NarGINA: Towards Accurate and Interpretable Children's Narrative Ability Assessment via Narrative Graphs

Anonymous ACL submission

Abstract

The assessment of children's narrative ability is crucial for diagnosing language disorders and planning interventions. Distinct from the typical automated essay scoring, this task focuses primarily on evaluating the completeness of narrative content and the coherence of expression, as well as the interpretability of assessment results. To address these issues, we propose a novel computational assessing framework NarGINA, under which the narrative graph is introduced to provide a concise and structured summary representation of narrative text, allowing for explicit narrative measurement. To this end, we construct the first Chinese children's narrative assessment corpus based on real children's narrative samples, and we then design a narrative graph construction model and a narrative graph-assisted scoring model to yield accurate narrative ability assessment. Particularly, to enable the scoring model to understand narrative graphs, we propose a multi-view graph contrastive learning strategy to pre-train the graph encoder and apply instruction-tuned large language models to generate scores. The extensive experimental results show that NarGINA can achieve significant performance improvement over the baselines, simultaneously possessing good interpretability. Our findings reveal that the utilization of structured narrative graphs beyond flat text is well suited for narrative ability assessment.

1 Introduction

A narrative can take several forms: recounting past experiences, retelling a previously heard or read story, or creating a composition (McCabe et al., 2008). Assessing narrative ability not only provides an objective measure of children's language development, but also plays a crucial role in the early diagnosis and intervention of language disorders (Pesco and Bird, 2016; Pico et al., 2021).

In the field of clinical linguistics, assessing narrative ability has been a major focus of research.

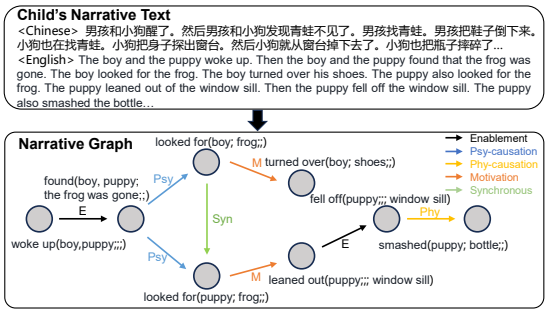

Figure 1: An example of narrative graph.

Studies typically analyze narratives from two perspectives: macrostructure (Blom and Boerma, 2016) and microstructure (Justice et al., 2006). As microstructural features are relatively easy to quantify, research has increasingly emphasized macrostructural coherence (Reese et al., 2011) and completeness (Kellas and Manusov, 2003). Causal networks (Trabasso and Sperry, 1985) are an important tool for assessing these aspects (Diehl et al., 2006; Torng and Sah, 2020), providing an intuitive representation of narrative macrostructure. However, researchers in this field generally rely purely on manual analyses of children's narrative samples, which poses a practical dilemma of being time-consuming and laborious; therefore, it is difficult to promote and apply in broader practices.

This paper focuses on automated assessment of children's narrative ability by exploring the forefront natural language processing (NLP) techniques. Outwardly, this task shares similarities with the multi-trait automated essay scoring (AES), which evaluates various essay genres across traits like content and language use. Some recent studies have applied autoregressive multi-trait score generation framework to leverage token generation probabilities (Do et al., 2024a,b). Nevertheless, compared to multi-trait AES, the automated assessment of children's narrative ability presents unique challenges in the following aspects: (1) the narra-

tive assessment task focuses primarily on evaluating the completeness of narrative content and the coherence of expression; (2) the assessment result of this task requires not only high accuracy, as well as the intuitiveness and interpretability, which are essential to provide actionable feedback for subsequent interventions. There has also been some sporadic research on this task. Hassanali et al. (2013) employed topic modeling to predict language disorders and coherence. Jones et al. (2019) simply used machine learning methods to score macrostructure. Obviously, these works have not presented effective solutions to the aforementioned challenges.

To address these issues, we propose the **Nar**rative **G**raph-based **I**nterpretable Children's **N**arrative Ability **A**ssessment (**NarGINA**) framework. To this end, we first introduce a **narrative graph** as a structured representation of narrative text, inspired by the causal networks in clinical linguistics (Torng and Sah, 2020). Though the causal network gives an intuitive representation of the input text, that structure simply considers clauses as nodes, which makes it difficult to clearly express the complex narrative content. Contrastively, in our narrative graph, nodes represent specific events, and edges capture event relations, such as various causal and synchronous connections (see an example in Figure 1). Compared to flat and unstructured narrative text, the narrative graph provides a concise summary representation, thus helping to explicitly measure and calculate the key narrative indicators such as completeness and coherence; meanwhile, the interpretability can also be naturally facilitated through the comparative analysis between the evaluation results and the gold-standard narrative graph [1].

Further, we design the computational framework NarGINA, based on the narrative graph, for assessing narrative ability. Unlike most existing AES systems that rely solely on feature learning from raw text, NarGINA evaluates the narrative quality mainly by examining the narrative graphs constructed from the input text, while also considering the raw text, as illustrated in Figure 2. To achieve this, we first establish a narrative graph annotation specification and then construct a Chinese narrative ability assessment corpus, incorporating macrostructure, microstructure, and psychological states. Next, we propose an automated narrative graph construction model and a narrative graph-assisted scoring model to yield accurate and interpretable narrative ability assessment. Particularly, to enable the scoring model to understand narrative graphs, we introduce a multi-view contrastive learning strategy to pre-train a graph encoder and apply instruction-tuned large language models (LLMs) to generate scores. Experimental results show that our approach significantly outperforms baselines in both performance and interpretability. The source code and corpus will be publicly available.

[1] See the definition in Section 3.3.

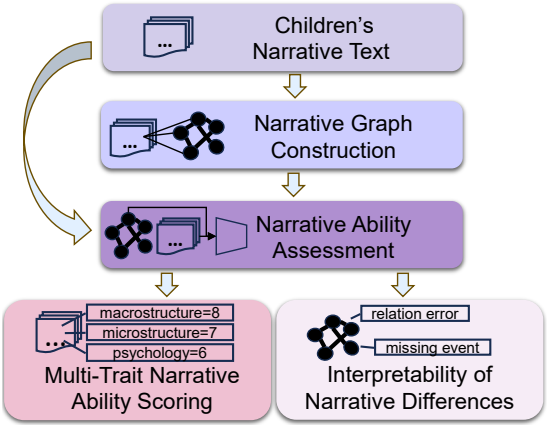

Figure 2: Overview of NarGINA framework.

In a nutshell, our contributions are as follows:

- We propose a novel method for automated children's narrative ability assessment, under which the narrative graph is innovatively introduced to explicitly measure the narrative quality, and then the narrative graph construction and scoring models are well designed.
- We introduce the first Chinese children's narrative corpus, by establishing a narrative graph specification, collecting real children's narrative samples and constructing a high-quality annotated dataset.
- Experimental results show impressive performance improvements along with interpretable scoring results.

2 Related Work

Narrative Ability Assessment Frameworks The story grammar model (Stein, 1979) and high-point analysis (Labov and Waletzky, 1967) provide the theoretical foundation for assessing macrostructure. The causal network (Trabasso and Sperry, 1985) has been used to assess narrative coherence by statistical features (Sah and Torng, 2015; Sah, 2013) and has also been applied in interventions

for reading difficulties (McMaster et al., 2014). MAIN (Gagarina et al., 2019) analyzed the portrayal of children's psychological states by internal state terms. Research on automated assessment remains relatively underexplored, with most methods focusing on detecting language disorders (Gabani et al., 2011) or classifying specific narrative traits, such as coherence (Hassanali et al., 2013). Recently, some studies have attempted to apply NLP techniques within manual assessment frameworks. Baumann et al. (2024) achieved the automated annotation of the story grammar structures in MAIN. However, these earlier studies have not provided a fully automated approach for comprehensively assessing narrative completeness and coherence, nor have they offered quantitative and interpretable results needed to inform subsequent interventions.

Graph-based Approaches for Text Assessment Graph-based methods have been applied to various tasks such as modeling mental states (Lee et al., 2021), event evolution (Yan and Tang, 2023), explainable causal reasoning (Du et al., 2021), and AES. In particular, Somasundaran et al. (2016) showed that graph properties (e.g., PageRank) derived from content words in essays can effectively model essay scores related to the quality of development. Another line of work constructed sentence-prompt graphs, where semantic similarity served as edge weights, to evaluate how well each sentence addresses the prompt (Bhatt et al., 2020). Yet, the graph structures in these prior studies were not designed for children's narrative assessment and therefore struggle to model the completeness and coherence of narratives.

3 Corpus Construction

The Chinese children's narrative assessment corpus comprises 543 annotated narrative texts, each paired with a narrative graph and scores for overall ability and three key traits: macrostructure, microstructure, and psychological states.

3.1 Data Collection

Instead of using the typical story-retelling task, we adopted a more challenging narrative generation task (Pearce et al., 2010) under the guidance of clinical linguistics experts. To collect narrative data, we used the book *Frog, Where Are You?* (Mercer, 1969), a wordless picture book widely used for assessing children's narrative ability (Reilly et al., 2004; Torng and Sah, 2020). Participants, aged 3 to 13, were independently asked to read the book and verbally narrate the story's events without any scripted guidance, ensuring that the narratives were based on their own interpretations and recollections of the visual cues. To establish a gold-standard narrative graph, we also collected 40 narrative samples from adults of normal intelligence, bringing the total corpus size to 543. All oral narratives were manually transcribed following CHILDES (MacWhinney, 2000) data procedures, and formatted in accordance with the CHAT (MacWhinney, 2017) guidelines (Appendix A.1).

3.2 Annotation Specification Design

Each transcribed narrative text is annotated with a narrative graph and scores for narrative ability.

Narrative Graph Annotation Figure 1 shows that a narrative graph consists of event nodes and event relation edges.

- **Event**: Unlike the predefined event types in corpora such as ACE 2005 (Walker et al., 2006), children's narrative expressions exhibit significant variability and diversity. Thus, we do not impose rigid event type constraints. An event is defined as a narrative element describing the story background, actions, or activities involving characters. We refer to the guidelines (LDC, 2005) and design a structured event representation for narrative text in the format: *Trigger (Subject; Object; Adverbial of Time; Adverbial of Place)*. If multiple arguments exist in the same slot, they should be separated by commas (,). Examples are provided in Appendix A.2.

- **Event Relation**: We adopt the causal relation classifications (*motivation*, *psychological causation*, *physical causation*, *enablement*) proposed in causal networks (Trabasso and Sperry, 1985) and further refine the relation definitions. Since the book *Frog, Where Are You?* contains several synchronous events, we incorporate *synchronous* relations, as defined in the Penn Discourse Treebank 3.0 (Webber et al., 2019). The Appendix A.3 provides detailed definitions and examples.

Narrative Ability Scoring Rubric We primarily focus on assessing the completeness and coherence of the narrative's macrostructure. A complete narrative should be clearly segmented in chronological or episodic order, demonstrate causal relations,

Dataset	#Documents	#Events	#Event Relations
ACE 2005	599	4090	-
Causal-TB	183	6811	5436
Event StoryLine	258	4732	12695
MAVEN-ERE	4480	103193	1290050
our corpus	546	20244	16390

Table 1: Comparison between our corpus and relevant corpora that contain events and event relations.

develop characters with emotional depth, express emotions and derive meaning, ensure coherence, and attribute responsibility to the characters in the story (Kellas and Manusov, 2003). Coherence is defined as the temporal and causal structure of a story (Karmiloff-Smith, 1985). For a more comprehensive assessment, the microstructure and psychological states are included in the scoring rubric. Each expert assigns scores ranging from 0 to 10 to each trait and the overall narrative ability.

3.3 Annotation Process

The annotation process consists of two stages: (1) in the narrative graph annotation stage, 14 trained annotators, divided into 7 pairs, independently annotated identical transcribed texts. The annotations were then compared and refined through consistency checks. (2) the narrative ability scoring stage requires expertise in children's language development and narrative ability. Hence, two experts with clinical or educational experience independently scored each sample. This dual annotation process helped to reduce subjective bias and improve reliability.

A **gold-standard narrative graph** was established through discussions among linguistic experts, based on adult narrative samples, and served as the benchmark for assessing children's narrative abilities. To improve efficiency, an annotation tool was developed (Appendix A.6).

3.4 Statistical Analysis

As shown in Table 1, we compare our corpus with existing relevant corpora, including ACE 2005, Causal-TB (Mirza et al., 2014), Event StoryLine (Caselli and Vossen, 2017), and MAVEN-ERE (Wang et al., 2022). In contrast, our corpus provides more comprehensive annotations, covering event triggers, arguments, and relations. Additionally, Table 2 presents our narrative graph statistics. The variation in edge counts across different types is due to the limited occurrences of physical causality and synchronous relations in *Frog, Where Are You?*. Differences in graph size reflect age-related differences in narrative completeness or potential language disorders. The score distribution is presented in Appendix A.7.

Statistics		
event node		17815
event relation edge	Synchronous	653
	Motivation	3356
	Psychological causation	1213
	Physical causation	384
	Enablement	10518
maximum graph	node	138
	edge	164
minimum graph	node	3
	edge	0

Table 2: Statistics of narrative graph features.

4 NarGINA

4.1 Overview

In the domain of automated narrative ability assessment, one of the main challenges is capturing both the structure and semantics of narrative texts, while also providing interpretability of the assessment results. For this reason, we introduce NarGINA. As illustrated in Figure 3, it consists of two stages: narrative graph construction and narrative ability scoring. First, NarGINA transforms narrative text into a structured graph, offering a concise summary representation that tackles key challenges in narrative modeling. Next, NarGINA integrates the narrative graph and the original text into the LLM, enabling scoring across multiple traits and providing interpretability analysis.

4.2 Narrative Graph Construction Model

In this section, we present the details of narrative graph construction, which consists of node and edge construction. The task mainly faces the following challenges: (1) children's oral narratives exhibit irregularities (e.g., missing grammatical components, repeated sentences, and scrambled word order) during node construction, requiring richer information supplementation; (2) data sparsity, due to the structural characteristics of narrative graphs, and missing information (e.g., missing nodes, missing triggers or arguments) caused by the irregularities, require fine-grained data and data augmentation during edge construction.

Narrative Nodes Construction To address these irregularities, we use the Universal Information

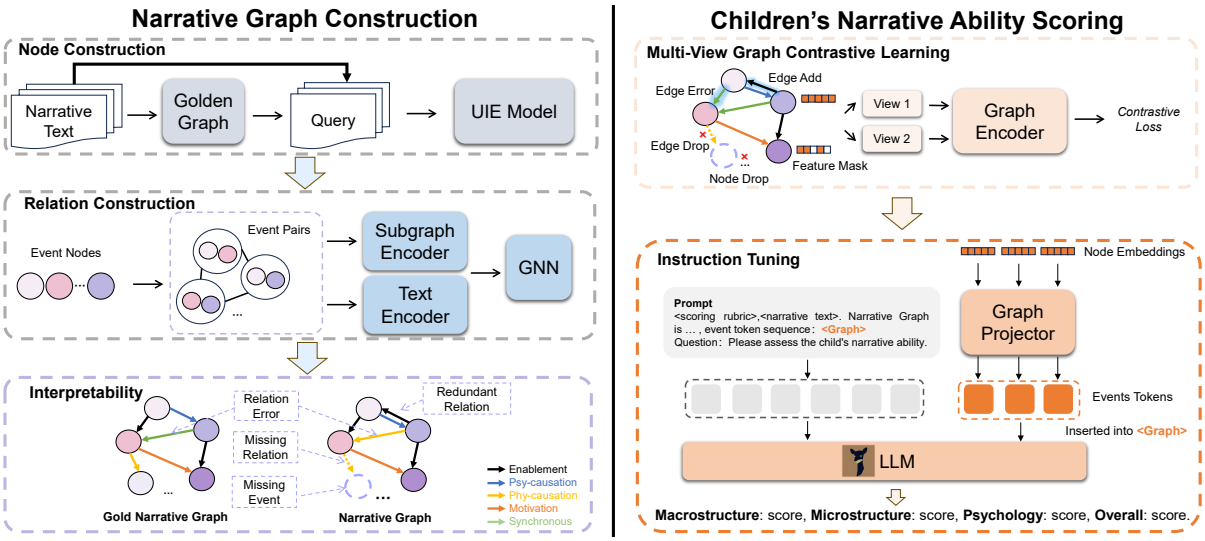

Figure 3: Illustration of the entire process for the proposed framework NarGINA.

Extraction (UIE) model to construct nodes and exploit its generalization ability (Lu et al., 2022) to extract richer information. Furthermore, we propose a retrieval-augmented strategy based on the gold-standard narrative graph G_{gold} and apply it to the our model to augment the data, illustrated as:

$$R = f(text, G_{gold}) \tag{1}$$

where R is the retrieved information, $text$ denotes the text of an event and $f(\cdot)$ denotes the text similarity matching function, which retrieves the text for the most relevant event in G_{gold} using a threshold of 0.9. Specifically, based on rexUIE (Liu et al., 2023a), we concatenate R with the input of rexUIE Q to generate an augmented input. Next, the augmented input is encoded by DeBERTa (He et al., 2021), generating the augmented embedding h_a:

$$h_a = DeBERTa(Q; R) \tag{2}$$

The final set of narrative nodes is denoted as $V_{narrative}$=$\{event_1, \ldots, event_s\}$, where $event$ is a structured representation consisting of a trigger and its arguments, s denotes the node count.

Narrative Edges Construction To address the issues of missing information and data sparsity, we encode narrative nodes using LLMs as GNN enhancers, extend the internal knowledge of LLMs to events, and construct superior graphs to enhance the input data. Besides, we construct subgraphs to supplement fine-grained dependency information. Specifically, based on OFA (Liu et al., 2024), we use the Llama2_13b (Touvron et al., 2023), without fine-tuning, to encode the superior graph, denoted as $G_{sup} = (V_{sup}, E_{sup})$, where the node set is denoted as $V_{sup} \subseteq \{(event_i; event_j) | i, j \in \mathbb{Z}^+, i \neq j\}$. Similarly, the edge set is denoted as $E_{sup} \subseteq \{(v_i, v_j) | v_i, v_j \in V_{sup}, \exists event_k \in (v_i \cap v_j), i, j, k \in \mathbb{Z}^+\}$. Due to the complexity of G_{sup}, we retain only the 10 nearest neighbor edges for each node. Furthermore, we use the multimodal model G2P2 (Wen and Fang, 2024) to encode the subgraph, denoted as G_{sub}= (V_{sub}, E_{sub}), where the node set V_{sub} consists of all event triggers and arguments, and the edge set E_{sub} consists of edges between the trigger and arguments in each event. The embeddings from the superior graph and the subgraph are then concatenated and ultimately fed into R-GCN (Schlichtkrull et al., 2018) for narrative edge classification. The set of narrative edges is denoted as $E_{narrative} \subseteq \{(event_i, re, event_j) | event_i, event_j \in V_{narrative}, re \in Re, i, j \in \mathbb{Z}^+\}$, where Re includes all edge types (Section 3.2).

Eventually, we get the narrative graph $G_{narrative} = (V_{narrative}, E_{narrative}, Re)$.

4.3 Narrative Graph-Assisted Scoring Model

There exist two main challenges: (1) narrative ability encompasses multiple traits, requiring the model to possess strong reasoning capabilities to capture cross-event logic; and (2) due to modality gaps, narrative graphs cannot be directly utilized in existing language model-based scoring methods.

Piper and Bagga (2024) demonstrated that fine-tuning LLMs can match the performance of GPT-4 on narrative understanding tasks, motivating us to integrate LLMs for narrative ability assessment and enhance their macrostructural modeling ability using narrative graphs. Although researchers have ex-

plored translating graph structures into natural language (Fatemi et al., 2024), such inputs tend to be verbose, potentially reducing LLMs' performance on downstream tasks (Chen et al., 2023). Graph-level tokenization (Chai et al., 2023) and node-level tokenization (Chen et al., 2024) address this issue but struggle to capture the complex logic of children's narratives and remain incompatible with the heterogeneity of narrative graphs. Hence, we integrate narrative graphs into LLMs using GNN, a graph projector, and instruction-tuning.

Multi-View Graph Contrastive Learning Unlike knowledge graphs, which represent entity relations, narrative graphs capture key storylines, causal dependencies, and shifts in psychological states. Moreover, the limited availability of labeled data hampers the generalization of supervised methods. Thus, we propose a multi-view graph contrastive learning strategy to learn unsupervised node representations. As shown in Figure 3, we generate multi-view graphs using strategies such as Node Drop and Edge Add, simulating issues like missing events and redundant causal relations. For graph encoding, we use the Graph Attention Network (GAT) (Veličković et al., 2017). By applying contrastive learning across these views, GAT enhances robustness against incompleteness, incoherence, and noise in graphs, while also improving its ability to capture event causality. To derive textual embeddings from the event and relation text, we apply Sentence Transformers (Reimers, 2019).

Given a narrative graph $G_{narrative}$, transformed from original text t, we apply random augmentation strategies to generate two augmented graphs G_1 and G_2, which are then encoded to generate node features h_v, $h_v^{(1)}$ and $h_v^{(2)}$. By optimizing the InfoNCE loss (Oord et al., 2018), we ensure that features of the same node in $h_v^{(1)}$ and $h_v^{(2)}$ are similar, while those of different nodes are distinct. After training, the final node features are represented as:

$$h_v = GAT(G_{narrative}) \tag{3}$$

Graph-Text Alignment To align data from text and graphs, we use MLP as the graph projector that maps node features h_v to the LLM's input dimensions, generating event tokens $e_v = MLP(h_v)$. Similar alignment methods are widely used in multimodal models (Liu et al., 2023b; Chen et al., 2024). The event tokens e_v are reordered based on the sequence of event occurrences.

Instruction Turning We fine-tune LLMs with specific instructions to effectively integrate narrative graph features for multi-trait scoring. Autoregressive score generation has been successfully applied to T5 (Raffel et al., 2020) for efficient multi-trait AES (Do et al., 2024a,b). Nevertheless, T5 adopts short prefix-tuning, which may pose challenges for directly integrating narrative graphs into the input. In contrast, LLMs support longer input sequences. Therefore, we define the scoring task as a question-answering (QA) task (Figure 3). Details of the QA instructions can be found in Appendix B.1. During preprocessing, the *<Graph>* tag in the prompt is replaced with e_v as input. The model then generates the scores as:

$$scores = LLM(prompt(e_v, t)) \tag{4}$$

For training, we fine-tune Vicuna_v1.5_7B (Chiang et al., 2023), keeping the graph projector trainable while freezing the graph encoder, which ensures robust graph features.

5 Experiments

5.1 Experimental Settings

We base our experiments on the Chinese children's narrative assessment corpus. The dataset is stratified across different total scores and divided into training (70%), validation (10%), and test (20%) sets. The detailed dataset split and key statistics are presented in Appendix A.8. For the narrative ability scoring task, we adopt Quadratic Weighted Kappa (QWK) (Cohen, 1968) to measure agreement between human annotations and model predictions. We train on four NVIDIA A40 GPUs. The narrative graph construction model employs full-parameter fine-tuning, while the scoring model uses LoRA (Hu et al., 2021) for parameter-efficient fine-tuning of Vicuna_v1.5_7B. These models are trained separately in a pipeline approach. Further implementation details are provided in Appendix B.2. All results are reported as averages.

5.2 Baselines

In the domain of automated assessment of children's narrative ability, to the best of our knowledge, there are almost no graph-based methods available for direct comparison. Therefore, we evaluate the following baseline models:

BERT Jones et al. (2019) applied BERT (Devlin, 2018) to score narrative macrostructure, focusing on story grammar components.

Model	Overall	Macro	Micro	Psych	Avg
Content words-based graph	0.537	0.494	0.605	0.439	0.519
Sentence similarity-based graph	0.651	0.600	0.670	0.522	0.611
BERT	0.680	0.635	0.664	0.539	0.629
ArTS-Vicuna_7B	0.745	0.734	0.707	0.550	0.684
NarGINA	**0.787**	**0.767**	**0.717**	**0.636**	**0.727**
NarGINA -w/o FT	0.688	0.685	0.673	0.488	0.634

Table 3: The QWK evaluation scores on our corpus. Macro: *Macrostructure*, Micro: *Microstructure*, Psych: *Psychological States*, Avg: *Average*, FT: *Fine-turning*.

Content words-based graph Somasundaran et al. (2016) constructed graphs where content words serve as nodes and sentence adjacency forms the edges, then extracted features to evaluate essays across multiple traits.

Sentence similarity-based graph Bhatt et al. (2020) constructed sentence-prompt graphs with semantic similarity as edge weights to derive features for overall essay scoring. We train separate models for each trait and discard features that were not applicable to Chinese.

ArTS-Vicuna_7B We extend the autoregressive score generation model ArTS (Do et al., 2024a) to Vicuna_v1.5_7B. We show the effectiveness of narrative graphs through comparative analysis.

5.3 Overall Performance

Table 3 reports the average QWK scores for NarGINA and the baseline approaches. We observe that our method outperforms the strongest baseline, ArTS-Vicuna_7B, by 4.3% in average QWK. It also exceeds all the other baselines on every trait, demonstrating the superiority of the proposed framework. Focusing on *macrostructure*, NarGINA achieves a 3.3% gain over ArTS-Vicuna_7B, suggesting that explicitly modeling key events and their relations via a narrative graph offers a richer representation of story structure and logic. For *psychological states*, the margin increases to 8.6%, showing the model's ability to capture more nuanced character portrayals. The improvement in *overall* further demonstrates NarGINA's ability to weigh all traits, providing a holistic assessment.

Notably, even without fine-tuning the LLM, narrative graph features generated by the pre-trained graph encoder and the lightweight graph projector can still effectively enhance narrative ability assessment, allowing the framework to perform well even in resource-constrained environments.

Model	ETE			EAE		
	P	R	F1	P	R	F1
Instruct-UIE	50.0	31.6	38.7	50.0	28.1	35.7
T5-UIE	61.9	68.0	64.8	62.7	69.0	65.7
rexUIE-G_{gold}	**72.4**	**73.3**	**72.8**	**76.5**	**76.4**	**76.4**

Table 4: Performances of narrative node construction. ETE denotes the event trigger extraction, EAE denotes the event argument extraction, and rexUIE-G_{gold} is our method to construct the narrative nodes.

Model	P	R	F1
RoBERTa-large	24.1	**80.3**	36.2
Vicuna_7B-FT	28.9	69.2	32.7
OFA-G_{sub}-G_{sup}	**73.3**	79.2	**75.3**

Table 5: Performances of narrative edge construction. OFA-G_{sub}-G_{sup} is our method for edge construction.

5.4 Narrative Graph Construction Analysis

In this section, we assess the quality of the narrative graphs generated by our framework, using Precision (P), Recall (R), and F1-score (F1) as evaluation metrics. We use T5-UIE (Lu et al., 2022) and Instruct-UIE (Wang et al., 2023) as baselines for narrative node construction, while using RoBERTa (Liu et al., 2019) and Vicuna_7B as baselines for narrative edge construction. Table 4 shows that our model rexUIE-G_{gold} outperforms the baselines across all metrics. Table 5 further demonstrates that our model OFA-G_{sub}-G_{sup} yields notable gains in precision and F1, while maintaining a recall comparable to the best result. Thus, the narrative graphs constructed by our framework capture events and their relations more accurately.

5.5 Ablation Study

Effect of LLM To evaluate NarGINA's effectiveness on different LLMs , we use Llama2_7b as the foundation model. As shown in Table 6, NarGINA-Llama2_7B, augmented by the narrative graphs, outperforms ArTS-Llama2_7B by 1.5% in average

Model	Overall	Macro	Micro	Psych	Avg
ArTS-Llama2_7B	0.736	0.725	0.708	0.527	0.674
NarGINA-Llama2_7B	0.750	0.759	0.690	0.555	0.689
NarGINA	**0.787**	**0.767**	**0.717**	**0.636**	**0.727**
-w/o graph encoder	0.709	0.724	0.689	0.556	0.669
-NG_TV	0.738	0.734	0.700	0.550	0.681

Table 6: Ablation study on key components for QWK performance.

QWK. It remains below the Vicuna-based model, presumably because Vicuna benefits from additional fine-tuning on Llama2, leading to stronger language modeling capacity. Overall, our framework improves performance across different LLMs and holds the potential for even greater improvements in larger models with more parameters.

Effect of Narrative Graph Construction Model To investigate our narrative graph construction model's contribution to scoring, we replace it with T5-UIE for node construction and Vicuna_7B-FT for edge construction, then feed the resulting graph into the scoring model (referred to as NG_TV). Table 6 shows that this approach results in a 4.6% decrease in average QWK scores, indicating that our approach can capture narrative events and relations, thus enhancing scoring performance more effectively.

Effect of Graph Encoder To verify the effectiveness of the graph encoder trained with multi-view graph contrastive learning, we adopt Sentence Transformers to derive node features directly from raw text, bypassing the structural modeling. As shown in Table 6, removal of the graph encoder leads to an average decrease of 5.8% in QWK scores across all traits. This is because the graph encoder effectively captures semantic and structural information in narrative graphs, thereby generating higher-quality node features.

5.6 Interpretability Analysis

Figure 4 illustrates a case study about an interpretable result, which helps intuitively identify deficiencies in the test sample's macrostructural completeness and coherence.

Missing Key Event The absence of events like *"woke up (boy,puppy;;;)"* and *"leaned out (puppy;;; window sill)"* makes the narrative less complete and also weakens the logical setup for subsequent events, reducing overall coherence.

Relation Errors and Redundancies Misrepresenting *looked for (boy; frog;;)* → *turned over (boy; shoes;;)* as psychological causation rather than motivation reveals the child's difficulties in establishing accurate causal relations during oral narrative. Missing events may lead to redundant causal relations, such as incorrectly associating *looked for (puppy; frog;;)* with *fell off (puppy;;; window sill)*. These issues suggest that the child, when narrating complex events, struggles to effectively structure causal relations, further impairing coherence.

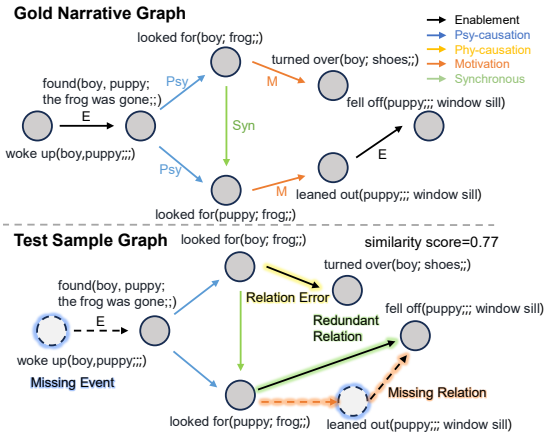

Figure 4: Example of Interpretability. We select a key segment due to the narrative graph's large scale.

Furthermore, the cosine similarity between the test sample and the gold-standard narrative graph, computed via global average pooling, serves as a quantitative indicator of the child's narrative ability.

6 Conclusion

In this paper, we focus on the automated assessment of children's narrative ability. We propose a novel computational assessing framework NarGINA that introduces the narrative graph to explicitly measure and calculate the key narrative indicators such as completeness and coherence. We construct the first Chinese children's narrative assessment corpus, and then propose the narrative graph construction model and a narrative graph-assisted scoring model. Experimental results demonstrate that NarGINA substantially outperforms the baselines, along with good interpretability. In particular, our findings reveal that the utilization of structured narrative graphs beyond flat text is well suited for narrative ability assessment. In future work, we will explore more effective narrative graph construction and scoring models to achieve better performance.

7 Limitations

Limited Materials and Forms Our data collection relies solely on the wordless picture book *Frog, Where Are You?*. While this material has been widely used in children's narrative research, the generalizability of our study to other forms (e.g., written stories, audiovisual content) remains unexamined. Future studies will expand to multiple materials and diverse genres to enhance the model's applicability across different narrative contexts.

Applicability in Resource-Constrained Environments Experimental results show that NarGINA performs well even without fine-tuning LLMs. Nevertheless, for clinicians and educators who lack stable access or sufficient computational resources, deploying and maintaining the framework may still pose significant challenges. Future research could explore models with fewer parameters or adaptive frameworks to reduce reliance on LLMs.

Need for Broader Real-World Validation Although we have conducted quantitative analyses and provided interpretable assessment results, there is still a lack of broader empirical research—such as large-scale user testing in real-world teaching or clinical settings.

8 Ethics Statement

Our work strictly follows the the ACL Code of Ethics.

For data collection (Section 3.1), we sampled data from children aged 3 to 13 and some undergraduate students. All child participants obtained parental consent, and all adult participants provided their own consent. Our corpus does not contain any protected information, and any potentially identifiable personal information has been anonymized. The anonymization method involves replacing personal names with identifiers in the format "Narrative-{index}".

For human annotation (Section 3.3), we recruited our annotators from the linguistics and computer science departments of our university to annotate graphs and invited two front-line teachers to annotate scores. Annotators were also paid above the minimum wage. The annotation does not involve any personally sensitive information. Additionally, we include comprehensive details about human annotation in Section 3.3. We present the instructions and screenshots of the interface for the human annotation in Appendix A.6. We inform the human annotators what the task is about and tell them that their responses will be used to assess the narrative ability using AI models.

We use the models and datasets when following their intended usage. We try our best to follow the ethical guidelines of ACL.

References

Timo Baumann, Korbinian Eller, and Natalia Gagarina. 2024. BERT-based annotation of oral texts elicited via multilingual assessment instrument for narratives. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 99–104, Miami, Florida, USA. Association for Computational Linguistics.

Reecha Bhatt, Malvik Patel, Gautam Srivastava, and Vijay Mago. 2020. A graph based approach to automate essay evaluation. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4379–4385. IEEE.

Elma Blom and Tessel Boerma. 2016. Why do children with language impairment have difficulties with narrative macrostructure? *Research in Developmental Disabilities*, 55:301–311.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. 2023. Graphllm: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*.

Runjin Chen, Tong Zhao, Ajay Kumar Jaiswal, Neil Shah, and Zhangyang Wang. 2024. LLaGA: Large Language and Graph Assistant. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2023. Exploring the Potential of Large Language Models (LLMs)in Learning on Graphs. *SIGKDD Explor.*, 25(2):42–61.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joshua J Diehl, Loisa Bennetto, and Edna Carter Young. 2006. Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of abnormal child psychology*, 34:83–98.

Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. ExCAR: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363, Online. Association for Computational Linguistics.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a Graph: Encoding Graphs for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Keyur Gabani, Thamar Solorio, Yang Liu, Khairun-nisa Hassanali, and Christine A. Dollaghan. 2011. Exploring a corpus-based approach for detecting language impairment in monolingual english-speaking children. *Artif. Intell. Medicine*, 53(3):161–170.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. Main: Multilingual assessment instrument for narratives–revised. *ZAS Papers in Linguistics*, 63:20–20.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2013. Using Latent Dirichlet Allocation for child narrative analysis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 111–115, Sofia, Bulgaria. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sharad Jones, Carly Fox, Sandra Gillam, and Ronald B Gillam. 2019. An exploration of automated narrative analysis via machine learning. *Plos one*, 14(10):e0224634.

Laura M. Justice, Ryan P. Bowles, Joan N. Kaderavek, Teresa A. Ukrainetz, Sarita L. Eisenberg, and Ronald B. Gillam. 2006. The Index of Narrative Microstructure: A Clinical Tool for Analyzing School-Age Children's Narrative Performances. *American Journal of Speech-Language Pathology*, 15(2):177–191.

Annette Karmiloff-Smith. 1985. Language and cognitive processes from a developmental perspective. *Language and cognitive processes*, 1(1):61–85.

Jody Koenig Kellas and Valerie Manusov. 2003. What's in a story? the relationship between narrative completeness and adjustment to relationship dissolution. *Journal of Social and Personal Relationships*, 20(3):285–307.

William Labov and Joshua Waletzky. 1967. Narrative analysis: Oral versions of personal experience. *Essays on the Verbal and Visual Arts*, pages 12–44.

LDC. 2005. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events.

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.

Chengyuan Liu, Fubang Zhao, Yangyang Kang, Jingyuan Zhang, Xiang Zhou, Changlong Sun, Kun Kuang, and Fei Wu. 2023a. RexUIE: A recursive method with explicit schema instructor for universal information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15342–15359, Singapore. Association for Computational Linguistics.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. 2024. One For All: Towards Training One Graph Model For All Classification Tasks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Brian MacWhinney. 2017. Tools for analyzing talk part 1: The chat transcription format. *Carnegie.[Google Scholar]*, 16.

Allyssa McCabe, Lynn Bliss, Gabriela Barra, and Mari-Beth Bennett. 2008. Comparison of Personal Versus Fictional Narratives of Children With Language Impairment. *American Journal of Speech-Language Pathology*, 17(2):194–206.

Kristen L McMaster, Christine A Espin, and Paul Van Den Broek. 2014. Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice*, 29(1):17–24.

Mayer Mercer. 1969. Frog, where are you.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Wendy M Pearce, Deborah GH James, and Paul F McCormack. 2010. A comparison of oral narratives in children with specific language and non-specific language impairment. *Clinical Linguistics & Phonetics*, 24(8):622–645.

Diane Pesco and Elizabeth Kay-Raining Bird. 2016. Perspectives on bilingual children's narratives elicited with the Multilingual Assessment Instrument for Narratives. *Applied Psycholinguistics*, 37(1):1–9.

Danielle L Pico, Alison Hessling Prahl, Christa Haring Biel, Amy K Peterson, Eric J Biel, Christine Woods, and Valentina A Contesse. 2021. Interventions designed to improve narrative language in school-age children: A systematic review with meta-analyses. *Language, Speech, and Hearing Services in Schools*, 52(4):1109–1126.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

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

Elaine Reese, Catherine A Haden, Lynne Baker-Ward, Patricia Bauer, Robyn Fivush, and Peter A Ornstein. 2011. Coherence of personal narratives across the lifespan: A multidimensional model and coding method. *Journal of cognition and development*, 12(4):424–462.

Judy Reilly, Molly Losh, Ursula Bellugi, and Beverly Wulfeck. 2004. "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome". *Brain and language*, 88(2):229–247.

N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Wen-Hui Sah. 2013. The Development of Coherence in Narratives: Causal Relations. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation, PACLIC 27, Taipei, Taiwan, November 21-24, 2013*. National Chengchi University, Taiwan.

Wen-hui Sah and Pao-chuan Torng. 2015. Narrative coherence of Mandarin-speaking children with high-functioning autism spectrum disorder: An investigation into causal relations. *First Language*, 35(3):189–212.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

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

NL Stein. 1979. An analysis of story comprehension in elementary school children. *New directions in discourse processing/Ablex*, 2:53–120.

Pao-Chuan Torng and Wen-Hui Sah. 2020. Narrative abilities of Mandarin-speaking children with and without specific language impairment: Macrostructure and microstructure. *Clinical linguistics & phonetics*, 34(5):453–478.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tom Trabasso and Linda L Sperry. 1985. Causal relatedness and importance of story events. *Journal of Memory and language*, 24(5):595–611.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. *CoRR*, abs/2304.08085.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Zhihao Wen and Yuan Fang. 2024. Prompt Tuning on Graph-Augmented Low-Resource Text Classification. *IEEE Trans. Knowl. Data Eng.*, 36(12):9080–9095.

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(2):206–221.

A Additional Details of Corpus

A.1 CHAT Format

The CHAT format document includes the basic transcription of the subject's speech, the header, and ancillary lines. The header records background information, such as the subject's and transcriber's personal details, testing date, transcription date, and other relevant data. The ancillary records document coding, evaluation, events, and other auxiliary information of interest to researchers. The main section marks phenomena such as word omissions, speech repetitions, and sentence corrections with special symbols. Hence, before starting the annotation process, the transcribed narrative text in the main section should be preserved, the header content in the CHAT format should be removed, and special marks for speech repetitions, sentence corrections, and other phenomena should be restored and processed.

A.2 Event Annotation

In children's narrative texts, children may describe the process in which the protagonist (such as a little boy or a puppy) searches for a frog, or the events involving other animals (like an owl, a bee, or a deer) that the protagonists encounter in the story. These descriptions are all considered event descriptions in the text during the annotation process. The components of an event include: the event trigger and the event arguments. The trigger is the predicate in the sentence, while the arguments are words such as the subject, object, and adverbial phrases.

As an example, consider the following sentence from the corpus: *"And my dog actually shook a nearby small tree."* Ideally, the annotator should label the event in this sentence as follows:

shook(dog;small tree;;)

In the case of the *"shook"* event, the *"dog"* serves as the agent, acting as the subject of the event, while the *"small tree"* functions as the patient, representing the object of the event.

A.3 Event Relation Annotation

The definitions of event relations are presented in Table 7.

Motivation In $Event_1$ and $Event_2$, $Event_1$ provides a goal-oriented direction for $Event_2$, thereby prompting the occurrence of the $Event_2$ action. This type of causality is referred to as motivation. Typically, $Event_1$ typically contains the goal information.

Example:

(1) *The little boy is looking for the frog. The little boy turns the boots over. [Reference narrative]*

$Event_1$: *looking for (the little boy; the frog; ;)* $\xrightarrow{M}$ $Event_2$: *turns over (the little boy; boots; ;)*

Explanation:

$Event_1$ expresses the little boy's goal of finding the frog, which motivates $Event_2$, *"turning the boots over."* The little boy turns the boots over because he wants to look inside for the frog.

Psychological causation In $Event_1$ and $Event_2$, the action in $Event_1$ triggers an internal reaction in $Event_2$. This type of causality is referred to as psychological causality. The internal reaction is understood as an internal state or psychological state, including various information related to desires, beliefs, thoughts, intentions, and emotions.

Example:

(2) *The next morning, when the boy and the dog woke up, they found the jar was empty. The little boy looked for the frog everywhere. [Reference narrative]*

$Event_1$: *found (they; the jar was empty; the next morning;)* $\xrightarrow{Psy}$ $Event_2$: *looked for (the little boy; the frog; ;)*

Explanation:

$Event_1$, *"found (they; the jar was empty; the next morning;)"*, triggers $Event_2$, *"looked for (the little boy; the frog; ;)"*. Here, the desire *" looked for the frog "* is the boy's internal psychological state, which motivates his action in $Event_2$.

Physical Causation Physical causation refers to the mechanical causal relation between objects and/or people in the real world. It indicates that $Event_1$ is sufficient to cause the occurrence of $Event_2$, without needing any background context.

Example:

(3) *The little dog accidentally fell from the windowsill. The jar broke. [Reference narrative]*

$Event_1$: *fell (The dog; ; ; the windowsill)* $\xrightarrow{Phy}$ $Event_2$: *broke (The jar; ; ;)*

Explanation:

When the dog falls, the jar inevitably breaks. This is consistent with our understanding of the real world, where the fall of the dog directly leads to the jar breaking. This represents physical causality.

Enablement A causality that satisfies the necessity criterion is called enablement. The necessity

relation	definition
synchronous	$Event_1$ and $Event_2$ have a certain degree of temporal overlap.
motivation	$Event_1$ provides a goal or motivation for $Event_2$, prompting the occurrence of the action in $Event_2$.
psychological causation	The action in $Event_1$ triggers an internal reaction in $Event_2$.
physical causation	Under the condition that all background story influences are excluded, $Event_1$ leads to $Event_2$ in a way that satisfies the condition of sufficiency, often governed by physical or natural laws.
enablement	A relation is classified as enablement if, through counterfactual inference, it does not meet the criteria for the other three types of causal relations

Table 7: Definitions of Event Relations. Detailed explanations of internal reactions and counterfactual inference can be found in A.4.

criterion means that if $Event_1$ does not occur, then $Event_2$ will not happen, which is a counterfactual reasoning argument (Appendix A.4). In enablement, the cause is necessary but not sufficient to trigger the result; it is a condition, not a causal reason in the strict sense.

Example:

(4) *The owl chased the little boy all the way. The little boy climbed onto the rock. [Reference narrative]*

$Event_1$:*chased (The owl; the little boy; ; all the way)* $\xrightarrow{E}$ $Event_2$:*climbed (The little boy; the rock; ;)*

Explanation:

If the owl had not chased the little boy, the boy would not have climbed the rock.

Synchronous When $Event_1$ and $Event_2$ describe sentences that indicate a certain level of temporal overlap between the events, expressed by terms like "at the same time" or "meanwhile", their relation is annotated as a synchronous relation.

Example:

(5) *While the little dog barked at the bees in the beehive, the little boy shouted at the hole in the ground. [Reference narrative]*

$Event_1$: *barked (The dog; the bees; ;)* $\xrightarrow{Syn}$ $Event_2$: *shouted (The little boy; ; ; the hole)*

Explanation:

The two events are connected by the temporal indicator "while…," indicating that the events happen at the same time.

A.4 Internal Reaction and Counterfactual Inference

Internal Reaction The "internal reaction" refers to the internal state or psychological states of a character, such as when *"discovering the frog is missing"* triggers the event of *"the boy searching for the frog."* This involves the character's internal psychological states of *"wanting to find the missing frog."*

Counterfactual Inference The counterfactual inference method refers to a reasoning approach where if $Event_1$ and $Event_2$ pass the test of "if $Event_1$ does not occur, $Event_2$ will not occur," then it is concluded that a causal relation exists between $Event_1$ and $Event_2$.

A.5 Scoring Rubric for Microstructure and Psychological States

Microstructure:

- Is the vocabulary rich and diverse? (Evaluate based on total word count and lexical variety.)
- Is the sentence structure complex? (Consider average sentence length and syntactical complexity.)
- Are rhetorical devices effectively used?

Narrative Psychological Expression:

- Are the characters' emotional expressions consistent with the development of the plot?
- Is there any portrayal of the characters' psychological states? (For children, basic emotional reactions are sufficient.)

A.6 The Annotation Tool

The annotation tool is custom-developed and iteratively implemented using the standard graphical user interface (GUI) library Tkinter, which is built into the Python environment. As a module natively

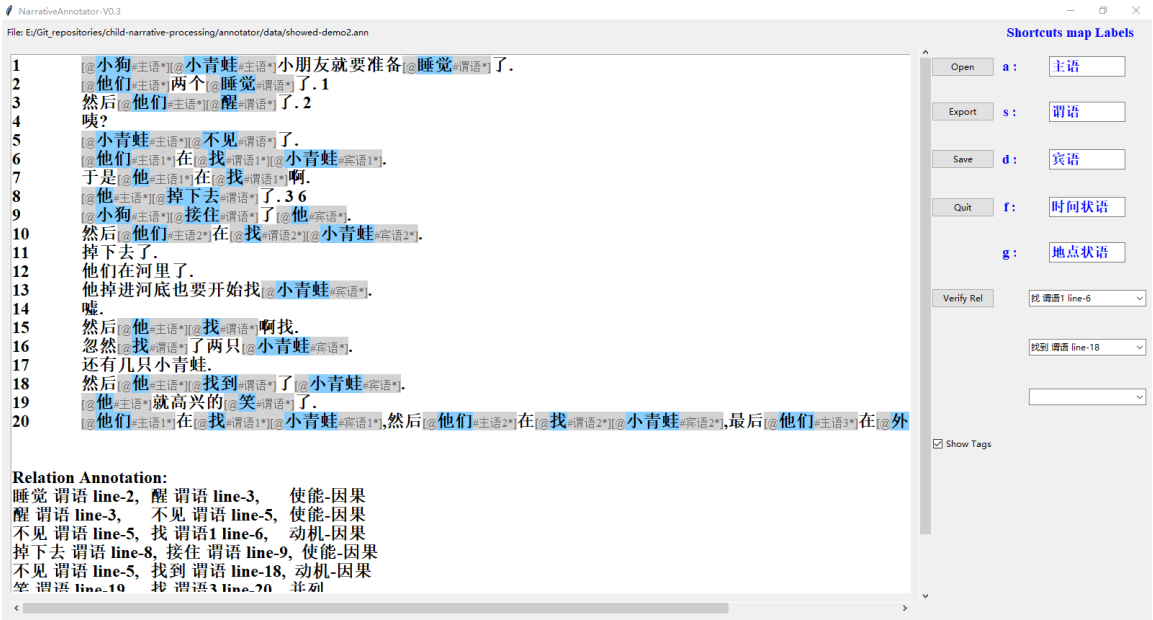

Figure 5: Corpus annotation tool interface.

supported by Python, the Tkinter library offers a high degree of compatibility and stability.

The user interface of the annotation tool is shown in Figure 5. When using the tool, annotators can mark the event triggers and arguments using shortcut keys. Additionally, the tool supports annotating the relations between events by selecting options from a dropdown menu. To avoid the special marks affecting the annotators' reading efficiency, the font size of these marks is reduced, and the background color of the arguments is differentiated for clarity. Annotators can also toggle the visibility of these marks by checking or unchecking the "show tags" checkbox.

The format for the event annotation information is as follows:

```
{
    "sentence_id_in_doc": 17,
    "sentence_text": "And they found the frog.",
    "event_mention": [
        {
            "trigger": {"mention": " found ", "role
                 ": "trigger", "start": 3, "end":
                 3},
            "arguments": [
                {"mention": " they ", "role": "
                     subject ", "start": 2, "end":
                     2},
                {"mention": " the frog ", "role": "
                     object ", start": 4, "end": 5}
            ]
        }
    ]
}
```

The format for the relation annotation information is as follows:

```
{
    "relation_type": "Motivation",
    "first_event": {
        "sentence_id_in_doc": 4,
        "sentence_text": "The frog was missing.",
        "event_mention": [
            {
                "trigger": {"mention": " was missing
                     ", "role": "trigger", "start":
                     3, "end": 4},
                "arguments": ["..."]
            }
        ]
    },
    "second_event": {
        "sentence_id_in_doc": 5,
        "sentence_text": "They were looking for the
             frog.",
        "event_mention": [
            {
                "trigger": {"mention": " were
                     looking for ", "role": "trigeer
                     ", "start": 2, "end": 4},
                "arguments": ["..."]
            }
        ]
    }
}
```

A.7 Scores Analysis

Figure 6 shows the score distributions for macrostructure, microstructure, narrative psychological state, and overall scores. The histograms suggest that while the majority of children's scores cluster around the middle-to-high range (roughly 6 to 8), there is also a noticeable spread on both the lower and higher ends. This indicates that, although many children demonstrate reasonably developed abilities in macrostructure, microstructure, and expression of psychological states, a signifi-

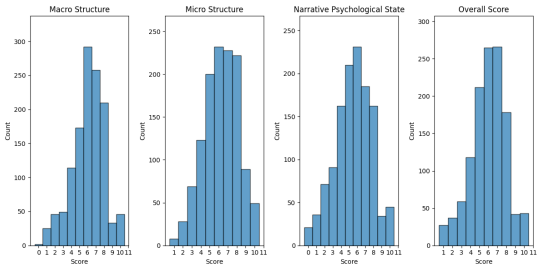

Figure 6: The statistical distribution of the scores.

Model	Overall	Macro	Micro	Psych	Avg
NarGINA-ST emb	**0.787**	**0.767**	**0.717**	**0.636**	**0.727**
NarGINA-LLM emb	0.730	0.720	0.719	0.578	0.687

Table 8: The QWK performance of the NarGINA framework using the LLM (Vicuna_7B) to encode the original event and relation text. ST: *Sentence Transformers*.

cant number exhibit either very strong or relatively weak skills in specific areas. For example, a subset of children score above 8, showing advanced narrative organization and use of language, while others fall below 6, suggesting areas in need of targeted interventions or additional practice.

A.8 Corpus Partition Statistics

The detailed dataset split and key statistics are presented in Table 9.

B Additional Details of Experiment Setup

B.1 Prompt Template

Table 10 presents the prompt templates used for instruction-tuning of LLMs.

B.2 Implementation Details

Narrative Ability Scoring Model We adopt Vicuna_v1.5_7B as our base model and fine-tune it using the `transformers.Trainer`. The evaluation strategy is set to epoch-based, with `per_device_train_batch_size` of 8, `per_device_eval_batch_size` of 4, and a total of 20 training epochs.

We utilize LoRA, a parameter-efficient fine-tuning method that significantly reduces both GPU memory usage and trainable parameters. In our experiments, we set the LoRA rank to 8, LoRA alpha to 16, dropout to 0.05, and use bfloat16 precision. The learning rate is fixed at 3e-4, the weight decay at 0.01, and the warmup ratio at 0.05.

For text generation, we configure both Vicuna_v1.5_7B and Llama2_7B with the following settings: `max_new_tokens=1024`, `temperature=0.2`, `top_p=1.0`, `num_beams=1`, `use_cache=True`, `do_sample=True`.

All fine-tuning and inference are conducted on four NVIDIA A40 GPUs, each equipped with 46 GB of memory.

Narrative Graph Construction Model For node construction, we adopt DeBERTa_v2 as our encoding model, which is fine-tuned using the `lightning.pytorch.Trainer`. The evaluation strategy is set to epoch-based, with `per_device_batch_size` of 16, and a total of 10 training epochs.

For edge construction, we adopt Llama2_13B as our encoding model, while R-GCN is the edge classification model, fine-tuned using the `lightning.pytorch.Trainer`. The evaluation strategy is also set to epoch-based, with `per_device_batch_size` of 16, and a total of 3 training epochs.

All fine-tuning and inference are conducted on four NVIDIA A40 GPUs, each equipped with 46 GB of memory.

C Addition Analysis

For Sentence Transformers, the embedding model is distiluse-base-multilingual-cased-v1. We directly apply it to convert the natural language to embeddings without fine-tuning it, avoiding the additional computational cost. We have considered using embeddings derived from the LLMs (Table 8). It is evident that the LLM-based encoded NarGINA has shown a decline across most traits, with only a slight improvement observed in the microstructure trait. This is due to the fact that LLMs possess a stronger ability to analyze the fine-grained and statistical features of language compared to Sentence Transformers.

	#Documents	#Sentences	#Events	#Arguments	#Event Relations
Train	380	12673	14232	23710	11542
Validation	55	1894	2113	3469	1565
Test	108	3248	3571	5819	3017
Total	543	17815	19916	32998	16124

Table 9: Corpus Partition Statistics.

	Prompt
Description	Your task is to assess a child's narrative ability on the book *Frog, Where Are You?*.
Scoring Criteria	Consider the following three traits, scoring each on a scale of 0-10 (integers): 1. Macrostructure - Does the story have a clear beginning, development, climax, and conclusion? - Is the overall structure coherent, with no abrupt jumps or unreasonable plot points? - Are the character actions logically connected by cause-and-effect relations? 2. Microstructure - Is the vocabulary used rich and diverse? (Refer to total word count and diversity of vocabulary) - Is the sentence structure complex? (Consider average sentence length and syntactical complexity) - Are rhetorical devices used? 3. Narrative Psychological States Expression - Are the character's emotional expressions consistent with the development of the plot? - Is there any psychological portrayal of the character? (For children, basic emotional reactions are sufficient) 4. Total Score - Finally, please weigh each trait's score and provide an overall score in the range of 0-10.
Task Data	1. This is a story told by a child: *<essay text>* 2. Narrative Graph - A narrative graph has been extracted from the essay, showing key events in the story and their causal relations,which can help you assess the organization and coherence of the macrostructure. - Each node represents an event, formatted as: verb (subject; object; adverbial of time; adverbial of place). - Edges represent relations between events, including synchronous, motivation, physical causality,psychological causality, and enablement. - Event token sequence: *<Graph>*
Output Format	Please provide your assessment in the following format: Macrostructure Score: *<macroscore>*, Microstructure Score:*<microscore>*, Psychological state Score: *<psych score>*, Total Score: *<total score>*
Question	Please assess the child's story in terms of the macrostructure,microstructure, and narrative psychological states.

Table 10: Prompt template.