LLM-Generated Stories for Students with Significant Cognitive Disabilities: Promise, Gaps, and Evaluation Framework

Pragati Maheshwary, Ananya Ganesh, Shamya Karumbaiah
University of Wisconsin-Madison
Madison, WI, USA
{pmaheshwary, aganesh27, shamya.karunbaiah}@wisc.edu

Abstract

Students with significant cognitive disabilities (SCD) require specially designed accessible stories for reading comprehension assessments, yet creating such content is labor-intensive and difficult to scale. This preliminary study investigates whether large language models (LLMs) can generate short accessible stories for alternate assessment system. Using an 8-fold cross-validation design, we generated 120 stories with GPT-4o via one-shot prompting with human-written exemplars and evaluated them against a test set comprising 7 expert-human written stories as baselines across three dimensions: simplicity, fluency & coherence, and thematic adherence. Cross-validation results show that generated stories meet surface-level simplicity targets, with approximately two-thirds falling within the human baseline range for readability metrics. However, generated stories exhibited a systematic coherence gap where only 5% fell within the human range for adjacent sentence similarity, a pattern consistent across all folds. Thematic adherence was moderate, with adequate diversity across stories. These findings suggest LLMs can serve as a drafting tool within accessible content generation pipelines, but human expert review remains essential to ensure coherence, testability, and alignment with quality standards required for high-stakes alternate assessments.

Keywords: accessible story generation, text simplification, cognitive disabilities, reference-free evaluation

1. Introduction

Students with significant cognitive disabilities (SCD) represent a heterogeneous population that includes learners with intellectual disability, autism, multiple disabilities and more, who have complex communication and language needs (Karvonen and Clark, 2019; Thurlow et al., 2016). Under the Individuals with Disabilities Education Act (IDEA) and the Every Student Succeeds Act (ESSA), states are required to include these students in accountability systems through alternate assessments based on alternate achievement standards (AA-AAS), with participation capped at approximately 1% of the total tested population (Thurlow et al., 2017). The Dynamic Learning Maps (DLM) alternate assessment system is one of the largest operational AA-AAS programs in the United States, serving approximately 90,000 students with SCD across more than 18 states (Karvonen and Clark, 2019; Karvonen et al., 2021). Within the DLM system, English Language Arts (ELA) assessments are designed for students to engage with short accessible stories and then respond to reading comprehension items aligned with *Essential Elements* (the alternate content standards that guide instruction and assessment for this population).

Creating accessible high-quality stories for this population is a significant challenge. Each story must adhere to accessibility criteria such as the use of simple vocabulary drawn from high-frequency word lists, short sentences with explicit referents, clear narrative structures that adhere to a singular theme, and minimal inference load and ambiguity. These criteria are grounded in what is known about the communication and language profiles of students with SCD, a substantial proportion of whom communicate primarily using one or two words, signs, or symbols at a time (Nash et al., 2016), and many of whom rely on augmentative and alternative communication (AAC) devices (Erickson and Geist, 2016). As a result, human-authored assessments undergo an extensive multi-stage development pipeline that includes initial drafting by trained item writers, peer review, multiple rounds of internal quality control by content and accessibility specialists, external review for content accuracy and bias & sensitivity, editorial review, and field testing before stories become operational (DLM Consortium, 2024). This process, while essential for ensuring quality and validity, is labor-intensive, time-consuming, and difficult to scale to meet the growing demand for diverse, engaging, and grade-appropriate content.

This exploratory study attempts to introduce LLMs as drafting tools that adhere to baseline quality controls to fast-track the multi-stage development pipeline for such stories by utilizing existing advancements in the field of text simplification and AI-assisted content generation. We investigate two primary questions: (1) How do LLMs perform at the task of short story generation that is inspired by existing literary sources but adapted for students with SCD? (2) How do LLM-generated stories for SCDs compare with expert-human written stories on established metrics from text simplification re-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 130–141
11 May 2026.

search? To answer these questions, we employ a cross-validation experimental design in which LLM-generated stories and human-written baselines are both scored using a multi-dimensional evaluation framework. The evaluation framework assesses stories across three key dimensions, (1) simplicity, (2) Fluency & Coherence, and (3) Thematic Adherence. This study makes two primary contributions. First, we introduce a reference-free evaluation framework with the aforementioned criteria for assessing LLM-generated accessible stories that matter for this specialized context. Second, we make a case for where LLMs as a drafting tool can fit in the pipeline for generating accessible stories using one-shot prompting with human-written exemplars, highlighting both their promise and gaps. Together, these contributions lay the groundwork for responsibly leveraging AI to supplement – not replace – human expertise in creating accessible educational content and assessment for students with significant cognitive disabilities.

2. Related Work

The broader field of text simplification has long sought to make written content more accessible to diverse reader populations. Traditional approaches relied on rule-based systems involving lexical substitution and syntactic simplification, as well as statistical methods that treated simplification as a form of monolingual translation using parallel corpora such as Wikipedia and Newsela (Xu et al., 2016). More recently, neural approaches, including sequence-to-sequence models and transformer-based architectures such as BERT, T5, and GPT variants, have substantially advanced the state of the art in text simplification (Alva-Manchego et al., 2020; Martin et al., 2022). These models have been evaluated using both traditional readability formulas such as Flesch-Kincaid Grade Level and Flesch Reading Ease, as well as task-specific metrics including SARI (Xu et al., 2016) for measuring simplification quality and BLEU and ROUGE for content preservation. Benchmarks such as ASSET (Alva-Manchego et al., 2020) and TurkCorpus have further supported systematic evaluation. Notably, Chamovitz and Abend (2022) demonstrated that incorporating cognitively motivated simplification operations, such as reducing syntactic complexity and resolving ambiguous references, can improve the quality of simplified text beyond what surface-level transformations achieve. At the document level, Vásquez-Rodríguez et al. (2023) have highlighted the importance of maintaining coherence when simplifying longer texts, showing that simplification must attend not only to sentence-level readability but also to the logical flow and connectedness of the overall text. However, the vast majority of this work has targeted general adult readers, second language learners, or individuals with low literacy levels. Very little simplification research has explored the needs of students with SCDs, whose reading and communication profiles differ from these other populations and require linguistic simplification and structural clarity that extend well beyond what standard simplification approaches typically produce (Yalon-Chamovitz, 2009).

Large language models (LLMs) have recently demonstrated strong performance across a wide range of natural language generation tasks, including creative writing, summarization, and text simplification through zero-shot and few-shot prompting (Brown et al., 2020; OpenAI, 2023). In the educational domain, there has been growing interest in using LLMs to support assessment development. Tan et al. (2025) provided a comprehensive review of automatic item generation techniques leveraging LLMs, documenting the rapid evolution of these approaches and noting both their promise for producing diverse item pools and the persistent challenges related to quality assurance and alignment with assessment specifications. Laverghetta Jr. and Licato (2023) showed that LLMs can generate items for cognitive assessments that approximate the psychometric properties of human-written items, and further developed a framework for using LLMs to generate and validate psychometric items, demonstrating the feasibility of integrating AI into the assessment development cycle. Beyond item generation, LLMs have also been explored for narrative content creation. Feng et al. (2025) developed a framework for generating social stories using LLMs, demonstrating their capacity to produce structured narratives that adhere to specific pedagogical conventions. Raffloer and Green (2025) investigated reader perceptions of AI-generated versus human-authored narratives, finding that while readers can sometimes detect differences, AI-generated stories can achieve comparable levels of narrative engagement under certain conditions. Despite these important advances in both AI-assisted item generation and narrative generation, the potential of LLMs to assist in the specific and labor-intensive task of drafting accessible stories for students with SCD has not yet been investigated.

3. Methods

This study employed a cross-validation experimental design to evaluate the quality of large language model (LLM)-generated accessible stories for students with cognitive disabilities. The design systematically compared AI-generated stories against human-written baseline stories using a quantitative evaluation framework inspired from some of the Dynamic Learning Maps (DLM) assessment criteria for accessible educational text.

3.1. Data

We manually extracted the reading comprehension stories written and evaluated by human experts from publicly available and retired DLM stories for Grade 11-12 English Language Arts [1]. These stories took inspiration from existing literary sources that are considered grade level appropriate readings recommend by State and Federal Educational Agencies. For the exploratory analysis in the current study, we focused only on generating and evaluating the story text and not on the questions/test items associated with the story for comprehension testing purposes. We ended up with 8 stories inspired from 3 source books, *The Great Gatsby, My Antonia and A White Heron*. Table 1 provides a detailed description about the stories associated with each book.

Table 1: Human Written Stories for Grade 11-12

Source Book	Story Title	Sentence Count	Word Count
My Antonia (MA)	Jim & Antonia	28	288
	Post Office	31	351
	The Garden	16	99
A White Heron (WH)	Mary & Martha	16	104
	The Businessman	22	129
	How about a Wig?	22	150
The Great Gatsby (GG)	The Valley of Ashes	20	195
	Nick Changes His Mind	26	157

3.2. Evaluation Framework

We developed a reference-free quantitative evaluation framework comprising three criteria partly aligned with DLM guidelines for accessible educational content (Table 2): Simplicity, Fluency & Coherence, and Thematic Adherence. The framework is reference-free by design for two reasons. First, the intended use case is to evaluate LLMs as drafting tools at the earliest stage of the novel content development pipeline, where a reference text does not exist. Second, because multiple distinct stories can be generated from the same source book, there is no one-to-one correspondence between human-written and LLM-generated texts that would make reference-based comparison meaningful (Belem et al., 2025).

We also adopted an automated quantitative framework as a necessary first step before human expert validation because collecting human judgments from accessibility specialists is time-consuming and expensive, and committing expert reviewer time is difficult to justify without first establishing that generated stories show sufficient baseline quality on well-validated metrics to warrant further review. The metrics selected for this framework are individually well-established in the text simplification and coherence literature (Kincaid et al., 1975; Vásquez-Rodríguez et al., 2023; Reimers and Gurevych, 2019). We note that these automated metrics do not capture all dimensions of quality that matter for DLM texts such as use of people-first language or fairness criteria that makes certain stories emphasizing biking over traveling in a car as unfair for students with motor disabilities. Therefore, human expert validation remains an essential next step in the broader scheme. The present framework is intended to provide a scalable, reproducible preliminary assessment that can identify systematic gaps before expert review resources are invested. The following subsections describe the operationalization of each criterion through the different metrics.

3.2.1. Simplicity

Readability was assessed using the Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) as the primary metric, computed as:

$$\text{FKGL} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59 \quad (1)$$

where ASL = average sentence length (words) and ASW = average syllables per word. The Flesch Reading Ease (FRE) score was retained as a supplementary measure:

$$\text{FRE} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}) \quad (2)$$

Average words per sentence and average syllables per word were also reported as interpretable diagnostics, as they are the direct inputs to both formulas.

3.2.2. Fluency and Coherence

Narrative coherence was assessed using three complementary measures. Entity continuity was computed as the proportion of adjacent sentence pairs sharing at least one proper noun (identified by capitalization) or personal pronoun, adapted from the entity-grid approach of (Vásquez-Rodríguez et al., 2023). Higher continuity scores indicate more referentially coherent narrative chains. Adjacent sentence similarity was measured as the mean TF-IDF cosine similarity between consecutive sentence pairs, capturing local topical flow. Sentence length standard deviation was used as a structural regularity indicator where high variance in sentence length beyond 5-12 words per sentence signals inconsistent adherence to accessibility constraints.

3.2.3. Thematic Adherence

Thematic adherence was operationalized with three different methodological approaches.

[1] Stories can be found at: https://monarchreader.com/home

Table 2: Evaluation Criteria, Metrics, and DLM Alignment

Dimensions	Metric	Interpretation Range	Purpose	DLM Alignment
Simplicity	Flesch-Kincaid Grade Level	Grade 0–16+ (lower = simpler)	Evaluates surface accessibility of syntax and vocabulary; lower grade level indicates simpler, more accessible text	Accessible Text Language: minimize inference load; Accessible Text Content: reduced depth, breadth, and complexity
	Flesch Reading Ease	0–100 (higher = easier)		
	Avg. words per sentence	Word count		
	Avg. syllables per word	Syllable count		
Fluency & Coherence	Entity Continuity Score	0–1 (higher = more coherent)	Assesses narrative logical flow, referential consistency, and structural regularity relative to DLM format constraints	Accessible Text Language: maintains logical structure and predictable sentence patterns
	Adjacent Sentence TF-IDF Similarity	0–1 (higher = smoother)		
	Sentence Length Standard Deviation	0–∞ (lower = more uniform)		
Thematic Adherence	Theme-Text Semantic Similarity	0–1 (higher = more coherent)	Measures whether each story's text semantically reflects its declared theme (embedding-based). Detects thematic collapse within a book's story set. Detects cross-book theme bleed within a fold	Instructional Relevance: preserves thematic intent and construct validity
	Within-Book Pairwise TF-IDF	0–1 (lower = more diverse)		
	Within-Fold Pairwise TF-IDF	0–1 (lower = more diverse)		

First, the generation prompt was designed to elicit an explicit theme statement from the LLM for each generated story, structured as "This [story/information text] explores [theme]." This approach was adopted because the source books (e.g., *My Antonia*) contain multiple themes across hundreds of pages, making direct comparison between a 15-30 sentence generated story and the full source text uninformative as a similarity signal. To assess whether each story's text reflect its declared theme, *single-story thematic consistency* was measured using embedding-based semantic similarity between the declared theme statement and the story text, computed as the cosine similarity between sentence embeddings produced by the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). Embedding-based similarity was preferred over lexical overlap measures because theme statements and story texts are expected to express the same meaning through different vocabulary – a story about perseverance will use words like "kept trying" and "finally succeeded" rather than the word "perseverance" itself, which lexical measures would penalize incorrectly.

Thematic diversity among the five stories generated for the same book within each fold was assessed by computing mean pairwise TF-IDF cosine similarity between all theme statements generated for the same target book within a fold. Low mean similarity indicates the LLM explored diverse themes naturally; high similarity indicates thematic collapse toward a dominant theme. This serves as both a quality indicator and a research finding about LLM behavior under one-shot prompting.

Within-fold thematic diversity was assessed to identify if themes overlap across source books within a fold, by computing mean pairwise TF-IDF cosine similarity across all 15 theme statements generated in a single fold, regardless of target book. If the LLM defaulted to the same thematic territory regardless of which book it was generating for, this would appear as high within-fold similarity. We chose TF-IDF cosine because comparison between theme statements concern lexical similarity — short, structurally parallel strings — rather than semantic coherence between a theme and a full story text.

Thematic adherence metrics were not computed for human-written stories, which were produced without LLM-declared theme statements. Human stories serve as a quality baseline for Simplicity, and Fluency & Coherence, but the Thematic Adherence criterion is specific to LLM-generated outputs.

3.3. LLM Text Generation

Stories were generated using OpenAI's GPT-4o model via the OpenAI API with temperature $= 0.4$

and $\texttt{max\_tokens} = 8000$. Each generation call used one-shot prompting, providing a single human-written text piece as an exemplar. The prompt instructed the model to generate five independent stories inspired by a specified target book, constrained to 15-30 sentences, using simple vocabulary and sentences, with characters and settings relating to the themes from a [target book]. The prompt additionally required the model to produce an explicit theme statement for each story which was used for thematic adherence evaluation as described in Section 3.2.3 (also see Appendix A).

3.4. Analysis

3.4.1. Baseline Characterization

Human-written texts were evaluated first to establish baseline distributions for all applicable metrics (i.e., Simplicity, and Fluency & Coherence). Descriptive statistics (means, standard deviations) were computed overall and by source book.

3.4.2. Cross-Validation Analysis

We implemented a multi-fold cross-validation design in which each human-written text served as the one-shot training exemplar exactly once. For each story fold:

- **Training set** ($n = 1$): One human text served as the exemplar in the generation prompt.
- **Test set** ($n = 7$): The remaining seven human stories served as quality baselines.
- **Generation**: The LLM produced five stories for each of the three source books, yielding 15 generated stories per fold.

This design yielded 120 total generated stories (8 folds $\times$ 3 books $\times$ 5 stories) while systematically varying the training exemplar to assess the influence of exemplar writing style and source book on generation quality.

For each of the 8 folds, the 15 generated stories were compared against the metric distributions of the 7 test stories in that fold. For each metric, we computed the proportion of generated stories whose value fell within the minimum–maximum range of the test stories. Because the training exemplar was excluded from the test set in each fold, this comparison is not contaminated by the story used to prompt the LLM. We then averaged these proportions across all folds to obtain a cross-validated estimate of how frequently generated stories achieve human-like metric values. We also computed the mean signed delta between each generated story's metric value and the test set mean to characterize the direction of any deviations. The proportion of generated stories falling within the test set range is reported overall, by training exemplar (at the fold level), by target book, and by prompting condition (same-source vs. cross-source). Mean signed deltas are reported at the overall level only. Thematic adherence is reported descriptively for generated texts only, as human stories were produced without LLM-declared theme statements.

4. Results

4.1. Human-Written Story Baseline Characteristics

Table 3 presents descriptive statistics for the eight human-written stories across the two evaluation criteria. These stories served as both the quality reference baseline and the one-shot training exemplars in the cross-validation design. Table 4 presents the same statistics aggregated by source book.

Simplicity. Human stories varied considerably in readability. FK Grade Level ranged from 1.97 (The Business Man) to 5.18 (Jim and Antonia), with a mean of 3.67 (SD = 1.04). This range is notable: even among texts approved for DLM Grade 11-12 administration, readability spans more than three grade levels, suggesting that the DLM accessible text standard accommodates substantial surface-level complexity variation. Flesch Reading Ease scores ranged from 72.52 to 91.36 (mean = 82.75), consistent with texts in the "fairly easy" to "easy" range. Average sentence length ranged from 5.86 to 11.32 words per sentence (mean = 7.85).

Fluency and Coherence. Entity continuity scores ranged from 0.10 *(How About a Wig?)* to 0.81 *(The Business Man)*, with a mean of 0.62 (SD = 0.24). The notably low entity continuity for *How About a Wig?* (0.10) reflects its narrative structure, which alternates between two characters rather than maintaining a single referential chain — a legitimate stylistic choice rather than a coherence failure. Adjacent sentence similarity ranged from 0.16 to 0.26 (mean = 0.20), indicating modest local topical continuity across all stories. Sentence length standard deviation ranged from 1.70 to 4.51, with longer stories *(The Post Office, Jim and Antonia)* showing greater structural variability.

Variation by Source Book. My Antonia stories had the highest mean FK Grade Level (4.17) and longest mean sentences (9.27 words), driven primarily by The Post Office and Jim and Antonia, which are the two longest and most complex stories in the corpus. Stories derived from A White Heron had the lowest mean FK Grade Level (3.18) and shortest sentences (6.39 words), as well as the lowest entity continuity (0.57), reflecting the shorter, more episodic narrative structure of that book's stories. The Great Gatsby stories fell between these two (FKGL = 3.65, 7.90 words/sentence). These between-book differences in the human baseline

Table 3: Human-Written Story Baseline Metrics

Story Title	Simplicity				Fluency & Coherence		
	FKGL	FRE	WPS	SPW	Ent. Cont.	Adj. Sim.	SL SD
Jim and Antonia	5.18	76.25	10.29	1.420	0.778	0.212	3.91
Mary and Martha	4.76	72.52	6.50	1.510	0.800	0.225	1.84
The Post Office	4.05	86.16	11.32	1.291	0.667	0.159	4.51
The Garden	3.27	82.63	6.19	1.394	0.600	0.259	1.94
Nick Changes His Mind	3.53	80.54	6.04	1.420	0.600	0.171	1.95
The Business Man	1.97	91.36	5.86	1.295	0.810	0.159	2.42
The Valley of Ashes	3.76	85.44	9.75	1.318	0.632	0.279	4.44
How About a Wig?	2.80	87.11	6.82	1.333	0.095	0.165	1.70
Mean	*3.67*	*82.75*	*7.85*	*1.373*	*0.623*	*0.204*	*2.84*
SD	*1.03*	*6.14*	*2.22*	*0.077*	*0.230*	*0.047*	*1.23*

FKGL = Flesch-Kincaid Grade Level; FRE = Flesch Reading Ease; WPS = Average Words per Sentence; SPW = Average Syllables per Word; Ent. Cont. = Entity Continuity; Adj. Sim. = Adjacent Sentence Similarity; SL SD = Sentence Length Standard Deviation.

are important context for interpreting the cross-validation results further.

4.2. LLM Generated Story Characteristics

Table 5 presents the descriptive statistics for the LLM Generated stories across two evaluation criteria, aggregated by target book. Table 6 presents the cross-validation results, reporting the percentage of generated stories in each fold whose metric values fell within the min-max range of the 7 test stories.

Simplicity. Generated stories were on average simpler than the human-written baseline. Mean FK Grade Level was 2.94 (SD = 1.19) compared to 3.67 (SD = 1.03) for human stories. Average words per sentence was 5.74 for generated stories as compared to 7.85 for human stories. Generated stories were also substantially more uniform in sentence length (SD = 1.25) than human stories (SD = 2.84), reflecting close adherence to the sentence length constraint in the prompt. By target book, *A White Heron* stories were simplest (FKGL = 2.71) and *My Antonia* stories most complex (FKGL = 3.24), mirroring the pattern in the human baseline (See Table 5).

Cross-validation results (Table 6) showed that simplicity metrics had the highest overlap with the test set range for each fold. For FK Grade Level, 67.5% of generated stories fell within the test set range (SD = 17.6 across folds), and similarly 67.5% for average syllables per word (SD = 10.9). Flesch Reading Ease showed 65.0% overlap (SD = 9.3). Average words per sentence had relatively lower overlap at 40.0%, with a mean delta of -2.10 words below the test set mean, confirming that generated sentences were consistently shorter than the simplest human stories in most folds. This metric also showed the highest variability across folds (SD = 32.1), ranging from 6.7% (*The Business Man*) to 93.3% (*The Post Office*), suggesting that exemplar sentence length strongly influenced generation output. Fold-level FKGL means ranged from 2.24 (*The Garden* as exemplar story) to 3.59 (*The Valley of Ashes* as exemplar story), indicating additional sensitivity to training exemplar complexity.

Fluency and Coherence. Generated stories showed lower coherence than human-written stories across all three measures. Entity continuity was 0.39 (SD = 0.29) for generated stories versus 0.62 (SD = 0.23) for human stories, and adjacent sentence similarity showed the largest gap (generated: 0.093; human: 0.204).

The cross-validation results quantify this gap more precisely. Adjacent sentence similarity had the lowest overlap with the test set range of any metric: only 5.0% of generated stories fell within the human range (SD = 7.8), with 0.0% overlap in five of eight folds (Table 6). Sentence length standard deviation showed similarly low overlap at 12.5% (SD = 15.7). Entity continuity showed moderate overlap at 64.2%, but with high variability across folds (SD = 28.7): overlap reached 93.3% when *Jim and Antonia* was the exemplar but dropped to 0.0% when *How About a Wig?* was the exemplar. This variability is driven by the unusual entity continuity profile of *How About a Wig?* (0.095), which alternates between two characters rather than maintaining a single referential chain; when this story is excluded from the test set (Fold 7), the test set range narrows substantially, causing all generated stories to fall outside it. The consistently near-zero overlap for adjacent sentence similarity across folds indicates that the coherence gap is a systematic limitation of the LLM's generation behavior rather than an artifact of particular training exemplars.

Thematic Adherence. Embedding-based se-

Table 4: Human-Written Story Baseline Metrics by Source Book

	Simplicity				Fluency & Coherence		
Source Book	FKGL	FRE	WPS	SPW	Ent. Cont.	Adj. Sim.	SL SD
A White Heron	3.18	83.66	6.39	1.379	0.568	0.183	1.99
My Antonia	4.17	81.68	9.27	1.368	0.682	0.210	3.45
The Great Gatsby	3.65	82.99	7.90	1.369	0.616	0.225	3.20

Stories per book: *A White Heron* (n = 3), *My Antonia* (n = 3), *The Great Gatsby* (n = 2).

Table 5: LLM-Generated Story Metrics by Training Exemplar and Target Book

	Simplicity				Fluency & Coherence			Theme
Target Book	**FKGL**	**FRE**	**WPS**	**SPW**	**Ent. Cont.**	**Adj. Sim.**	**SL SD**	**Sem. Sim.**
A White Heron	2.71	86.53	6.11	1.349	0.333	0.110	1.275	0.370
My Antonia	3.24	82.08	5.77	1.406	0.322	0.079	1.251	0.422
The Great Gatsby	2.86	84.11	5.35	1.387	0.529	0.089	1.224	0.385
Mean	*2.94*	*84.24*	*5.74*	*1.380*	*0.395*	*0.093*	*1.250*	*0.392*
SD	*1.19*	*8.89*	*0.95*	*0.109*	*0.285*	*0.039*	*0.376*	*0.091*

Sem. Sim. = Theme-Text Semantic Similarity; Each cell is contains mean values computed across all 40 stories (8 folds $\times$ 5 stories) generated for each target book.

mantic similarity between each story's declared theme statement and its text was moderate, with a mean of 0.39 (SD = 0.09, range: 0.20-0.64). This indicates that generated stories generally reflected their declared themes at the semantic level. Similarity was comparable across target books: A White Heron (mean = 0.37), The Great Gatsby (mean = 0.39), and My Antonia (mean = 0.42).

Within-book theme diversity showed that the five stories generated for the same book within a fold were moderately distinct from one another, with mean pairwise TF-IDF similarity of 0.44 (SD = 0.13, range: 0.25-0.82) across all book-fold combinations. This indicates the LLM explored different thematic territory across the five stories rather than collapsing to a single dominant theme, though some theme overlap was present. The maximum pairwise similarity between any two theme statements within a book-fold combination reached 1.00 in two instances (both involving *My Antonia* as the target book), indicating occasional identical theme statements.

Within-fold theme diversity showed that the 15 stories generated across all three target books within a fold were distinct from one another (mean pairwise TF-IDF = 0.40, SD = 0.07), indicating that theme generation was driven more by story-level variation than by target book. Maximum pairwise similarity within a fold reached 1.00 in four of eight folds, again reflecting occasional identical theme statements, but mean similarity remained consistently low across folds (range: 0.33-0.53), suggesting no systematic cross-book theme bleed.

Same-Source vs. Cross-Source Prompting. Same-source prompting (i.e., the source book of the exemplar matching the target book) produced stories that more frequently fell within the test set range across most metrics. The largest differences were observed for entity continuity (82.5% of same-source stories vs. 55.0% of cross-source stories in the test set range), average words per sentence (60.0% vs. 30.0%), and adjacent sentence similarity (12.5% vs. 1.2%). Same-source prompting also consistently produced lower FK Grade Level than cross-source prompting for all three target books.

The effect on coherence was inconsistent across books. For *A White Heron*, same-source prompting produced higher entity continuity than cross source (0.416 vs. 0.283), suggesting that a same-book exemplar better scaffolded referential coherence. For *The Great Gatsby*, the pattern reversed: cross-source prompting produced higher entity continuity (0.566 vs. 0.418). Theme semantic similarity was comparable across both conditions for all three books (differences $\leq$ 0.06), indicating that exemplar source had no meaningful effect on thematic coherence.

5. Discussion

The simplicity results confirm that one-shot prompting with explicit sentence length and vocabulary constraints can reliably produce accessible text, consistent with prior findings that LLMs respond well to structural specifications in educational content generation (Laverghetta Jr. and Licato, 2023). Cross-validation showed that approximately two-thirds of generated stories fell within the test set range for simplicity metrics, and that fold-level FKGL means varied depending on the training exemplar. The sensitivity to exemplar complexity sug-

Table 6: Cross-Validation: Percentage of Generated Stories Within Test set Range, by Fold

		Simplicity				Fluency & Coherence		
Fold	**Training Exemplar Used**	**FKGL**	**FRE**	**WPS**	**SPW**	**Ent. Cont.**	**Adj. Sim.**	**SL SD**
0	Jim and Antonia	73.3	66.7	80.0	66.7	93.3	0.0	13.3
1	Mary and Martha	60.0	60.0	20.0	46.7	80.0	6.7	0.0
2	The Post Office	73.3	60.0	93.3	60.0	86.7	0.0	46.7
3	The Garden	53.3	66.7	26.7	80.0	60.0	0.0	0.0
4	Nick Changes His Mind	80.0	73.3	13.3	66.7	66.7	0.0	0.0
5	The Business Man	33.3	46.7	6.7	80.0	60.0	13.3	13.3
6	The Valley of Ashes	86.7	73.3	26.7	66.7	66.7	20.0	6.7
7	How About a Wig?	80.0	73.3	53.3	73.3	0.0	0.0	20.0
Mean		67.5	65.0	40.0	67.5	64.2	5.0	12.5
SD		17.6	9.3	32.1	10.9	28.7	7.8	15.7

Each cell reports the percentage of 15 generated stories in that fold whose metric value falls within the min–max range of the 7 test set stories. Mean and SD are computed across the 8 fold-level percentages.

gests that strategic exemplar selection could serve as a practical lever for targeting specific readability bands in operational use. However, the uniformity of generated sentence lengths (SD = 1.25 vs. 2.84 for human stories) suggests that the model might have interpreted the prompt's sentence length constraint as a target rather than a ceiling, producing text that is structurally monotonous. Whereas, human-written DLM stories varied sentence length deliberately to maintain reader engagement and signal narrative transitions—properties that matter for students who rely on predictable but not rigid textual patterns (Erickson and Geist, 2016). Future prompt designs should distinguish between a maximum sentence length and the expectation of natural variation within that bound.

The drop in coherence for LLM-generated stories in comparison to human-authored stories is the most consequential finding: only 5.0% of generated stories achieved adjacent sentence similarity within the test set range, and this near-zero overlap held across folds (0.0% in five of eight). This does not mean that 95% of generated stories are entirely incoherent; rather, it means that the degree to which consecutive sentences share overlapping vocabulary and content was consistently lower than what human experts produced, suggesting that LLM-generated stories tend to introduce new referents between sentences more than human-written DLM stories. However, students with SCD referential continuity is not a stylistic preference but a comprehension necessity. These readers depend on explicit cues such as repeated character names, pronoun chains, and topical bridges between sentences to track who is doing what across a story. When those cues are absent, even individually simple sentences become difficult to integrate into a coherent mental model of the narrative. Incorporating explicit coherence instructions into the generation prompt, following cognitively motivated generation strategies (Chamovitz and Abend, 2022)(e.g., requiring that each sentence share at least one referent with its predecessor, or that character names be reused across non-adjacent sentences) could address this gap without sacrificing simplicity.

Thematic adherence was moderate, with generated stories generally reflecting their declared themes at the semantic level and maintaining adequate within-book diversity. However, the occasional production of identical theme statements within a fold (4 of 8 folds) suggests that the model's theme generation draws from a constrained latent space of "accessible story themes" rather than engaging deeply with the source material's thematic range. In practice, this means that while individual stories may appear thematically appropriate, a set of five stories for the same book may lack the substantive differentiation needed to support distinct assessment items targeting different reading comprehension skills. Human review at the theme-selection stage, or prompt modifications that provide explicit thematic anchors drawn from different chapters or subplots of the source book, could mitigate this collapse.

6. Conclusion

In conclusion, our findings suggest that LLMs can serve as a useful drafting tool within the DLM development pipeline, but not as a replacement for human authoring. Generated stories would need revision by accessibility specialists to address coherence before entering the review stages that are fundamental to DLM quality assurance (DLM Consortium, 2024). This aligns with the broader consensus that human-in-the-loop oversight is essential for AI-generated educational content in high-stakes contexts (Clark et al., 2025; Tan et al., 2025). Future work should target the coherence gap through prompt engineering (e.g., explicit instructions for entity reuse and topical connectivity), incorporate expert human review alongside automated metrics,

expand the pipeline to additional grade levels and text types, and evaluate whether students and educators perceive meaningful differences between human-written and LLM-generated stories in assessment contexts.

7. Ethical Considerations and Limitations

Several limitations warrant careful consideration. First, the human baseline comprised only 8 stories compared to 120 LLM-generated stories, creating an asymmetry that limits the precision of comparison. Second, the evaluation framework used in this study has not been reviewed by DLM operational staff, nor have the AI-generated stories themselves been evaluated through DLM's internal review processes. Two criteria central to DLM's own quality standards are absent from our framework: (a) every DLM text must contain sufficient testable points aligned to a specific node to support the development of approximately five distinct items measuring the same skill (because we primarily focused on assessing LLMs' story generation capabilities), and (b) Bias, sensitivity and people-first language flags (which we plan to address in future work that extends the evaluation framework). It is important to note that even though we propose the use of readability formulas such as Flesch-Kincaid, DLM relies more heavily on expert human judgment which capture nuances missed by automated metrics. Consequently, the differences observed in our automated metrics should not be interpreted as evidence that LLM-generated stories surpass human-written DLM texts in quality. Rather, the differential may reflect that the human-written texts met a categorically different and more comprehensive standard of quality than what our automated metrics capture. Third, we only experiment with one model (OpenAI's GPT-4o) for story generation and assumed that the LLM had prior knowledge of the canonical source books used in this study, an assumption that may not hold across models or when extending this work to less widely known, multilingual, or culturally specific literary sources. Fourth, the generation setup is intentionally minimal — one-shot prompting with a single human-written exemplar due to limited availability of human written examples which limits the strength of conclusions that can be drawn about LLM capability for this task more broadly. Future work can compare our results with alternative prompting strategies such as few-shot prompting, chain-of-thought, constrained decoding, or structured prompting.

Given these limitations, a critical next step is to have the generated stories evaluated by human experts using both the automated metrics proposed here and DLM's own review criteria, including testability, accessibility, and bias and sensitivity standards. Such a validation study would allow assessment of the reliability and validity of the proposed evaluation framework, determine whether the automated metrics align with expert judgment, and establish whether LLM-generated stories can meet the comprehensive quality standards required for operational use. More broadly, ethical deployment of AI-generated content in high-stakes assessment contexts demands that generated texts undergo the same rigorous multi-stage review process applied to human-authored stories, ensuring that no content reaches students without thorough human oversight.

8. Lay Summary

Reading comprehension tests for students with significant cognitive disabilities(SCD) require specially written short stories that use simple words, short sentences, and clear, easy-to-follow narratives. Writing these stories takes a great deal of time and expertise. This study explored whether large language models could help by generating first drafts of these stories. We gave (GPT-4o) a single example of a human-written story and asked it to produce new short stories inspired by well-known books like "The Great Gatsby" and "My Antonia", but written in a simple and accessible way. We then measured how well the AI-generated stories compared to stories written by human experts across three criteria: (1) how easy they were to read, (2) how well each sentence connected to the next, and (3) whether the stories stayed focused on a single theme. We found that the AI was good at keeping the language simple, but consistently struggled to write sentences that flowed naturally from one to the next in the way that human experts did — a feature that is especially important for students with SCD who rely on clear, predictable language to understand what they read. Stories generally stayed on topic, though the AI sometimes repeated the same themes across different stories. Overall, our findings suggest that AI can be a useful starting point in the story-writing process, but that human expert review and revision remain essential before these stories could be used in real assessments.

9. Acknowledgments

This research was made possible through the generous Research Fellowship support of Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas.

10. References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics.

Catarina G. Belem, Parker Glenn, Alfy Samuel, Anoop Kumar, and Daben Liu. 2025. Readability reconsidered: A cross-dataset analysis of reference-free metrics. *arXiv preprint arXiv:2510.15345*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Elad Chamovitz and Omri Abend. 2022. Cognitive simplification operations improve text simplification. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 241–265.

A. K. Clark, A. Hirt, D. Whitcomb, W. J. Thompson, M. Wine, and M. Karvonen. 2025. Artificial intelligence in science and mathematics assessment for students with disabilities: Opportunities and challenges. *Education Sciences*, 15(2):233.

DLM Consortium. 2024. 2023–2024 technical manual—dlm alternate assessment system. Technical report, University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Karen Erickson and Lori A. Geist. 2016. The profiles of students with significant cognitive disabilities and complex communication needs. *Augmentative and Alternative Communication*, 32(3):187–197.

Yi Feng, Mingyang Song, Jiaqi Wang, Zhuang Chen, Guanqun Bi, Minlie Huang, Liping Jing, and Jian Yu. 2025. Ss-gen: a social story generation framework with large language models. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Meagan Karvonen and A. K. Clark. 2019. Students with the most significant cognitive disabilities who are also english learners. *Research and Practice for Persons with Severe Disabilities*, 44(2):71–86.

Meagan Karvonen, A. K. Clark, C. Carlson, S. Wells Moreaux, and J. Burnes. 2021. Approaches to identification and instruction for students with significant cognitive disabilities who are english learners. *Research and Practice for Persons with Severe Disabilities*, 46(4):223–239.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Millington, TN, Research Branch.

Anthony Laverghetta Jr. and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 414–428. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Brooke Nash, A. K. Clark, and Meagan Karvonen. 2016. First contact: A census of students taking the dynamic learning maps alternate assessment. Technical report, University of Kansas, Center for Educational Testing and Evaluation.

OpenAI. 2023. Gpt-4 technical report.

G. Raffloer and Melanie C. Green. 2025. Of love & lasers: Perceptions of narratives by ai versus human authors. *Computers in Human Behavior: Artificial Humans*, 5:100168.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

B. Tan, N. Armoush, E. Mazzullo, O. Bulut, and M. Gierl. 2025. A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.

Martha L. Thurlow, S. S. Lazarus, E. D. Larson, D. A. Albus, K. K. Liu, and E. Kwong. 2017.

Alternate assessments for students with significant cognitive disabilities: Participation guidelines and definitions. Technical report, University of Minnesota, National Center on Educational Outcomes.

Martha L. Thurlow, Y. Wu, Rachel F. Quenemoen, and E. Towles. 2016. Characteristics of students with significant cognitive disabilities. Technical report, National Center and State Collaborative.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Shira Yalon-Chamovitz. 2009. Invisible access needs of people with intellectual disabilities: A conceptual model of practice. *Intellectual and Developmental Disabilities*, 47(5):395–400.

A. Story Generation Prompt

The following system and user prompts were used to generate stories via the OpenAI API (GPT-4o, temperature = 0.4). The `{target_book}` and `{n}` placeholders are filled at runtime per fold.

System Prompt

You are an English Language Arts story writer producing accessible short stories for Grade 11–12 students with cognitive disabilities.

Rules:

- Produce short stories in plain, simple sentences inspired by the book `{target_book}`, but adapted for students.
- Each story should be between 15–30 sentences long.
- Use simple vocabulary.
- Characters and settings should relate to themes from `{target_book}`.
- For each story, provide a theme statement in the form: *"This story explores [theme]."*
- The theme statement must be reflected consistently throughout the story text.
- Output MUST be a single JSON array of exactly `{n}` objects with keys: `story_id`, `book_inspiration`, `grade`, `story_text`, `title`, `theme_statement`.
- Do NOT include any prose, explanations, or code fences. JSON array only.

User Prompt

`EXEMPLAR={exemplar_json}`

Now generate `{n}` independent stories inspired by "`{target_book}`" following the exemplar structure and style constraints. Return ONLY a JSON array with `{n}` story objects. No wrapper. No extra text.

B. Example Stories

B.1. Human-Written Story

Source Book: *My Antonia*
Story Title: Jim and Antonia
Grade: 11–12

1. Jim was a boy.
2. Jim was ten years old.
3. Jim moved to Nebraska to stay with the Burdens.
4. The Burdens were Jim's grandparents.
5. The Shimerdas were the Burdens' new neighbors.
6. The Shimerdas did not have a garden yet.
7. Jim's grandmother decided to pack food into an old wagon and take it to the Shimerdas.
8. Jim and his grandmother packed potatoes and pork to take to the Shimerdas.
9. Jim and his grandmother packed bread, butter, and pumpkin pie to take to the Shimerdas.
10. Jim and his grandmother rode in the wagon to the Shimerdas' house.
11. Jim and his grandmother arrived at the Shimerdas' house.
12. It was a small house built into the side of a hill.
13. Jim and his grandmother met Mr. and Mrs. Shimerda and their daughter Antonia.
14. The Shimerdas did not speak very much English.
15. Jim needed someone to tell him what the Shimerdas said.
16. Jim and Antonia played together in a field while Jim's grandmother stayed with Mr. and Mrs. Shimerda.
17. Antonia wanted to learn the English words for things that she saw.
18. Antonia pointed to the things she wanted to learn how to say.
19. Jim told Antonia the English words for the things she pointed to.
20. Antonia pointed to the sky, a tree, and Jim's eyes.
21. Jim told her the English words for all the things she pointed to.
22. Antonia learned many new English words while she played with Jim.
23. When it was time for Jim and his grandma to leave, Mr. Shimerda asked Jim's grandmother for a favor.
24. Mr. Shimerda asked Jim's grandmother to teach Antonia how to speak English.
25. Jim's grandmother said she would.
26. Antonia was excited to learn more English words.
27. Jim was excited to have a new friend.

B.2. LLM-Generated Story

Example 1

Target Book: *A White Heron*
Story Title: Sylvia and the Heron
Grade: 11–12
Theme Statement: This story explores the importance of protecting nature.
Training Exemplar: Jim and Antonia

1. Sylvia lived with her grandmother in the woods.
2. Sylvia loved the woods and the animals that lived there.
3. One day, Sylvia met a man in the woods.
4. The man was a hunter. He was looking for a white heron.
5. The man asked Sylvia if she had seen the heron.
6. Sylvia knew where the heron lived.
7. The man offered Sylvia money to help him find the heron.
8. Sylvia liked the man, but she loved the heron more.
9. Sylvia climbed a tall tree to see the heron's nest.
10. Sylvia saw the heron flying in the sky.
11. Sylvia felt happy watching the heron.
12. Sylvia decided not to tell the man where the heron was.
13. Sylvia wanted to protect the heron.
14. The man left the woods without finding the heron.
15. Sylvia was glad she kept the heron safe.

Example 2

Target Book: *My Antonia*
Story Title: The New Home
Grade: 11–12
Theme Statement: This story explores family togetherness.
Training Exemplar: Jim and Antonia

1. Anna and her family moved to a new place.
2. Their new home was on a big farm.
3. Anna's father worked hard to build their house.
4. Anna's mother planted a garden.
5. Anna helped her mother plant seeds in the soil.
6. Anna liked to watch the plants grow.
7. Anna's little brother played in the fields.
8. Anna and her brother found many bugs and flowers.
9. Anna's family did not have much money.
10. Anna's mother made clothes from old fabric.
11. Anna learned how to sew from her mother.
12. Anna made a dress for her little brother.
13. Anna's father taught her how to care for the animals.
14. Anna fed the chickens and cows every day.
15. Anna loved her new home on the farm.
16. Anna and her family worked hard together.
17. Anna felt happy and safe with her family.
18. Anna knew they would be okay in their new home.

Example 3

Target Book: *The Great Gatsby*
Story Title: Nick and the Party
Grade: 11–12
Theme Statement: This story explores social connections.
Training Exemplar: Jim and Antonia

1. Nick was a quiet man.
2. Nick lived next to Jay.
3. Jay invited Nick to a party.
4. Nick was excited to go.
5. The party had bright lights.
6. There was music and dancing.
7. Nick met many people.
8. Nick talked to a man named Tom.
9. Tom was loud and funny.
10. Nick liked talking to Tom.
11. Nick also met Daisy.
12. Daisy was kind and friendly.
13. Nick felt happy at the party.
14. Nick made new friends.
15. Nick hoped to go to more parties.