Are NLP Models Good at Tracing Thoughts: An Overview of Narrative Understanding

Lixing Zhu[a*] **Runcong Zhao**[a*] **Lin Gui**[a] **Yulan He**[ab†]
[a]Department of Informatics, King's College London
[b]The Alan Turing Institute, UK
{Lixing.Zhu,Runcong.Zhao,Lin.Gui,Yulan.He}@kcl.ac.uk

Abstract

Narrative understanding involves capturing the author's cognitive processes, providing insights into their knowledge, intentions, beliefs, and desires. Although large language models (LLMs) excel in generating grammatically coherent text, their ability to comprehend the author's thoughts remains uncertain. This limitation hinders the practical applications of narrative understanding. In this paper, we conduct a comprehensive survey of narrative understanding tasks, thoroughly examining their key features, definitions, taxonomy, associated datasets, training objectives, evaluation metrics, and limitations. Furthermore, we explore the potential of expanding the capabilities of modularized LLMs to address novel narrative understanding tasks. By framing narrative understanding as the retrieval of the author's imaginative cues that outline the narrative structure, our study introduces a fresh perspective on enhancing narrative comprehension.

1 Introduction

When reading a narrative, it is common for readers to analyze the author's cognitive processes, including their knowledge, intentions, beliefs, and desires (Castricato et al., 2021; Kosinski, 2023). In general, narrative is a medium for personal experiences (Somasundaran et al., 2018). Although Large Language Models (LLMs) have the capability to generate grammatically coherent texts, their ability to accurately capture the author's thoughts, such as the underlying skeletons or outline prompts devised by the authors themselves (Mahowald et al., 2023), remains questionable. This is supported by cognitive research that bilingual individuals tend to convey more precise thoughts compared to monolingual English speakers (Chee, 2006). The potential deficiency in tracing thoughts within narratives would hinder the practical application of narrative understanding and, thereby preventing readers from fully understanding the true intention of the authors.

Narrative understanding has been explored through various approaches that aim to recognize thoughts within narratives (Mostafazadeh et al., 2020; Kar et al., 2020; Lee et al., 2021; Sang et al., 2022). Still, these approaches are often fragmented, focusing on diverse tasks scattered across multiple datasets, obfuscating the fundamental elements (e.g., the characters, the events and their relationships) of the narrative structure (Ouyang and McKeown, 2014, 2015; Cutting, 2016). To address this gap, in this paper, we lay the foundation and provide a comprehensive synthesis of the aforementioned narrative understanding tasks. We start with the key features of this genre, followed by a formal definition of narrative understanding. We then present a taxonomy of narrative understanding tasks and their associated datasets, exploring how these datasets are constructed, the training objectives, and the evaluation metrics employed. We proceed to investigate the limitations of existing approaches and provide insights into new frontiers that can be explored by leveraging current modularized LLMs (e.g., GPT with RLHF (Ouyang et al., 2022)), with a particular focus on potential new tasks.

To sum up, our work firstly aligns disparate tasks with the LLM paradigm, and categorizes them based on the choices of context and input-output format. Then it catalogues datasets based on the established taxonomy. Subsequently, it introduces Bayesian prompt selection as an alternative approach to define the task of narrative understanding. Finally, it outlines open research directions.

2 Definition of Narrative Understanding

Narrative texts possess distinct characteristics, which are different from other forms of discourse. Elements such as point of view, salient characters,

*Equal contribution.
†Corresponding author.

and events, which are associated or arranged in a particular order (Chambers and Jurafsky, 2008; Ouyang and McKeown, 2015; Piper et al., 2021), giving rise to a cohesive story synopsis known as the plots (Hühn et al., 2014). Its scope spans across various genres, including novels, fiction, films, theatre, and more, within the domain of literary theory (Genette, 1988). Although it is unnecessary to endorse a particular narrative theory, some elements are commonly encountered in comprehension. For example, readers have to understand the causation or relationship that goes beyond a timeline and delve into the relationships between the characters (Worth, 2004). From a model-theoretic perspective, narrative understanding can be described as a process through which the audience perceives the narrator's constructed plot or thoughts (Czarniawska, 2004; Castricato et al., 2021).

To this end, we define narrative understanding as the process of reconstructing the writer's creative prompts that sketch the narrative structure (Ouyang and McKeown, 2014; Fan et al., 2019). In line with Brown et al. (2020), we adopt the practice of using the descriptions of NLP tasks as context to accommodate different paradigms. Additionally, we employ the LLaMA (Touvron et al., 2023) taxonomy to dichotomize this data-oriented task as either multiple-choice or free-form text completion. Let $\{x_n, y_n\}_{n=1}^{N}$ denote the dataset where $x_{1:N}$ are narratives, and $y_{1:N}$ are the annotated sketches, narrative understanding aims to predict Y given X by optimizing $p_\theta(y_{1:N}|x_{1:N}, \text{context})$. Existing literature can be roughly categorized based on the format of y_n and how context is described. To align with the classical NLP taxonomy, we specify context as a single prompt from either `Reading Comprehension` (Section 2.1), `Summarisation` (Section 2.2), or `Question Answering` (Section 2.3). In the rest of this section, the context will be further elaborated and specialized in more narrowly-defined tasks to refine the taxonomy, resulting in the reformatting of y_n accordingly.

2.1 Narrative Reading Comprehension

In machine reading comprehension, the context prompt is instantiated as a single prompt of "selecting which option is consistent with the story", and y_n is structured as categorical label(s) that correspond to the available options.

Narrative Consistency Check involves determining whether an assertion aligns with the narrative or contradicts it. This task encompasses various scopes, ranging from the entire narrative structure (Ouyang and McKeown, 2014) to discourse structure (Mihaylov and Frank, 2019) and ultimately, the constituents of the narrative, such as agents and events (Piper et al., 2021; Wang et al., 2021).

For example, Granroth-Wilding and Clark (2016) designed a Multiple Choice Narrative Cloze (MCNC) prediction task, where stories are structured as a sequence of events. Each event is represented by a 3-tuple, which comprises the verb lemma, the grammatical relation, and the associated entity. They aimed to predict the subsequent event from a given set of options, framed in the context of story cloze. Furthermore, Chaturvedi et al. (2017) extended this prediction task to encompass the prediction of a story ending based on its existing content. Similarly, the ROC story cloze task (Mostafazadeh et al., 2016), addressed by Cai et al. (2017), involves choosing the most plausible ending. There are various approaches developed for story ending prediction, such as the incorporation of commonsense knowledge (Li et al., 2018b), utilization of skip-thought embeddings (Srinivasan et al., 2018), entity-driven recurrent networks (Henaff et al., 2017; Liu et al., 2018), scene structure (Tian et al., 2020), centrality or salience of events (Zhang et al., 2021), and contextualized narrative event representation (Wilner et al., 2021), respectively. Simple and well-established, the Story Cloze Test does not cover the core aspects of narrative structure, though. Roemmele and Gordon (2018a) introduced an advancements in this task by predicting causally related events in stories using the Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011) dataset. Each instance in the COPA dataset contains three sentences: a *Premise*, *Alternative 1* and *Alternative 2*, with the *Premise* describing an event and the *Alternatives* proposing the plausible cause or effect of the event. Building upon this, Qin et al. (2019) aligned the ROC story cloze and COPA dataset with HellaSwag (Zellers et al., 2019) and introduced the counterfactual narrative reasoning. This task involves re-writing the story to restore narrative consistency. Their proposed TimeTravel dataset features 29,849 counterfactual revisions to initial story endings. Ippolito et al. (2020) further

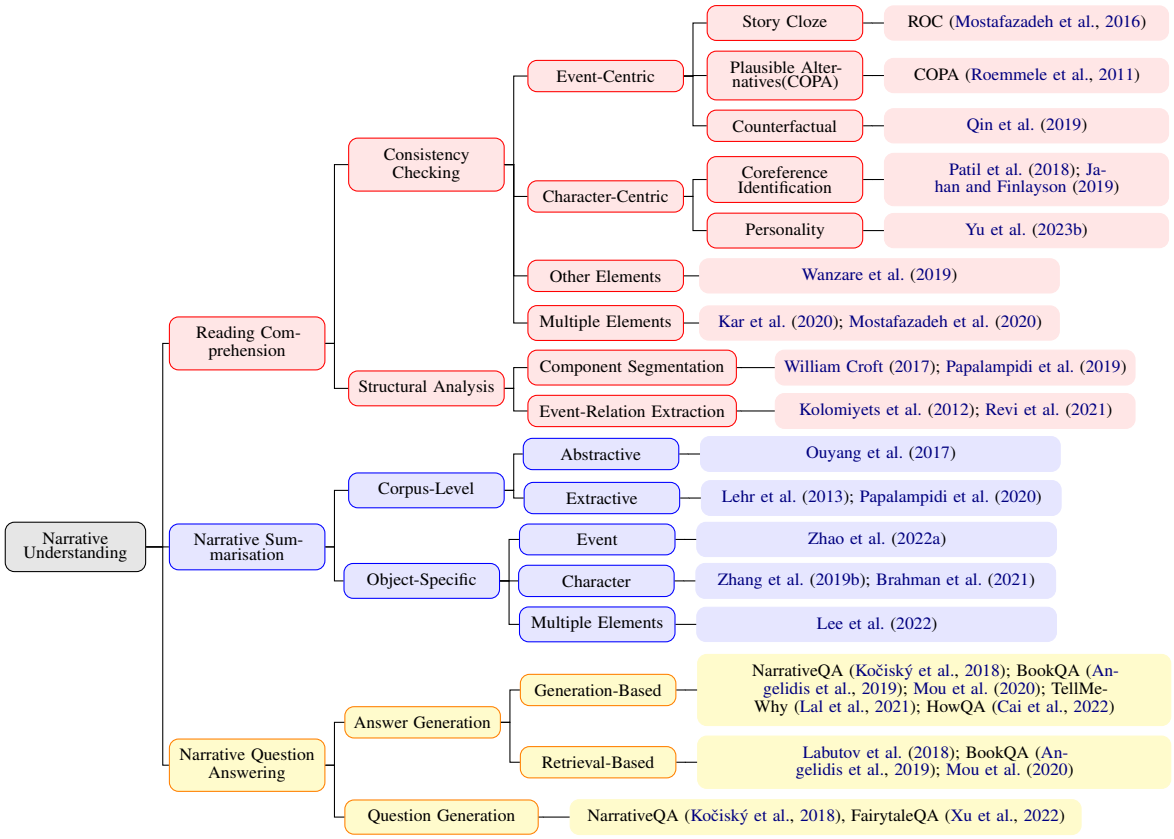

Figure 1: Typology of Narrative Understanding. Some literature sources are repeated since they contain both types of datasets or input-output schemes.

expanded the scope of the story cloze task to the entire narrative and proposed a sentence-level language model for consecutive multiple-choice prediction among candidate sentences on the Toronto Book Corpus (Zhu et al., 2015).

In addition to event prediction, Wanzare et al. (2019) introduced the concept of prototypical events, referred to as scenarios, to incorporate essential commonsense knowledge in narratives. They also introduced a benchmark dataset for scenario detection. Building upon their work, Situated Commonsense Reasoning (Ashida and Sugawara, 2022) aimed at deriving possible scenarios following a given story. Other attributes, such as a movie's success and event salience, have been explored in studies (Kim et al., 2019; Otake et al., 2020; Wilmot and Keller, 2021).

In line with the event consistency, characters, also known as participants, play a crucial role in linking narratives (Patil et al., 2018; Jahan and Finlayson, 2019). Patil et al. (2018) employed Markov Logic Network to encode linguistic knowledge for the identification of aliases of participants in a narrative. They defined participants as entities and framed the task as Named Entity Recognition (NER) and dependency parsing. In contrast, a recent approach introduced the TVShowGuess dataset, which simplified speaker guessing as multiple-choice selection (Sang et al., 2022). However, it is difficult for some narrative-specific NLP tasks, e.g., NER, to determine whether labelling 'a talking cup' as a 'person' is appropriate. To mitigate this, some approaches take a character-centric perspective. Brahman et al. (2021) introduced two new tasks: Character Identification, which assesses the alignment between a character and an unidentified description, and Character Description Generation, which emphasizes generating summaries for individual characters. Other work targeted personality prediction (Yu et al., 2023b), or used the off-the-shelf LLM to perform role extraction (Stammbach et al., 2022). To delve into the psychology of story characters and understand the causal connections between story events and the mental states of characters, Rashkin et al. (2018) introduced a dataset, StoryCommonsense, which contains the annotations of motivations and emotional reactions of story characters.

While much existing work on narrative understanding focuses on specific aspects, Kar et al.

(2020) considered the multifaceted features of narratives and created a multi-label dataset from both plot synopses and movie reviews. Chaturvedi et al. (2018) identified narrative similarity in terms of plot events, and resemblances between characters and their social relationships. Mostafazadeh et al. (2020) built the GeneraLized and COntextualized Story Explanations (GLUCOSE) dataset from children's stories and focused on explaining implicit causes and effects within narratives. It includes events, locations, possessions, and other attributes of the curated claims within the stories.

Structural Analysis of Events: Plot and Storyline Extraction The narrative structure encompasses a sequence of events that shape a story and define the roles of its characters (Hearst, 1997; Cutting, 2016). Unlike narrative consistency checking, which focuses on elementary consistency, this task involves two key objectives. First, it aims to extract a clear and coherent timeline that underlies the narrative's progression. Second, it aims to sequence the relationship of key factors to construct the plot of the narrative (Kolomiyets et al., 2012).

In the first line of approaches, Ouyang and McKeown (2015) considered significant shifts, referred to as Turning Points, in a narrative. These turning points represent the reportable events in the story. Li et al. (2018a) divided a typical story into five parts: *Orientation*, *Complicating Actions*, *Most Reportable Event*, *Resolution* and *Aftermath*, which are annotated in chronological order, capturing the temporal progression of the story. The temporal relationships within the narrative can be extracted and structured into a database of temporal events (Yao and Huang, 2018). In a similar vein, Papalampidi et al. (2019) strived to identify turning points by segmenting screenplays into thematic units, such as setup and complications. Anantharama et al. (2022) developed a pipeline approach that involves event triplet extraction and clustering to reconstruct a time series of narrative clusters based on identified topics.

Based on the types of event relations such as temporal, causal, or nested, storylines can be organised as timelines (Ansah et al., 2019; yang Hsu et al., 2021), hierarchical trees (Zhu and Oates, 2012), or directed graphs (Norambuena and Mitra, 2021; Yan and Tang, 2023). Current approaches often construct storylines using stated timestamps (Ansah et al., 2019; Revi et al., 2021). However, challenges arise in narratives where time details may be vague or absent. To address this issue, William Croft (2017) proposed to decompose the storyline by considering individual temporal subevents for each participant in a clausal event, which interact causally. Bamman et al. (2020) focused on resolving coreference in English fiction and presented the LitBank to resolve the long-distance within-document mentions. Building upon this work, Yu et al. (2023a) released a corpus of fiction and Wikipedia text to facilitate anaphoric reference discovery. Yan et al. (2019) introduced a more complex structure called Functional Schema, which utilizes language models, to reflect how storytelling patterns make up the narrative. Mikhalkova et al. (2020) introduces the Text World Theory (Werth, 1999; Wang et al., 2016) to regulate the structured annotations of narratives. This annotation scheme, profiling the world in narrative, is expanded in (Levi et al., 2022) by adding new narrative elements. Situated reasoning datasets, such as *Moral Stories* (Emelin et al., 2021), target the branching developments in narrative plots, specifically focusing on if-else scenarios. Tools have also been created for annotating the semantic relations among the text segments (Raring et al., 2022).

In addition to the aforementioned approaches for plot or storyline construction, several joint models have been developed to simultaneously uncover key elements and predict their connections. A notable work is PlotMachines (Rashkin et al., 2020) which involves an outline extraction method for automatic constructing the outline-story dataset. In another study, Lee et al. (2021) employed Graph Convolutional Networks (GCN) to predict entities and links on the StoryCommonsense and DesireDB datasets (Rahimtoroghi et al., 2017).

2.2 Narrative Summarisation

Narrative summarisation, often referred to as Story Retelling (Lehr et al., 2013), can be specified by restricting y_n to be a paraphrase that captures the essense of the original literature. Similar to other summarisation tasks, narrative summarisation can be extractive or abstractive, depending on whether the paraphrase is text snippets directly extracted from the story or is generated from input text.

Early tasks, such as Automated Narrative Retelling Assessment (Lehr et al., 2013), primarily focused on recapitulation story elements in the form of a tagging task. Subsequently, Narrative Summarisation Corpora (Ouyang et al., 2017; Papalampidi et al., 2020) were developed to facilitate

more comprehensive understandings of narratives. The former is designed for abstractive summarization, while the latter is intended for informative sentence retrieval, taking into account the inherent narrative structure. IDN-Sum (Revi et al., 2020) provides a unique view of summarisation within the context of Interactive Digital Narrative (IDN) games. Recent work has proposed benchmarks that require machines to capture specific narrative elements, e.g., synoptic character descriptions (Zhang et al., 2019b) and story rewriting anchored in six dimensions (Lee et al., 2022). In the same vein, Goyal et al. (2022) collected span-level annotations based on discourse structure to evaluate the coherence of summaries for long documents. Brahman et al. (2021) presented a character-centric view by introducing two new tasks: Character Identification and Character Description Generation. More recently, a benchmark called NarraSum (Zhao et al., 2022a) has been developed for large-scale narrative summarisation, encompassing salient events and characters, albeit without explicit framing.

2.3 Narrative Question Answering

Answering implicit, ambiguous, or causality questions from long narratives with diverse writing styles across different genres requires a deep level of understanding (Kalbaliyev and Sirts, 2022). From the task perspective, Narrative QA can be categorized based on either the format of y_n, or the task prompt that is referred to as context. The format of y_n could be categorical, where an answer is provided as a span specified by starting-ending positions. Alternatively, it can be free-form text that is generated. The context could be specified as answer selection/generation or question generation.

Numerous research works have focused on providing accurate answers to curated questions, with a specific focus on event frames (Tozzo et al., 2018), or questions related to external commonsense knowledge (Labutov et al., 2018). They all fall into the category of retrieval-based QA, where relevant information is selected from narratives. In contrast, the NarrativeQA (Kočiský et al., 2018) dataset took a different approach by instructing the annotators to ask questions and express answers in their own words after reading an entire long document, such as a book or a movie script. This resulted in high-quality questions designed by human annotators, and human-generated answers. The dataset further provided supporting snippets from human-written abstractive summaries and the original story.

To effectively handle long context, Tay et al. (2019) introduced a Pointer-Generator framework to sample useful excerpts for training, and chose between extraction and generation for answering. Meanwhile, the BookQA (Angelidis et al., 2019) approach targeted *Who* questions in the NarrativeQA corpus by leveraging BERT to locate relevant content. Likewise, Mou et al. (2020) proposed a two-step approach which consists of evidence retrieval to build a collection of paragraphs and a question-answering step to produce an answer given the collection. Mou et al. (2021) surveyed open-domain QA techniques and provided the Ranker-Reader solution, which improves upon the work of Mou et al. (2020) with a newer ranker and reader model.

Unlike pipeline approaches, Mihaylov and Frank (2019) converted the free-text answers from NarrativeQA into text-span answers and used the span answers as labels for training and prediction. Other attempts have been made to adapt NarrativeQA for extractive QA. For example, Frermann (2019) modified the dataset into an extractive QA format suitable for passage retrieval and answer span prediction. On the other hand, the TellMeWhy (Lal et al., 2021) dataset combined the commonsense knowledge and the characters' motivations in short narratives when designing the questions, presenting a new challenge for answering why-questions in narratives. Kalbaliyev and Sirts (2022) reviewed the WhyQA challenges, and Cai et al. (2022) collected how-to questions from *Wikihow* articles to build the HowQA dataset, which serves as a testbed for a Retriever-Generator model.

Generating meaningful questions is an important aspect of human intelligence that holds great educational potential for enhancing children's comprehension and stimulating their interest (Yao et al., 2022; Zhang et al., 2022a). Early work in this area focused on generating questions of multiple choice word cloze from children's Books, targeting named entities, nouns, verbs and prepositions (Hill et al., 2016). Other studies designed commonsense-related questions from narratives in simulated world (Labutov et al., 2018). To enhance children's learning experiences, the FairytaleQA (Xu et al., 2022) dataset was created for question generation (QG) tasks, covering seven types of narrative elements or relations. The dataset was used in the work of (Yao et al., 2022), which employed a

pipeline approach comprising heuristics-based answer generation, BART-based question generation, and DistilBERT-based candidate ranking. Zhao et al. (2022b) experimented with a question generation model, which incorporated a summarisation module, using the same FairytaleQA dataset. Additionally, the StoryBuddy (Zhang et al., 2022a) system provided an interactive interface for parents to participate in the process of generating question-answer pairs, serving an educational purpose.

A major taxonomical concern in tasks or datasets employing similar input-output schemes, particularly NarrativeQA (Kočiský et al., 2018) with the objective of answering questions relating to narrative consistency, is the potential overlap between finely-detailed tasks as defined in our taxonomy. For example, depending on the instructions and prompts given, the scope of QA can encompass more specific tasks, including narrative summarization (Who is Charlie?), and narrative generation (What would happen after the Princess marries the Prince?). Despite some similarities, the majority of instruction templates are indistinguishable (Sanh et al., 2022). Therefore, they are considered as equivalent tasks.

3 Dataset

We have conducted a review of datasets (Table 1-3 in Appendix) that are either designed for, or applicable to, tasks related to narrative understanding. There are three major concerns:

Data Source Due to copyright restrictions, most datasets focus on public domain works. This limitation has rendered valuable resources, such as the Toronto Book Corpus (Zhu et al., 2015), unavailable. To overcome this challenge, diverse sources have been explored, such as plot descriptions from movies and TV episodes (Frermann et al., 2018; Zhao et al., 2022a). However, there remains a need for datasets that cover specialized knowledge bases for specific worldviews in narratives, such as magical worlds, post-apocalyptic wastelands, and futuristic settings. One potential data source that could be utilized for this purpose is TVTropes [1], which provides extensive descriptions of character traits and actions.

Data Annotation The majority of existing datasets contain short stories, consisting of only a few sentences, which limits their usefulness for understanding complex narratives found in books and novels. Due to high annotation costs, there is a lack of sufficiently annotated datasets for these types of materials (Zhu et al., 2015; Bandy and Vincent, 2021). Existing work (Frermann et al., 2018; Chaudhury et al., 2020; Kryscinski et al., 2022) employed available summaries and diverse meta-information from books, movies, and TV episodes to generate sizable, high-quality datasets. Additionally, efficient data collection strategies, such as game design (Yu et al., 2023a), character-actor linking for movies (Zhang et al., 2019b), and leveraging online reading notes (Yu et al., 2023b), can be explored to facilitate the creation of datasets.

Data Reuse Despite the availability of high-quality annotated datasets for various narrative understanding tasks, there is limited reuse of these datasets in the field. Researchers often face challenges in finding suitable data for their specific tasks, which leads to the creatation of their own new and costly datasets. Some chose to build a large, general dataset (Zhu et al., 2015; Mostafazadeh et al., 2020), while others chose to gradually annotate the same corpus over time (Frermann et al., 2018). Some made use of platforms such HuggingFace for data sharing (Huang et al., 2019), or provided dedicated interfaces for public access (Koupaee and Wang, 2018). To facilitate data reuse and address the challenges associated with finding relevant data, the establishment of an online repository could prove beneficial.

4 Evaluation Methods

For classification tasks, such as multiple-choice QA and next sentence prediction, accuracy, precision, recall, and the F1 score serve as the most suitable evaluation metrics. Here we mainly discuss evaluation metrics for free-form or generative tasks, crucial yet challenging for narrative understanding. Traditional metrics based on N-gram overlap like BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2003), METEOR (Lavie and Agarwal, 2007), CIDER (Vedantam et al., 2015), WMD (Kusner et al., 2015), and their variants, such as NIST (Doddington, 2002), BLEUS (Lin and Och, 2004), SacreBLEU (Post, 2018) remain popular due to their reproducibility. However, their correlation with human judgment is weak for certain tasks (Novikova et al., 2017; Gatt and Krahmer, 2018).

Alternative content overlap measures have been proposed, such as PEAK (Yang et al., 2016), which

[1] http://tvtropes.org

compares weighted Summarization Content Units, and SPICE (Anderson et al., 2016), which evaluates overlap by parsing candidate and reference texts into scene graphs representing objects and relations. The advent of BERT introduced a new approach to evaluation, based on contextualized embeddings, as seen in metrics such as BERTScore (Zhang et al., 2019a), MoverScore (Zhao et al., 2019), and BARTScore (Yuan et al., 2021). Task-specific evaluations using BERT-based classifiers, such as FactCC (Kryscinski et al., 2020) for summary consistency and QAGS (Wang et al., 2020) to determine if answers are sourced from the document, have been explored.

Recently, LLMs have also been employed for evaluation tasks. For instance, CANarEx (Ananthara ma et al., 2022) assesses time-series cluster recovery from GPT-3 generated synthetic data. UniEval (Zhong et al., 2022) uses a pretrained T5 model for evaluation score computation. GPTScore (Fu et al., 2023) leverages zero-shot capabilities of ChatGPT for text scoring, while G-Eval (Liu et al., 2023) applies LLMs within a chain-of-thoughts framework and form-filling paradigm for output quality assessment.

To facilitate tracking the research progress, we list the performance of current SoTA models on representative benchmarks in Table 4 in the Appendix. These benchmarks have been categorized into specific tasks based on the instruction templates outlined in Section 2. It is worth noting that methodological development or result analysis is not the focus of this survey. Sections 2.1-2.3 are intended to serve as timelines depicting the evolution of different research directions and task comparisons. Further, Section 3 offers a review of the diverse datasets and their respective formats.

5 Applications

Narrative Assessment The "good at language -> good at thought" fallacy (Mahowald et al., 2023) has spurred the application of narrative understanding to the automatic assessment of student essays. Studies on automatic story evaluation (Chen et al., 2022; Chhun et al., 2022) reveal that the referenced metrics, e.g., BLEU and ROUGE scores, deviate from human preferences such as aesthetic and intrigue. This calls for the identification of narrative elements and their relations. For example, Somasundaran et al. (2016) builds a graph of discourse proximity of essay concepts to predict the essay quality w.r.t. the development of ideas and exemplification. Somasundaran et al. (2018) annotates multiple dimensions of narrative quality, such as narrative development and narrative organization, to combat the scarcity of scored essays. Such narrativity is also evaluated in (Steg et al., 2022) with a focus on detecting the cognitive aspects, i.e., suspense, curiosity, and surprise. Other sub-tasks, such as comment generation, are studied in (Lehr et al., 2013; Zhang et al., 2022b).

Story Infilling As mentioned in Section 2.1, the story cloze task selects a more coherent ending based on the story context. The story infilling completes a story in a similar way that sequences of words are removed from the text to create blanks for a replacement (Ippolito et al., 2019). Mori et al. (2020) makes a step forward in detecting the missing or flawed part of a story. The proposed method predicts the positions and provides alternative wordings, which serves as a writing assistant. It is worth mentioning that narrative generation intersects with this task with the key difference that story infilling aims to comprehend the narratives and generate minor parts to complete the story. Wider applications of this task are auxiliary writing systems such as an educational question designer (Zhang et al., 2022a) and a creative writing supporter (Roemmele and Gordon, 2018b).

Narrative Understanding vs. Narrative Generation The main distinction we make between narrative understanding and generation is that the latter aims to produce longer sequences conditioned on prototypical story snippets. An epitome is story generation conditioned upon prompts (Fan et al., 2019), where the prompts draft the action plan and reserve the placeholders for generative models to complete. Unlike Story Infilling, the out-of-distribution narrative elements and their relations, e.g., plot structures (Goldfarb-Tarrant et al., 2020), novel plots (Ammanabrolu et al., 2019), creative interactions (DeLucia et al., 2021), and interesting endings (Gupta et al., 2019), are to be generated, which poses the main challenge in story generation (Goldfarb-Tarrant et al., 2019). Other literature focuses on the significant enrichment of details to a brief story skeleton (Zhai et al., 2020). Latent discrete plans illustrated by thematic keywords are posited (Peng et al., 2018), and latent variable models are leveraged to steer the generation (Jhamtani and Berg-Kirkpatrick, 2020). Commonsense reasoning also needs to be pondered for storytelling

everyday scenarios (Mao et al., 2019; Xu et al., 2020). Evaluation criteria are adjusted accordingly to measure the aesthetic merit and correlate with human preferences (Akoury et al., 2020; Chhun et al., 2022). In this sense, narrative generation is the opposite task of narrative understanding with a key emphasis on generating intriguing plots and rich details based on a small corpus. Despite being the opposite in the model-theoretic view, narrative generation needs to comply with certain restrictions, e.g., pragmatics, coreference consistency (Clark et al., 2018), long-range cohesion (Zhai et al., 2019; Goldfarb-Tarrant et al., 2020), and adherence to genres (Alabdulkarim et al., 2021), to name a few.

6 Challenges and Future Directions

Prompt Tuning and Author Prompt Recapitulation In lieu of the prompt-driven nature of LLMs that elicits tasks with task descriptions or few-shot examples rather than task-specific fine-tuning (Brown et al., 2020; Sanh et al., 2022; Touvron et al., 2023; Tay et al., 2023; Anil et al., 2023), we propose a unified approach for narrative understanding. This approach involves a single supervised training process, denoted as $p_\theta(y_{1:N}|x_{1:N}, \mathsf{context})$, where $\mathsf{context}$ represents a prompt of task description or few-shot examples, and $y_{1:N}$ denotes predictable annotations produced by crowd-sourced platforms such as Amazon Mechanical Turk (Mostafazadeh et al., 2016; Papalampidi et al., 2020; Lal et al., 2021). While the framework is universally applicable, the practice of relying on annotators to infer authors' thoughts or construct skeleton prompts is inefficient, inaccurate and unreliable (Mahowald et al., 2023), Direct consultation with the authors is often impractical, particularly for posthumous masterworks. As suggested in the model-theoretic view (Castricato et al., 2021), a gap needs to be closed between the narrator and the audience to overcome the reader's uncertainty and other environmental limitations.

In this regard, we envisage a Bayesian perspective (Lyle et al., 2020) for the recovery of the author's thoughts. Let $p_\phi(\mathsf{narrative}|\mathsf{sketch}, \mathsf{task})$ denote the oracle model that simulates the author's composition process, where the author fills in the conceived skeleton out of their own initiative (where sketch is the skeleton prompts and task is the task description). The recapitulation of the author's prompt can then be expressed as $\mathrm{argmax}_{\mathsf{sketch}}\ p_\phi(\mathsf{narrative}|\mathsf{sketch})$. Previous efforts have formulated the objective as predicting the narrative elements or the narrative structure (as discussed in Section 2), given the intractability of the likelihood of generating the narrative, $p_\phi(\mathsf{narrative}|\mathsf{sketch}, \mathsf{task})$. It is also viable, however, to probe and optimize sketch directly, considering the ability of LLMs to compute the error in close proximation. Hence, $p_\phi(\mathsf{narrative}|\mathsf{sketch})$ could be derived as the marginal likelihood $\Sigma_{\mathsf{task}} p_\phi(\mathsf{narrative}, \mathsf{task}|\mathsf{sketch})$, and its expectation form $\mathbb{E}_{p(\mathsf{task}|\mathsf{sketch})} p_\phi(\mathsf{narrative}|\mathsf{sketch}, \mathsf{task})$ can be approximated by sampling an appropriate task. The dependence between task and sketch aligns well with the composition practice of choosing a suitable genre to match the author's creative intent. In more complex cases where the narrator is known to have employed particular narratology techniques (e.g., Flashback (Han et al., 2022)), task and sketch can be parameterized by generative models (e.g., LM-driven prompt engineers (Zhou et al., 2023)), which can be tuned by a prompt base through gradient descent optimization.

Interactive Narrative LLMs, with their ability to carry out numerous language processing tasks effectively in a zero-shot manner (Shen et al., 2023), are paving the way towards a future of immersive and interactive narrative environments. These environments could resemble the dynamic storylines experienced by individuals, as depicted in the TV series "Westworld".

Agent Recent studies (Park et al., 2023; AutoGPT, 2023) have shown promising results in using a database to store an agent's experiences and thought processes, effectively serving as a personality repository. By retrieving relevant memories from this database and incorporating them into prompts, LLMs can be guided to predict behaviors, thereby producing human-like responses. However, extracting comprehensive character-centric memory from narratives, encompassing aspects such as "Who", "When", "Where", "Action", "Feeling", "Causal relation", "Outcome", and "Prediction" (Xu et al., 2022), remain largely unexplored. Current studies primarily focus on simple corpora and there is ample room for further investigation in this area.

Environment Creating immersive and interactive environments for users and agents presents several challenges, primarily due to three key factors: **(1) Environment Extraction**. Character locations and environments, often vaguely defined unless crucial

to the plot, have to be clarified. Most works rely on pre-built sandbox environments (Côté et al., 2018; Hausknecht et al., 2020; Park et al., 2023) to address this issue. However, challenges remain in extracting and representing the environment accurately. **(2) Environment Generation**. Interactive narratives aim to provide users with greater freedom, but this poses the challenge of automatically generating reasonable and coherent details within the narrative's world. It is crucial to maintain consistency and engagement in storytelling, despite varying user inputs and directions. **(3) Environment Update**. Agents' text commands may change the state of the world, requiring accurate and cost-effective updates. Current systems update environment states using predefined rules (Côté et al., 2018; Wang et al., 2022). However, using LLMs to derive and generate narrative environments challenges the use of predefined rules, making efficient and large-scale environment updates a future research direction.

Open World Knowledge Incorporating external knowledge and commonsense has been a long-standing challenge in both dataset construction and model design (Zellers et al., 2019; Wanzare et al., 2019; Mikhalkova et al., 2020; Ashida and Sugawara, 2022). Efforts to address this challenge have been made over the years, with the emergence of LLMs providing a source potential of commonsense knowledge (Bosselut et al., 2019; Petroni et al., 2019). Notably, the Text World Theory (TWT) (Werth, 1999; Gavins, 2007) has been leveraged to provide world knowledge relevant to everyday life, which is simulated through natural language descriptions (Labutov et al., 2018). Similarly, in (Mikhalkova et al., 2020), a text world framework is established, in which the world-building elements (e.g., characters, time and space) are annotated to enhance readers' perception.

The text world entails nuanced knowledge derived from the interplay of various elements. However, such supervision provided by the static world is somewhat limited, as the reward is implicit and the model needs to extrapolate from the annotations to exploit the world knowledge. In contrast, the open world (Raistrick et al., 2023) presents an ideal source of supervisor, where the model can be rewarded with incentives derived from world mechanisms (Assran et al., 2023) that synergistically complements the audience model with the everyday commonsense. To this end, the RLHF (Reinforcement Learning with Human Feedback) (Ouyang et al., 2022) strategy could be applied to the open-world system, which enables the iterative training of the narrative understanding process.

7 Conclusion

In this paper, we have systematically examined the emerging field of narrative understanding, cataloguing the approaches and highlighting the unique challenges it poses along with the potential solutions that have emerged. We have emphasized the crucial role of LLMs in advancing narrative understanding. Our intention is for this survey to serve as a thorough guide for researchers navigating this intricate domain, drawing attention to both the commonalities and unique aspects of narrative understanding in relation to other NLP research paradigms. We aspire to bridge the gap between existing works and potential avenues for further development, thus inspiring meaningful and innovative progress in this fascinating field.

Limitations

This paper provides an overview of narrative understanding tasks, drawing inspiration from computational narratology (Matthews et al., 2003) and exploring potential new directions. However, it does not delve into broader concepts of cognitive computational theory, such as the theory of mind (Happé, 1994), the philosophy of reading (Mathies, 2020) and pedagogics (Nicolopoulou and Richner, 2007). Therefore, this survey does not incorporate cognitive-theoretic insights into the underlying mechanisms that contribute to the success of models. Another major limitation is the lack of discussion of methodological improvements, as the focus of the research progression is primarily centered around the tasks. Additionally, this survey does not explore narrative generation tasks.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and helpful suggestions. This work was funded by the UK Engineering and Physical Sciences Research Council (grant no. EP/T017112/1, EP/T017112/2, EP/V048597/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1, EP/V020579/2).

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: Challenges and attempts. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara Martin, and Mark Riedl. 2019. Guided neural language generation for automated storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 46–55, Florence, Italy. Association for Computational Linguistics.

Nandini Anantharama, Simon Angus, and Lachlan O'Neill. 2022. CANarEx: Contextually aware narrative extraction for semantically rich text-as-data applications. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3551–3564, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of Machine Translation Summit IX: Papers*, Amsterdam, The Netherlands. Springer.

Stefanos Angelidis, Lea Frermann, Diego Marcheggiani, Roi Blanco, and Lluís Màrquez. 2019. Book QA: Stories of challenges and opportunities. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 78–85, Hong Kong, China. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Jeffery Ansah, Lin Liu, Wei Kang, Selasie Kwashie, Jixue Li, and Jiuyong Li. 2019. A graph is worth a thousand words: Telling event stories using timeline summarization graphs. In *Proceedings of the World Wide Web Conference*, page 2565–2571, San Francisco, CA, USA. Association for Computing Machinery.

Mana Ashida and Saku Sugawara. 2022. Possible stories: Evaluating situated commonsense reasoning under multiple possible scenarios. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3606–3630, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, Vancouver, Canada. IEEE Computer Society.

AutoGPT. 2023. Auto-gpt: An autonomous gpt-4 experiment.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2017. Embracing data abundance: Booktest dataset for reading comprehension.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, 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*, page 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, online. Curran Associates, Inc.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pengshan Cai, Mo Yu, Fei Liu, and Hong Yu. 2022. Generating coherent narratives with subtopic planning to answer how-to questions. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 26–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending:strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622, Vancouver, Canada. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2021. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, page 77–86, Vancouver, Canada. Association for Computational Linguistics.

Louis Castricato, Stella Biderman, David Thue, and Rogelio Cardona-Rivera. 2021. Towards a model-theoretic view of narratives. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 95–104, Virtual. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have I heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 673–678, New Orleans, Louisiana. Association for Computational Linguistics.

Atef Chaudhury, Makarand Tapaswi, Seung Wook Kim, and Sanja Fidler. 2020. The shmoop corpus: A dataset of stories with loosely aligned summaries.

Michael W.L. Chee. 2006. Dissociating language and word meaning in the bilingual brain. *Trends in Cognitive Sciences*, 10(12):527–529.

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games. In *Proceedings of the Computer Games Workshop at IJCAI 2018*, Stockholm, Sweden. IJCAI Press.

James E Cutting. 2016. Narrative theory and the dynamics of popular movies. *Psychonomic Bulletin & Review*, 23(1):1713–1743.

Barbara Czarniawska. 2004. Narratives in social science research.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846, Vancouver, Canada. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, page 138–145, San Francisco, United States. Morgan Kaufmann Publishers Inc.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Lea Frermann. 2019. Extractive NarrativeQA with heuristic pre-training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 172–182, Hong Kong, China. Association for Computational Linguistics.

Lea Frermann, Shay B. Cohen, and Mirella Lapata. 2018. Whodunnit? crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, 6:1–15.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of AI Research*, 61(1):65–170.

Joanna Gavins. 2007. *Text World Theory: An Introduction*. Edinburgh University Press.

Gérard Genette. 1988. *Narrative Discourse Revisited*. Cornell University Press.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 89–97, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2727–2733. AAAI Press.

Prakhar Gupta, Vinayshekhar Bannihatti Kumar, Mukul Bhutani, and Alan W Black. 2019. WriterForcing: Generating more interesting story endings. In *Proceedings of the Second Workshop on Storytelling*, pages 117–126, Florence, Italy. Association for Computational Linguistics.

Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. 2022. Go back in time: Generating flashbacks in stories with event temporal prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1450–1470, Seattle, United States. Association for Computational Linguistics.

Francesca G. E. Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2):129–154.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7903–7910, New York, New York, USA. AAAI Press.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Peter Hühn, Jan Christoph Meister, John Pier, and Wolf Schmid, editors. 2014. *Handbook of Narratology*. De Gruyter, Berlin, München, Boston.

Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. 2019. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43, Minneapolis, Minnesota. Association for Computational Linguistics.

Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. 2020. Toward better storylines with sentence-level language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7472–7478, Online. Association for Computational Linguistics.

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

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2020. Narrative text generation with a latent discrete plan. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3637–3650, Online. Association for Computational Linguistics.

Emil Kalbaliyev and Kairit Sirts. 2022. Narrative why-question answering: A review of challenges and datasets. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 520 – 530, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sudipta Kar, Gustavo Aguilar, Mirella Lapata, and Thamar Solorio. 2020. Multi-view story characterization from movie plot synopses and reviews. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5629–5646, Online. Association for Computational Linguistics.

You Jin Kim, Yun Gyung Cheong, and Jung Hoon Lee. 2019. Prediction of a movie's success from plot summaries using deep learning models. In *Proceedings of the Second Workshop on Storytelling*, pages 127–135, Florence, Italy. Association for Computational Linguistics.

Tomáš Kočiský, 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.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea. Association for Computational Linguistics.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 9332–9346, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, page 957–966, Lille, France. JMLR.org.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 2018. Multi-relational question answering from narratives: Machine reading and reasoning in simulated worlds. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 833–844, Melbourne, Australia. Association for Computational Linguistics.

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*, page 5043–5054, Online. Association for Computational Linguistics.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, page 228–231, Prague, Czech Republic. Association for Computational Linguistics.

I-Ta Lee, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Modeling human mental states with an entity-based narrative graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4926, Online. Association for Computational Linguistics.

Yoonjoo Lee, Tae Soo Kim, Minsuk Chang, and Juho Kim. 2022. Interactive children's story rewriting through parent-children interaction. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants*, pages 62–71, Dublin, Ireland. Association for Computational Linguistics.

Maider Lehr, Izhak Shafran, Emily Prud'hommeaux, and Brian Roark. 2013. Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–220, Atlanta, Georgia. Association for Computational Linguistics.

Effi Levi, Guy Mor, Tamir Sheafer, and Shaul Shenhav. 2022. Detecting narrative elements in informational text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1755–1765, Seattle, United States. Association for Computational Linguistics.

Boyang Li, Beth Cardier, Tong Wang, and Florian Metze. 2018a. Annotating high-level structures of short stories and personal anecdotes. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association (ELRA).

Qian Li, Ziwei Li, Jin-Mao Wei, Yanhui Gu, Adam Jatowt, and Zhenglu Yang. 2018b. A multi-attention based neural network with external knowledge for story ending predicting task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1754–1762, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, page 150–157. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, page 501–507, Geneva, Switzerland. International Committee on Computational Linguistics.

Fei Liu, Trevor Cohn, and Timothy Baldwin. 2018. Narrative modeling with memory chains and semantic supervision. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–284, Melbourne, Australia. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Clare Lyle, Lisa Schut, Robin Ru, Yarin Gal, and Mark van der Wilk. 2020. A bayesian perspective on training speed and model selection. In *Advances in Neural Information Processing Systems*, volume 33, pages 10396–10408. Curran Associates, Inc.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. Improving neural story generation by targeted common sense grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5988–5993, Hong Kong, China. Association for Computational Linguistics.

Susanne Mathies. 2020. The simulated self – fiction reading and narrative identity. *Philosophia*, 48(1):325–345.

Alastair Matthews, Jan Christoph Meister, and Marie-Laure Ryan. 2003. *Computing Action: A Narratological Approach*.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.

Elena Mikhalkova, Timofei Protasov, Polina Sokolova, Anastasiia Bashmakova, and Anastasiia Drozdova. 2020. Modelling narrative elements in a short story: A study on annotation schemes and guidelines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 126–132, Marseille, France. European Language Resources Association.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. Inscript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, page 3485–3493, Portorož, Slovenia. Association for Computational Linguistics.

Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. 2020. Finding and generating a missing part for story completion. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 156–166, Online. International Committee on Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4586, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James F. Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, page 46–51, Valencia, Spain. Association for Computational Linguistics.

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. *Transactions of the Association for Computational Linguistics*, 9:1032–1046.

Xiangyang Mou, Mo Yu, Bingsheng Yao, Chenghao Yang, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2020. Frustratingly hard evidence retrieval for QA over books. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 108–113, Online. Association for Computational Linguistics.

Ageliki Nicolopoulou and Elizabeth S. Richner. 2007. From actors to agents to persons: The development of character representation in young children's narratives. *Child Development*, 78(2):412–429.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2273–2284, Florence, Italy. Association for Computational Linguistics.

Brian Felipe Keith Norambuena and Tanushree Mitra. 2021. Narrative maps: An algorithmic approach to represent and extract information narratives. *Proceedings of the ACM on Human-Computer Interaction*, 4(228):1–333.

Jekaterina Novikova, Ondˇrej Dusˇek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Carolin Odebrecht, Lou Burnard, and Christof Schöch. 2021. [link].

OpenAI. 2023. Gpt-4 technical report.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. Mcscript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics*, page 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Takaki Otake, Sho Yokoi, Naoya Inoue, Ryo Takahashi, Tatsuki Kuribayashi, and Kentaro Inui. 2020. Modeling event salience in narratives via barthes' cardinal functions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1784–1794, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jessica Ouyang, Serina Chang, and Kathy McKeown. 2017. Crowd-sourced iterative annotation for narrative summarization corpora. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling reportable events as turning points in narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158, Lisbon, Portugal. Association for Computational Linguistics.

Jessica Ouyang and Kathy McKeown. 2014. Towards automatic detection of narrative structure. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 4624–4631,

Reykjavik, Iceland. European Language Resources Association.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, New Orleans, LA, USA. Curran Associates, Inc.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online. Association for Computational Linguistics.

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*, page 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia Pennsylvania. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. Association for Computing Machinery.

Sangameshwar Patil, Sachin Pawar, Swapnil Hingmire, Girish Palshikar, Vasudeva Varma, and Pushpak Bhattacharyya. 2018. Identification of alias links among participants in narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 63–68, Melbourne, Australia. Association for Computational Linguistics.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186–191, Belgium, Brussels. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, page 152–160, Portland, Oregon, USA. Association for Computational Linguistics.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369, Saarbrücken, Germany. Association for Computational Linguistics.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. 2023. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641. IEEE Computer Society.

Michael Raring, Malte Ostendorff, and Georg Rehm. 2022. Semantic relations between text segments for semantic storytelling: Annotation tool - dataset - evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4923–4932, Marseille, France. European Language Resources Association.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4274–4295, Online. Association for Computational Linguistics.

Ashwathy T. Revi, Stuart E. Middleton, and David E. Millard. 2020. Idn-sum: A new dataset for interactive digital narrative extractive text summarisation. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, page 1–12, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ashwathy T. Revi, Stuart E. Middleton, and David E. Millard. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. 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*, Seattle, Washington, USA. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95. AAAI Press.

Melissa Roemmele and Andrew Gordon. 2018a. An encoder-decoder approach to predicting causal relations in stories. In *Proceedings of the First Workshop on Storytelling*, pages 50–59, New Orleans, Louisiana. Association for Computational Linguistics.

Melissa Roemmele and Andrew Gordon. 2018b. Linguistic features of helpfulness in automated support for creative writing. In *Proceedings of the First Workshop on Storytelling*, pages 14–19, New Orleans, Louisiana. Association for Computational Linguistics.

Melissa Roemmele, Kyle Shaffer, Katrina Olsen, Yiyi Wang, and Steve DeNeefe. 2023. Ablit: A resource for analyzing and generating abridged versions of english literature. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. 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, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the Tenth International Conference on Learning Representations*, Virtual. ICLR Press.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. 2020. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1978, Online. Association for Computational Linguistics.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578, Osaka, Japan. The COLING 2016 Organizing Committee.

Siddarth Srinivasan, Richa Arora, and Mark Riedl. 2018. A simple and effective approach to the story cloze test. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 92–96, New Orleans, Louisiana. Association for Computational Linguistics.

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

Max Steg, Karlo Slot, and Federico Pianzola. 2022. Computational detection of narrativity: A comparison using textual features and reader response. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–114, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA. IEEE Computer Society.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931, Florence, Italy. Association for Computational Linguistics.

Zhixing Tian, Yuanzhe Zhang, Kang Liu, Jun Zhao, Yantao Jia, and Zhicheng Sheng. 2020. Scene restoring for narrative machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3063–3073, Online. Association for Computational Linguistics.

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

Alex Tozzo, Dejan Jovanović, and Mohamed Amer. 2018. Neural event extraction from movies description. In *Proceedings of the First Workshop on Storytelling*, pages 60–66, New Orleans, Louisiana. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA. IEEE Computer Society.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5008–5020, Online. Association for Computational Linguistics.

Jing Wang, Yufang Ho, Zhijie Xu, Dan McIntyre, and Jane Lugea. 2016. The visualisation of cognitive structures in forensic statements. In *2016 20th International Conference Information Visualisation (IV)*, pages 106–111, Lisbon, Portugal. IEEE Computer Society.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11279 – 11298, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shichao Wang, Xiangrui Cai, HongBin Wang, and Xiaojie Yuan. 2021. Incorporating circumstances into narrative event prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4840–4849, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lilian Diana Awuor Wanzare, Michael Roth, and Manfred Pinkal. 2019. Detecting everyday scenarios in narrative texts. In *Proceedings of the Second Workshop on Storytelling*, pages 90–106, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Paul Werth. 1999. *Text Worlds: Representing Conceptual Space in Discourse*. Pearson Education Limited, New York, USA. 2010.

Michael Regan William Croft, Pavlína Pešková. 2017. Integrating decompositional event structures into storylines. In *Proceedings of the Events and Stories in the News Workshop*, page 98–109, Vancouver, Canada. Association for Computational Linguistics.

David Wilmot and Frank Keller. 2021. Memory and knowledge augmented language models for inferring salience in long-form stories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 851–865, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative embedding: Re-Contextualization through attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah E Worth. 2004. Narrative understanding and understanding narrative. *Contemporary Aesthetics*, 2.

Wenhan Xiong, Anchit Gupta, Shubham Toshniwal, Yashar Mehdad, and Wen tau Yih. 2022. Adapting pretrained text-to-text models for long text sequences.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby 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*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Xinru Yan, Aakanksha Naik, Yohan Jo, and Carolyn Rose. 2019. Using functional schemas to understand social media narratives. In *Proceedings of the Second Workshop on Storytelling*, pages 22–33, Florence, Italy. Association for Computational Linguistics.

Zhihua Yan and Xijin Tang. 2023. Narrative graph: Telling evolving stories based on event-centric temporal knowledge graph. *Journal of Systems Science and Systems Engineering*, 32:206–221.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. Peak: Pyramid evaluation via automated knowledge extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, page 2673–2680, Phoenix, Arizona. AAAI Press.

Chi yang Hsu, Yun-Wei Chu, Ting-Hao Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. In *Findings of the Association for Computational Linguistics*, page 4443–4453, Bangkok, Thailand. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Wenlin Yao and Ruihong Huang. 2018. Temporal event knowledge acquisition via identifying narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 537–547, Melbourne, Australia. Association for Computational Linguistics.

Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Garcia, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2023a. Aggregating crowdsourced and automatic judgments to scale up a corpus of anaphoric reference for fiction and Wikipedia texts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 767–781, Dubrovnik, Croatia. Association for Computational Linguistics.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023b. Personality understanding of fictional characters during book reading. In *Proceedings of 61st Annual Meeting of the Association for Computational Linguistics*, page 14784–14802, Toronto, Canada. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Fangzhou Zhai, Vera Demberg, and Alexander Koller. 2020. Story generation with rich details. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2346–2351, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fangzhou Zhai, Vera Demberg, Pavel Shkadzko, Wei Shi, and Asad Sayeed. 2019. A hybrid model for globally coherent story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 34–45, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artziv. 2019a. Bertscore: Evaluating text generation with bert. In *Proceedings of the Seventh International Conference on Learning Representations*, New Orleans, United States. ICLR Press.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019b. Generating character descriptions for automatic summarization of fiction. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA. AAAI Press.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1418–1428, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022a. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian, and Minlie Huang. 2022b. Automatic comment generation for Chinese student narrative essays. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 214–223, Abu Dhabi, UAE. Association for Computational Linguistics.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022a. NarraSum: A large-scale dataset for abstractive narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 563–578, Hong Kong, China. Association for Computational Linguistics.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022b. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5073–5085, Dublin, Ireland. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023 – 2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Xianshu Zhu and Tim Oates. 2012. Finding story chains in newswire articles. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, page 93–100, Las Vegas, Nevada, USA. IEEE Computer Society.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 19–27, Santiago, Chile. IEEE Computer Society.

Setting For Datasets Related To Narrative Understanding Tasks			
Dataset	Reference	Task	Evaluation
Toronto Book Corpus	Zhu et al. (2015)	Narrative Understanding	-
BookCorpusOpen	Bandy and Vincent (2021)	Narrative Understanding	-
ELTeC	Odebrecht et al. (2021)	Narrative Understanding	-
GLUCOSE	Mostafazadeh et al. (2020)	Reading Comprehension	BLEU scores + Human
ROCStories	Mostafazadeh et al. (2016)	Consistency Checking: Event-Centric	Story Cloze Test
ROCStories Winter 2017	Mostafazadeh et al. (2017)	Consistency Checking: Event-Centric	Story Cloze Test
Possible Stories	Ashida and Sugawara (2022)	Consistency Checking: Event-Centric	Accuracy
COPA	Roemmele et al. (2011)	Consistency Checking: Plausible Alternatives	Accuracy
TIMETRAVEL	Qin et al. (2019)	Counterfactual	Similarity Metrics + Human[1]
HellaSwag	Zellers et al. (2019)	Counterfactual	Accuracy
StoryCommonsense	Rashkin et al. (2018)	Consistency Checking: Character-Centric	F-score + Explanation score[2]
TVShowGuess	Sang et al. (2022)	Consistency Checking: Character-Centric	Accuracy
PERSONET	Yu et al. (2023b)	Consistency Checking: Character-Centric	Accuracy
LitBank	Bamman et al. (2020)	Coreference Identification	F-score
Phrase Detectives	Yu et al. (2023a)	Coreference Identification	Accuracy
-	Wanzare et al. (2019)	Consistency Checking: Other Elements	F-score
SNaC	Goyal et al. (2022)	Consistency Checking: Multiple Elements	F-score
-	Pustejovsky and Stubbs (2011)	Structural Analysis	Accuracy
InScript	Modi et al. (2016)	Structural Analysis	-
Hippocorpus	Sap et al. (2020)	Structural Analysis	Narrative Flow + Event contains
ESC v0.9	Caselli and Vossen (2021)	Structural Analysis	F-score
DesireDB	Rahimtoroghi et al. (2017)	Event-Relation Extraction	F-score
Moral Stories	Emelin et al. (2021)	Event-Relation Extraction	F-score + Similarity Metrics[3]
CSI	Frermann et al. (2018)	Corpus-Level Summarisation	F-score
Shmoop	Chaudhury et al. (2020)	Corpus-Level Summarisation	Accuracy
NovelChapter	Ladhak et al. (2020)	Corpus-Level Summarisation	Similarity Metrics
BookSum	Kryscinski et al. (2022)	Corpus-Level Summarisation	ROUGE-n, BERTScore, SummaQA
NARRASUM	Zhao et al. (2022a)	Corpus-Level Summarisation	ROUGE-n, SummaC
IDN-Sum	Revi et al. (2020)	Corpus-Level Summarisation	ROUGE-1 + F-score
ABLIT	Roemmele et al. (2023)	Corpus-Level Summarisation	ROUGE-1 + F-score[4]
CMU Movie Summary	Bamman et al. (2013)	Character-Centric Summarisation	Variation of information + Purity score
-	Zhang et al. (2019b)	Character-Centric Summarisation	Recall@K
LiSCU	Brahman et al. (2021)	Character-Centric Summarisation	Accuracy
BookTest	Bajgar et al. (2017)	Story Cloze	Accuracy
MCTest	Richardson et al. (2013)	Answer Generation	Accuracy
Children's Book Test	Hill et al. (2016)	Answer Generation	Accuracy
MovieQA	Tapaswi et al. (2016)	Answer Generation	Accuracy
WikiHow	Koupaee and Wang (2018)	Answer Generation + Summarisation	METEOR
MCScript2.0	Ostermann et al. (2019)	Answer Generation	Accuracy
Cosmos QA	Huang et al. (2019)	Answer Generation	Accuracy
NarrativeQA	Kočiský et al. (2018)	Narrative Question Answering	Similarity Metrics
TellMeWhy	Lal et al. (2021)	Narrative Question Answering	Similarity Metrics
FairytaleQA	Xu et al. (2022)	Narrative Question Answering	ROUGE-L F1 score

Table 1: Settings for datasets related to narrative understanding tasks, including references, the tasks the dataset was created for, and the evaluation methods. Below are some supplementary information that cannot fit in the table: [1] Similarity Metrics (e.g. BLUE-4, ROUGE-L, BERT, BERT-FT, Word Mover's Similarity, Sentence + Word Mover's Similarity), complemented by human evaluation using Likert scale scores; [2] F-score for category labels, Vector average and extrema score for annotation explanations; [3] Accuracy and F1 score for classification, as well as similarity metrics for generation tasks; [4] ROUGE-1 precision score between spans and F-score for sentence labels.

Domain Information On Datasets Related To Narrative Understanding Tasks				
Dataset	Domain	Dataset Size	Average Text Length	Language
Toronto Book Corpus	Romance, Historical, Adventure, etc.	11,038 books	∼6,704 sentences	English
BookCorpusOpen	Romance, Historical, Adventure, etc.	17,868 books	-	English
ELTeC	-	1,250 novels	-	8 Languages[1]
GLUCOSE	Commonsense stories (ROCStories)	4,881 stories	5 sentences	English
ROCStories	Commonsense Stories	49,255 stories	5 sentences	English
ROCStories Winter 2017	Commonsense Stories	98,159 stories	5 sentences	English
Possible Stories	Short story with multiple endings	1,313 passages	46.3 tokens	English
COPA	Choice Of Plausible Alternatives	1K questions	1 sentence	English
TIMETRAVEL	Commonsense stories (ROCStories)	29,849 story rewritings	5 sentences	English
HellaSwag	Commonsense stories (SWAG)	70K passage	1 sentence	English
StoryCommonsense	Commonsense stories (ROCStories)	15K stories	5 sentences	English
TVShowGuess	Scripts of TV series	318 characters	137,568 tokens	English
PERSONET	Novel	33 books	11,876 sentences	English, Chinese
LitBank	Fiction	100 fictions	2,105.3 tokens	English
Phrase Detectives	Fiction and Wikipedia	805 documents	1,712.4 tokens	English
Wanzare et al. (2019)	Blog (Spinn3r)	504 stories	35.74 sentences	English
SNaC	LLMs generated book/movie summaries	150 books	41 sentences	English
Pustejovsky and Stubbs (2011)	-	183 articles	-	English
InScript	Commonsense stories (given scenarios)	910 stories	12.4 sentences	English
Hippocorpus	Stories of imaged/recalled events	6,854 Stories	17.6 sentences	English
ESC v0.9	ECB+ corpus[2]	258 documents	-	English
DesireDB	Blog (Spinn3r)	3,680 instances	-	English
Moral Stories	Social Norms, Morality/Ethics	12K stories	-	English
CSI	Crime Drama	39 episodes (59 cases)	689 sentences per case	English
Shmoop	Novels, plays, short stories[3]	231 stories	112,080 tokens	English
NovelChapter	Novel	4,383 chapters	5,165 words	English
BookSum	Plays, short stories, novels (Gutenberg)	405 books	112,885.15 tokens	English
NARRASUM	Plot descriptions of Movie/TV episodes	122K narratives	786 tokens	English
IDN-Sum	Narrative game scripts	8 IDN episodes	3250 sentences	English
ABLIT	Novels (Gutenberg)	868 chapters	154.1 sentences	English
CMU Movie Summary	Movie plot summaries	42,306 movies	176 words [4]	English
Zhang et al. (2019b)	Romance, Werewolf, etc. (Wattpad)	1,036,965 stories	15,600 words	English
LiSCU	Educational stories	1,220 books	1431.2 tokens [5]	English
BookTest	Books (Gutenberg)	14,140,82 questions	522 tokens	English
MCTest	Books (Gutenberg)	500 stories + 2,000 questions	212 words [6]	English
Children's Book Test	Books (Gutenberg)	108 books + 687,343 questions	462.7 /30.7 words	English
MovieQA	Movie scripts	14,944 questions	9.3 words	English
WikiHow	HowWiki website	230,843 articles	579.8 tokens	English
MCScript2.0	Short stories around everyday scenarios	3,487 texts + 19,821 questions	164.4 / 8.2 tokens	English
Cosmos QA	Paragraph + Questions [7]	35,600 (paragraphs + questions)	69.4 /10.3 tokens	English
NarrativeQA	Books (Gutenberg), movie scripts	1572 documents + 46,765 questions	61,472 / 9.8 tokens	English
TellMeWhy	Commonsense stories (ROCStories)	9,636 stories + 30,519 questions	5 sentences	English
FairytaleQA	Classic fairytale stories	278 stories + 10,580 questions	1401.3 / 3.3 tokens	English

Table 2: Domain information on datasets related to narrative understanding tasks, including data domain, dataset size, average text length (per narrative or character), and the language used. Below are some supplementary information that cannot fit in the table: [1] 8 Europoean Language including Czech, German, English, French, Hungarian, Polish, Portuguese and Slovenian; [2] ECB+ corpus focuses on calamity events, such as shooting and accidents; [3] Short story sources include Shmoop website and Gutenberg; [4] The median length is 176 words since no average text length is provided; [5] The length pertains to book summaries; [6] Text length ranges from 150-300 words; [7] Questions relate to causes, effects, facts, and counterfactuals.

Annotation Information On Datasets Related To Narrative Understanding Tasks			
Dataset	Total Number of Annotations	Annotation Type	Annotation Procedure
Toronto Book Corpus	-	-	-
BookCorpusOpen	-	-	-
ELTeC	-	-	-
GLUCOSE	∼670K	Commonsense Causal Knowledge	Human-Crowdsourced (MTurk)
ROCStories	-	Causal + Temporal Span	Human-Crowdsourced (MTurk)
ROCStories Winter 2017	-	Causal + Temporal Span	Human-Crowdsourced (MTurk)
Possible Stories	8,885 ending + 4,533 questions	Alternative Ending + Causal Question	Human-Crowdsourced (MTurk)
COPA	2 Alternatives for each question	the more Plausible Alternatives	Human
TIMETRAVEL	81,407 counterfactual branch	Counterfactual Rewritings	Human-Crowdsourced (MTurk)
HellaSwag	70K Answers[1]	Counterfactual Reasoning	Machine-generated
StoryCommonsense	55,747 w/motiv + 104,930 w/emot	Motivations + Emotional Reactions	Human
TVShowGuess	12,413 scene	Character Facts	Human (2 experts)
PERSONET	140,268[2]	Personalities Traits	Automatic collection + Human
LitBank	29,103 mentions	Anaphoric Reference + Entity Category	Human (3 experts)
Phrase Detectives	282,558 mentions	Anaphoric Reference	Human-Crowdsourced[3]
Wanzare et al. (2019)	10,754 sentences[4]	Scenarios + Segmentation	Human(4 student assistants)
SNaC	9.6K Span	Coherence Error Span + Type	Human[5]
Pustejovsky and Stubbs (2011)	-	Temporal Span	Human (3 students)
InScript	62,062[6]	Script Structure	Human-Crowdsourced (MTurk)
Hippocorpus	-	Human Recalled Events	Human-Crowdsourced (MTurk)
ESC v0.9	9169 relations[7]	Event Relation + Temporal Span	Human (2 students)
DesireDB	3,680	Desire Expressions[8]	Human-Crowdsourced (MTurk)
Moral Stories	24K action + 48K consequence	Story Segment + Sentence Categories	Human-Crowdsourced (MTurk)
CSI	Story Segment + Sentence Categories	Factual/Structural Metadata[9]	Human (3 students)
Shmoop	7,234 summaries for 7,234 chapters	Segmentation + chapter-level summaries	Automatic collection[10]
NovelChapter	8,088 chapter/summary pairs	Chapter-level summaries	Automatic collection + Human written[11]
BookSum	405 summaries	Paragraph, chapter, book-level summaries	Automatic alignment + Human inspection
NARRASUM	122K summaries	Book-level summaries	Automatic alignment + Human inspection
IDN-Sum	10k summaries for 10k documents	Interactive narratives summaries	Automatic collection[12]
ABLIT	868	Paragraph-level abridged texts	Automatic alignment + Human written[13]
CMU Movie Summary	29,802 characters[14]	Character metadata	Automatic matching
Zhang et al. (2019b)	18,100 characters	Character Metadata + Tropes	Automatic extraction + Human[15]
LiSCU	9499[16]	Character Description	Automatic creation + Human evaluation[17]
BookTest	141,408,250 options	Cloze-form	Automatic creation
MCTest	8000[18]	Multiple-choice	Human-Crowdsourced (MTurk)
Children's Book Test	10 choices for each question	Multiple-choice	Automatic creation
MovieQA	74,720 answers	Multiple-choice	Human
WikiHow	230,843 summaries	Subtopics + Free-form Answer	Automatic collection
MCScript2.0	2 choices for each question	Answer Generation + Multiple-choice	Human-Crowdsourced (MTurk)
Cosmos QA	4 choices for each question	Multiple-choice	Human-Crowdsourced (MTurk)
NarrativeQA	46,765 answers	Free-form Answer	Human-Crowdsourced (MTurk)
TellMeWhy	3 answers for each question	Free-form Answer	Human-Crowdsourced (MTurk)
FairytaleQA	10,580	Answer + Ground-truth Question Pairs	Human-5 postgraduate students

Table 3: Annotation information on datasets related to narrative understanding tasks, including the total number of annotations, annotation type, and annotation procedure (expert vs. crowdsourced, human vs. automatic). Below are some supplementary information that cannot fit in the table: [1] Adversarial wrong answers for each passage; [2] 140,268 traits are derived from 110,114 notes, which were automatically collected from reading apps. These notes were written by the app's users; [3] The data was sourced via a game-with-a-purpose approach; [4] Annotators labelled a total of 504 documents, which comprised 10,754 sentences. A label for a scenario could be assigned from one of the 200 predefined scenarios or marked as "None" for sentences that didn't fit any scenario; [5] Both expert evaluators (3 experts) and human crowdsourcing through MTurk were used for annotation; [6] Stories were annotated across 10 distinct scenarios. Verbs and noun phrases were labelled with event and participant types, respectively. The text also includes coreference annotations. [7] The dataset includes 6,904 temporal relations and 2,265 explanatory relations.; [8] It has gold standard labels for identifying statements of desire, spans of evidence supporting the fulfillment of the desire, and annotations indicating whether the stated desire is fulfilled based on the narrative context; [9] References to the mentioned perpetrator and relation to previous cases; [10] Paired summaries and narrative texts sourced from websites; [11] Summaries and chapter pairs were automatically collected from online study guides, written by experts; [12] Paired summaries and narrative texts sourced from websites; [13] The dataset has an automatically aligned abridged version, which is written by a single human author.; [14] Characters are matched to actors with a public date of birth; [15] Annotations are collected through questionnaires to 100 authors; [16] 1708 literature summaries and 9499 character descriptions; [17] 3 judges were asked to evaluate the quality of the description, focusing on fact coverage and task difficulty; [18] There are 4 questions associated with each story, and each question offers 4 answer choices.

Task	Representative Dataset	Best Model	Result
StoryCloze	StoryCloze (Mostafazadeh et al., 2016)	FLAN 137B zero-shot (Wei et al., 2022)	93.4 Accuracy
CounterfactualReasoning	Hellaswag (Zellers et al., 2019)	GPT-4 (OpenAI, 2023)	95.3 Accuracy
NarrativeSummarization	BookSum (Kryscinski et al., 2022)	BART-LS (Xiong et al., 2022)	38.5 Rouge-1
NarrativeQA	NarrativeQA (Kočiský et al., 2018)	Masque (Nishida et al., 2019)	59.87 Rouge-L
NarrativeQA	Children's Book Test (Hill et al., 2016)	NSE (Dhingra et al., 2017)	71.9 Accuracy

Table 4: Performance of the most recent models on representative datasets. The results are extracted from their respective papers.