Interpreting Themes from Educational Stories

Yigeng Zhang[1], Fabio A. González[2], Thamar Solorio[1,3]
[1]University of Houston, Houston, USA
[2]Universidad Nacional de Colombia, Bogotá, Colombia
[3]MBZUAI, Masdar City, United Arab Emirates
[1]{yzhang168,tsolorio}@uh.edu, [2]fagonzalezo@unal.edu.co

Abstract

Reading comprehension continues to be a crucial research focus in the NLP community. Recent advances in Machine Reading Comprehension (MRC) have mostly centered on literal comprehension, referring to the surface-level understanding of content. In this work, we focus on the next level - interpretive comprehension, with a particular emphasis on inferring the themes of a narrative text. We introduce the first dataset specifically designed for interpretive comprehension of educational narratives, providing corresponding well-edited theme texts. The dataset spans a variety of genres and cultural origins and includes human-annotated theme keywords with varying levels of granularity. We further formulate NLP tasks under different abstractions of interpretive comprehension toward the main idea of a story. After conducting extensive experiments with state-of-the-art methods, we found the task to be both challenging and significant for NLP research. The dataset and source code have been made publicly available to the research community at https://github.com/RiTUAL-UH/EduStory.

Keywords: Corpus, Document Classification, Text Categorization, Question Answering

1. Introduction

Reading and understanding are fundamental aspects of human intellectual activity. Reading comprehension is one of the many abilities that AI is anticipated to develop on par with humans. The NLP community has dedicated substantial efforts to machine reading comprehension (MRC) research, resulting in significant advancements in models' reading capabilities. From an educational research standpoint, reading comprehension is divided into three levels: literal comprehension, inferential/interpretive comprehension, and critical/evaluative comprehension ((Herber, 1978), further developed by (Vacca and Vacca, 1998)). The first level involves understanding direct and explicit information extracted from a text, including facts, vocabulary, events, and other stated information. The second level demands that readers make inferences from contextual information, such as deducing cause and effect or determining the main idea. The third level transcends the text, requiring readers to integrate their own opinions and critically analyze the content or assess a viewpoint.

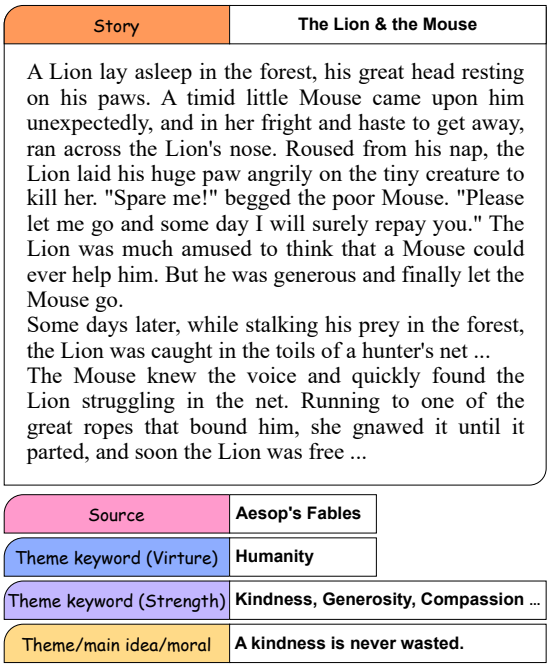

Figure 1: An example of theme interpretation.

Current NLP research does not explicitly regard reading comprehension from different levels or distinguish between them, with the majority of MRC research focusing on the literal level (Richardson et al., 2013; Kočiský et al., 2018; Saha et al., 2018). However, in real-world learning environments, mere word decoding and literal matching are inadequate. Recognizing the inherent meaning of a text or its implied information remains an area of ongoing study. This work concentrates on a novel research problem: interpreting themes from text using NLP methods. This topic falls within the second level, interpretive comprehension. A theme goes beyond a simple summary of the story's plot or character actions. Instead, it reflects deeper insights and conveys the key message that is implied within the context. This complexity requires that NLP models not only process the context but also make inferences and interpret the theme or main idea, which is often not explicitly stated in the text.

To further explore and gain empirical knowledge

LREC-COLING 2024, pages 9190–9203
20-25 May, 2024.

on this research problem, we choose educational stories as our context. These narrative texts, such as fables and folktales, often convey a lesson via a series of events with a clear consequence. These stories are widely embraced by individuals from relatively diverse cultural backgrounds and knowledge levels and are commonly used as children's bedtime reading. For each story, the theme sentence(s) (main idea/lesson/moral/meaning) is often provided by the story's author or editor. We use our best efforts to gather educational stories and create a dataset of high-quality English story-theme pairs from various sources and cultural backgrounds. Figure 1 depicts an exemplar story with its attributes.

Existing language resources for narrative comprehension, such as those presented in (Xu et al., 2022) and (Zhao et al., 2023), have been designed primarily for explicit and implicit question answering and they do not focus on the comprehension of story themes. Additionally, these datasets tend to be not only limited in size but also lack diversity in their sources. To address this gap, we probe further into the story content and characterize the challenge of theme interpretation across various NLP abstractions. Given the challenge of interpretive comprehension, we outline tasks according to their levels of difficulty. First, we propose to investigate **theme identification**. Educational story themes are categorized based on values from positive psychology, character strengths, and virtues. The task is formulated as story classification at the theme keyword level, such as wisdom and integrity. Next, we examine **story-theme matching**, where a story is given, and its theme sentence must be found within a collection of theme sentences, or vice versa. This task involves story-theme or theme-story retrieval. Additionally, we investigate **story reading comprehension on themes**. Similar to typical MRC or Q&A tasks, we design multiple-choice problems on themes given a story. Finally, we conduct exploratory research on **theme generation**. By leveraging recent advances in pretrained large language models (LLMs), we explore the capability of generating accurate theme text from a given story.

To assess how challenging the proposed theme interpreting tasks are, we designed and conducted experiments using different machine learning (ML)-based methods, covering both conventional ML models and large language model (LLM)-based techniques. Experimental results on the classification, text retrieval, and MRC tasks show that interpreting themes from narrative text is still challenging even with state-of-the-art LLM-based methods. We further use human judges to evaluate the LLM-generated theme sentences. The evaluation shows the strong capability of state-of-the-art LLMs to a certain extent, however, LLMs are far from perfect at interpreting a story theme that human judges can easily understand.

In sum, the contribution of this work is summarized as follows:

- Our work serves as an initial call to the community, urging further exploration and reflection on MRC issues from different levels. Specifically, we introduce the concept of *theme interpretation* as a task in NLP, framed within the context of inferential/interpretive reading comprehension.
- We formulate the task comprehensively from various research aspects of NLP and provide extensive empirical research and analysis.
- We publish the first dataset in theme interpreting for the community[1], which offers rich value for further investigation and development.

2. *EduStory*: the dataset

To highlight the importance of theme interpretation and establish a benchmark, we introduce *EduStory*, the first dataset specifically created for interpretive/inferential comprehension of themes in narrative text. We use educational stories from different eras and cultural backgrounds as the context and their corresponding themes. In this work, we surveyed various types of stories and went through multiple stages of data collection and annotation.

2.1. Educational stories

Educational stories utilized in creating the dataset are those that employ narratives to illustrate a point or teach a lesson to the reader. These narratives are typically written in plain language and clearly depict the characters' actions. Lessons are often conveyed to readers through positive or negative outcomes corresponding to character movements, further presenting educational main ideas, such as the importance of being kind to others and the harm of dishonesty. The main ideas of the stories are hardly stated directly in the context. Readers must look beyond the literal words and employ reasoning skills to comprehend and extract insights from the narrative using their knowledge and common sense. This story collection covers a wide range of literary genres, which are not limited to the following:

- **Fables:** Fables are tales that mainly employ anthropomorphic animals as characters, placed in fantastical scenarios that teach ethical lessons. The purpose of this genre is to offer moral guidance through captivating narratives.

[1] https://github.com/RiTUAL-UH/EduStory.

Wisdom and Knowledge 228					**Humanity** 59		
Creativity	Curiosity	Judgment	Learning	Perspective	Love	Kindness	Social Intelligence
3	0	26	0	199	16	31	12
Transcendence 25					**Justice** 22		
Appreciation	Gratitude	Hope	Humor	Spirituality	Citizenship	Fairness	Leadership
1	20	2	0	2	17	4	1
Courage 57				**Temperance** 60			
Bravery	Persistence	Integrity	Vitality	Forgiveness	Humility	Prudence	Self-Regulation
5	23	29	0	2	25	17	16

Table 1: Distribution of character strengths and virtues across themes. *Learning* represents *Love of Learning*, *Forgiveness* represents *Forgiveness and Mercy*, *Humility* represents *Humility and Modesty*, and *Appreciation* represents *Appreciation of Beauty and Excellence*.

- **Folk Stories:** Folk stories often originate in a particular culture or region. They are spread among people over generations as a unique medium of education. Folktales, legends, fairy tales, and more usually belong to this category and they typically have an educational message.
- **Idiom Stories:** Idiom stories illuminate the origins or interpretations of idioms from a specific language or culture. These narratives often illustrate a memorable event to convey a moral lesson, thereby demonstrating the integral connection between language and morality.
- **Miscellaneous:** This category comprises stories that weave educational narratives, although they may not conform to a specific genre. It includes ones that may be called moral tales, success stories, and inspiring stories.

These stories may portray scenes of enlightenment or individuals who have made notable achievements. The educational theme is often delivered by a key character in a specific scene or the narratives emphasizing the admirable qualities of successful individuals, thereby inspiring readers towards virtues such as hard work and open-mindedness.

2.2. Data collection

We use our best efforts to collect story-theme pairs with free access to the Internet. The educational story search includes but is not limited to, fables, moral stories, folk stories (folktales/legends), children's stories, and inspiring stories. All stories are written in plain language and designed to deliver educational value. More importantly, each story is accompanied by a piece of text that reveals the theme or main idea. While we limited our search to stories written in English, we tried to collect stories from various cultural backgrounds and different ages aiming for a relatively diverse representation. Our efforts resulted in a collection of 580 story-theme pairs, with their sources recorded. We further manually filtered out the stories with overlapping storylines and main ideas but with rather different narratives, and this resulted in 451 unique story-theme pairs. Nevertheless, we retained the 129 pairs with duplicates, recognizing their value as diverse, human-crafted language resources.

2.3. Theme keywords and annotation

Since the stories are composed of educational values and aim to teach people, we propose organizing the stories by specific human values conveyed in the themes. For example, when parents tell the story of *the shepherd and the wolf*, they expect their child to learn the importance of honesty and develop a sense of integrity. Educators, such as parents and teachers, hope that children will learn and build good character traits from stories, which require them to interpret the main idea of the story. We, therefore, introduce the taxonomy of favorable traits from positive psychology. In foundational work by Peterson and Seligman (Peterson et al., 2004), they loosely categorize character virtues into six categories with specific character strengths. We designed an annotation plan based on this principled taxonomy.

We use a hierarchical annotation and auditing scheme to label the theme keywords for the stories. First, two annotators are asked to read and understand the six character virtues and 24 strengths in the book. Afterwards, they work individually to read the stories and assign virtue and strength labels, basing their decisions on their best interpretation of the theme from the narrative and the original theme sentences. The first round of annotation results in a Cohen's Kappa score of $\kappa = 0.30$. As expected, this indicates a lower level of annotation agreement, reflecting the subjective nature of human theme interpretation in educational stories. We further introduce an iterative auditing process to gain additional views on theme interpretation. The first auditor independently reviews every story where the two annotators disagree on the theme keywords. The auditor then determines the most fitting label for the story. This 'auditor-decided' label is then used

Dataset	Context type	Number of articles	Task	Answer	Level of RC
MCTest (Richardson et al., 2013)	Narrative	500	Fact check	In context	Literal
SQuAD (Rajpurkar et al., 2016)	Informational	536	Fact check	In context	Literal
NarrativeQA (Kočiský et al., 2018)	Narrative	1572	Fact check	In context	Literal
DuoRC (Saha et al., 2018)	Narrative	7680	Fact check	In context	Literal
FairytaleQA (Xu et al., 2022)	Narrative	278	Mixed QA	In context/implicit	Literal/interpretive
StoryQA (Zhao et al., 2023)	Narrative	148	Mixed QA	In context/implicit	Literal/interpretive
EduStory (this work)	Narrative	580/451	Theme interpretation	Implicit	Interpretive

Table 2: Comparison across datasets with different contexts and levels of reading comprehension.

as the gold label for the story theme. If neither of the annotations satisfies the auditor, the auditor will provide a third opinion on the story theme. Next, a second auditor gets involved and focuses on disagreements between the two annotators and the first auditor. The second auditor either makes their best judgment among the three keywords or offers a fourth opinion. If there is still any disagreement, we go through this iterative process by introducing new auditors to determine the gold label. In our practice, two auditors can successfully resolve disagreements. It is important to note that these resulting annotations should not be viewed as the 'gold standard' for story interpretation. We will release annotations from all annotators and auditors to showcase the diversity of interpretations, which can serve as useful indicators for further studies. More details about the human annotators, auditors, and judges can be found in Appendix section 9.1.

2.4. Data analysis

The distribution of themes is uneven based on the annotation. More than half of the themes belong to *Wisdom* because many of these stories teach readers a specific life lesson rather than a definitive virtue or strength, such as *integrity* and *humility*. A detailed category distribution is shown in Table 1.

The stories are mostly short in length. The average length of all stories is 284 words while the median length is 201. A majority (82%) of the stories have less than 400 words. A detailed text length distribution is shown in Figure 2.

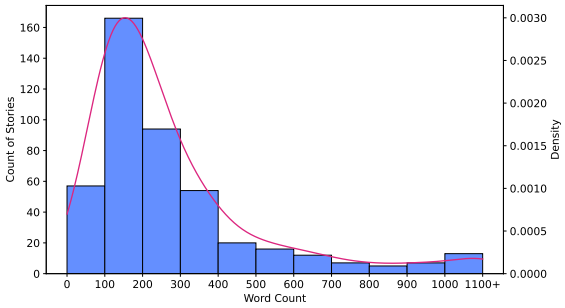

Figure 2: The distribution of word count of the story collection.

One of the standout features of this dataset is the diverse cultural origins of the stories. Firstly, we sourced and curated stories from various versions of *Aesop's Fables* found online, which resulted in over half of our collection tracing back to ancient Europe. It is important to acknowledge that not every fable associated with Aesop may be his original work. Many stories attributed to him may have unclear authorship. Given the missing evidence and the challenges of individual verification, we tentatively accept any source that labels a story as one of *Aesop's Fables*. Therefore, we use the term *Ancient Europe* as a loose categorization to denote the source of these stories. In addition, we have managed to gather educational stories from ancient China and India, including fables, folk tales, and idiom stories. We have also made our best effort to source other educational narratives from the open internet, including children's stories, contemporary inspirational stories, and success stories of celebrated individuals who have achieved significant accomplishments. Figure 3 shows a detailed proportional representation of stories' cultural origins.

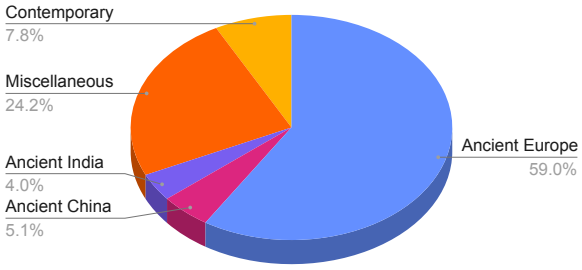

Figure 3: The statistical plot of the cultural origins of the stories in the dataset.

2.5. Comparing to relevant MRC datasets

MRC and QA have been important and popular research topics in NLP and there are many existing language resources. Existing narrative MRC datasets focus on finding specific facts and inferences in one story or plot. Typical questions such as *"What did James do after he ordered the fries?"*, *"Why was the Boy so greedy?"* consists of the majority of the reading comprehension problems in the dataset. In table 2, we compare *EduStory* to several relevant datasets to give comprehensive information on the positioning of this work.

The primary objective of *EduStory* is to evaluate the NLP capability of interpreting themes in narrative text. Although the question is as simple as *"What is the main idea of the story?"*, the answer can be hardly found in the context unless one integrates information across the narrative and uses intrinsic knowledge to make inferences. This task strictly lies at the second level of reading comprehension, which means it requires more comprehensive reasoning abilities than fact-checking in context.

2.6. Value of further development

The *EduStory* dataset holds diverse possibilities for further development. From an NLP research perspective, additional annotations can be applied to design other MRC questions, including but not limited to literal matching and other inferential understanding. Simultaneously, *EduStory* also provides resources for story generation studies based on educational themes, both from keywords and theme sentences. For educational research, this dataset can serve as a benchmark for student comprehension evaluation. Educators can leverage AI methods to compose new stories or new questions for teaching purposes. *EduStory* also contributes to the fields of positive psychology and moral education. Our annotation serves as a useful reference for educators to select suitable teaching contexts. For instance, educational stories in this dataset are relatively diverse, and discovering encouragement and punishment plots and consequences is a direction for developing constructive teaching strategies and storytelling AI applications.

3. Theme keyword identification

The task of theme keyword identification is to assign a predefined theme keyword from a collection to a piece of the story. We introduce theme keywords of educational values from positive psychology. We have annotations on 6 higher-level classes of character virtues and 24 fine-grained character strengths as discussed in Section 2. The task is formulated as a typical multi-class text classification problem. More formally, in a collection of N labeled story-theme keyword pairs:

$S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$,

where x_i is one piece of the story and y_i is a theme keyword from a set of K classes. The learning objective is either to build up adequate decision boundaries of different classes or to produce the desirable answer from model generation.

3.1. Experiments

The goal of the theme keyword identification experiment is to evaluate the performance of various supervised learning-based text classification models using different textual feature representations. We use the macro F1 score as the classification performance metric. We apply various supervised learning-based text classification models using different textual feature representations:

	Virtue		Strength	
	Dev	Test	Dev	Test
TF-IDF	15.4	11.2	4.8	6.5
BOWV	14.7	18.0	4.8	4.5
TextCNN	15.1	13.4	5.1	9.5
BERT	**22.9**	**21.5**	**17.1**	**11.6**
Flan-T5	19.9	14.5	13.8	5.9

Table 3: Performance comparison on theme keyword identification task across different methods. Metric: macro F1 score.

TF-IDF: For the sparse vector representation method, we compute TF-IDF vectors to represent each story passage and then apply a linear SVM to perform classification.

Bag-of-word-vectors: For dense vector representation, distributed word vector representations are used to vectorize each word. We take the average vector as a story passage representation and use SVM as the classifier. Here we use GloVe (Pennington et al., 2014) as word vector representations.

TextCNN: We use a sequence of dense vectors to represent a passage and a convolutional neural network (CNN) to extract features and a linear layer to perform classification. Here we use the TextCNN (Kim, 2014) model.

BERT: Pretrained Transformer-based language models have been proven to be effective in various NLP tasks. We choose BERT (Devlin et al., 2019) as one of the experiment models.

Prompt tuning using LMs: Prompting has shown great potential in recent research. We apply the instruction-finetuned T5 model, Flan-T5 (Chung et al., 2022), to perform classification with task-specific prompt tuning. We manually designed a classification template, in which we list all the keywords as textual prompting and add a prompt sentence that is the best theme keyword to describe the story. Then finetune the model to conditionally generate the correct theme keyword.

The experiments of classification on character virtues and strengths are carried out separately. Table 3 presents the classification performance in F1 scores for each method. This work's experiment and implementation details can be found in Appendix section 9.2.

3.2. Discussion

All models tested showed low F1 scores in theme identification tasks. This can potentially be attributed to two main factors. Firstly, the limited number of training instances may not provide the models with sufficient information to learn effectively. More importantly, the ambiguous interpretation of theme keywords presents a significant challenge. Achieving consensus, even among human evaluators, can be difficult due to the inherent subjectivity of theme interpretation.

4. Story-theme matching

This task is to match one story with the correct theme sentence or vice versa. We formulate it as a text retrieval problem. Take story-theme matching as an example: when given one story, the retrieval model should find the best matching theme sentence out of a collection of all theme sentences. In this setting, a story will act as a query and the themes are the documents. Since there is no ground truth for relevancy scores between every possible story-theme combination, we simply take the original story-theme pair as the only correct match. Therefore, the task is formally described as follows: Given a story (query) q_i and a collection of N theme sentences (documents) $D = \{d_1, d_2, ..., d_i, ..., d_N\}$, the learning objective is to rank the correct theme sentence d_i to the top among every other theme sentence in D, according to its best relevance to the story (query). Similarly, we switch the story-theme pairs as document-query when using one theme to retrieve the correct story.

4.1. Experiments

In our benchmark analysis, we evaluate the performance of various text retrieval techniques, including conventional and deep methods. Given that we assume a single story has the ground truth theme as its only matching pair and we rank the complete theme sentence set, we employ the Mean Reciprocal Rank (MRR) of the ground truth theme in the retrieval results as our evaluation metric. We use the following calculation, $MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{r_i^d}$., where Q is the story collection that acts as the queries in the retrieve attempts and r is the ranked position of the gold theme d for one story q.

In this case, a higher ranking r_i (represented by smaller numerical values) is considered better. The same setting is applied to theme-story retrieval.

BM25: We use the widely-used BM25 algorithm (Jones et al., 2000) as ranking baseline.

Dense passage retriever (DPR): We use the architecture of DPR (Karpukhin et al., 2020) for the retrieval task: two different BERT encoders are applied for a story (as a query) and a theme sentence (as a document) respectively. A dot-product similarity is used as a ranking metric. The objective is to learn vector representations such that matched story-theme pairs will have higher similarity scores.

Sentence-BERT: Sentence-BERT (Reimers and Gurevych, 2019) applies to the BERT model in a Siamese network structure. The story and theme sentences are processed independently through the model to get embeddings for similarity calculation.

MPnet: MPnet (Song et al., 2020) is trained on both masked and permuted language modeling and shows stronger capability in semantic representations. We apply MPnet in the bi-encoder architecture for this retrieval task. We train the bi-encoder dense retrievers using contrastive loss with negative samples, i.e., irrelevant theme sentences from other stories.

Cross-Encoder: We also experiment with a cross-encoder (CE) as a matching scorer. We use MiniLM (Wang et al., 2020) as our backbone model and perform binary relevance classification between combinations of all the story-theme pairs.

Table 4 shows the average rank of the correct theme for a story or the correct story for a given theme in the collection.

	Story-theme		Theme-story	
	Dev	Test	Dev	Test
BM25	0.35	0.28	0.22	0.14
DPR	0.49	0.31	0.52	0.33
BERT	0.46	0.24	0.46	0.28
MPnet	**0.65**	**0.40**	**0.60**	**0.43**
MiniLM $_{CE}$	0.42	0.27	-	-

Table 4: Retrieval performance (MRR) of different models on story and theme pairs.

4.2. Discussion

In this task, bi-encoders demonstrate substantial efficacy, even under the experimental assumption that each narrative is associated with a single correct theme. The challenges faced by retrieval models appear comparable in both story-theme and theme-story matching. Much like the theme keyword identification task, this task also implies the ambiguity inherent in story interpretation. It is conceivable that a theme (or story) deemed relevant may rank higher than the gold answer.

5. Story reading comprehension on themes

Different from many MRC tasks which may contain various contexts and questions, the reading comprehension task on themes is solely targeted

at understanding the main idea of the given context. So the question can be uniquely designed like *"What is the main idea of this story?"*, or can be simply omitted. Meanwhile, typical MRC tasks may include span extraction, cloze test, multiple choice, etc., however, story themes are not explicitly reflected in the story context, so problem forms like span extraction and cloze are not applicable. In this work, we design multiple choice problems for reading comprehension on themes: given a story x, the model is supposed to identify the correct theme y out of a collection of one correct answer and N distractor themes $\{d_1, d_2, ..., d_N\}$.

5.1. Experiments

To gain a better understanding of the challenges with the reading comprehension of themes in multiple-choice settings, we utilize pretrained Transformer models to address the multiple-choice problems and assess their performance using accuracy as the evaluation metric.

BERT and MPnet: We use the pretrained LMs as in the previous experiments. The training of BERT and MPnet follows the same schema: given the story x and a set of options $A = \{a_1, a_2, ..., a_k, ..., a_{N+1}\}$, where a_k is either the correct theme sentence y or any distractor d_i. The input is formed as the concatenation of the context and one option theme $(x \bigoplus a_i)$. Each of the options will be encoded with the story context and finally produce a set of contextualized representations by taking the *[CLS]* tokens. All of the representations should go through a linear classifier to determine the final answer probability $p(a_1, ..., a_{N+1}|x)$ using the cross-entropy loss.

Flan-T5: For the Text-To-Text multi-task model, Flan-T5, we simply define the input text template as $(x \bigoplus A)$ where every a_i in A is slightly modified by adding a prefix index letter (e.g., A, B, C, ...). Further, the model is trained to generate the correct option letter.

We employ three different strategies for selecting distractors as answer candidates:

- **Random**: randomly sample distractor candidates from the entire collection of themes;
- **From different virtue class**: select distractor candidates from themes belonging to different virtue categories than the context story;
- **From same virtue class**: choose distractor candidates from themes that fall within the same virtue category as the context story.

Table 5 gives the prediction performance comparison of different models under different distractor selection settings.

	Diff Virtue		Random		Same Virtue	
	Dev	Test	Dev	Test	Dev	Test
BERT	**0.67**	**0.58**	**0.74**	**0.68**	**0.72**	**0.64**
MPnet	0.37	0.26	0.52	0.29	0.37	0.36
Flan-T5	0.56	0.46	0.50	0.41	0.56	0.46

Table 5: Multiple-choice question answering experiment with prediction accuracy score reported.

5.2. Discussion

In the question-answering task, pretrained LMs demonstrate their ability to leverage learned information from finetuning, with cross-encoding prediction offering the best performance in this context. We also observe that distractors from the different virtue categories evidently pose a challenge for the best-performing model's predictions, while those from the same categories make the task easier, yet still remain challenging overall. A common observation from both the matching and multiple-choice Q&A tasks is that LLMs can identify relevant answers but still fall short of perfect performance. This leaves significant room for further exploration in this area.

6. Theme generation

We investigate theme interpretation as a text-generation task. We apply three large-scale LLMs for theme generation on ten held-out stories. The generation adopts a normalized language model prompting schema: A prefix *'Story: '* is added to the story content and a suffix task description and prompt *The main idea of this story is:* .

Flan-T5: We finetune the Text-To-Text language model with story-theme pairs using the template above.

OPT-175B: The OPT-175B model (Zhang et al., 2022) is pretrained on large-scale open-access datasets and has comparable prompting performance to the GPT-3 (Brown et al., 2020).

ChatGPT: OpenAI ChatGPT is an online chatbot service with strong general NLP capabilities. The backbone model is trained with human instructions in a prompt and finetuned with human feedback (Ouyang et al., 2022). The version of ChatGPT we experiment with is from February 2023.

Since there is no golden rule to validate the correctness of the generated theme, we use human evaluation on the results. Ten stories are selected as hold-out samples and not used for finetuning. Three human judges are presented with several themes, and asked to perform two tasks after reading a story:

1. Evaluate how reasonable each theme sounds and assign a score. The question for the human judges:

	Model generated theme			Model generated and original theme			
Source	Flan-T5	OPT-175B	ChatGPT	Flan-T5	OPT-175B	ChatGPT	Original
Human eval score	11	23	**50**	13	20	**46**	36
Best interpretation	7%	17%	**77%**	0%	10%	**50%**	40%

Table 6: Human evaluation of theme generation experiments. We implement two scenarios for human judgment: in the first, only model-generated themes are presented to the human evaluators; in the second, the original theme is included alongside the model-generated ones. A model can achieve a maximum human evaluation score of 60. The best interpretation is expressed as a percentage of the total votes from human judges.

`Give the rating scores for each main idea based on your judgment. Ratings: 2. Reasonable 1. Somewhat reasonable 0. Not reasonable.`
Given three judges, a maximum of two points that can be earned for each theme, and ten stories in total, the highest score a model can achieve is $3 \times 2 \times 10 = 60$.

2. Make a single choice on the best-generated theme sentence among the candidates. The question given to the judges is:
`Which answer is the best? <Multiple choice (single answer) question>.`
Each judge gets one vote per story to identify the best theme. We then calculate the percentage of votes each model received relative to the total number of votes to determine its performance.

During the evaluation, each theme's corresponding model name is hidden from the judges, and the order of presentation is randomized. The evaluation results can be found in Table 6. We categorize the experiment based on whether the original theme sentence accompanies the generated theme text pieces. Our aim is to determine if human judges might prefer the generated themes, providing a comparative view of the quality of the generated text.

6.1. Discussion

In the generation task, we are impressed by the remarkable performance of ChatGPT. The human evaluation shows that it has the capability to produce better quality interpretations than the original theme sentences. The generated interpretations and explanations provide insights into the training strategies used by the model.

A potential reason why human evaluators sometimes favor ChatGPT-generated themes over the original ones could be the inherent simplicity and ambiguity of some original themes. These original themes, often sourced from ancient literature, tend to use concise, philosophical wording that can be challenging to understand for speakers with elementary-level English proficiency. In contrast, ChatGPT articulates in straightforward language and provides a more comprehensive interpretation, which may resonate more with contemporary readers. To ensure a fair comparison, we further refined our experiment, directing ChatGPT to generate a single sentence similar to the original themes. The further designed prompt looks like: `<Story> Please tell me the main idea of this story. Limit your answer to one single sentence.` Table 7 displays the results from this adjusted experimental setup.

	Model generated and original theme			
Source	Flan-T5	OPT-175B	ChatGPT	Original
Human eval score	12	18	34	**44**
Best interpretation	3%	13%	30%	**53%**

Table 7: Human judgment on original theme with model-generated themes under one-sentence restriction.

The results indicate that while ChatGPT outperforms other LLM methods, it does not always secure the top preference from human judges. One prominent issue is that when restricted to interpreting the theme in a single sentence, the LLM sometimes defaults to summarizing the story. This shortcoming remains even when the only change in the experimental conditions is the added instruction to provide a one-sentence response. Examples of generated themes and the human-selected best themes are detailed in the Appendix section 9.3. These findings highlight potential limitations in LLMs' capabilities to offer concise theme interpretations. This observation points to a novel research question for future exploration: how can we ensure that LLMs consistently produce quality and reliable theme interpretations within a constrained context window?

7. Conclusion

In this work, we first emphasized the importance of advancing NLP models beyond literal comprehension to address more nuanced aspects of reading comprehension, specifically interpretive compre-

hension. This work served as an initial call for the research community to reflect on and further investigate machine reading comprehension (MRC) problems. We introduced theme interpretation as an NLP task in the context of inferential and interpretive reading comprehension, formulating the task comprehensively from various NLP research perspectives and providing extensive empirical research and analysis. This work builds upon existing MRC research, offering a new perspective by establishing an evaluation criterion for assessing the capability of an NLP system to reason about themes within a narrative content piece. In the future, we plan to expand the dataset for various research purposes and explore LLM behavior in constrained settings for theme interpretation.

Limitations

Limitations of the dataset: This dataset includes only English versions of stories and corresponding theme texts. The stories from different origins are not balanced in this collection. We acknowledge that many stories originate in Western culture, and contemporary editions might emphasize Western narratives. The sources of the stories are relatively diverse, but they are not balanced and are lacking stories from regions such as Oceania and Africa. Future users of this dataset should be mindful of this limitation in the context of inclusivity and diversity. Despite our best efforts to collect as many story-theme pairs as possible from the open Internet, the literature available for this research topic extends beyond what we've managed to include. We will put effort into further development in future research and invite the community to contribute.

Limitations of the usage of LLMs: We acknowledge that while current LLMs have limitations, they remain highly effective for executing NLP tasks. Existing research suggests that various prompting and instruction techniques can further enhance the zero-shot capabilities of LLMs for specific tasks. In this work, we choose simple, conversational prompts and instructions, similar to a natural conversation and question-answering, leaving further curated prompting and instruction techniques for future research.

Limitations of the annotation process and human evaluation: The annotations in the dataset are the result of human effort. We recognize and respect the reality that different people can interpret the same thing in diverse ways. Therefore, our annotations serve as references and are used as experimental ground truths, rather than being considered as universal, absolute truths. We release labels from all annotators and audits with their different opinions for further investigation, not limited to the NLP community.

Ethics Statement and Broader Impact

Ethical concerns on the dataset: We used our best efforts to collect educational stories that are freely accessible on the Internet. A considerable portion of these narratives originate from classical literary works, including *Aesop's Fables* and *Pañcatantra*. Some of this content may contain values and perspectives that are incompatible with modern sensibilities. Specifically, we discovered that a small number of stories (less than 3%) feature explicit, questionable content, such as gender and racial biases as well as discrimination against specific groups of people. Additionally, many fables contain stereotypical portrayals of characters like the 'evil wolf' or 'cunning fox,' which could potentially mislead audiences, particularly children, by oversimplifying the image of specific entities or individuals. We also source stories from contemporary success and entrepreneurial narratives. These narratives might reflect utilitarian values or materialism, and their use for educational purposes should be approached with caution. In our research, we have endeavored to minimize the risk associated with the use of this data by manually labeling and categorizing stories that contain questionable content or potentially raise ethical concerns.

Ethical concerns regarding the methodology: We use large-scale machine learning models (e.g., LLMs) to address the issue of theme interpretation. Present machine learning models carry the risk of inadvertently learning undesired patterns (such as biases) from the narratives in the training data and reproducing these during inference. In potential applications of this research, machine learning models may consequently generate biases originating from the training process. Additionally, the theme sentences provided by authors or editors may not represent diverse perspectives. Given the constraints of the model's training framework, the model may generate main ideas based solely on previously learned resources and knowledge. This could potentially restrict a user's perspective or reinforce the inherent bias in a real-world human-AI interaction scenario.

Broader impact: The broader impact of this work includes theme understanding in other contexts, such as dialogue and video, paving the way for potential research topics and applications, such as automatic lecture interpretation and meeting summarization.

Acknowledgements

We thank Tamzid Alam from the University of Houston, Hao Guo from Chalmers University of Technology, and Chaoxian Qi from the University of Houston for their help in data annotation. We would like