or The Whale	2	51	9911
Resurrection	2	38	9760

Table 9: Detailed information of books included in our PERSONET.



(a) Dantès and Albert from *The Count of Monte Cristo*

(b) Frollo and Quasimodo from *Notre-Dame de Paris*

Figure 7: Plots of sentiments of characters' traits along time.

Category: *(1) multiple possible answers given the snippet without history*
Target Character: *Dantès* **Groundtruth Trait:** *simple*
Distractors: *insincere, dirty, impressionable, loquacious*
Snippet: *i am the abbe faria, and have been imprisoned as you know in this chateau d ' if since the year 1811 ; previously to which i had been confined for three years in the fortress of fenestrelle. in the year 1811 i was transferred to piedmont in france. it was at this period i learned that the destiny which seemed subservient to every wish formed by napoleon, had bestowed on him a son, named king of rome even in his cradle. i was very far then from expecting the change you have just informed me of ; namely, that four years afterwards, this colossus of power would be overthrown. then who reigns in france at this moment — napoleon ii.? " " no, louis xviii. " ... dantes ' whole attention was riveted on a man who could thus forget his own misfortunes while occupying himself with the destinies of others. " yes, yes, " continued he, " ' twill be the same as it was in england. after charles i., cromwell ; after cromwell, charles ii., and then james ii., and then some son - in - law or relation, some prince of orange, a stadtholder who becomes a king. then new concessions to the people, then a constitution, then liberty. ah, my friend! " said the abbe, turning towards dantes, and surveying him with the kindling gaze of a prophet, " you are young, you will see all this come to pass. ..."*

Category: *(2) plot cannot be understood without history*
Target Character: *The elder* **Groundtruth Trait:** *intelligent*
Distractors: *confident, breezy, single-minded, decadent*
Snippet: *the servant soon returned. the decanter and the glass were completely empty. noirtier made a sign that he wished to speak. " why are the glass and decanter empty? " asked he ; " valentine said she only drank half the glassful. " the translation of this new question occupied another five minutes. " i do not know, " said the servant, " but the housemaid is in mademoiselle valentine ' s room : perhaps she has emptied them. " " ask her, " said morrel, translating noirtier ' s thought this time by his look. the servant went out, but returned almost immediately. " mademoiselle valentine passed through the room to go to madame de villefort ' s, " said he ; " and in passing, as she was thirsty, she drank what remained in the glass ; as for the decanter, master edward had emptied that to make a pond for his ducks. " noirtier raised his eyes to heaven, as a gambler does who stakes his all on one stroke. from that moment the old man ' s eyes were fixed on the door, and did not quit it. it was indeed madame danglars and her daughter whom valentine had seen ; they had been ushered into madame de villefort ' s room, who had said she would receive them there. that is why valentine passed through her room, which was on a level with valentine ' s, and only separated from it by edward ' s. the two ladies entered the drawing - room with that sort of official stiffness which preludes a formal communication. among worldly people manner is contagious. madame de villefort received them with equal solemnity. valentine entered at this moment, and the formalities were resumed. ...*

Category: *(3) The current snippet can be associated to some previous plot where the character demonstrates a personality trait*
Target Character: *Esmeralda* **Groundtruth Trait:** *simple*
Distractors: *rational, mature, emotional, egocentric*
Snippet: *is she to be hung yonder? " " fool! t'is here that she is to make her apology in her shift! the good god is going to cough latin in her face! that is always done here, at midday. if 'tis the gallows that you wish, go to the greve. " " i will go there, afterwards. " " tell me, la boucanbry? is it true that she has refused a confessor? " " it appears so, la bechaigne. " " you see what a pagan she is! " "'tis the custom, monsieur. the bailiff of the courts is bound to deliver the malefactor ready judged for execution if he be a layman, to the provost of paris ; if a clerk, to the official of the bishopric. " " thank you, sir. " " oh, god! " said fleur - de - lys, " the poor creature! " this thought filled with sadness the glance which she cast upon the populace. the captain, much more occupied with her than with that pack of the rabble, was amorously rumpling her girdle behind. she turned round, entreating and smiling. " please let me alone, phoebus! if my mother were to return, she would see your hand! " at that moment, midday rang slowly out from the clock of notre - dame. a murmur of satisfaction broke out in the crowd. the last vibration of the twelfth stroke had hardly died away when all heads surged like the waves beneath a squall, and an immense shout went up from the pavement, the windows, and the roofs, " there she is! " fleur - de - lys pressed her hands to her eyes, that she might not see. " charming girl, " said phoebus, " do you wish to withdraw? " " no, " she replied ...*

Table 10: Example of cases that require history information to solve. The red texts are the underlined text of the notes that used to construct the labeled instance. In the first example, according to the snippet, both *simple* and *impressionable* are possible traits to explain the character's behavior. Only from the history that *Dantes* is a brave and determined person, we can select *simple* as the correct answer. In the second example, only when the readers know that *Noirtier* (*The elder*) aims to help Valentine get immunity from the poisoned juicy, they can understand the character's wisdom. In the third example, *Esmeralda* is not present. However, the scene of love between Phoebus and Fleur-de-Lys is quite similar to her story with Phoebus, illustrating that she was easily deceived by the man.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8.

☑ A2. Did you discuss any potential risks of your work?
Section 8.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. We create artifacts by ourselves.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 7. We will release our dataset for public research with CC-BY 4.0.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. We create artifacts by ourselves.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 4.1. We anonymized the data with user information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1 and Appendix C.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.4.

C ☑ Did you run computational experiments?

Section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 6.1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6.1.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6.2. We report mean-std results with 5 runs.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 1 and Section 4.2 Step 1.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4.2 Step 1.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. The data is from social network thus the study follows IRB.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 4.2 Step 1.