When Humans Can't Agree, Neither Can Machines: The Promise and Pitfalls of LLMs for Formative Literacy Assessment

Owen Henkel
University of Oxford
owen.henkel@education.ox.ac.uk

Kirk Vanacore
Cornell University
kpv27@cornell.edu

Bill Roberts
Legible Labs
bill@legiblelabs.com

Abstract

Story retell assessments provide valuable insights into reading comprehension but face implementation barriers due to time-intensive administration and scoring. This study examines whether Large Language Models (LLMs) can reliably replicate human judgment in grading story retells. Using a novel dataset we conduct three complementary studies examining LLM performance across different rubric systems, agreement patterns, and reasoning alignment. We find that LLMs (a) achieve near-human reliability with appropriate rubric design, (b) perform well on easy-to-grade cases but poorly on ambiguous ones, (c) produce explanations for their grades that are plausible for straightforward cases but unreliable for complex ones, and (d) different LLMs display consistent "grading personalities" (systematically scoring harder or easier across all student responses). These findings support hybrid assessment architectures where AI handles routine scoring, enabling more frequent formative assessment while directing teacher expertise toward students requiring nuanced support.

1 Introduction

Story retell tasks offer unique advantages for assessing reading comprehension, requiring students to actively reconstruct understanding rather than simply recognize correct answers. Despite their pedagogical value, implementation faces significant barriers: administering, transcribing, and scoring individual responses is time-intensive, particularly for teachers managing large classes in resource-constrained environments.

Recent advances in Large Language Models (LLMs) present opportunities to address these barriers while maintaining assessment quality. This study examines whether LLMs can reliably replicate human judgment in grading story retells and, critically, whether they can identify cases requiring human attention. Such capabilities could enable hybrid assessment systems that automate routine scoring while preserving teacher expertise for complex decisions.

We address three interconnected questions: (1) To what extent can LLMs replicate human judgments about story retell quality across different rubric systems? (2) What patterns emerge in model-human agreement, particularly for cases humans find ambiguous? (3) How well do LLM explanations correspond with human reasoning?

Using 95 student story retells from Ghana, we examine these questions through three complementary studies. Study 1 establishes baseline performance across rubric types. Study 2 investigates agreement patterns and model "grading personalities." Study 3 explores the relationship between explanation quality and scoring accuracy.

Our findings have direct implications for educators considering AI-supported assessment tools, providing evidence for how such technologies might enhance rather than replace human judgment in literacy assessment, particularly in contexts where frequent formative assessment is essential but difficult to implement.

2 Prior Work

2.1 Story Retell as Reading Comprehension Assessment

Story retelling provides a unique window into reading comprehension by requiring students to actively reconstruct narratives rather than simply recognize correct answers. The cognitive demands of retelling—drawing upon memory for factual details, generating inferences to fill gaps, and reconstructing events in sequence—mirror authentic comprehension processes (Reed and Vaughn, 2012; Wilson et al., 1985). This active recall requirement distinguishes retelling from recognition-based assessments, potentially providing richer insights into student understanding.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 69–78

Research on retell effectiveness has yielded mixed but generally positive findings. Reed and Vaughn (2012)'s review of 54 studies found moderate correlations between retell scores and standardized comprehension measures across grade levels. However, some studies report inconsistent relationships with reading abilities (Hagtvet, 2003; Marcotte and Hintze, 2009), suggesting retelling may capture distinct aspects of comprehension not fully reflected in traditional assessments.

2.2 Scoring Approaches and Challenges

Multiple scoring methods exist, each with inherent trade-offs. Idea-unit analysis divides passages into weighted narrative elements, enabling granular assessment but requiring text-specific development that limits cross-story comparison (Maria, 1990). Component-based scoring evaluates narrative elements (characters, setting, plot) more generalizably but faces reliability challenges, with researchers observing inconsistent inter-rater agreement (Klesius and Homan, 1985).

Holistic scoring assigns overall quality ratings, balancing efficiency with detail but introducing subjectivity that can compromise reliability without careful rubric design and rater calibration. Word-count measures offer objectivity and ease of automation but may reward verbosity over comprehension quality, as critics note students could manipulate metrics without demonstrating understanding (Altwerger et al., 2007; Goodman, 2006).

2.3 Formative Assessment in Literacy Contexts

Black and Wiliam (1998)'s seminal meta-analysis established formative assessment as one of education's most powerful interventions, demonstrating effect sizes between 0.4 and 0.7 standard deviations. Building on Sadler (1989)'s framework—understanding quality standards, comparing work against standards, and possessing gap-closing strategies—formative assessment becomes "formative" when evidence actively adapts instruction to meet student needs.

In literacy contexts, formative assessment plays a critical role in comprehension development. The comprehensive nature of reading assessment demands substantial time investment that conflicts with classroom constraints, particularly given increasing student-teacher ratios.

Story retelling emerges as a particularly powerful formative tool, demonstrating moderate correlations with other comprehension measures while showing stronger relations to authentic literacy tasks than traditional assessments. The interactive nature provides diagnostic capabilities revealing thinking strategies inaccessible through traditional measures.

2.4 Large Language Models and Educational Assessment

Recent advances in Large Language Models present distinctive capabilities for assessment support. Unlike rigid scoring systems, LLMs demonstrate capacity to generalize to new tasks with minimal examples, completing assessments through prompt modification rather than retraining (Ouyang et al., 2022). However, Schneider et al. (2023) caution that readiness for independent grading remains uncertain given the complexity of human narrative interpretation. This study examines whether these capabilities can be systematically applied to story retell assessment within educational contexts.

3 Dataset and Methods

3.1 Dataset Description

The ROARS dataset comprises responses from 130 Ghanaian adolescent students who read one of two 400-word fictional stories and completed comprehension tasks including story retell. Of these, 95 students completed the retell task, with remaining responses left blank. All retells were transcribed verbatim and word counts recorded. This dataset provides a diverse context for examining AI-assisted assessment capabilities in a Global South educational setting.

3.2 Human Rating Process and Rubric Development

Three human raters evaluated the 95 story retells using three distinct rubric systems adapted from literature. All raters held Master of Education degrees and had classroom teaching experience, providing professional expertise in literacy assessment. Ground truth scores were determined by averaging ratings and rounding to the nearest whole number. This averaging process itself highlights inherent assessment ambiguity—unanimous agreement occurred in only 66% of cases.

The adapted rubrics are presented in Appendix **??**.

Examples of different rater's scoring by rubric			
Retell 1			
lucy was a girl who like learning around she round through the country stole that night allways when everyone including the sheeps and lambs were as sleep. lucy helped to save the shepherd when the shepherd got a broke in his leg.			
	2-class	3-class	5-class
Rater 1	0	1	1
Rater 2	0	1	2
Rater 3	0	1	2
Ground Truth	0	1	2
Retell 2			
Lucy was different from all the other sheep right from the start. One day something terrible happened. The shepherd fell over and broke his leg...			
	2-class	3-class	5-class
Rater 1	1	2	4
Rater 2	1	2	3
Rater 3	1	1	4
Ground Truth	1	2	4

Table 1: Examples of different rater's scoring by rubric

3.3 LLM Assessment Methodology

For the automated grading component, we used GPT-4 (GPT-4o-2024-05-13) with carefully designed prompts that replicated the human rating context. The prompts instructed the model to act as a literacy teacher evaluating reading comprehension through story retell assessment. Each prompt included: the complete rubric with detailed scoring criteria, the original story text for context, instructions to provide only the numeric score output, and role-based framing to establish appropriate assessment perspective.

4 Study 1: LLM Replication of Human Judgments

4.1 Inter-Rater Agreement Analysis

Before examining LLM performance, we first established baseline human inter-rater agreement to understand the inherent reliability of the assessment task. Analysis of average ratings revealed systematic differences between raters. Rater 3 consistently awarded higher scores than Raters 1 and 2 across rubrics, suggesting more lenient grading standards. For the two-class rubric, average scores were 0.17, 0.21, and 0.37 respectively (scale 0-1). Similar patterns emerged for three-class (0.37, 0.36, 0.59; scale 0-2) and five-class rubrics (1.41, 1.14, 1.22; scale 0-4).

Inter-rater reliability varied substantially across rubric types. For the binary rubric, Fleiss' kappa was 0.56, indicating moderate agreement. Agreement improved markedly for the three-class rubric (Kendall's W = 0.81) and further for the five-class rubric (Kendall's W = 0.85). As validation, pairwise Cohen's kappa averages showed the same progression: 0.60 (two-class), 0.74 (three-class), and 0.78 (five-class).

This pattern suggests that more granular rubrics enable higher inter-rater reliability, possibly because they provide clearer distinctions between performance levels. The binary forced choice between "bad" and "good" may inadequately capture the complexity of student responses, leading to inconsistent judgments when responses fall near the decision boundary.

Prediction	Prec.	Rec.	F1	Supp.
Bad Retell (0)	0.86	1.00	0.93	74
Good Retell (1)	1.00	0.25	0.40	16
Average	0.89	0.87	0.83	90

Table 2: Two-class rubric performance (LWK/QWK = 0.35)

4.2 LLM Performance

We evaluated GPT-4's ability to score student retells using the same rubrics as human raters. The model received simple prompts explaining the task, relevant rubric, original story, and student response, then provided numeric scores. This straightforward approach established baseline capabilities without sophisticated prompt engineering.

4.2.1 Two-Class Rubric Results

The model's performance on the binary rubric revealed significant challenges, as shown in Table 2.

The model showed bias toward the "Bad Retell" category, correctly identifying all poor responses but capturing only 25% of good retellings. This conservative grading produced perfect precision for "Good Retell" (no false positives) but poor recall. The LWK of 0.35 indicates low agreement with human consensus, substantially below the human inter-rater agreement of 0.56.

4.2.2 Three-Class Rubric Results

Performance improved dramatically with the three-class rubric, as shown in Table 3.

The model excelled at identifying bad retellings (94% recall, 95% precision) and showed improved recognition of mediocre responses (81% recall). However, it remained conservative with "Good

Prediction	Prec.	Rec.	F1	Supp.
Bad Retell (0)	0.95	0.94	0.94	64
Mediocre (1)	0.57	0.81	0.67	16
Good Retell (2)	1.00	0.25	0.40	8
Average	0.89	0.85	0.84	88

Table 3: Three-class rubric performance (QWK = 0.78)

Prediction	Prec.	Rec.	F1	Supp.
Bad (0)	0.68	0.91	0.78	23
Poor (1)	0.77	0.68	0.72	34
Mediocre (2)	0.65	0.58	0.61	19
Acceptable (3)	0.50	0.38	0.43	7
Good (4)	1.00	1.00	1.00	1
Average	0.69	0.69	0.68	84

Table 4: Five-class rubric performance (QWK = 0.82)

Retell" classifications, capturing only 25% despite perfect precision. The QWK of 0.78 approaches the human inter-rater agreement of 0.81, suggesting near-human reliability.

4.2.3 Five-Class Rubric Results

The five-class rubric yielded the highest agreement levels, as shown in Table 4.

While individual category performance varied, the QWK of 0.82 nearly matches human agreement (0.85). The model showed strongest performance at the extremes—identifying clearly bad (91% recall) and the single good retelling (100% recall)—with more uncertainty in middle categories. This pattern mirrors human rating behavior, where edge cases between adjacent categories prove most challenging.

4.3 Comparative Analysis

The progression of model-human agreement across rubrics closely parallels the pattern in human inter-rater reliability:

Rubric	Model-Human	Human-Human
2-Class	0.35	0.56
3-Class	0.78	0.81
5-Class	0.82	0.85

Table 5: Agreement comparison across rubric types

This parallel suggests that model performance is fundamentally constrained by the same factors affecting human reliability. The poor performance on binary classification appears to stem from the rubric's inadequacy rather than model limitations. When provided with sufficiently granular evaluation criteria, LLMs approach human-level reliability.

4.4 Implications

These findings demonstrate that LLMs can achieve near-human reliability in story retell assessment when provided with appropriate rubric structures. The critical factor appears to be rubric design rather than model capability. Binary classifications prove problematic for both humans and machines, while detailed rubrics enable consistent evaluation.

The model's conservative grading tendency—high precision but lower recall for positive categories—may actually benefit educational applications. False positives (incorrectly identifying poor comprehension as good) pose greater instructional risks than false negatives, as they could lead teachers to overlook students needing support. The model's bias toward identifying weaknesses aligns with formative assessment goals of catching students who need help.

The strong performance on five-class rubrics (QWK = 0.82) suggests AI assessment has reached practical viability for supporting classroom instruction. However, this performance depends critically on well-designed evaluation criteria that provide sufficient granularity to capture meaningful distinctions in student performance.

5 Study 2: Model-Human Agreement Patterns

5.1 Research Questions and Approach

Building on Study 1's finding that models approach human reliability with detailed rubrics, Study 2 investigates deeper patterns in model-human agreement. Specifically, we examine how rating consistency relates to assessment uncertainty and whether models exhibit systematic grading tendencies similar to human raters.

We expanded our analysis to include Claude Sonnet 4 and Gemini 2.0 Flash alongside GPT-4, testing each at three temperature settings (0, 0.5, 1.0) to examine consistency. This provided nine model configurations plus three human raters for comparison. Given Study 1's poor results with binary classification, we focused exclusively on the three-class rubric.

We analyze these data in two ways: first by examining the differences in Cohen's Kappa when the human raters agreed and disagreed, and whether

or not the human raters score the comprehension activity as 1. Next we evaluate the consistency of raters directional bias (i.e., whether they gave higher or lower grades than average) by estimating a multilevel regression model with random intercepts of the students who was being graded and raters (humans and models). Then we compared these random intercepts to see whether their were patterns of rater bias within and between models.

Figure 1 presents Cohen's kappa values for pairwise comparisons between human raters and AI models. Across all ratings (human and model), agreement was moderate to high, with kappas ranging from .50 to 1.0. The highest agreement occurred within models, indicating that temperature differences between 0 and 1 do not introduce meaningful variability in coding.

5.2 Key Findings

Cohen's kappa values for pairwise comparisons between human raters and AI models revealed moderate to high agreement, with kappas ranging from .50 to 1.0. The highest agreement occurred within models, indicating that temperature differences between 0 and 1 do not introduce meaningful variability in coding. By contrast, the lowest agreements were observed between human coders and the models.

When all human coders agreed, the kappa values between models and human coders increased to between .66 and .80. Alternatively, when at least one coder disagreed, the models showed low inter-rater reliability with the human raters, with kappas ranging from 0.15 to 0.45. The same pattern emerged when comparing cases in which no human coder gave a score of 1 (i.e., raters were confident the student either did or did not comprehend the text) versus cases in which at least one human rater assigned a score of 1 (i.e., at least one human was uncertain about whether the student comprehended the text). These findings suggest that the models align more closely with human judgments when humans themselves are more certain of the outcome.

It is noteworthy that two of the human raters (Rater 1 and Rater 2) tended to agree even in uncertain cases. Their kappas were 0.65 when at least one human rater disagreed and 0.70 when at least one human rater gave a score of 1. However, the models still tended to diverge more from these raters under uncertain circumstances. This suggests that even when humans reach consensus in difficult cases, the models may continue to struggle to

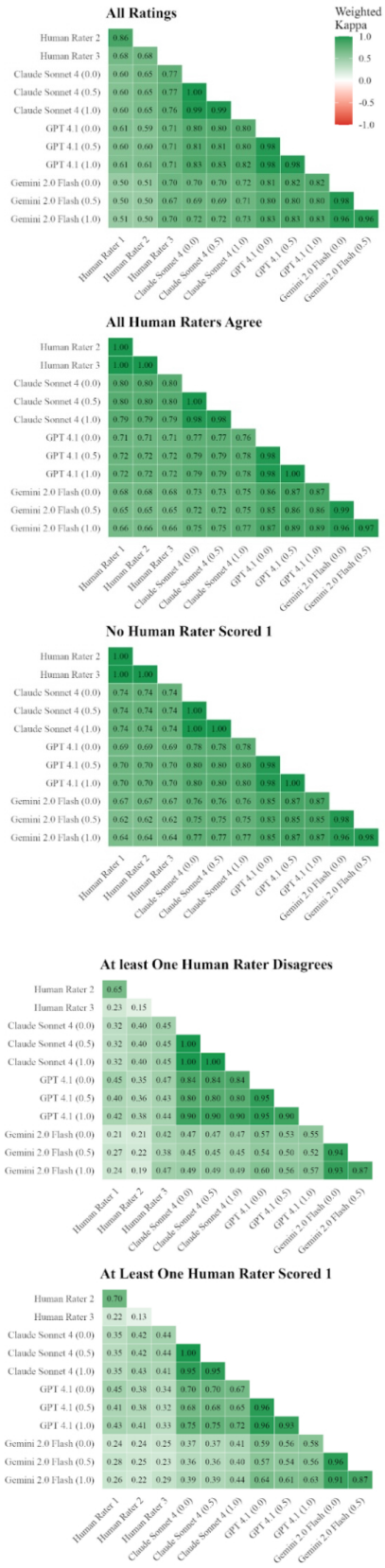

Figure 1: Pairwise weighted kappa heatmaps comparing agreement among human raters and AI models across multiple rating conditions.

align with them, potentially because the models are sensitive to the uncertainty reflected in these cases.

Furthermore, even when the humans were uncertain, the models maintained high internal consistency, as indicated by strong within-model reliability. For example, when at least one human rater disagreed or assigned a retell a score of 1, the Claude Sonnet 4 models consistently exhibited very high agreement within model ($\kappa = .95$–1.0).

Figure 2: Random intercept estimates for human and AI raters across groups.

Figure 2 presents the random intercepts from a multilevel regression model predicting ratings by rater. The model included random intercepts for each retell to account for performance-related differences across students. Thus, the random intercepts for each rater can be interpreted as systematic deviations from the overall mean rating (i.e., an intercept of 0 indicates that the rater's judgments consistently align with the grand mean).

Random intercepts from a multilevel regression model predicting ratings by rater revealed systematic grading tendencies. Models exhibited consistent levels and directions of bias relative to the mean. Specifically, Gemini tended to assign higher scores than all other raters, Claude tended to assign lower scores, and GPT produced ratings closest to the mean.

Human raters showed more variation: two raters (Human Rater 1 and Human Rater 2) consistently assigned lower scores, while one rater (Human Rater 3) tended to assign higher scores. This pattern aligns with the inter-rater reliability findings, which showed that Human Raters 1 and 2 had higher reliability with each other compared to Human Rater 3.

5.3 Implications for Hybrid Assessment

These findings suggest that AI models may be particularly effective for evaluating clear-cut cases, where human raters also show high certainty and agreement. In contrast, more difficult-to-evaluate student responses—those requiring additional context, nuanced interpretation, or expert teacher judgment—could be flagged for human review. Leveraging AI confidence scores to identify such cases can support hybrid assessment approaches that combine the efficiency of automated scoring with the depth and expertise of human evaluation. This layered approach ensures that straightforward judgments are handled quickly and consistently, while complex or ambiguous cases receive the careful consideration of trained educators.

6 Study 3: Analysis of Grading Rationales [Exploratory]

6.1 Research Questions

While Studies 1 and 2 established that models can achieve human-level scoring reliability, a critical question remains: Do models reach correct answers through human-like reasoning? Understanding whether models identify the same strengths and weaknesses that teachers recognize has profound implications for using AI-generated feedback in formative assessment. This exploratory study examines whether model explanations align with human reasoning and whether such alignment predicts scoring accuracy.

6.2 Methods

One human rater (former classroom teacher, M.Ed.) provided written justifications for all 94 story retell scores using the three-class rubric. We then modified prompts for Claude Sonnet 4, GPT-4.1, and Gemini 2.0 Flash to request explanations alongside scores.

We assessed explanation similarity based on conceptual alignment rather than exact wording, examining: identification of similar strengths or weaknesses, reference to comparable story elements or gaps, assessment of narrative flow and comprehension quality, overall evaluation tone (weak/medium/strong). For example, human noting "lacks essential narrative elements" and model stating "missing key story components" were considered conceptually similar. This approach fo-

cused on substantive agreement rather than linguistic matching.

6.3 Results

6.3.1 Quantitative Context: When Models Succeed and Struggle

Before examining reasoning quality, we established when models achieve accurate scoring. Analysis revealed no instances of maximal disagreement among humans (0 vs 2 scores), suggesting disagreements occur at category boundaries rather than reflecting fundamental assessment differences.

For the approximately two-thirds of cases (62 out of 94) where all three human raters assigned identical scores, we considered these as potentially "easy to grade" responses—those with clear indicators of quality that multiple raters could consistently identify. For these unanimous cases, the human ground truth score was simply the agreed-upon score. For the remaining one-third of cases (32 out of 94) where human raters showed some level of disagreement, we considered these as potentially ambiguous or difficult-to-assess responses. For these non-unanimous cases, the human ground truth was determined by majority vote.

We then compared these human ground truth scores against the model consensus scores (determined by majority vote across Claude Sonnet 4, GPT-4.1, and Gemini 2.0 Flash) to assess how model performance varies based on the inherent difficulty of the assessment task.

Agreement Level	Cases	Direct	QWK
Unanimous	62 (66%)	82.3%	0.808
Non-Unanimous	32 (34%)	50.0%	0.516

Table 6: Model consensus performance by human agreement

The results reveal a striking pattern: when human raters unanimously agree on a score, the model consensus achieves strong agreement (QWK = 0.808) with the human judgment. However, when human raters disagree—suggesting inherent ambiguity in the student response—model performance drops substantially to moderate agreement (QWK = 0.516), with direct agreement falling to chance levels (50%). This pattern suggests that models excel at identifying clear-cut cases but struggle with the same ambiguous responses that challenge human raters.

6.3.2 Explanation-Score Alignment Analysis

Given that models achieve high accuracy on clear cases but struggle with ambiguous ones, we examined whether the reasoning behind their scores aligns with human thinking. Do models identify the same strengths and weaknesses that teachers recognize, even when they arrive at the correct score?

Table 7 presents the relationship between explanation similarity and scoring accuracy across all three models:

Model	Similar		Different	
	Match	No Match	Match	No Match
Claude	87.0%	13.0%	66.7%	33.3%
GPT-4.1	81.3%	18.7%	65.2%	34.8%
Gemini	69.5%	30.5%	77.1%	22.9%

Table 7: Relationship Between Explanation Similarity and Score Agreement

Claude and GPT-4 demonstrate strong alignment: when their explanations resemble human reasoning, scores match 81-87% of the time. This high precision suggests that explanation similarity could serve as a confidence indicator for automated scoring. However, it's important to note that similar explanations occurred in only about 50% of cases for Claude and GPT-4, while Gemini achieved 62.8% explanation similarity.

The moderate occurrence of similar explanations (50-63% across models) reveals an important insight: many accurate scores emerge through different reasoning paths than humans employ. This suggests that models may identify alternative but potentially valid indicators of comprehension quality that differ from traditional human assessment approaches.

Gemini presents an interesting anomaly—achieving the highest rate of similar explanations but showing the weakest correlation between explanation similarity and score agreement (69.5%). This pattern suggests that surface-level explanation similarity may not always indicate deep alignment in assessment reasoning, and that the quality of reasoning alignment may be more important than the quantity.

6.3.3 Explanation Similarity as a Trust Signal

To evaluate whether explanation similarity could serve as a practical indicator of scoring reliability in operational systems, we calculated performance

metrics treating explanation similarity as a predictor of score agreement:

Model	Precision	Recall	F1	FPR
Claude Sonnet 4	87.0%	55.6%	0.68	13.0%
GPT-4.1	81.3%	56.5%	0.67	18.7%
Gemini 2.0	69.5%	60.3%	0.65	30.5%

Table 8: Performance Metrics for Using Explanation Similarity to Predict Score Agreement

The high precision for Claude and GPT-4 (>80%) suggests that when these models "speak the same language" as human raters, their scores are generally trustworthy. The low false positive rates (13-19% for Claude and GPT-4) indicate they rarely provide human-like explanations for incorrect scores—a desirable property for building educator trust.

However, the moderate recall values (55-60%) reveal that many correct scores emerge through different reasoning paths. This asymmetry has practical implications: similar explanations strongly predict reliable scores, but divergent explanations don't necessarily indicate unreliability. Models may identify alternative but valid indicators of comprehension quality that human raters don't typically consider.

6.4 Implications and Limitations

These findings suggest limited utility for direct student feedback from model explanations. While models can identify obvious strengths and weaknesses in clear-cut cases, their explanations for ambiguous responses—where students most need guidance—prove unreliable. The moderate overall explanation similarity (50-63%) indicates models identify alternative but potentially valid comprehension indicators that humans don't typically consider. This could enrich assessment if properly understood but requires careful interpretation.

The finding that models reach correct scores through different reasoning paths reinforces that AI assessment should complement rather than replace human evaluation. Models may notice patterns humans miss, but their reasoning remains opaque and potentially misleading, particularly for challenging cases.

This exploratory analysis has significant limitations. Single human rater evaluation limits generalizability. Subjective determination of explanation "similarity" introduces potential bias. The specific task and rubric may not represent broader assessment contexts. Despite limitations, consistent patterns across scoring and explanation analysis suggest current language models can handle routine assessment but shouldn't be trusted with generating feedback for ambiguous responses. The relationship between model reasoning and human judgment merits systematic study with multiple raters across diverse contexts.

7 Discussion and Conclusion

7.1 Key Findings and Implications

This research demonstrates that Large Language Models can achieve near-human reliability in story retell assessment, but with critical nuances that guide implementation. The convergence of model performance (QWK = 0.82) with human inter-rater reliability (0.85) represents practical viability, yet this aggregate metric masks important performance stratification.

The most significant finding is the dramatic performance difference based on case ambiguity. When human raters unanimously agree—approximately 66% of cases—models achieve 82% direct agreement. For the 34% generating human disagreement, model performance drops to chance levels (50%). This natural segmentation suggests a clear division of labor: AI handles routine cases while humans address ambiguous responses requiring professional judgment.

The discovery of consistent "grading personalities" across model families has important implementation implications. Claude's systematic strictness, Gemini's leniency, and GPT's moderation persist across temperature settings, indicating these are fundamental model characteristics. Schools must be aware of these tendencies to avoid inadvertently advantaging or disadvantaging students through model selection.

Rubric design emerges as foundational for both human and AI reliability. The progression from poor binary classification to strong five-class performance underscores that technology amplifies rather than compensates for assessment design quality. The conditional reliability of model explanations—strong for clear cases but unreliable for ambiguous ones—limits their utility for direct student feedback.

These findings collectively support hybrid assessment architectures that leverage respective strengths. For teachers managing 25+ students, automating the 66% of clear-cut cases could en-

able weekly rather than monthly retell assessments, dramatically increasing formative data availability while redirecting teacher expertise toward students most needing support.

7.2 Limitations and Future Directions

This exploratory study examined specific models, rubrics, and student populations. Generalization requires systematic investigation across diverse educational contexts. The single human rater providing explanations limits reasoning analysis conclusions. Future research should examine longitudinal impacts on learning outcomes and develop robust methods for uncertainty detection beyond simple confidence scores.

7.3 Conclusion

Large Language Models can reliably support story retell assessment when implemented thoughtfully within hybrid human-AI systems. The technology has reached sufficient maturity for practical application, but success depends on understanding both capabilities and limitations. By handling routine assessment tasks, AI can free teachers to focus on complex pedagogical decisions that truly require professional judgment. The goal isn't to automate education but to enhance human connections at its heart, providing teachers with better tools for understanding and supporting student learning.

Limitations

This study has several important limitations. First, our dataset consists of only 95 student responses from a specific educational context in Ghana, which may not generalize to other populations or educational settings. Second, while we tested three prominent LLMs, the rapid pace of model development means our findings may not apply to newer or different architectures. Third, our analysis of grading rationales relied on a single human rater's judgments, limiting the generalizability of our conclusions about explanation quality. Fourth, we examined only story retell tasks, and performance may differ for other literacy assessment types. Finally, our study is cross-sectional and cannot address the long-term impacts of AI-assisted assessment on student learning outcomes or teacher practices. Future work should address these limitations through larger, more diverse datasets, longitudinal studies, and multiple raters for explanation analysis.

Ethics Statement

This research was conducted with appropriate ethical oversight and student data was anonymized prior to analysis. We acknowledge several ethical considerations: First, automated assessment systems risk perpetuating or amplifying biases present in training data or human rating patterns. Second, over-reliance on AI assessment could diminish valuable teacher-student interactions that occur during traditional assessment. Third, the use of student data from Ghana raises questions about technological colonialism and the appropriateness of applying Western assessment frameworks in diverse cultural contexts. We advocate for AI assessment tools as supplements rather than replacements for human judgment, emphasizing transparency about system limitations and maintaining teacher agency in all assessment decisions. Any deployment should involve stakeholder consultation, particularly in Global South contexts, to ensure cultural appropriateness and educational benefit.

Acknowledgements

We thank the students and teachers who participated in this study, as well as the human raters who provided expert assessments.

References

Bess Altwerger, Nancy Jordan, and Nancy Rankie Shelton. 2007. *Rereading Fluency: Process, Practice, and Policy*. Heinemann, Portsmouth, NH.

Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.

Kenneth S Goodman. 2006. *The Truth About DIBELS: What It Is – What It Does*. Heinemann, Portsmouth, NH.

Bente E Hagtvet. 2003. Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills as well as phonological skills. *Reading and Writing: An Interdisciplinary Journal*, 16(6):505–539.

Janell P Klesius and Susan P Homan. 1985. A validity and reliability update on the informal reading inventory with suggestions for improvement. *Journal of Learning Disabilities*, 18(2):71–76.

Alicia M Marcotte and John M Hintze. 2009. Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, 47(5):315–335.

Katherine Maria. 1990. *Reading Comprehension Instruction: Issues and Strategies*. York Press, Parkton, MD.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Deborah K Reed and Sharon Vaughn. 2012. Retell as an indicator of reading comprehension. *Scientific Studies of Reading*, 16(3):187–217.

D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144.

J Schneider, B Schenk, C Niklaus, and M Vlachos. 2023. Towards LLM-based autograding for short textual answers. arXiv preprint arXiv:2309.11508.

R M Wilson, Linda B Gambrell, and W R Pfeiffer. 1985. The effects of retelling upon reading comprehension and recall of text information. *The Journal of Educational Research*, 78(4):216–220.