

Benchmarking of Automated Assessment of K-5 Student Narratives Using Large Language Models

Saman Enayati^{*}, Elmira Talebianaraki^{*}, Ziyu Yang & Slobodan Vucetic

Department of Computer and Information Sciences

Temple University

Philadelphia, PA, USA

{saman.enayati, elmira.talebianaraki, zyyang, slobodan.vucetic}@temple.edu

Trina D. Spencer

Department of Applied Behavioral Science

University of Kansas

Lawrence, KS, USA

trinaspencer@ku.edu

Abstract

Assessing narrative language is essential for evaluating early literacy development, but manual scoring of children’s narratives is labor-intensive and requires expert training. In this work, we study whether large language models (LLMs) can reliably assess narrative structure in K–5 student stories, focusing on rubric-based scoring of discourse elements such as character, setting, and problem. We present a systematic evaluation of LLM-based scoring across two expert-annotated datasets of transcribed oral narratives, comparing fine-tuning and few-shot prompting approaches across a range of modern LLM architectures. Using human inter-rater reliability (IRR) as a reference, we find that even smaller-sized fine-tuned LLMs achieve near-human Quadratic Weighted Kappa (QWK) accuracy on in-distribution narratives, while few-shot prompting is less accurate. However, fine-tuned models do not generalize robustly and exhibit substantial degradation on out-of-sample story types. We find that few-shot prompting is competitive in low-resource settings, while fine-tuning is preferable when sufficient labeled data is available. This underscores a need for further research in low-resource automated narrative assessment.

1 Introduction

A narrative is a telling or retelling of causally and temporally related events, such as the characters, problem, attempts to resolve it, emotional responses, and the outcome. Narratives serve as an important marker of linguistic and cognitive development. A substantial body of research has shown that early oral narrative skills are strong predictors of later reading comprehension and writing ability (Bishop & Edmundson, 1987; Dickinson & McCabe, 2001; Snow et al., 2007; Kim et al., 2015). In educational settings, the assessment of children’s narrative skills is important in monitoring language development and screening for developmental language disorders. One widely used approach to assessing narratives involves evaluating macrostructural features, such as the presence and clarity of canonical discourse elements: Character, Setting, Problem, Plan, Action, Consequence, and Ending. This type of scoring, often guided by story grammar rubrics, provides a structured way to assess narrative coherence and complexity (Greenhalgh & Strong, 2001).

^{*}Equal contribution.

Narrative assessment is time-consuming and requires substantial expertise. Teachers and specialists must be trained to recognize the story features and apply the rubrics consistently. The effort required for narrative assessment is often beyond what educators can accommodate in their busy schedules, and this challenge has motivated research into automated assessment systems. Early efforts focused on rule-based systems such as CLAN (MacWhinney & Snow, 1985), and LLUNA (Fox et al., 2022), using metrics derived from word counts and lexical and morphological variables. Traditional machine learning approaches such as random forests have also been employed for automated assessment (Somasundaran et al., 2015) using expert-crafted linguistic features. More recently, deep learning approaches have been applied to narrative scoring tasks and have shown improved performance over traditional approaches (Jones et al., 2019; Fernandez et al., 2022).

Large language models (LLMs), particularly instruction-tuned models such as the GPT family (Brown et al., 2020; Ouyang et al., 2022), have enabled new approaches to educational assessment by supporting both few-shot prompting (FSP), a term often used interchangeably with in-context learning (ICL) in the literature, and task-specific fine-tuning. These models have been successfully applied to a range of educational tasks, including automated essay scoring and short-answer grading (Mizumoto & Eguchi, 2023). Prior work in automated essay scoring (AES) has explored both encoder-based and generative approaches for multi-trait scoring, including fine-tuned BERT models (Fernandez et al., 2022), sequence-to-sequence methods such as T5 (Do et al., 2024a), and rationale-augmented or reinforcement learning-based pipelines that aim to align predictions with scoring rubrics (Chu et al., 2024; Do et al., 2024).

These advances suggest that LLMs may be capable of assessing more complex aspects of student language, including narrative discourse. However, relatively little is known about how well current models can evaluate structured narrative elements in children’s oral storytelling, particularly in settings with limited labeled data and variation across narratives. In this work, we address this gap by systematically evaluating LLM-based approaches to scoring narrative discourse elements in K–5 student narratives.

We evaluate a range of model families, including encoder-based models (e.g., BERT, RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), ModernBERT (Warner et al., 2024)), sequence-to-sequence models (FLAN-T5 (Chung et al., 2024)), and decoder-only large language models (e.g., GPT-3.5/4, LLaMA (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), and DeepSeek-R1 (Guo et al., 2025)). For decoder-only LLMs, we consider both FSP approaches and task-specific fine-tuning. As a traditional baseline, we include a Random Forest model trained on linguistic features. To contextualize model performance, we also report human inter-rater reliability (IRR), which serves as an approximate upper bound on achievable agreement.

Our evaluation is conducted on two datasets of transcribed oral narratives from K–5 students, annotated using established rubric-based scoring schemes. We consider both in-sample and out-of-sample settings to assess generalization across story types. Performance is measured using Quadratic Weighted Kappa (QWK), a standard metric for agreement in educational assessment. In addition, we examine the impact of training data size and prompting strategies, enabling a comparison between data-efficient few-shot prompting and fully supervised fine-tuning approaches.

This work provides a systematic study of LLM-based narrative assessment in a low-resource educational setting. Our contributions are threefold: (1) a unified evaluation framework that compares multiple model families under consistent conditions and against human agreement; (2) an analysis of the trade-offs between few-shot prompting and fine-tuning, including their data efficiency; and (3) empirical evidence that, while LLMs can achieve near-human agreement on in-sample data, their performance degrades under distribution shift, highlighting limitations for real-world deployment.

Table 1: Definition of rubrics and their score ranges for each dataset

Rubrics	Definition	Alien Data	ALPS Data
Character (Char)	Who in the story acts as the agent	0–3	0–3
Setting (Sett)	Time/place where story takes place	0–3	0–3
Initiating Event (Prob)	Event/problem that causes the story	0–3	0–4
Plan (Plan)	Idea character has to fix the problem	0–3	NA
Action (Act)	The action taken by character	0–3	NA
Plan & Attempt (Att)	Plan/action taken by character	NA	0–4
Consequence (Con)	Event caused by character’s action	0–3	0–4
Ending (End)	Action by character at the end	NA	0–2
Emotion (Emo)	Character’s feeling about the problem	NA	0–3

2 Task and Experimental Setup

2.1 Narrative Story Datasets

We evaluate model performance on two datasets of transcribed oral narratives produced by K–5 students: the Academic Language of Primary Students (ALPS) dataset (Spencer & Staff, 2023) and the Alien Story dataset (Jones et al., 2019). These datasets differ in elicitation protocols, scoring rubrics, and narrative diversity, allowing us to assess model performance across varied narrative assessment conditions. Each narrative is assessed by trained human raters using rubric-based scoring schemes that assign discrete scores to a set of discourse elements (e.g., Character, Setting, Problem, Consequence). These elements capture key aspects of narrative macrostructure and are commonly used in educational assessment.

2.1.1 ALPS Dataset

The ALPS dataset contains 2,695 narratives produced by students in kindergarten through third grade. Narratives are elicited using picture-based prompts. There were nine sets of three pictures, referred to as story types. Researchers would randomly pick a picture set and display the three related pictures to a student. They would either model a grade-appropriate story and ask the student to retell it or ask the student to generate a story about the pictures without modeling (Kirby et al., 2021). All narratives are transcribed and assessed by human experts using the Narrative Language Measures (NLM) framework (Petersen & Spencer, 2012) with discourse elements defined in Table 1. A subset of 823 narratives were assessed by two different researchers, enabling estimation of inter-rater reliability.

2.1.2 Alien Story Dataset

The Alien Story dataset consists of 414 narratives obtained from K-5 students (Jones et al., 2019). The narratives were generated in response to prompts from the Test of Narrative Language-2 (TNL) (Gillam & Pearson, 2004). The stories were elicited by showing a single image depicting an alien family landing in the park. Narratives were transcribed and scored by human experts following the Monitoring Indicators of Scholarly Language (MISL) rubric (Gillam et al., 2017) with six narrative elements from Table 1. All narratives were independently annotated by two experts.

2.2 Problem Statement

We formulate narrative assessment as a multi-output prediction task: given a transcribed narrative, the model predicts a vector of rubric-based scores $r = \{r_1, r_2, \dots, r_m\}$, where m is the number of discourse elements. This formulation reflects the structured nature of narrative assessment, where multiple interdependent elements (e.g., problem, action, consequence) must be evaluated jointly. Model outputs are constrained to valid score ranges defined by the corresponding rubrics.

2.3 Models for Narrative Assessment

2.3.1 Random Forest with Coh-Metrix Features (CMRF)

As a classical baseline, we train a Random Forest regression model on 108 linguistic features extracted using Coh-Metrix (McNamara et al., 2014). These features capture lexical, syntactic, and discourse-level properties of the narrative. Separate regressors are trained for each discourse element. This baseline, referred to as CMRF, was proposed in (Somasundaran et al., 2015) and serves as a lower bound comparison for modern language models.

2.3.2 Encoder-Based Models

We fine-tune encoder-based transformer models, including BERT-large (Devlin et al., 2018), RoBERTa-large (Liu et al., 2019), DeBERTa-v3-large (He et al., 2021), and ModernBERT-large (Warner et al., 2024), to predict rubric scores. For each discourse element, a regression head is applied to the final representation, and models are trained using mean squared error loss. Predictions are rounded to the nearest valid score.

2.3.3 Sequence-to-Sequence Model

We evaluate FLAN-T5-large (Chung et al., 2024), an instruction-tuned sequence-to-sequence model, by framing narrative assessment as a text generation task. Given a narrative and scoring prompt, the model generates structured outputs describing scores for each discourse element. Numerical scores are extracted via post-processing.

2.3.4 Decoder-Only Language Models

We fine tune several decoder-only instruction-tuned models using an autoregressive objective, including GPT-3.5, GPT-4o-mini, LLaMA 3 (8B), Mistral-7B-Instruct-v0.3, Qwen3.5-4B, and DeepSeek-R1-Distill-Qwen-7B. To improve fine-tuning efficiency, we use Low-Rank Adaptation (LoRA) (Hu et al., 2022), updating a small set of adapter parameters rather than the full model. All models are trained to jointly predict all rubric scores and outputs are post-processed to extract numerical predictions.

In addition to fine-tuning, we also evaluated GPT-3.5, GPT-4o-mini, GPT-4o under FSP paradigm, where models are prompted with task instructions and possibly a small number of labeled examples, without updating model parameters. For FSP, we experiment with variations in prompt design, including the number of exemplars and output formats.

The choice between fine-tuning and FSP depends on the availability of labeled examples. Few-shot prompting is appealing when there are only a few labeled examples or even when the labeled examples are unavailable. Alternatively, fine-tuning is appealing when the labeled dataset is sufficiently large, where "sufficiently large" depends on the application.

2.4 Evaluation Metric

We evaluate model predictions using *Quadratic Weighted Kappa* (QWK), a standard reliability metric in educational assessment. It compares predictions to the true labels, adjusts for agreement that could happen by chance, and rewards close mistakes more than large ones. It is especially suitable for ordinal data, such as rubric scores, where the magnitude of disagreement matters. QWK of 1 indicates perfect agreement, while 0 indicates agreement equivalent to random guessing.

Since both ALPS and Alien dataset have narratives scored by two human annotators, we can establish an upper-bound on model accuracy as the human *inter-rater reliability* (IRR) calculated as QWK between two trained human annotators. The primary coder's annotations serve as the ground truth and the secondary coder's as the prediction. We compute element-wise and macro-averaged QWK across rubric dimensions.

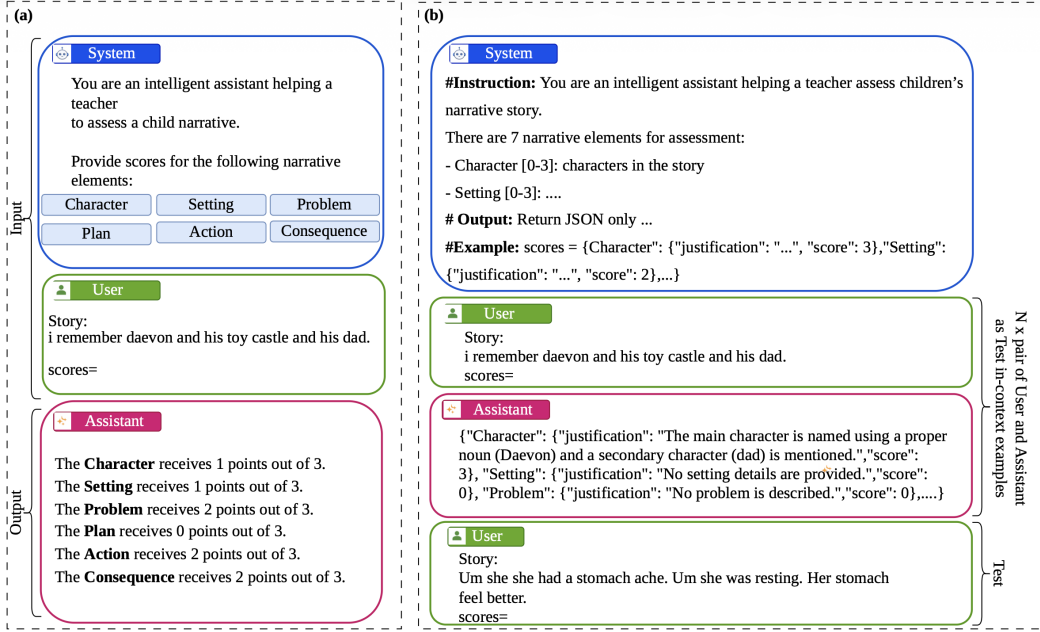


Figure 1: Comparison of GPT fine-tuning and FSP on ALPS data. (a) Input-output training example for fine-tuning. (b) Default prompt used for FSP on ALPS data.

3 Experiments

3.1 Experimental Protocol

3.1.1 Data Splits

Using ALPS data, we evaluate model performance on both in-sample and out-of-sample settings to assess generalization across narrative types. We partition 2,695 narratives across nine story types in ALPS dataset into: 1) an in-sample subset, consisting of narratives from six story types, used for training (1,686), validation (186), and testing (573); and 2) an out-of-sample test set, consisting of 250 narratives from the remaining three story types, which are not observed during training. This setup enables evaluation of model generalization to unseen narrative types.

To study data efficiency, we also construct a low-resource setting using 10% of the ALPS training data (168 examples). In this setting, model selection is performed via 5-fold cross-validation rather than a held-out validation set.

For the Alien Story dataset (414 narratives), we use a random split of 331 training, 33 validation, and 50 test examples. As all narratives correspond to a single story type, evaluation on this dataset reflects in-sample performance.

In both datasets, a subset of narratives is annotated by multiple raters, allowing computation of human inter-rater reliability (IRR) as an approximate upper bound on model performance.

The summary statistics for both datasets are shown in Table 4 in the Appendix.

3.1.2 Model Training

All models are trained on their respective training sets using consistent input-output formats based on rubric scoring, as explained in the previous section. Hyperparameter settings for training and fine-tuning are as follows: **CMRF**: Random Forest with 100 trees, default scikit-learn settings. **Encoder model fine-tuning**: learning rate $5e-5$, batch size 8, warmup steps 25, linear scheduler, 30 epochs. **FLAN-T5 fine-tuning**: learning rate $1e-4$, batch

size 32, max sequence length 512. **Decoder-only LLMs:** (LLaMA 3, Mistral-7B, Qwen3.5, DeepSeek-R1): fine-tuned with LoRA, rank 16, $\alpha = 32$, dropout 0.05, learning rate $5e-6$, batch size 1 with gradient accumulation (8 steps), 30 epochs, early stopping patience 2. **GPT Fine-Tuning:** We fine-tuned GPT-3.5 and GPT-4o-mini models. We used OpenAI manual (OpenAI, 2024) defaults, temperature 0, stop token “##”. The specific input-output format for GPT fine-tuning is illustrated in Figure 1. **GPT Few-Shot Prompting:** temperature 0, max output length 512. The best model checkpoint is selected based on its performance on validation set.

3.1.3 Few-Shot Prompting

We evaluate FSP using GPT-3.5, GPT-4o-mini, and GPT-4o. Prompts consist of task instructions, a set of labeled exemplars, and a target narrative to be scored. We explore two different ways to present exemplars; one is just providing the stories with their assigned scores, and another is providing scoring justifications together with the scores. ALPS and Alien datasets do not come with any scoring justifications. Thus, we randomly selected 20 narratives from the training data and asked an expert on our team to provide brief rubric-grounded explanations for the already assigned human scores. These justifications were used solely as prompt exemplars and were written after the human scores had already been assigned and therefore did not influence the gold-standard annotations used for training or evaluation. The FSP prompt with justifications is illustrated in Figure 1 in the main text and Figure 4 in the Appendix.

By default, we use five exemplars sampled from the training set, each annotated with rubric scores and brief justifications. As a baseline, we also explored the zero-shot approach where only the task instructions are provided through the system prompt, without providing any exemplars. We also perform ablation experiments with variations in prompt design, including: number of exemplars (0, 1, 5, 10, 20), exemplar selection strategy, output format, and inclusion of chain-of-thought reasoning. For efficiency, ablation experiments are conducted on a subset of the ALPS test set (150 in-sample and 150 out-of-sample examples). For the Alien dataset, we use a fixed prompt configuration without ablations.

3.2 Results

3.2.1 In-Sample Performance

Table 2 summarizes model performance on the ALPS in-sample test set (the table shows only 4 of 7 discourse elements that are shared with the Alien dataset. The full Table 13 is included in the Appendix). It includes performance of models trained or fine-tuned on the full training dataset, 10% of the training dataset, and performance of FSP of GPT models. In this subsection, we will focus on in-sample performance of models fine-tuned on the full ALPS training dataset, which includes 1,686 narratives for training and another 168 for validation. In-sample means that the models were tested on test narratives from the same 6 story types that were seen during training.

Across model families, language models fine-tuned on the full training data achieve moderate to strong agreement with human raters. The best performing fine-tuned ModernBERT, FLAN-T5, and DeBERTa-v3 and slightly less accurate Mistral, gpt-4o-mini, Llama-3, RoBERTa, BERT, gpt-3.5, and Qwen3.5 all approach human inter-rater reliability (IRR). In particular, these models achieve macro-averaged QWK scores within a few points of the human agreement level, indicating that many LLM-based systems can approximate expert scoring under in-distribution conditions. In contrast to LLMs, the feature-based Random Forest baseline performs substantially worse, highlighting the importance of semantic and contextual modeling for narrative assessment.

Interestingly, in our experiments, encoder-based models are more accurate than decoder-only models. This could probably be explained by their smaller sensitivity to the choice of fine-tuning hyperparameters. We note that we did not have computational resources to perform exhaustive hyperparameter search and mostly relied on common defaults.

Table 2: QWK accuracy of models on in-sample (573) and out-of-sample (250) ALPS narratives. "***" indicates models fine-tuned on 10% (168) of the training data. FSP uses 5 exemplars.

model	in-sample - ALPS					out-of-sample - ALPS				
	Char	Sett	Prob	Con	Avg	Char	Sett	Prob	Con	Avg
IRR	98.1	86.8	75.1	66.6	80.0	93.6	83.5	77.8	57.9	72.4
fine-tuning										
CMRF	48.3	41.2	12.0	32.7	32.1	38.9	18.3	10.7	32.4	23.4
CMRF**	38.1	29.1	0.0	24.6	22.1	38.6	13.5	3.2	27.7	22.1
BERT-large	92.6	82.4	69.5	56.3	73.5	79.9	65.6	50.3	48.3	57.0
BERT-large**	72.9	50.7	39.4	25.0	42.9	37.9	34.2	14.3	26.0	26.1
DeBERTa-v3-large	91.9	86.2	72.5	71.0	78.3	86.8	71.1	60.0	66.1	64.7
DeBERTa-v3-large**	77.3	69.3	25.3	43.7	50.9	77.0	57.9	18.7	31.3	41.4
RoBERTa-large	93.0	84.0	65.1	64.7	75.1	84.2	65.2	61.6	52.4	56.4
RoBERTa-large**	77.0	68.9	51.7	56.2	59.9	66.9	38.7	11.6	37.7	42.5
ModernBERT-large	92.0	87.4	71.6	72.9	79.3	81.1	72.0	52.6	60.8	63.5
ModernBERT-large**	82.5	59.7	33.7	46.8	50.3	75.9	30.8	24.0	33.3	36.9
FLAN-T5-large	93.4	88.9	70.1	70.2	78.5	79.9	70.1	64.1	60.2	65.2
FLAN-T5-large**	87.9	73.8	56.8	53.3	66.1	84.2	59.3	24.4	33.6	50.2
Mistral-7B	90.6	83.7	72.6	71.3	77.8	83.3	65.4	58.4	57.1	63.1
Mistral-7B**	80.9	48.5	55.5	56.1	58.8	78.0	45.7	32.8	54.8	48.8
Qwen3.5-4B	88.9	74.9	61.1	63.6	72.7	83.0	55.2	50.2	57.5	60.6
Qwen3.5-4B**	67.1	52.1	65.1	52.8	54.9	62.0	38.3	41.4	53.8	48.8
DeepSeek-R1 (7B)	87.4	73.9	57.5	58.2	66.8	79.6	53.3	46.5	57.8	55.6
DeepSeek-R1 (7B)**	66.5	52.0	41.8	45.9	46.4	60.8	45.1	36.2	39.8	40.7
Llama-3-Instruct (8B)	89.8	81.6	66.9	67.4	75.2	80.4	66.8	56.8	52.7	63.0
Llama-3-Instruct (8B)**	76.3	69.2	57.6	52.3	62.1	69.9	36.7	39.7	45.7	46.2
gpt-3.5	90.3	84.0	62.1	64.9	72.8	78.5	71.7	56.7	61.9	62.2
gpt-3.5**	88.3	74.5	61.0	61.7	70.3	81.2	61.9	47.6	54.6	59.8
gpt-4o-mini	91.3	86.6	67.5	67.8	77.0	80.0	71.3	62.4	63.3	65.7
gpt-4o-mini**	89.3	80.7	68.2	68.3	74.7	83.8	63.3	60.8	63.5	64.9
few-shot prompting										
gpt-3.5	77.0	53.4	31.5	38.2	49.5	68.7	28.9	39.1	44.3	43.0
gpt-4o-mini	80.4	60.8	60.0	68.4	60.3	69.9	44.4	47.3	61.8	51.9
gpt-4o	82.7	67.3	68.1	68.4	67.3	77.5	51.4	54.5	60.4	55.2

3.2.2 Out-of-Sample Generalization

To evaluate model robustness to change in distribution, we measure model performance on ALPS narratives on three out-of-sample story types that were not observed during training (Table 2). All models exhibit a decline of at least 10% in out-of-sample QWK compared to the in-sample setting. This gap suggests that high agreement with human raters in controlled settings does not necessarily translate to robust performance under distribution shift.

Interestingly, we observe that fine-tuned encoder-based models exhibited a larger reduction in accuracy on out-of-sample data than the fine-tuned GPT models. As a result, gpt-4o-mini emerges as the most accurate out-of-sample model. This behavior is not surprising considering that pretrained gpt-4o-mini model is superior to pretrained encoder-based models used in our study.

3.2.3 Data Efficiency and Learning Paradigms

In this subsection we focus on understanding the impact of training data size and learning paradigm by comparing models trained on the full ALPS dataset with those trained on its 10% random subset (Table 2).

Table 3: QWK accuracies of models on 50 Alien test stories. Fine-tuned models use 331 training instances; FSP uses 5 exemplars.

	Char	Sett	Prob	Con	Avg QWK
<i>IRR</i>	95.6	60.1	54.7	41.0	55.5
<i>fine-tuning</i>					
<i>CMRF</i>	33.0	18.8	34.3	26.7	27.0
<i>BERT-large</i>	93.8	26.8	39.6	31.7	48.5
<i>DeBERTa-v3-large</i>	97.8	34.0	37.8	47.9	52.9
<i>RoBERTa-large</i>	96.7	40.6	48.9	47.0	52.6
<i>ModernBERT-large</i>	95.5	14.3	37.7	42.0	44.4
<i>FLAN-T5-large</i>	95.5	26.8	44.2	42.3	44.5
<i>Mistral-7B</i>	76.9	15.2	46.3	60.0	47.0
<i>Qwen3.5-4B</i>	75.8	29.0	43.3	47.0	47.5
<i>DeepSeek-R1 (7B)</i>	72.4	34.8	46.1	44.7	45.1
<i>Llama3-Instruct (8B)</i>	67.0	39.4	26.6	51.8	45.6
<i>gpt-3.5</i>	93.9	39.8	45.1	53.1	57.4
<i>gpt-4o-mini</i>	97.7	41.3	50.0	49.3	54.9
<i>few-shot prompting</i>					
<i>gpt-3.5</i>	36.0	12.1	35.8	35.9	32.5
<i>gpt-4o-mini</i>	59.8	20.1	63.8	53.3	48.5
<i>gpt-4o</i>	59.0	29.3	56.5	57.8	50.0

Not surprisingly, reducing the amount of training data leads to a consistent decrease in performance across all model families. Interestingly, the degradation is much more significant in encoder-only models than in gpt models, where QWK decreased only slightly. As a result, gpt-4o-mini fine-tuned on 10% of training data achieves QWK of 74.7, while the best fine-tuned encoder-only RoBERTa-large achieves QWK of only 59.9. It appears that large pre-trained decoder-only models can benefit from relatively small fine-tuning training data, while the smaller models require much larger fine-tuning training data to reach state of the art performance.

FSP offers a competitive alternative when labeled data is extremely scarce. Table 2 lists GPT FSP results when only 5 exemplars of labeled narratives with justifications are available. The resulting QWK are smaller than for fine-tuned GPT models, but the difference is not extremely large.

Table 3 shows QWK for all models on the Alien dataset (full Table 15 with all 6 discourse elements is in the Appendix), which includes only a single narrative type and no out-of-sample test split. All models are trained on 331 training and 33 validation narratives, which is comparable in size to the 10% ALPS training dataset. The results are consistent with the 10% ALPS results, in that fine-tuned GPT models are the most accurate. Interestingly, gpt-4o FSP model that sees only 5 exemplars is more accurate than all but two (RoBERTa-large, DeBERTa-v3-large) non-GPT fine-tuned models.

Together, these results highlight the complementary strengths of FSP and fine-tuning, and suggest that the choice of approach should depend on the availability of labeled data and the desired level of robustness. The results suggest that while FSP is viable in low-resource settings, supervised fine-tuning provides more reliable performance when a moderate to large training dataset is available.

3.2.4 Few-Shot Prompt Design Ablations

To understand how prompt construction affects few-shot prompting performance, we conducted several ablation studies described in more detail in the Appendix. We varied the number and selection of exemplars and prompts. We observed that selecting exemplars

randomly is sufficiently successful. We observed that using 5 or more exemplars is better than using single or no exemplars in the prompt. Five exemplars resulted in comparable accuracy to 10 and 20 exemplars. Given the cost of prompting, 5 exemplars appears to be a good choice. We also observed that including justifications for each score consistent with the chain-of-thought approaches resulted in improved accuracy compared to using exemplars without justifications. Finally, we observed that FSP is superior to zero-shot learning that relies only on instructing how to assess the narratives rather than demonstrating it through examples.

These results confirm that FSP for educational assessment tasks is highly sensitive to prompt engineering. Since the ablation experiments were conducted on subsets of the test data, we view these findings as exploratory rather than definitive.

3.3 Cost Analysis

In addition to accuracy, we analyze the monetary cost associated with different modeling approaches. Table 12 in the Appendix presents the specific dollar cost of training each model on the full ALPS training set for one epoch and testing on the full ALPS test set. For API-based models, cost is estimated based on token usage during both inference and fine-tuning, following provider pricing at the time of experimentation. For locally trained models on an NVIDIA RTX A5000 GPU, we use an on-demand rate of \$1.006 per hour, which reflects the current cost of renting the same GPU through Amazon cloud services.

Our results reveal a trade-off between cost and performance. Fine-tuned LLMs achieve the highest accuracy but require substantially more resources due to the need for supervised training on labeled data. In contrast, FSP avoids training costs but incurs higher per-instance inference cost, as each prediction requires processing multiple exemplars within the prompt.

Despite these differences, FSP remains a practical option in low-resource settings, where labeled data is limited and fine-tuning may not be feasible. However, as the number of predictions increases, the cumulative cost of FSP can exceed that of fine-tuned models, particularly when using larger models with longer prompts.

Among the evaluated approaches, smaller fine-tuned models (e.g., FLAN-T5 and encoder-based models) offer a favorable balance between cost and performance, achieving competitive accuracy at significantly lower computational expense. These findings suggest that model selection for narrative assessment should consider not only accuracy but also deployment constraints, such as data availability, inference scale, and computational budget.

4 Conclusions

In this work, we evaluated the use of LLMs for automated assessment of narrative discourse in K–5 student stories. Across two datasets and multiple model families, we found that fine-tuned LLMs can achieve strong agreement with human raters, approaching inter-rater reliability on in-distribution narratives. However, this performance does not generalize robustly as all models deteriorate when evaluated on out-of-sample narrative types. We also observed that few-shot prompting provides a flexible and data-efficient alternative when labeled data is scarce, while fine-tuning yields more reliable performance when training data is plentiful. Future work should focus on improving robustness under distribution shift and exploring methods that better capture and utilize relationships among discourse elements such as problem, plan, consequence, and ending. Several discourse elements that proved challenging for automated scoring also exhibited relatively low human inter-rater reliability. Future work should investigate whether these difficulties arise primarily from limitations of current models or from ambiguity in how certain discourse elements are defined and assessed.

5 Limitations

Our study has several limitations that warrant consideration.

First, our evaluation focuses on relatively short oral narratives produced by students in lower school. It remains to be seen how well current models scale to longer, more complex stories or to writing from older students. Additionally, our datasets are drawn from U.S.-based, English-speaking populations using fixed picture prompts. Thus, generalization to other languages and educational contexts remains an open question. In settings where labeled data are unavailable or expensive to obtain, our low-resource results suggest that few-shot prompting may provide a practical alternative to task-specific fine-tuning, but this remains to be studied in multilingual narrative assessment.

Second, our models are trained and evaluated against specific scoring rubrics (NLM and MISL). Their adaptability to alternative assessments remains an open question.

Third, we did not have access to student demographic information, which would be useful in revealing different types of biases in the narratives and their human and automatic assessments.

Fourth, while our results indicate usefulness of few-shot prompting in low-resource regimes, they also reveal fragility: performance varies substantially depending on prompt design, number and type of examples, and presence of rubric-based justifications. This sensitivity may limit the reliability of few-shot prompting in uncontrolled environments.

Finally, several of our strongest-performing models such as GPT-4o and GPT-4o-mini are proprietary and accessed via closed APIs. This constrains transparency, reproducibility, and accessibility for practitioners in low-resource settings. Exploring open-weight alternatives remains an important direction.

Despite these limitations, our work provides evidence that fine-tuned LLMs can approach or exceed human-level scoring accuracy on narrative discourse tasks, even with limited training data. Future work should explore model calibration, fairness, and transfer learning across rubrics and populations, and integrate these systems into end-to-end educational tools, including speech-to-text pipelines.

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Dorothy Vera Margareth Bishop and Andrew Edmundson. Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders*, 52(2):156–173, 1987.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. Rationale behind essay scores: Enhancing S-LLM’s multi-trait essay scoring with rationale generated by LLMs. *arXiv preprint arXiv:2410.14202*, 2024.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- David K Dickinson and Allyssa McCabe. Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research & Practice*, 16(4):186–202, 2001.

- Heejin Do, Yunsu Kim, and Gary Lee. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, 2024.
- Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk, and Andrew Lan. Automated scoring for reading comprehension via in-context BERT tuning. In *International Conference on Artificial Intelligence in Education*, pp. 691–697. Springer, 2022.
- Carly Fox, Sharad Jones, Sandra Laing Gillam, Megan Israelsen-Augenstein, Sarah Schwartz, and Ronald Bradley Gillam. Automated progress-monitoring for literate language use in narrative assessment (LLUNA). *Frontiers in Psychology*, 13:894478, 2022.
- Ronald Bradley Gillam and Nils A Pearson. *Test of Narrative Language*. Pro-Ed, 2004.
- Sandra Laing Gillam, Ronald B Gillam, Jamison D Fargo, Abbie Olszewski, and Hugo Segura. Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. *Communication Disorders Quarterly*, 38(2):96–106, 2017.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Kellie S Greenhalgh and Carol J Strong. Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools*, 32(2):114–125, 2001.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.
- Sharad Jones, Carly Fox, Sandra Gillam, and Ronald B Gillam. An exploration of automated narrative analysis via machine learning. *PLOS ONE*, 14(10):e0224634, 2019.
- Young-Suk Kim, Stephanie Al Otaiba, and Jeanne Wanzek. Kindergarten predictors of third grade writing. *Learning and Individual Differences*, 37:27–37, 2015.
- Megan S Kirby, Trina D Spencer, and Yi-Jui Iva Chen. Oral narrative instruction improves kindergarten writing. *Reading & Writing Quarterly*, 37(6):574–591, 2021.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Brian MacWhinney and Catherine Snow. The child language data exchange system. *Journal of Child Language*, 12(2):271–295, 1985.
- Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, 2014.
- Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023.

- OpenAI. OpenAI platform documentation: Fine-tuning. <https://platform.openai.com/docs/guides/fine-tuning/hyperparameters>, 2024. Accessed: 30 April 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Douglas B Petersen and Trina D Spencer. The narrative language measures: Tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education*, 19(4):119–129, 2012.
- Catherine E Snow, Michelle V Porche, Patton O Tabors, and Stephanie Ross Harris. *Is Literacy Enough? Pathways to Academic Success for Adolescents*. Paul H. Brookes Publishing Co., 2007.
- Swapna Somasundaran, Chungmin Lee, Martin Chodorow, and Xinhao Wang. Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 42–48, 2015.
- Trina D. Spencer and Staff. Academic language of primary students (ALPS), 2023. Available at: <https://nyu.databrary.org/volume/1632> (Accessed: 27 February 2024).
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Alenty Wettig, Quentin Vallat, Rémi Louf, and Max Bartolo. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

A Appendix

A.1 Basic Statistics for ALPS and Alien Story Data

Table 4: Statistics of the ALPS and Alien datasets

	Alien	ALPS
Story Types	1	9
Number of story types for training	1	6
Number of training instances	331	1686
Number of validation instances	33	186
Number of testing instances	50	823
In-sample test instances [story types]	50 [1]	573 [6]
Out-sample test instances [story types]	NA	250 [3]

A.2 Ablation study of few-shot prompting on ALPS data

To explore the effectiveness of various prompt engineering techniques, we conducted experiments on several zero-shot and few-shot prompting settings using *GPT-4o* and *GPT-4o-mini* models. QWK was measured on 150 narratives randomly chosen from the in-sample ALPS test set and 150 narratives randomly chosen from the out-of-sample ALPS test set. Our evaluation focused on **exemplar selection strategy**, **number of exemplars** and **prompting**.

A.2.1 Ablation study of exemplar selection

We used the prompt illustrated in the right panel of Figure 1 with five exemplars. The ablation experiments varied how those five exemplars were selected from the available set of 20 candidate exemplars with justifications from ALPS training data:

- **Random Selection:** Five exemplars were selected randomly.
- **Quantile-Based Selection:** Candidate exemplars were sorted by their total score. One exemplar was randomly chosen from each fifth of the sorted list.
- **Single-Story Type:** All 5 exemplars were taken from a single story type.

Table 5 shows QWK accuracies and their standard deviations on 150 ALPS in-sample test narratives by the few-shot prompting of GPT-4o model. Table 6 shows QWK accuracies of the same model on 150 ALPS out-of-sample test narratives. Tables 7 and 8 are the same as Tables 5 and 6, but using GPT-4o-mini model. The results indicate that random story selection might be superior to the other two options. Selection of the exemplars from a single story type appears to be inferior, which can be explained by the lower diversity of assessed narratives. We notice that the QWK differences between the three selection strategies are small compared to their standard deviations. Therefore, we view these findings as exploratory rather than definitive.

A.2.2 Ablation study of exemplar numbers

We used the prompt illustrated in Figure 1 with $N = 1, 5, 10, 20$ in-context exemplars randomly selected from the 20 candidate exemplars. We evaluated both GPT-4o (Table 9) and GPT-4o-mini (Table 10) models on 150 in-sample ALPS test examples. We could observe that $N = 1$ appears inferior to other choices of N . The differences between $N = 5, 10, 20$ choices appeared to be statistically insignificant on our test data. This result justifies choice of $N = 5$ because it results in smaller prompts and reduced inference costs.

A.2.3 Ablation study of alternative prompt configurations

We refer to the default prompt format shown in Fig 1 with 5 randomly selected exemplars as *In-Context with Justification + Scores*. We examined the following alternative configurations:

- **Zero-Shot (No Flowchart):** This configuration includes only the task instructions through the system prompt, without providing any exemplars.
- **Zero-Shot (Flowchart):** The same as the previous non-exemplar setting, but the system prompt includes a point-by-point guide for how teachers should score a narrative, following a structured flowchart (See Figure 3).
- **In-Context with Scores Only:** In this variation, the scores for exemplars were provided without justifications (See Figure 2).
- **In-Context with Scores + Justification** refers to the setup where the justification is provided first, followed by the score, which is the opposite of the default.

Table 11 presents QWK accuracies of GPT-4o with different prompt configurations on 150 ALPS in-sample test examples. Among the tested setups, the *In-Context with Scores + Justification* configuration achieved the highest overall performance, with an average QWK score of 67.3%, followed by the default *In-Context with Justification + Scores* configuration with 65.7%. Following these, the *Zero-Shot (Flowchart)* configuration emerged as the best exemplar-free setup, achieving an overall QWK score of 56.4%. However, the instruction for this configuration is longer than the default one, which makes the experiments more costly.

The results also underscore the critical role of providing reasoning in prompts. This is evident from the performance of the *In-Context with Scores Only* configuration, which resulted in a significant drop in overall QWK score to 53.6%, marking a 12.1% decrease compared to the *In-Context with Justification + Scores* configuration. These findings highlight the importance of incorporating detailed reasoning and task demonstrations in prompt design to enhance model performance across narrative scoring tasks.

We also examine how the GPT model’s performance is affected by the order of score and justification in both the output and in-context examples. As shown in Table 11, placing the score before the justification yields slightly better results.

Table 5: Performance of GPT-4o with few-shot prompting on 150 in-sample ALPS narratives.

Configuration	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWK
Random [5]	82.0 $\pm$ 3.0	67.3 $\pm$ 4.2	64.2 $\pm$ 10.1	69.5 $\pm$ 6.6	69.7 $\pm$ 7.1	37.7 $\pm$ 6.5	69.8 $\pm$ 5.1	65.8 $\pm$ 3.1
Quantile [5]	83.1 $\pm$ 2.5	59.6 $\pm$ 5.3	67.8 $\pm$ 10.3	68.6 $\pm$ 6.7	61.9 $\pm$ 7.2	37.0 $\pm$ 6.2	79.1 $\pm$ 3.9	65.3 $\pm$ 3.0
Single Story Type [5]	61.5 $\pm$ 3.8	62.1 $\pm$ 4.8	64.3 $\pm$ 10.0	61.2 $\pm$ 7.2	56.4 $\pm$ 8.4	32.3 $\pm$ 6.0	72.6 $\pm$ 4.7	58.6 $\pm$ 3.1
IRR	96.7 $\pm$ 1.4	88.1 $\pm$ 2.7	72.7 $\pm$ 9.3	82.2 $\pm$ 4.5	68.4 $\pm$ 7.0	61.7 $\pm$ 6.4	83.2 $\pm$ 4.4	79.0 $\pm$ 2.7

Table 6: Performance of GPT-4o with few-shot prompting on 150 out-of-sample ALPS narratives.

Configuration	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWK
Random [5]	75.2 $\pm$ 3.5	52.1 $\pm$ 6.1	62.4 $\pm$ 7.9	63.1 $\pm$ 6.6	61.5 $\pm$ 7.0	24.2 $\pm$ 6.2	39.1 $\pm$ 6.5	53.9 $\pm$ 2.9
Quantile [5]	77.7 $\pm$ 3.3	50.0 $\pm$ 6.2	54.0 $\pm$ 8.7	65.1 $\pm$ 6.6	62.0 $\pm$ 6.7	24.4 $\pm$ 5.9	51.1 $\pm$ 6.9	54.9 $\pm$ 2.8
Single Story Type [5]	58.2 $\pm$ 4.4	42.7 $\pm$ 6.3	62.3 $\pm$ 7.7	65.2 $\pm$ 6.6	63.7 $\pm$ 7.3	22.1 $\pm$ 6.0	47.6 $\pm$ 7.4	51.7 $\pm$ 3.1
IRR	95.7 $\pm$ 1.5	85.8 $\pm$ 3.9	82.2 $\pm$ 4.8	56.7 $\pm$ 8.2	52.2 $\pm$ 6.8	52.1 $\pm$ 6.6	70.8 $\pm$ 6.4	70.8 $\pm$ 2.3

Table 7: Performance of GPT-4o-mini with few-shot prompting on 150 in-sample ALPS narratives.

Configuration	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWK
Random [5]	79.9 $\pm$ 3.1	60.4 $\pm$ 4.9	49.4 $\pm$ 10.8	62.3 $\pm$ 6.0	58.9 $\pm$ 7.3	40.7 $\pm$ 6.9	66.2 $\pm$ 4.9	59.7 $\pm$ 3.1
Quantile [5]	85.0 $\pm$ 2.2	61.2 $\pm$ 5.5	49.3 $\pm$ 10.4	58.2 $\pm$ 6.6	58.6 $\pm$ 6.5	42.5 $\pm$ 6.6	57.4 $\pm$ 6.0	58.9 $\pm$ 3.1
Single Story Type [5]	83.8 $\pm$ 2.3	59.2 $\pm$ 5.0	46.5 $\pm$ 9.1	57.7 $\pm$ 7.0	64.5 $\pm$ 5.9	37.5 $\pm$ 6.2	55.0 $\pm$ 5.2	57.8 $\pm$ 3.0
IRR	96.7 $\pm$ 1.5	88.1 $\pm$ 2.7	72.7 $\pm$ 10.0	82.2 $\pm$ 4.6	68.4 $\pm$ 7.1	61.7 $\pm$ 6.7	83.2 $\pm$ 4.2	79.0 $\pm$ 2.7

Table 8: Performance of GPT-4o-mini with few-shot prompting on 150 out-of-sample ALPS narratives.

Configuration	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWK
Random [5]	76.8 $\pm$ 3.8	48.0 $\pm$ 6.2	45.8 $\pm$ 9.1	61.2 $\pm$ 6.3	65.3 $\pm$ 6.3	21.2 $\pm$ 6.3	52.9 $\pm$ 6.6	53.0 $\pm$ 2.9
Quantile [5]	75.1 $\pm$ 3.4	48.1 $\pm$ 5.6	47.8 $\pm$ 8.5	54.5 $\pm$ 6.3	53.4 $\pm$ 6.2	25.3 $\pm$ 6.7	48.2 $\pm$ 6.5	50.3 $\pm$ 2.9
Single Story Type [5]	75.0 $\pm$ 3.3	49.7 $\pm$ 5.6	51.5 $\pm$ 8.1	62.5 $\pm$ 6.0	57.7 $\pm$ 7.2	26.4 $\pm$ 6.2	40.4 $\pm$ 6.1	51.9 $\pm$ 2.8
IRR	95.7 $\pm$ 1.5	85.8 $\pm$ 3.7	82.2 $\pm$ 5.0	56.7 $\pm$ 7.8	52.2 $\pm$ 7.1	52.1 $\pm$ 7.2	70.8 $\pm$ 6.0	70.8 $\pm$ 2.3

Table 9: Performance of few-shot prompting for GPT-4o model based on different number of in-context samples chosen randomly. The results are reported on 150 in-sample narratives from the ALPS dataset.

# of In-Context Samples	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWK
1	79.3 $\pm$ 2.9	70.0 $\pm$ 4.9	54.1 $\pm$ 9.5	77.5 $\pm$ 4.7	66.4 $\pm$ 5.8	25.1 $\pm$ 5.5	67.7 $\pm$ 4.8	62.9 $\pm$ 2.7
5	84.7 $\pm$ 2.8	66.5 $\pm$ 4.4	62.9 $\pm$ 10.6	70.0 $\pm$ 6.4	67.9 $\pm$ 7.0	42.5 $\pm$ 6.4	75.4 $\pm$ 4.2	67.1 $\pm$ 3.1
10	82.4 $\pm$ 2.9	62.9 $\pm$ 4.8	67.0 $\pm$ 10.5	62.0 $\pm$ 7.4	70.0 $\pm$ 6.9	35.1 $\pm$ 6.4	72.5 $\pm$ 4.9	64.6 $\pm$ 3.0
20	88.5 $\pm$ 2.6	69.2 $\pm$ 4.2	53.4 $\pm$ 12.3	61.8 $\pm$ 8.1	64.6 $\pm$ 7.2	41.2 $\pm$ 6.4	84.2 $\pm$ 3.3	66.1 $\pm$ 3.5

Table 10: Performance of few-shot prompting for GPT-4o-mini model based on different number of in-context samples chosen randomly. The results are reported on 150 in-sample narratives from the ALPS dataset.

# of In-Context Samples	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWK
1	72.3 $\pm$ 3.0	43.2 $\pm$ 5.8	47.4 $\pm$ 8.1	54.9 $\pm$ 7.2	62.5 $\pm$ 6.2	23.7 $\pm$ 5.3	51.4 $\pm$ 5.2	50.8 $\pm$ 3.5
5	84.7 $\pm$ 2.4	56.9 $\pm$ 5.3	47.1 $\pm$ 9.0	59.1 $\pm$ 7.0	59.8 $\pm$ 7.4	38.1 $\pm$ 5.9	46.7 $\pm$ 5.8	56.1 $\pm$ 3.4
10	90.8 $\pm$ 1.7	48.7 $\pm$ 5.0	46.7 $\pm$ 11.4	51.2 $\pm$ 8.1	65.4 $\pm$ 6.3	39.0 $\pm$ 6.4	78.5 $\pm$ 4.1	60.0 $\pm$ 3.2
20	82.7 $\pm$ 2.4	57.3 $\pm$ 5.2	56.1 $\pm$ 10.8	55.0 $\pm$ 8.2	63.7 $\pm$ 7.0	42.3 $\pm$ 6.6	75.5 $\pm$ 4.5	61.8 $\pm$ 3.4

Table 11: Experiments comparing different prompting strategies for few-shot prompting with GPT-4o on 150 ALPS in-sample narratives.

Configuration	Char	Sett	Prob	Att	Cons	End	Emo	Avg QWk
Zero-shot (No FlowChart)	34.1 $\pm$ 6.0	61.6 $\pm$ 4.7	59.5 $\pm$ 9.6	64.5 $\pm$ 6.6	58.1 $\pm$ 7.7	28.6 $\pm$ 6.3	49.0 $\pm$ 5.6	50.8 $\pm$ 3.5
Zero-shot (FlowChart)	86.4 $\pm$ 2.8	58.5 $\pm$ 5.4	39.8 $\pm$ 8.2	55.8 $\pm$ 7.3	50.9 $\pm$ 8.0	25.1 $\pm$ 6.2	78.5 $\pm$ 4.5	56.4 $\pm$ 3.0
In-Context with Scores Only	63.2 $\pm$ 4.4	64.2 $\pm$ 4.1	56.4 $\pm$ 9.9	56.9 $\pm$ 7.3	64.0 $\pm$ 5.8	31.0 $\pm$ 5.9	39.3 $\pm$ 4.6	53.6 $\pm$ 3.4
In-Context with Justification + Scores	84.9 $\pm$ 3.0	70.3 $\pm$ 3.7	61.6 $\pm$ 8.9	63.0 $\pm$ 6.5	67.9 $\pm$ 5.8	45.4 $\pm$ 6.2	66.7 $\pm$ 4.8	65.7 $\pm$ 2.9
In-Context with Scores + Justification	86.3 $\pm$ 2.8	67.8 $\pm$ 3.6	67.0 $\pm$ 10.3	77.5 $\pm$ 5.4	68.6 $\pm$ 7.2	34.2 $\pm$ 6.4	69.5 $\pm$ 4.8	67.3 $\pm$ 2.9

System
#Instruction you are an intelligent assistant and your task is to help a teacher to assess the children narrative story. There are 7 narrative elements required for assessment, each have different score range and definition as detailed below: - Character [0-3]: specify the characters in the story) - Setting [0-3]: explain time, activity, location at the beginning of the story - Problem [0-4]: explain the main character's problem - Plan_Attempt [0-4]: what did the main character do to solve the problem - Consequence [0-4]: what happened after the attempt, or what did the secondary character do - Ending [0-2]: what the main character did at the end of the story - Emotion [0-3]: how did the main character feel #Output Format The output MUST be in JSON format, without any markdown formatting. The score MUST be within the corresponding range for each element. For example: scores= {"Character":3, "Setting":2, "Problem":4, "Plan_Attempt":1, "Consequence":0, "Ending":1, "Emotion":2}

Figure 2: An example of the system prompt for ALPS data without assessment justifications.

System
#Instruction You are an intelligent assistant and your task is to help a teacher to assess the children's narrative story. There are 7 narrative elements required for assessment, each with a different score range and definition as detailed below: - "Character [0-3]": specify the characters in the story 3 if a proper noun for the main character and a general noun for the secondary character is stated. 2 if a proper noun for the main character is stated in the story and no mention of a secondary character. 1 if only a general noun is used for the main character. 0 if no character is stated, or only pronouns are used for the main character. - "Setting [0-3]": explain time, activity, and location at the beginning of the story 3 if all of the time, activity, and location are stated in the story. 2 if only two of them are stated. 1 if only one of them is stated. 0 if the story does not have setting information. - "Problem [0-4]": explain the main character's problem 4 if the story has two complete and clear problems. 3 if the story has one complete and clear problem. 2 if the story has an incomplete or unclear problem. 1 if the problem is implied from the story. 0 if there is no problem stated in the story. - "Plan_Attempt [0-4]": what did the main character do to solve the problem 4 if the story has two complete and clear attempts to solve the problem by the main character or storyteller. 3 if the story has one complete and clear attempt by a secondary character. 2 if the story has an incomplete or unclear attempt by any character. 1 if the plan or action is not related to the problem. 0 if there is no attempt to solve the problem. - "Consequence [0-4]": what happened after the attempt, or what did the secondary character do? 4 if the story has two complete and clear consequences related to the problem or attempt. 3 if the story has one complete and clear consequence related to the problem or attempt. 2 if the story has an incomplete or unclear consequence related to the problem or attempt. 1 if the consequence-like information is not related to the problem or attempt. 0 if there is no consequence in the story. - "Ending [0-2]": what the main character did at the end of the story 2 if there is a complete and clear ending. 1 if the ending is incomplete and unclear. 0 if there is no ending in the story. - "Emotion [0-3]": how did the main character feel at the end of the story 3 if two or more emotions are explicitly mentioned in the story. 2 if one emotion is explicitly mentioned in the story. 1 if a general emotion or emotional behavior (e.g., crying) is stated. 0 if no emotion is stated. #Output Format The output MUST be in JSON format, without any markdown formatting. You need to provide a short justification for the provided score for all elements. The scores MUST be within the corresponding range for each element. For example: <pre>scores= {"Character": {"justification": "short reason", "score": 3} , "Setting": {"justification": "short reason", "score": 2}, ...}</pre>

Figure 3: An example of system prompt with detailed scoring flowchart for zero-shot scenario on ALPS data.

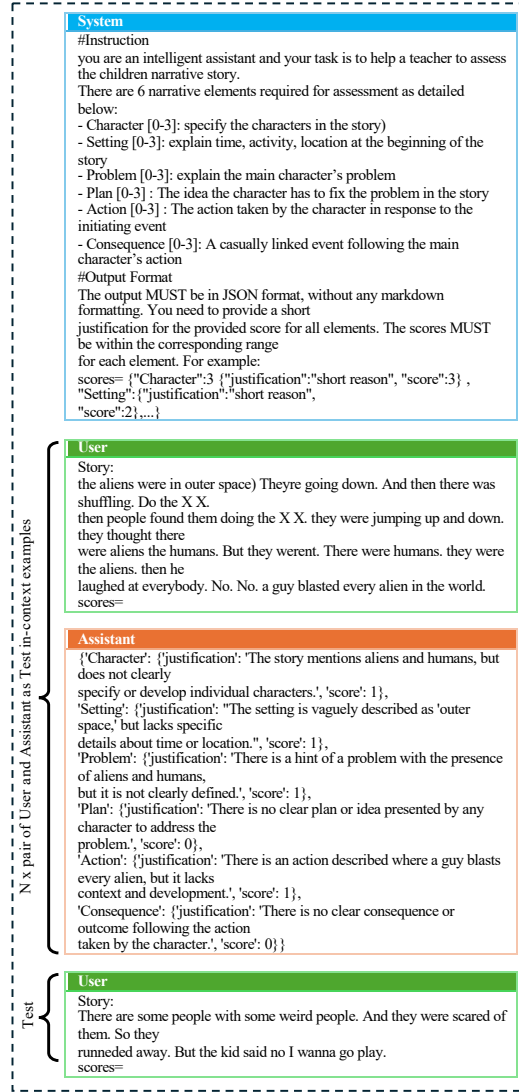


Figure 4: An example of few-shot prompting prompt for Alien story dataset. Only one exemplar is shown.

Table 12: Model training and testing cost.

model type	model	train cost (USD)	test cost (USD)
other models (fine-tuning)			
	BERT-large	0.0783	0.0027
	DeBERTa-v3-large	0.1462	0.0031
	RoBERTa-large	0.0849	0.0029
	ModernBERT-large	0.1374	0.0033
	FLAN-T5-large	0.0587	0.0096
	Mistral-7B	0.6124	0.1089
	Qwen3.5-4B	0.3912	0.0873
	DeepSeek-R1 (7B)	0.6543	0.1247
	Llama3-Instruct	0.5420	0.0961
GPT-4o			
	GPT-4o (fine-tuning)	8.4755	0.5008
	Zero-shot (no flowchart)	0	1.6065
	Zero-shot (flowchart)	0	2.2813
	FSP [1] (justification + scores)	0	2.0452
	FSP [5] (justification + scores)	0	3.8106
	FSP [10] (justification + scores)	0	6.0117
	FSP [20] (justification + scores)	0	10.4163
	FSP [5] (scores only)	0	1.1681
GPT-4o-mini			
	GPT-4o-mini (fine-tuning)	0.9249	0.0318
	Zero-shot (no flowchart)	0	0.0928
	Zero-shot (flowchart)	0	0.1306
	FSP [1] (justification + scores)	0	0.1286
	FSP [5] (justification + scores)	0	0.2293
	FSP [10] (justification + scores)	0	0.3660
	FSP [20] (justification + scores)	0	0.6279
	FSP [5] (scores only)	0	0.0701

Table 13: QWK for 573 in-sample narratives from the ALPS dataset. "***" indicates models fine-tuned on 10% (168) of the training data.

ALPS (in-sample)	Char	Sett	Prob	Att	Con	End	Emo	Avg
IRR	98.1	86.8	75.1	83.5	66.6	61.5	88.4	80.0
Fine-tuning								
CMRF	48.3	41.2	12.0	34.0	32.7	38.1	18.3	32.1
CMRF**	38.1	29.1	0.0	25.5	24.6	30.6	6.6	22.1
BERT-large	92.6	82.4	69.5	70.7	56.3	59.4	83.8	73.5
BERT-large**	72.9	50.7	39.4	41.1	25.0	38.7	32.2	42.9
DeBERTa-v3-large	91.9	86.2	72.5	80.1	71.0	60.1	86.4	78.3
DeBERTa-v3-large**	77.3	69.3	25.3	55.6	43.7	26.0	59.1	50.9
RoBERTa-large	93.0	84.0	65.1	71.0	64.7	59.2	88.4	75.1
RoBERTa-large**	77.0	68.9	51.7	63.5	56.2	32.6	69.7	59.9
ModernBERT-large	92.0	87.4	71.6	79.0	72.9	64.1	88.0	79.3
ModernBERT-large**	82.5	59.7	33.7	51.5	46.8	31.2	46.6	50.3
FLAN-T5-large	93.4	88.9	70.1	77.0	70.2	63.5	86.1	78.5
FLAN-T5-large**	87.9	73.8	56.8	57.6	53.3	51.9	81.7	66.1
Mistral-7B	90.6	83.7	72.6	76.5	71.3	61.6	88.1	77.8
Mistral-7B**	80.9	48.5	55.5	57.4	56.1	38.9	74.4	58.8
Qwen3.5-4B	88.9	74.9	61.1	76.3	63.6	59.1	85.4	72.7
Qwen3.5-4B**	67.1	52.1	65.1	48.3	52.8	36.1	62.8	54.9
DeepSeek-R1 (7B)	87.4	73.9	57.5	60.9	58.2	45.7	83.7	66.8
DeepSeek-R1 (7B)**	66.5	52.0	41.8	48.6	45.9	31.7	38.4	46.4
Llama-3-Instruct (8B)	89.8	81.6	66.9	77.1	67.4	57.4	86.0	75.2
Llama-3-Instruct (8B)**	76.3	69.2	57.6	56.4	52.3	46.8	76.2	62.1
gpt-3.5	90.3	84.0	62.1	68.9	64.9	54.3	85.1	72.8
gpt-3.5**	88.3	74.5	61.0	73.5	61.7	49.8	83.1	70.3
gpt-4o-mini	91.3	86.6	67.5	77.1	67.8	62.8	85.7	77.0
gpt-4o-mini**	89.3	80.7	68.2	79.3	68.3	50.8	86.1	74.7
Few-shot prompting								
gpt-3.5	77.0	53.4	31.5	44.5	38.2	32.1	69.9	49.5
gpt-4o-mini	80.4	60.8	60.0	68.5	68.4	32.8	51.4	60.3
gpt-4o	82.7	67.3	68.1	77.1	68.4	37.0	70.7	67.3

Table 14: QWK for 250 out-of-sample narratives from the ALPS dataset. "****" indicates models fine-tuned on 10% (168) of the training data.

ALPS (out-sample)	Char	Sett	Prob	Att	Con	End	Emo	Avg
IRR	93.6	83.5	77.8	61.7	57.9	53.6	77.9	72.4
Fine-tuning								
CMRF	38.9	18.3	10.7	36.7	32.4	19.1	7.8	23.4
CMRF**	38.6	13.5	3.2	34.9	27.7	24.1	12.3	22.1
BERT-large	79.9	65.6	50.3	47.1	48.3	37.9	69.6	57.0
BERT-large**	37.9	34.2	14.3	38.3	26.0	27.2	4.8	26.1
DeBERTa-v3-large	86.8	71.1	60.0	55.9	66.1	39.4	73.4	64.7
DeBERTa-v3-large**	77.0	57.9	18.7	38.9	31.3	19.6	46.2	41.4
RoBERTa-large	84.2	65.2	61.6	48.6	52.4	11.3	71.8	56.4
RoBERTa-large**	66.9	38.7	11.6	49.0	37.7	29.9	63.4	42.5
ModernBERT-large	81.1	72.0	52.6	56.0	60.8	48.2	74.0	63.5
ModernBERT-large**	75.9	30.8	24.0	42.6	33.3	20.2	31.3	36.9
FLAN-T5-large	79.9	70.1	64.1	62.3	60.2	47.8	72.0	65.2
FLAN-T5-large**	84.2	59.3	24.4	47.4	33.6	36.5	66.0	50.2
Mistral-7B	83.3	65.4	58.4	59.4	57.1	43.2	75.1	63.1
Mistral-7B**	78.0	45.7	32.8	50.8	54.8	21.1	58.2	48.8
Qwen3.5-4B	83.0	55.2	50.2	58.7	57.5	46.8	72.8	60.6
Qwen3.5-4B**	62.0	38.3	41.4	57.6	53.8	40.9	47.5	48.8
DeepSeek-R1 (7B)	79.6	53.3	46.5	58.8	57.8	28.4	64.9	55.6
DeepSeek-R1 (7B)**	60.8	45.1	36.2	42.7	39.8	27.4	33.1	40.7
Llama-3-Instruct (8B)	80.4	66.8	56.8	62.0	52.7	51.5	70.7	63.0
Llama-3-Instruct (8B)**	69.9	36.7	39.7	47.4	45.7	26.4	57.5	46.2
gpt-3.5	78.5	71.7	56.7	60.7	61.9	41.9	64.1	62.2
gpt-3.5**	81.2	61.9	47.6	59.3	54.6	46.7	67.2	59.8
gpt-4o-mini	80.0	71.3	62.4	63.7	63.3	45.5	73.9	65.7
gpt-4o-mini**	83.8	63.3	60.8	60.5	63.5	49.3	73.0	64.9
Few-shot prompting								
gpt-3.5	68.7	28.9	39.1	33.8	44.3	26.3	60.6	43.0
gpt-4o-mini	69.9	44.4	47.3	63.1	61.8	28.4	48.6	51.9
gpt-4o	77.5	51.4	54.5	61.1	60.4	27.9	53.7	55.2

Table 15: QWK accuracies of models on 50 Alien test stories. Fine-tuned models use 331 training instances; FSP uses 5 exemplars.

	Char	Sett	Prob	Plan	Act	Con	Avg QWK
<i>IRR</i>	95.6	60.1	54.7	40.0	41.7	41.0	55.5
<i>fine-tuning</i>							
<i>CMRF</i>	33.0	18.8	34.3	14.9	34.0	26.7	27.0
<i>BERT-large</i>	93.8	26.8	39.6	39.2	59.6	31.7	48.5
<i>DeBERTa-v3-large</i>	97.8	34.0	37.8	46.2	53.7	47.9	52.9
<i>RoBERTa-large</i>	96.7	40.6	48.9	47.4	35.3	47.0	52.6
<i>ModernBERT-large</i>	95.5	14.3	37.7	28.0	49.0	42.0	44.4
<i>FLAN-T5-large</i>	95.5	26.8	44.2	12.3	45.8	42.3	44.5
<i>Mistral-7B</i>	76.9	15.2	46.3	29.4	54.2	60.0	47.0
<i>Qwen3.5-4B</i>	75.8	29.0	43.3	38.7	51.3	47.0	47.5
<i>DeepSeek-R1 (7B)</i>	72.4	34.8	46.1	-8.0	68.0	44.7	45.1
<i>Llama3-Instruct (8B)</i>	67.0	39.4	26.6	46.1	42.6	51.8	45.6
<i>gpt-3.5</i>	93.9	39.8	45.1	61.6	50.8	53.1	57.4
<i>gpt-4o-mini</i>	97.7	41.3	50.0	47.0	43.9	49.3	54.9
<i>few-shot prompting</i>							
<i>gpt-3.5</i>	36.0	12.1	35.8	30.8	44.3	35.9	32.5
<i>gpt-4o-mini</i>	59.8	20.1	63.8	46.4	47.7	53.3	48.5
<i>gpt-4o</i>	59.0	29.3	56.5	39.6	57.7	57.8	50.0