# Leveraging Large Language Models for Automated Assessment of Elementary Students' Block-Based Narrative Programs

Anisha Gupta
Robert Monahan
North Carolina State University
Raleigh, North Carolina, USA
agupta44@ncsu.edu
rpmonaha@ncsu.edu

Jessica Vandenberg
Andy Smith
North Carolina State University
Raleigh, North Carolina, USA
jvanden2@ncsu.edu
pmsmith4@ncsu.edu

Rasha Elsayed
Kimkinyona Fox
WestEd
San Francisco, California, USA
relsaye@wested.org
kfox@wested.org

James Minogue
Kevin Oliver
North Carolina State University
Raleigh, North Carolina, USA
jminogu@ncsu.edu
kmoliver@ncsu.edu

Aleata Hubbard Cheuoua
Cathy Ringstaff
WestEd
San Francisco, California, USA
ahubbar@wested.org
cringst@wested.org

Bradford Mott
North Carolina State University
Raleigh, North Carolina, USA
bwmott@ncsu.edu

## Abstract

Recent years have seen increasing awareness of the need to engage young learners in computational thinking (CT). Integrating digital storytelling, where students create short narratives, and CT offers significant potential for promoting interdisciplinary learning for students; however, it is critical to provide both teachers and students with automated support. A promising approach for enabling support is to leverage advances in Large Language Models (LLMs), which have demonstrated considerable potential for assessing both programming and natural language artifacts. In this work, we investigate the capabilities of LLMs to automatically assess student-created block-based programs developed using a narrative-centered learning environment that engages upper elementary students (ages 9 to 11) in learning CT and physical science through the creation of interactive science narratives. Using the narrative programs created by 28 students, we explore the efficacy of LLMs to assess the programs across two dimensions.

## CCS Concepts

• **Social and professional topics** → **K-12 education**; • **Computing methodologies** → **Natural language processing**.

## Keywords

K-12 education, Natural language processing

**ACM Reference Format:**
Anisha Gupta, Robert Monahan, Jessica Vandenberg, Andy Smith, Rasha Elsayed, Kimkinyona Fox, James Minogue, Kevin Oliver, Aleata Hubbard Cheuoua, Cathy Ringstaff, and Bradford Mott. 2024. Leveraging Large Language Models for Automated Assessment of Elementary Students' Block-Based Narrative Programs. In *Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 2 (SIGCSE Virtual 2024), December 5–8, 2024, Virtual Event, NC, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3649409.3691089


*SIGCSE Virtual 2024, December 5–8, 2024, Virtual Event, NC, USA*

ACM ISBN 979-8-4007-0604-2/24/12
https://doi.org/10.1145/3649409.3691089

## 1 Introduction

InfuseCS is a narrative-centered learning environment designed to support interdisciplinary learning for students, while promoting mastery in digital storytelling, science, and computational thinking [3]. InfuseCS features an overarching backstory about a group of scientists stranded on a remote island. Students explore the island while learning physical science concepts, then as a culminating activity create an interactive narrative using the characters and science concepts they have encountered. Building on lessons from similar systems targeting this age-range [4], the narrative is created using a custom block-based programming interface embedded in the learning environment (Figure 1, left).

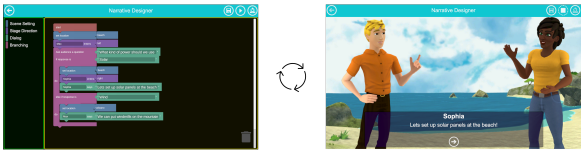


**Figure 1: InfuseCS learning environment.**

The story creation task produces block-based programs containing large amounts of natural language text, which has not been a part of traditional automated programming assessment. To leverage large language models (LLMs) abilities to understand natural language data, students' programs were converted to a textual format and evaluated using Meta's Llama 3. Prior work has shown that LLMs can be effectively used for automated grading, including essay assessments, [2]. The evaluations in this work align with our prior efforts to automatically assess student narrative programs [3] using traditional NLP techniques, and assess the stories across two dimensions: *Story Structure*, and *Science Concepts*.

Students were asked to write a choose-your-own story based on science concepts, using a block-based programming interface. They modified the following starter story: "You picked the first response You picked the second response". Your task is to rate students' stories on a scale of 0 to 2, based on their [science concepts/story quality]. Here is the rubric:

- **0:** [Rubric for score 0]
- **1:** [Rubric for score 1]
- **2:** [Rubric for score 2]

Review the following examples for context:
Example - [Sample story 1] Science concepts rating: 0
Example - [Sample story 2] Science concepts rating: 1
Example - [Sample story 3] Science concepts rating: 2

**Now rate this specific story:**
Story to be rated - "You picked the first response You picked the second response"
**Rating Task:** Directly provide a rating for the above story only, using the exact following format:
"[Science concepts/Story quality] rating: [0/1/2]; explanation: [limit explanation to 20 words]".
**Important: Include only the rating and explanation in your response. Do not add a preface, postscript, or any additional comments.**

**Figure 2: LLM prompt template using few-shot learning.**

## 2 NLP Analysis and Results

With the aim of providing real-time support to students by providing personalize feedback on their stories, it is necessary to create automated techniques capable of evaluating the "quality" of their narrative programs. Manual evaluation, capable of providing more nuanced analysis of the story plot and structure, is time and effort intensive, and not currently suitable for real-time feedback. Our previous work introduced AI-driven automated assessment techniques to evaluate students' stories on overall story quality and the inclusion of science concepts [3]. For each of these evaluation dimensions, human annotators labeled each story on a scale of 0 to 2. A story was rated 0 on overall story rating if no modifications were made to the starter story, or the additions were nonsensical (e.g., 'fjioau'), 1 if the starter story was modified but the story is incomplete, and 2 if the story is complete and logical. The story was rated 0 on inclusion of science concepts if no science concepts were included, 1 if the science concepts were included but used inaccurately, and 2 if science concepts are included in the narrative and used accurately.

Our investigation utilizes pilot data from 26 upper elementary students in California and North Carolina following an IRB-approved protocol. Students spent approximately 4 hours with the learning environment over multiple days, with the last half of the interaction focused on creation and refinement of their narrative programs. Our prior work used various representations of students' stories, including BoW and pre-trained word embeddings (ELMo, BERT). The best results for evaluating overall story quality were achieved using BERT embedding representation of the stories, with 78.57% accuracy and 0.65 Cohen's $\kappa$, whereas science was best evaluated using ELMo embeddings, achieving an accuracy of 57.14% and Cohen's $\kappa$ of 0.34. Given we only had 26 unique stories in our dataset, the scoring models were evaluated using leave-one-out cross validation. While the results were promising, these models rely on human-coded labels for training. Moreover, the performance of these models is strictly bound by the size of the training dataset. Given the advent of LLMs and their promise in zero shot and few shot learning, it could enable automatic evaluation of students' narratives with reduced human coding.

To evaluate the effectiveness of using LLMs for rating students' narratives, we use Meta's Llama 3 [1](pre-trained version with 8B parameters) as our generative model, since at the time of our evaluation it is the most capable openly available LLM. The final prompt used is shown above in 2, and consists of a description of the learning platform and the evaluation rubric. We then evaluate model performance in both zero shot as well as few shot settings. For few shot learning, we randomly sample one instance of each label category from the remaining dataset (excluding the current sample) and include these in the prompt template (Figure 2). Table 1 presents the results for both zero shot and few shot learning, compared against results from our prior evaluations with BoW and word embeddings with classification accuracy and Cohen's $\kappa$ as our evaluation metrics. We observe that zero shot learning using LLM performed worse than traditional machine learning classifiers using word embeddings for both science rating (39.29% accuracy, 0.12 Cohen's $\kappa$) as well as overall story quality (46.43% accuracy, 0.23 Cohen's kappa) assessment. However, in a few shot learning setting, the LLM outperformed word embedding-based models for science rating with 64.29% accuracy and 0.47 Cohen's $\kappa$, and achieved comparable performance for overall story rating, with 75% accuracy and 0.61 Cohen's $\kappa$.

**Table 1: Performance of Different Models on Story Quality and Science Concepts.**

| Model | Story Quality | | Science Concepts | |
|---|---|---|---|---|
| | Accuracy | $\kappa$ | Accuracy | $\kappa$ |
| SVM with ELMo | 64.00% | 0.39 | 57.14% | 0.34 |
| SVM with BERT | 78.57% | 0.65 | 46.43% | 0.18 |
| Llama3 (zero shot) | 46.43% | 0.23 | 39.29% | 0.12 |
| Llama3 (few shot) | 75.00% | 0.61 | 64.29% | 0.47 |

The results indicate that LLMs are more effective for evaluating students' inclusion of Science Concepts, while performing roughly the same on evaluating Story Quality compared to our previously used NLP techniques. Additionally, providing the pre-trained LLMs with a few examples across all possible label categories greatly increases model performance as compared to prompts that do not.

## Acknowledgments

This research was supported by the National Science Foundation (NSF) through grants DRL-1921495 and DRL-1921503. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

## References

[1] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Cyril Chhun, Fabian M Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics* 12 (2024), 1122–1142.

[3] Anisha Gupta, Andy Smith, Jessica Vandenberg, Rasha ElSayed, Kimkinyona Fox, James Minogue, Aleata Hubbard Cheuoua, Kevin Oliver, Cathy Ringstaff, and Bradford Mott. 2023. Fostering Interdisciplinary Learning for Elementary Students Through Developing Interactive Digital Stories. In *International Conference on Interactive Digital Storytelling*. Springer, 50–67.

[4] Caitlin Kelleher, Randy Pausch, and Sara Kiesler. 2007. Storytelling alice motivates middle school girls to learn computer programming. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1455–1464.