A Multidimensional Evaluation of LLM-Generated Learning Materials for Greek Primary School Learners

Eleftherios Kalogeros[1], Sofia Stamou[1], Markella Kolocheri[1], Thaleia-Dionysia Kolovou[2], Manolis Gergatsoulis[1], Matthew Damigos[1]

[1]Laboratory on Digital Libraries and Electronic Publishing,
Department of Information Science, Ionian University
kalogero@ionio.gr, stamou@ionio.gr, markellakol9121999@gmail.com
manolis@ionio.gr, mgdamigos@ionio.gr

[2]Primary School of Giannades, Corfu, Greece
thalia.k3@gmail.com

Abstract

Large language models (LLMs) could support computer-assisted learning by dynamically generating educational texts adapted to different learner groups. Before such material can be used in learner-facing applications, however, linguistic accessibility, semantic content alignment, and factual reliability need to be evaluated jointly. This study evaluates 16 LLMs on the generation of Greek educational texts for primary-school learners, using Grade 3 students and a cultural heritage topic on Cycladic figurines as a controlled case study. The generated texts were assessed across repeated runs using four complementary dimensions: readability and linguistic complexity, grade-specific lexical accessibility using HelexKids, semantic content alignment with a Grade 3 textbook reference using BGE-M3 embeddings, and factual accuracy based on a teacher-scored rubric. Correlation analysis was additionally used to examine relationships among the evaluation dimensions. The results show substantial differences across models and indicate that lexical accessibility, semantic alignment, and factual accuracy capture complementary rather than interchangeable properties of generated learning material. These findings support multi-criteria evaluation when LLM-generated texts are considered for adaptive and learner-facing educational applications.

1 Introduction

Large language models (LLMs) create new opportunities for dynamically generating and adapting educational materials for learners of different ages, grade levels, and language abilities. In computer-assisted learning environments, such models could support the creation of informational texts tailored to particular learner groups. However, fluent generation alone is not sufficient for educational use. Generated material should also be linguistically appropriate for the intended learners, semantically aligned with the intended instructional content, and factually reliable.

Age-adapted versions of the same informational content can support a variety of educational uses. They may be used in textbooks, supplementary learning resources, educational websites, digital learning environments, and learner-facing applications. Museums and cultural heritage institutions provide one concrete setting in which this type of adaptation is useful, since the same information may need to be presented differently to visitors of different ages and educational levels. Recent work has also explored the use of LLMs to generate and adapt cultural heritage descriptions for different audiences (Dror et al., 2024).

The broader aim of this work is to investigate how LLM-generated educational texts can be evaluated before their use as learning materials for primary-school learners. In the present study, Grade 3 students are used as a case study, and Cycladic figurines are selected as the target topic to represent a cultural heritage learning scenario.

The study follows a three-step process. First, each LLM is prompted to generate a short educational text in Greek on a predefined topic and to include a set of required content elements. Second, the generated texts are evaluated using complementary measures of readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment with a textbook refer-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 160–172, November 19–20, 2026, University of Tartu Library

ence, and factual accuracy. Third, the results are compared across models and repeated generations in order to examine both performance and consistency.

From an ICALL perspective, age- or level-adapted text generation could support the dynamic creation of learning materials tailored to learner characteristics. Such use, however, requires evaluation of linguistic accessibility, semantic content alignment, and factual reliability rather than reliance on fluency or readability alone. The present study examines these requirements through a controlled Greek primary-school case study.

Rather than treating any single linguistic or content-based measure as an indicator of educational suitability, the present study examines these dimensions jointly under a controlled Greek generation setting. The contribution is therefore a multidimensional evaluation of age-adapted learning materials, rather than the introduction of a new readability or simplification method.

This work makes three main contributions. First, it provides a multidimensional evaluation framework for LLM-generated age-adapted learning materials in Greek, considering linguistic complexity, grade-specific lexical accessibility, semantic content alignment, and factual accuracy. Second, it compares 16 open-weight and proprietary LLMs across repeated generations, allowing both output quality and consistency to be examined. Third, it investigates the relationships among evaluation dimensions, showing that linguistic accessibility, semantic alignment, and factual accuracy capture complementary properties of generated educational text.

The study addresses three research questions. First, it examines how different LLMs compare when generating Grade 3 Greek educational texts in terms of readability, lexical accessibility, semantic content alignment, and factual accuracy (RQ1). Second, it investigates how consistent these evaluation outcomes are across repeated generations (RQ2). Third, it explores whether models that produce more lexically accessible texts also achieve higher factual accuracy and semantic content alignment (RQ3).

2 Related Work

Recent work has investigated whether large language models can generate or adapt educational content to the reading level of younger learners. Valentini et al. (2023) evaluated LLM-generated children's stories with respect to lexical and readability levels and found that current models do not consistently restrict their vocabulary to levels appropriate for younger age groups. More recently, Oh et al. (2026) introduced a framework for generating grade-specific educational content across multiple educational levels. Their results showed that explicit adaptation to grade level can improve alignment with learners' expected comprehension level while preserving response accuracy. These studies motivate the need to evaluate age-appropriate generation using multiple linguistic and content-oriented criteria rather than relying only on model fluency.

LLMs have also been examined as tools for educational text simplification. Watanabe (2025) evaluated LLM-based rewriting of educational texts using measures of linguistic sophistication, cohesion, and readability, showing that successful simplification depends on both the model and the prompting strategy. Day et al. (2025) similarly evaluated the simplification of informational passages to targeted grade levels and considered both readability accuracy and semantic consistency. Their findings showed substantial variability across models and prompting approaches, particularly when adapting material to lower grade levels. Together, these studies highlight the difficulty of simplifying educational text while maintaining the intended informational content.

In cultural heritage settings, LLMs have increasingly been explored as tools for adapting and personalizing information for different audiences. Dror et al. (2024) investigated the use of LLMs to support the generation of cultural heritage descriptions tailored to different users. Dibitonto et al. (2026) used GPT-4 to generate museum audio guides adapted to different visitor profiles and evaluated the resulting content through user studies and expert assessment. Their findings highlighted both the potential of personalized narratives and the importance of factual accuracy, domain terminology, and curatorial supervision. Ferrato (2025) further explored the use of LLMs as personalized guides in cultural heritage learning environments, where information can be adapted to different visitor profiles and learning needs.

Stamou et al. (2027) examined LLM-based narrative generation from structured cultural heritage data represented using CIDOC CRM. They

compared 34 freely available LLMs on their ability to transform RDF descriptions into coherent museum-oriented narratives and evaluated the outputs using complementary measures of semantic similarity, content coverage and focus, BERTScore-based alignment, and readability. Their study further illustrates the value of multidimensional evaluation for LLM-generated cultural heritage text, although it focuses on the verbalization of structured knowledge rather than on age-adapted learning materials.

Existing work has primarily focused on English educational texts, general text simplification, or the generation and personalization of cultural heritage information for different audiences. The present study extends these research directions to the generation of Greek educational texts for primary-school learners. Specifically, we use Grade 3 students and a cultural heritage topic as a case study and compare multiple open-weight and proprietary LLMs across readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment, factual accuracy, and generation consistency.

3 Problem Formulation and Evaluation Framework

The task considered in this study is the generation of an age-appropriate educational text from a predefined topic and a set of required content elements. The generated text should be understandable for the intended learner group while accurately representing the essential information associated with the topic.

Evaluating this type of output requires more than a single measure. A text may use simple and familiar vocabulary but omit important information. In contrast, another text may be factually accurate but remain linguistically complex for the intended audience. Similarly, a text may show substantial semantic overlap with a reference passage without being fully correct in all of its factual details.

For this reason, the evaluation is organized around four complementary dimensions. Readability and linguistic complexity characterize whether the linguistic form of a generated text may be unnecessarily demanding for the target learners. Lexical accessibility examines whether the vocabulary corresponds to the intended grade level. Semantic content alignment assesses the extent to which the generated text reflects the intended instructional content represented by the reference passage. Finally, factual accuracy examines whether the required information is represented correctly and whether the generated material can be considered reliable for educational use.

These dimensions are examined separately rather than combined into a single overall score. This makes it possible to identify different strengths and weaknesses across models and to determine whether strong performance in one dimension is accompanied by strong performance in others. In particular, a model that produces linguistically accessible text may not necessarily produce semantically aligned or factually reliable content.

Repeated generations are additionally used to examine the consistency of model behaviour across outputs. This is important for learner-facing applications, since the suitability of a model should not depend on a single generation. The overall framework therefore considers both the quality of individual outputs and the stability of model performance across repeated runs.

4 Experimental Setup

This section describes the models, generation task, and run configuration used in the evaluation. All models were tested under the same task specification in order to support a consistent comparison of their ability to generate Grade 3 educational texts in Greek.

4.1 Model Selection

The evaluation included 16 large language models from multiple model families, covering both open-weight and proprietary systems and a broad range of model scales and architectural configurations. The selected models include both dense and mixture-of-experts (MoE) architectures, with publicly disclosed parameter counts ranging from 4B to 117B. Three proprietary systems—GPT-5.4, Claude Sonnet 5, and Gemini 3.8 Flash—were also included. Models with reasoning capabilities were allowed to use reasoning when supported, reflecting their intended operational configuration. For MoE models, total and active parameter counts were treated separately rather than interpreted as directly equivalent to the parameter counts of dense models. Table 1 presents the selected models and their main characteristics.

Table 1: Large language models included in the evaluation.

Model	Type	Total Params.	Active Params.
Qwen3 8B	Open-weight / Dense	8.2B	8.2B
Qwen3 14B	Open-weight / Dense	14.8B	14.8B
Qwen3 32B	Open-weight / Dense	32.8B	32.8B
Qwen3 30B-A3B	Open-weight / MoE	30.5B	3.3B
Llama 3.1 8B	Open-weight / Dense	8B	8B
Llama 3.3 70B	Open-weight / Dense	70B	70B
Gemma 3 4B	Open-weight / Dense	4B	4B
Gemma 3 12B	Open-weight / Dense	12B	12B
Gemma 3 27B	Open-weight / Dense	27B	27B
Mistral Small 3.2 24B	Open-weight / Dense	24B	24B
DeepSeek V4.1 Flash	Open-weight / MoE	N/D	N/D
GPT-OSS 20B	Open-weight / MoE	21B	3.6B
GPT-OSS 120B	Open-weight / MoE	117B	5.1B
GPT-5.4	Proprietary	N/D	N/D
Claude Sonnet 5	Proprietary	N/D	N/D
Gemini 3.8 Flash	Proprietary	N/D	N/D

Note. N/D indicates information not publicly disclosed. For mixture-of-experts (MoE) models, total and active parameter counts are reported separately where available.

4.2 Generation Task

All models received the same prompt in Greek. The task requested a 90–110-word informational text about Cycladic figurines for Grade 3 students aged 8–9 years. For transparency, an English translation of the prompt is reproduced below; the original Greek version was used in all experiments.

> Write a short informational text in Greek for Grade 3 primary-school students aged 8–9 years about Cycladic figurines. The text should be approximately 90–110 words long and explain what Cycladic figurines are, what material they were made from, what they usually represent, and where they have been found. Use clear, natural, age-appropriate language and historically accurate information. Return only the final continuous text, without a title, bullet points, or a numbered list.

The generated texts were subsequently evaluated from several complementary perspectives. Readability and linguistic-complexity measures were used to describe the linguistic form of each text, while Grade 3 HelexKids-based measures were used to estimate lexical accessibility for the target learners. BGE-M3 cosine similarity was used to assess semantic content alignment with the corresponding textbook passage, and a predefined factual rubric was used to evaluate whether the required content was correctly represented. Repeated generation runs were additionally used to examine the consistency of these measures across outputs.

4.3 Run Configuration

Each model was evaluated across four runs. For models supporting temperature control, Run 1 used $T = 0$ with seed 42 and served as the low-temperature reference condition. This setting was intended to reduce generation variability, but Run 1 is not assumed to be strictly deterministic, since hosted LLM systems may exhibit nondeterminism even under nominally deterministic generation settings (Atıl et al., 2025). Runs 2–4 used $T = 0.7$ with seeds 101, 202, and 303, respectively. Top-p was fixed at 1.0. Temperature and seed parameters were sent only when supported by the corresponding model.

GPT-5.4 and Claude Sonnet 5 did not expose temperature control through the used API interface and were therefore evaluated under their native generation settings. Their Run 1 results are retained for completeness in the general model comparison but are not interpreted as deterministic. These models are excluded from analyses specifically concerning the effect of temperature.

Reasoning was enabled for models that supported it, while the reasoning content itself was not included in the evaluation. A maximum completion budget of 4,000 tokens was used for all models. Temporary API failures, including rate limits, timeouts, connection errors, server-side errors, and empty outputs, were retried automatically up to three times.

5 Evaluation Metrics

The generated texts were evaluated across four dimensions: readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment, and factual accuracy. Together, these measures provide a multidimensional assessment of the suitability and accuracy of the generated educational texts.

Table 2: Abbreviations used for linguistic-complexity and readability-related indicators.

Abbreviation	Indicator
W/S	Words per sentence
Easy/100	Easy words per 100 words
MultiSyl/100	Multisyllabic words per 100 words
Guiraud R	Guiraud's lexical-diversity index
Affix/100	Words with prefixes/suffixes per 100 words
Middle/100	Middle-voice forms per 100 words
Conj/100	Conjunctions per 100 words
LAdv/100	Learned adverbial forms per 100 words
Part/100	Participles per 100 words

5.1 Linguistic Complexity and Readability-Related Indicators

Readability was assessed using the online readability tool provided by the Centre for the Greek Language (Centre for the Greek Language, 2025; Mikros and Voskaki, 2021). Each generated text was processed with the tool, and the resulting linguistic indicators were recorded for analysis.

The tool was originally developed for automatic readability analysis of Modern Greek and uses a range of stylometric and linguistic indicators associated with text difficulty (Mikros and Voskaki, 2021). In the present study, we use the individual linguistic indicators reported by the tool rather than its overall CEFR-level prediction.

Words per sentence (W/S) represents the mean number of words per sentence. Easy words per 100 words (Easy/100) represents the proportion of words classified as easy according to the lexical classification implemented by the readability tool; this classification was used as provided by the tool and was not modified in the present study. Words with more than two syllables per 100 words (MultiSyl/100) captures the frequency of multisyllabic words. Guiraud's R (Guiraud R) measures lexical diversity as the number of distinct lexical types divided by the square root of the total number of word tokens.

The tool also reports normalized frequencies per 100 words for several morphological and morphosyntactic features. These include words with prefixes or suffixes (Affix/100), middle-voice morphology (Middle/100), conjunctions (Conj/100), learned adverbial forms (LAdv/100), and participles (Part/100).

Together, these measures capture complementary lexical, morphological, and morphosyntactic characteristics associated with text readability and linguistic complexity. Because the indicators describe different properties of the text, they were interpreted separately rather than combined into a single composite readability score.

These indicators are treated as linguistic features associated with readability rather than as direct or interchangeable measures of readability itself. Their relationship with perceived text difficulty is not assumed to be linear, and lower values are not uniformly interpreted as preferable across all indicators. The measures are therefore examined individually rather than combined into a single readability score.

5.2 Grade-Specific Lexical Accessibility

Lexical accessibility was assessed using the Grade 3 HelexKids frequency list (Terzopoulos et al., 2017). HelexKids is a grade-based lexical frequency database derived from 116 primary-school textbooks used in Greece and Cyprus. The complete resource contains 1,355,265 tokens and 68,692 non-lemmatized word types across Grades 1–6, while the Grade 3 component contains 21,193 word types. The resource therefore provides a curriculum-based estimate of the written vocabulary encountered by Greek-speaking primary-school learners.

Matching was performed at the surface-form level without lemmatization. Both the generated texts and the HelexKids entries were normalized to Unicode NFC and lowercased before matching, while Greek accents were retained. Punctuation was excluded during tokenization. As a result, inflected forms were treated as distinct lexical items and were counted as matches only when the corresponding surface form was present in the Grade 3 lexicon.

Lexical coverage (Coverage) represents the percentage of text tokens found in the Grade 3 HelexKids lexicon. Mean log frequency per million (Mean logF/m) represents the average log-transformed frequency per million of the matched words. Higher Coverage indicates that a larger proportion of the generated text consists of vo-

cabulary attested in the Grade 3 lexicon, whereas higher Mean logF/m indicates that the matched words are, on average, more frequent within that lexical resource.

The out-of-vocabulary rate (OOV) was calculated as $100 - \text{Coverage}$ and was retained as a complementary descriptive measure. Because Coverage and OOV are exact complements, OOV was not treated as an independent evaluation dimension.

5.3 Semantic Content Alignment

Semantic content alignment with the corresponding textbook reference (Papadimitriou and Savvaki, 2026) was assessed using cosine similarity between text embeddings generated with BAAI's BGE-M3 model (Chen et al., 2024). Higher cosine similarity values indicate greater semantic overlap between the generated text and the textbook reference (Papadimitriou and Savvaki, 2026). The measure was used to assess semantic content alignment rather than lexical or surface-form similarity.

5.4 Factual Accuracy

Factual accuracy was evaluated using a topic-specific rubric comprising K core factual criteria. Each criterion (F_i) was scored as 0 (incorrect), 1 (partially correct), or 2 (correct). The overall Factual Accuracy Score was normalized to a 0–1 scale as

$$\text{Factual Accuracy Score} = \frac{\sum_{i=1}^{K} F_i}{X}, \qquad X = 2K,$$

where X represents the maximum attainable score for the corresponding rubric. Higher values indicate greater factual accuracy.

For example, for the topic of Cycladic figurines, four core factual criteria were evaluated: (F1) what Cycladic figurines are, (F2) their main material, (F3) what they typically represent, and (F4) where they have been found. Thus, $K = 4$ and $X = 8$, giving

$$\text{Factual Accuracy Score} = \frac{F_1 + F_2 + F_3 + F_4}{8}.$$

The factual criteria were manually scored by a primary-school teacher according to the predefined rubric, with each criterion rated as incorrect, partially correct, or correct. The generated texts were presented to the evaluator in randomized order without information about the model or generation condition associated with each output.

5.5 Correlation Analysis

To examine relationships among the evaluation dimensions, Spearman's rank correlation coefficient (ρ) was calculated using model-level mean values across Runs 2–4. Correlations were examined between the two HelexKids lexical accessibility measures (Coverage and Mean logF/m), semantic content alignment measured using BGE-M3 cosine similarity, and Factual Accuracy.

Spearman correlation was selected because of the small number of models ($n = 16$) and the bounded and potentially non-normal distribution of several evaluation measures. The analysis was exploratory, and correlation coefficients were interpreted together with their corresponding p-values. Given the exploratory nature of the analysis, reported p-values are unadjusted and should be interpreted cautiously.

6 Experimental Results

This section presents the evaluation results across four dimensions: readability and linguistic complexity, Grade 3 lexical accessibility, semantic content alignment, and factual accuracy. Run 1 results are reported separately from Runs 2–4 in order to examine both model performance and run-to-run consistency.

The generation prompt requested texts of approximately 90–110 words. Across the 64 generated texts, 32 outputs (50.0%) satisfied this range, 24 (37.5%) were shorter than 90 words, and 8 (12.5%) exceeded 110 words. Generated texts ranged from 9 to 128 words. Length compliance was not used as an exclusion criterion, and all valid generations were retained in the evaluation.

6.1 Linguistic Complexity and Readability-Related Indicators

Table 3 presents the readability and linguistic complexity measures obtained for Run 1. Considerable variation was observed across models. Mean sentence length (W/S) ranged from 6.67 words for Llama 3.1 8B to 19.60 words for Gemma 3 12B. The proportion of easy words (Easy/100) ranged from 40.63 to 55.77, with Gemma 3 27B producing the highest value.

Gemma 3 12B exhibited the lowest proportion

of multisyllabic words (MultiSyl/100 = 29.59) and the lowest frequency of words containing prefixes or suffixes (Affix/100 = 20.41). Lexical diversity, measured using Guiraud's R, ranged from 5.06 for Llama 3.1 8B to 8.38 for GPT-OSS 120B. The remaining morphological and morphosyntactic measures also differed across models, indicating variation not only in vocabulary and sentence length but also in morphological and syntactic complexity.

Across Runs 2–4, the linguistic and readability-related indicators were not equally stable across models. Table 4 summarizes the core readability measures across Runs 2–4. The Gemma family generally exhibited comparatively accessible lexical characteristics. Gemma 3 4B achieved the highest mean Easy/100 value (51.18 ± 2.08), while Gemma 3 12B exhibited the lowest mean MultiSyl/100 value (34.07 ± 2.53). GPT-OSS 120B showed the highest mean Guiraud's R (8.09 ± 0.16).

Table 5 reports the corresponding morphological and morphosyntactic measures. Gemma 3 27B exhibited the lowest mean Affix/100 value (23.22 ± 0.61). Differences were also observed in middle-voice morphology, conjunction use, learned adverbial forms, and participles, confirming that the models differed in multiple dimensions of linguistic complexity.

Variation was also observed in lexical diversity. GPT-OSS 120B exhibited the highest mean Guiraud's R across the repeated runs (8.09 ± 0.16), whereas Llama 3.1 8B produced the lowest mean value (5.16 ± 1.91). DeepSeek V4.1 Flash showed particularly stable lexical diversity across repeated generations (7.85 ± 0.09). In contrast, Llama 3.3 70B exhibited substantial variability in the proportion of multisyllabic words (43.90 ± 9.76).

Overall, the models exhibited different linguistic profiles across the individual indicators, and differences on one measure did not necessarily correspond to similar patterns on the others. Because the indicators capture distinct lexical, morphological, and syntactic properties and do not share a common direction of desirability, they were interpreted separately rather than combined into a single composite score.

6.2 Grade-Specific Lexical Accessibility

Table 6 presents the grade-specific lexical accessibility results obtained using the Grade 3 HelexKids frequency list. Coverage represents the percentage of tokens occurring in the Grade 3 lexicon, whereas Mean logF/m reflects the mean log-transformed frequency per million of the matched words. Higher values on both measures indicate greater correspondence with the Grade 3 lexical environment.

In Run 1, lexical coverage ranged from 75.00% for Llama 3.1 8B to 98.96% for DeepSeek V4.1 Flash. The textbook reference, which was used as a descriptive benchmark but was not included as a model in the table, yielded a coverage of 87.13% and a Mean logF/m of 3.667. Most models therefore produced Run 1 texts with lexical coverage above the textbook benchmark. However, the two HelexKids measures did not always vary in parallel. For example, DeepSeek V4.1 Flash achieved very high coverage (98.96%) but a Mean logF/m of 3.698, only slightly above the textbook value. In contrast, Llama 3.1 8B exhibited the lowest coverage (75.00%) but the highest Run 1 Mean logF/m (4.101). This indicates that the words successfully matched in the HelexKids lexicon could be highly frequent even when a comparatively large proportion of the complete text remained unmatched.

Across Runs 2–4, lexical accessibility was relatively stable for several models but more variable for others. Among models for which $T = 0.7$ was controlled, Gemma 3 27B achieved the highest mean coverage across Runs 2–4 ($97.75 \pm 2.56\%$), followed by Gemma 3 12B ($97.02 \pm 1.62\%$) and DeepSeek V4.1 Flash ($96.30 \pm 3.21\%$). In contrast, Llama 3.1 8B showed the greatest variation in coverage ($88.12 \pm 10.82\%$), indicating substantial run-to-run differences in the proportion of vocabulary matched by the Grade 3 HelexKids lexicon.

Mean logF/m was generally more stable than coverage. Among the repeated runs, Gemma 3 4B (3.979 ± 0.050) and GPT-OSS 20B (3.950 ± 0.033) exhibited particularly high mean values, while Gemma 3 12B showed very little variation across runs (3.805 ± 0.004). These results suggest that lexical coverage and lexical frequency capture complementary properties: a model may use a large proportion of Grade 3-attested vocabulary without necessarily selecting the most frequent words within that vocabulary.

Overall, most models produced texts with HelexKids coverage above the 87.13% observed

Table 3: Readability and linguistic complexity for Run 1. For compact presentation, shortened column labels are used; Table 2 provides the corresponding full abbreviations and metric definitions.

Model	W/S	Easy	MultiSyl	Guiraud R	Affix	Middle	Conj	LAdv	Part
Qwen3 8B	13.20	40.91	46.97	7.14	37.88	4.55	7.58	4.55	1.52
Qwen3 14B	15.40	41.56	42.86	7.41	35.06	9.09	10.39	1.30	1.30
Qwen3 32B	13.00	46.15	37.18	7.47	26.92	6.41	8.97	1.28	1.28
Qwen3 30B-A3B	12.83	42.86	44.16	7.29	37.66	3.90	7.79	5.19	0.00
Llama 3.1 8B	6.67	50.00	37.50	5.06	32.50	7.50	10.00	0.00	0.00
Llama 3.3 70B	15.33	52.17	41.30	5.46	30.43	6.52	13.04	2.17	0.00
Gemma 3 4B	16.00	53.13	34.38	7.51	21.09	1.56	13.28	0.78	0.00
Gemma 3 12B	19.60	53.06	29.59	7.58	20.41	2.04	14.29	1.02	2.04
Gemma 3 27B	17.33	55.77	38.46	6.86	23.08	2.88	12.50	0.96	0.96
Mistral Small 3.2 24B	15.33	44.57	39.13	7.61	30.43	4.35	14.13	1.09	0.00
DeepSeek V4.1 Flash	10.67	40.63	38.54	7.86	33.33	2.08	8.33	1.04	2.08
GPT-OSS 20B	15.57	43.12	46.79	7.57	28.44	6.42	14.68	0.00	0.00
GPT-OSS 120B	12.88	40.78	44.66	8.38	35.92	4.85	7.77	2.91	0.00
GPT-5.4	15.17	45.05	42.86	7.02	32.97	1.10	13.19	3.30	1.10
Claude Sonnet 5	17.17	46.60	39.81	7.39	27.18	0.97	12.62	1.94	0.97
Gemini 3.8 Flash	17.33	47.12	45.19	7.84	33.65	0.96	9.62	0.96	0.96

Table 4: Core readability measures across Runs 2–4. Values are reported as mean ± standard deviation.

Model	W/S	Easy/100	MultiSyl/100	Guiraud R
Qwen3 8B	14.38 ± 4.43	43.17 ± 4.91	39.85 ± 0.96	7.22 ± 0.15
Qwen3 14B	14.99 ± 3.57	42.33 ± 3.30	44.77 ± 2.52	7.39 ± 0.27
Qwen3 32B	15.78 ± 2.13	45.19 ± 3.15	37.30 ± 4.40	7.49 ± 0.26
Qwen3 30B-A3B	14.90 ± 1.17	44.67 ± 2.59	37.59 ± 1.55	7.53 ± 0.08
Llama 3.1 8B	10.43 ± 1.29	45.63 ± 8.60	41.92 ± 2.18	5.16 ± 1.91
Llama 3.3 70B	12.15 ± 2.70	44.65 ± 5.37	43.90 ± 9.76	6.14 ± 1.10
Gemma 3 4B	14.35 ± 2.43	51.18 ± 2.08	37.23 ± 4.27	7.98 ± 0.22
Gemma 3 12B	16.86 ± 2.25	46.84 ± 1.19	34.07 ± 2.53	7.65 ± 0.23
Gemma 3 27B	15.91 ± 0.94	51.02 ± 1.33	36.62 ± 3.41	7.60 ± 0.45
Mistral Small 3.2	14.06 ± 2.27	45.71 ± 1.62	43.73 ± 0.86	7.22 ± 0.28
DeepSeek V4.1 Flash	13.56 ± 2.02	43.00 ± 1.74	38.43 ± 1.67	7.85 ± 0.09
GPT-OSS 20B	16.66 ± 3.27	45.45 ± 3.06	37.56 ± 3.45	7.36 ± 0.33
GPT-OSS 120B	13.69 ± 1.91	40.05 ± 2.19	46.65 ± 4.04	8.09 ± 0.16
GPT-5.4	15.06 ± 1.87	46.82 ± 0.60	39.76 ± 1.86	7.40 ± 0.43
Claude Sonnet 5	17.65 ± 1.30	44.43 ± 2.27	40.65 ± 3.98	7.62 ± 0.27
Gemini 3.8 Flash	16.22 ± 0.54	42.49 ± 2.90	46.63 ± 3.23	8.01 ± 0.39

Table 5: Morphological and morphosyntactic measures across Runs 2–4. Values are reported as mean ± standard deviation.

Model	Affix/100	Middle/100	Conj/100	LAdv/100	Part/100
Qwen3 8B	35.05 ± 1.29	3.05 ± 0.87	11.27 ± 0.89	1.72 ± 0.68	0.88 ± 0.77
Qwen3 14B	33.78 ± 5.32	5.69 ± 1.65	9.82 ± 0.89	1.23 ± 0.08	0.84 ± 0.73
Qwen3 32B	30.75 ± 3.54	3.19 ± 0.60	10.77 ± 1.77	1.20 ± 0.05	1.62 ± 0.77
Qwen3 30B-A3B	30.74 ± 4.69	3.20 ± 1.05	9.90 ± 2.18	3.13 ± 1.72	0.69 ± 1.19
Llama 3.1 8B	28.66 ± 5.76	4.89 ± 4.23	11.62 ± 2.65	1.72 ± 1.86	0.00 ± 0.00
Llama 3.3 70B	32.26 ± 8.85	5.65 ± 1.48	8.08 ± 4.94	2.06 ± 1.02	0.44 ± 0.76
Gemma 3 4B	25.34 ± 2.25	2.90 ± 0.47	10.56 ± 1.72	1.58 ± 0.79	0.53 ± 0.46
Gemma 3 12B	25.03 ± 2.13	2.66 ± 2.08	12.75 ± 2.73	1.33 ± 0.53	1.67 ± 0.59
Gemma 3 27B	23.22 ± 0.61	3.15 ± 0.52	12.32 ± 0.33	1.14 ± 0.49	1.43 ± 0.51
Mistral Small 3.2	29.72 ± 3.01	2.73 ± 0.90	12.03 ± 3.31	1.16 ± 0.12	0.42 ± 0.72
DeepSeek V4.1 Flash	34.06 ± 3.56	3.25 ± 2.00	12.07 ± 0.89	2.14 ± 1.12	2.14 ± 1.12
GPT-OSS 20B	31.61 ± 5.77	5.55 ± 1.45	10.48 ± 1.17	1.33 ± 1.47	0.00 ± 0.00
GPT-OSS 120B	35.18 ± 2.24	7.17 ± 2.82	9.76 ± 3.12	2.16 ± 1.18	1.46 ± 1.27
GPT-5.4	32.00 ± 2.62	1.75 ± 0.57	11.25 ± 0.66	2.45 ± 0.48	1.06 ± 0.05
Claude Sonnet 5	29.84 ± 1.37	1.42 ± 1.65	10.41 ± 1.72	2.32 ± 0.38	0.36 ± 0.62
Gemini 3.8 Flash	36.33 ± 5.55	1.04 ± 1.81	8.25 ± 1.98	1.38 ± 0.63	1.36 ± 0.54

for the textbook reference in at least one generation condition. This result should be interpreted specifically as greater lexical correspondence with the Grade 3 HelexKids corpus rather than as evidence that the generated texts were superior to the textbook overall. Lexical accessibility constitutes only one dimension of educational text quality and should therefore be considered together with read-

Table 6: Grade-specific lexical accessibility according to the Grade 3 HelexKids frequency list. Values for Runs 2–4 are reported as mean ± standard deviation.

	Run 1		Runs 2–4	
Model	Coverage (%)	Mean logF/m	Coverage (%)	Mean logF/m
Qwen3 8B	78.79	3.878	89.58 ± 3.58	3.846 ± 0.092
Qwen3 14B	88.31	3.876	90.11 ± 3.73	3.850 ± 0.086
Qwen3 32B	87.34	3.837	92.41 ± 0.50	3.897 ± 0.133
Qwen3 30B-A3B	88.31	3.755	90.82 ± 1.18	3.885 ± 0.090
Llama 3.1 8B	75.00	4.101	88.12 ± 10.82	3.883 ± 0.102
Llama 3.3 70B	89.13	4.074	91.79 ± 1.46	3.821 ± 0.247
Gemma 3 4B	96.88	3.998	92.62 ± 2.51	3.979 ± 0.050
Gemma 3 12B	95.92	3.921	97.02 ± 1.62	3.805 ± 0.004
Gemma 3 27B	98.08	3.994	97.75 ± 2.56	3.901 ± 0.055
Mistral Small 3.2 24B	90.22	3.958	89.82 ± 3.41	3.911 ± 0.167
DeepSeek V4.1 Flash	98.96	3.698	96.30 ± 3.21	3.798 ± 0.095
GPT-OSS 20B	90.83	3.838	86.68 ± 5.36	3.950 ± 0.033
GPT-OSS 120B	89.32	3.621	87.98 ± 1.14	3.675 ± 0.079
GPT-5.4	97.80	3.758	98.95 ± 1.08	3.783 ± 0.020
Claude Sonnet 5	93.20	3.851	93.36 ± 1.10	3.778 ± 0.115
Gemini 3.8 Flash	95.19	3.796	92.14 ± 1.42	3.755 ± 0.016

ability, semantic content alignment, and factual accuracy.

6.3 Semantic Content Alignment

Table 7 presents the cosine similarity between each generated text and the textbook reference using BGE-M3 embeddings. Higher values indicate greater semantic content alignment with the reference text.

In Run 1, cosine similarity ranged from 0.710 for Gemma 3 4B to 0.821 for Qwen3 8B. Qwen3 14B (0.811) and GPT-5.4 (0.800) also showed relatively high semantic alignment with the textbook reference.

Across Runs 2–4, Mistral Small 3.2 24B achieved the highest mean cosine similarity (0.794 ± 0.009), followed by Qwen3 8B (0.783 ± 0.006) and DeepSeek V4.1 Flash (0.781 ± 0.021). Most models showed relatively small run-to-run variation, although Llama 3.1 8B exhibited greater variability (0.707 ± 0.078).

Overall, the results show that semantic alignment with the textbook reference varied across models and across repeated generations. These scores should be interpreted as measures of semantic content overlap rather than factual correctness or overall educational quality.

6.4 Factual Accuracy

Table 8 presents the normalized Factual Accuracy Scores for Run 1 and Runs 2–4. Scores range from 0 to 1, with higher values indicating greater factual accuracy.

Substantial differences were observed across models. In Run 1, DeepSeek V4.1 Flash, GPT-5.4, Claude Sonnet 5, Gemini 3.8 Flash, and Qwen3 14B achieved the maximum score of 1.00. By contrast, Llama 3.1 8B and GPT-OSS 20B obtained markedly lower scores of 0.50 and 0.25, respectively.

Across Runs 2–4, DeepSeek V4.1 Flash, GPT-5.4, and Gemini 3.8 Flash obtained mean factual-accuracy scores of 1.000 ± 0.000. Claude Sonnet 5 also showed consistently high performance (0.958 ± 0.072). Several other models exhibited moderate run-to-run variation, including Qwen3 30B-A3B (0.833 ± 0.144), GPT-OSS 120B (0.792 ± 0.144), and Qwen3 32B (0.750 ± 0.125). The lowest mean performance across Runs 2–4 was observed for Llama 3.1 8B (0.208 ± 0.191) and GPT-OSS 20B (0.292 ± 0.191).

Table 7: Semantic content alignment with the textbook reference using BGE-M3 cosine similarity. Values for Runs 2–4 are reported as mean ± standard deviation.

Model	Run 1	Runs 2–4
Qwen3 8B	0.821	0.783 ± 0.006
Qwen3 14B	0.811	0.777 ± 0.010
Qwen3 32B	0.790	0.731 ± 0.031
Qwen3 30B-A3B	0.758	0.768 ± 0.024
Llama 3.1 8B	0.767	0.707 ± 0.078
Llama 3.3 70B	0.778	0.774 ± 0.038
Gemma 3 4B	0.710	0.744 ± 0.010
Gemma 3 12B	0.733	0.709 ± 0.019
Gemma 3 27B	0.718	0.705 ± 0.018
Mistral Small 3.2 24B	0.787	0.794 ± 0.009
DeepSeek V4.1 Flash	0.776	0.781 ± 0.021
GPT-OSS 20B	0.747	0.771 ± 0.020
GPT-OSS 120B	0.764	0.764 ± 0.042
GPT-5.4	0.800	0.776 ± 0.009
Claude Sonnet 5	0.770	0.746 ± 0.008
Gemini 3.8 Flash	0.763	0.735 ± 0.011

Table 8: Factual Accuracy Scores by model. Run 1 values are reported separately; values for Runs 2–4 are reported as mean ± standard deviation.

Model	Run 1	Runs 2–4
Qwen3 8B	0.625	0.583 ± 0.072
Qwen3 14B	1.000	0.833 ± 0.072
Qwen3 32B	0.875	0.750 ± 0.125
Qwen3 30B-A3B	0.750	0.833 ± 0.144
Llama 3.1 8B	0.500	0.208 ± 0.191
Llama 3.3 70B	0.875	0.875 ± 0.000
Gemma 3 4B	0.625	0.625 ± 0.125
Gemma 3 12B	0.750	0.833 ± 0.072
Gemma 3 27B	0.875	0.917 ± 0.072
Mistral Small 3.2 24B	0.875	0.917 ± 0.072
DeepSeek V4.1 Flash	1.000	1.000 ± 0.000
GPT-OSS 20B	0.250	0.292 ± 0.191
GPT-OSS 120B	0.875	0.792 ± 0.144
GPT-5.4	1.000	1.000 ± 0.000
Claude Sonnet 5	1.000	0.958 ± 0.072
Gemini 3.8 Flash	1.000	1.000 ± 0.000

Overall, the factual-accuracy results revealed a clear separation between models that consistently reproduced the required core information and models that introduced substantial factual errors or omitted key facts. Importantly, high lexical accessibility did not always correspond to high factual accuracy. For example, GPT-OSS 20B achieved relatively high HelexKids lexical coverage but substantially lower factual-accuracy scores. This finding supports the use of factual accuracy as a separate evaluation dimension rather than inferring content reliability from lexical accessibility alone.

6.5 Relationships Between Evaluation Dimensions

Table 9 presents Spearman correlations between lexical accessibility, semantic content alignment, and factual accuracy using model-level means across Runs 2–4. Coverage showed a positive association with factual accuracy ($\rho = 0.648$, $p = 0.007$). In contrast, Coverage was not significantly associated with BGE-M3 cosine similarity ($\rho = -0.262$, $p = 0.327$).

Mean logF/m showed a negative association with factual accuracy ($\rho = -0.502$, $p = 0.048$), whereas its association with BGE-M3 cosine similarity was negligible ($\rho = -0.050$, $p = 0.854$).

BGE-M3 cosine similarity showed only a weak positive association with factual accuracy ($\rho = 0.191$, $p = 0.479$), which was not statistically significant. This suggests that greater semantic overlap with the textbook reference does not necessarily correspond to higher accuracy on the predefined factual criteria.

Overall, the correlation results suggest that lexical accessibility, semantic content alignment, and factual accuracy capture partly distinct properties of the generated texts rather than interchangeable indicators of educational suitability.

Table 9: Spearman correlations between evaluation measures using model-level means across Runs 2–4 ($n = 16$).

Relationship	ρ	p
Coverage – Factual Accuracy	0.648	0.007
Coverage – BGE-M3 cosine	-0.262	0.327
Mean logF/m – Factual Accuracy	-0.502	0.048
Mean logF/m – BGE-M3 cosine	-0.050	0.854
BGE-M3 cosine – Factual Accuracy	0.191	0.479

7 Discussion

The results show that the evaluated LLMs differed substantially in their ability to generate Grade 3 educational texts in Greek. No single model showed consistently strong performance across all evaluation dimensions, confirming that the suitability of generated learning material is multidimensional. Readability, lexical accessibility, semantic content alignment, and factual accuracy therefore need to be considered jointly rather than inferred from a single metric.

The readability analysis revealed clear differences in sentence length, lexical difficulty, lexical diversity, and morphological complexity. The Gemma 3 family generally exhibited a comparatively accessible profile on several readability indicators, particularly Easy/100, MultiSyl/100, and Affix/100. However, values suggesting greater accessibility on individual indicators did not imply uniformly lower linguistic complexity across all metrics. This supports the decision to interpret the individual indicators separately rather than combining them into a single readability score.

Semantic content alignment also varied across models, while most systems showed relatively limited run-to-run variation. Importantly, higher BGE-M3 similarity should be interpreted as greater semantic overlap with the textbook reference rather than as evidence of factual correctness. The weak and non-significant association between BGE-M3 similarity and factual accuracy further supports this distinction. Semantic overlap with the textbook reference and correctness of the required factual elements therefore appear to capture different properties of the generated texts.

The HelexKids analysis further showed that many models produced vocabulary with high correspondence to the Grade 3 lexical resource. Nevertheless, lexical accessibility did not necessarily imply factual reliability. GPT-OSS 20B, for example, produced relatively accessible vocabulary while obtaining substantially lower factual-accuracy scores. This illustrates an important distinction for learner-facing material: text can appear linguistically appropriate for the target grade while still omitting or incorrectly representing important content.

The practical implication is that lexical or readability-based screening cannot replace independent content validation in an LLM-based learning-material generation pipeline.

The correlation analysis provides further evidence that the evaluation dimensions should not be treated as interchangeable. Lexical Coverage was positively associated with factual accuracy, but it was not positively associated with semantic content alignment. Mean logF/m also did not show a clear positive relationship with semantic alignment. These findings suggest that linguistic accessibility alone cannot serve as a general proxy for the overall educational suitability of generated text. In particular, a model that uses grade-appropriate vocabulary is not necessarily producing content that is semantically aligned or factually reliable.

Factual accuracy showed substantial differences across models. DeepSeek V4.1 Flash, GPT-5.4, and Gemini 3.8 Flash achieved consistently high factual scores across repeated generations, whereas Llama 3.1 8B and GPT-OSS 20B showed considerably lower performance. The repeated-run analysis also revealed differences in stability across models, indicating that evaluation based on a single generated output may provide an incomplete picture of model behaviour.

These findings have implications for LLM-supported learning-material generation and ICALL-oriented applications. Selecting a model only on the basis of readability or lexical accessibility may be insufficient, since linguistically accessible outputs may still omit, distort, or incorrectly represent important educational content. A practical generation pipeline should therefore consider linguistic adaptation, semantic content alignment, and factual validation as separate but complementary requirements. The observed run-to-run variation further suggests that models intended for learner-facing use should be evaluated across multiple outputs rather than through a single generation.

Overall, the results indicate that LLMs can generate Greek educational texts with characteristics that are relevant to primary-school learners, but educational suitability cannot be established through linguistic simplicity alone. Multi-criteria evaluation is therefore necessary when generated texts are intended for learning materials or adaptive educational environments. The Grade 3 cultural heritage task used in this study provides a controlled case through which these relationships can be examined, while broader generalization requires evaluation across additional topics, grade levels, and educational contexts.

8 Limitations and Future Work

The present study has several limitations. First, the evaluation focuses on a single Grade 3 cultural heritage topic, and the findings should therefore not be assumed to generalize to other subjects, age groups, or educational contexts. The study also uses a single prompt formulation. Although keeping the prompt fixed supports a controlled comparison across models, alternative prompt wording or prompting strategies may produce different linguistic, semantic, or factual profiles. In addition, the results reflect the specific model versions and API configurations available at the time of the experiments and should therefore be interpreted as a snapshot of model behaviour rather than as permanent characteristics of the corresponding model families.

Second, factual accuracy was assessed by a single primary-school teacher, so inter-rater reliability could not be estimated. Third, semantic content alignment was measured automatically using BGE-M3 embeddings and was not independently validated against human judgments for the Greek texts used in this study. Moreover, only three repeated generations were used for the run-to-run analysis, so the reported variability should be interpreted descriptively rather than as a precise estimate of each model's generation distribution. Finally, the factual-accuracy rubric focused on the predefined core content elements and did not separately assess all additional factual claims that may have appeared in the generated texts.

Future work should extend the evaluation to

multiple curricular and cultural heritage topics, additional primary-school grade levels, and alternative prompt formulations. Larger numbers of repeated generations would also support more robust estimates of model variability. Further evaluation should involve multiple independent educators, include human assessment of semantic content alignment and age appropriateness, and examine unsupported or incorrect factual claims beyond the predefined rubric. Because the present study targets L1 Greek primary-school learners, the applicability of these evaluation dimensions to L2 learners should also be examined separately using appropriate proficiency-oriented resources and learner populations.

9 Conclusion

This study evaluated 16 large language models on the generation of age-adapted Greek learning materials, using Grade 3 students and a cultural heritage topic as a controlled case study. The models were assessed across readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment, factual accuracy, and consistency across repeated generations.

The results demonstrate that the suitability of generated educational material is multidimensional: strong performance on one evaluation dimension does not necessarily imply strong performance on others. In particular, linguistic accessibility alone should not be treated as a proxy for semantic alignment or factual reliability. These findings support the use of multi-criteria evaluation before LLM-generated texts are integrated into learner-facing or adaptive educational applications. Future work should extend the evaluation to additional topics and grade levels and include multiple independent educators and broader human evaluation.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: OpenAI ChatGPT was used as an assistive tool for language editing, improving clarity and readability, and refining the organization and presentation of the manuscript. It was also used to support the review of wording and methodological descriptions. All scientific content, experimental procedures, results, analyses, and interpretations were reviewed and verified by the authors. No AI tool was used to generate or alter the reported experimental results.

References

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Nondeterminism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pages 135–148. Association for Computational Linguistics.

Centre for the Greek Language. 2025. Readability software for modern Greek texts. Portal for the Greek Language. Accessed 15 September 2026.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie L. Day, Jacapo Cirica, Steven R. Clapp, Veronika Penkova, Amy E. Giroux, Abbey Banta, Catherine Bordeau, Poojitha Mutteneni, and Ben D. Sawyer. 2025. Evaluating GenAI for simplifying texts for education: Improving accuracy and consistency for enhanced readability. ArXiv:2501.09158.

Massimiliano Dibitonto, Alessio Ferrato, Carla Limongelli, and Olga Concetta Patroni. 2026. Museum audio guides generation using visitor categories and large language models. *Multimedia Systems*, 32:439.

Rotem Dror, Daniel Hutchinson, Mason A. Jones, Tsvi Kuflik, and Victoria Van Hyning. 2024. The curator's helper. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 496–504. Association for Computing Machinery.

Alessio Ferrato. 2025. Large language models to enhance learning in cultural heritage. volume 2590 of *Communications in Computer and Information Science*, pages 458–463. Springer.

George Mikros and Rania Voskaki. 2021. A modern Greek readability tool: Development of evaluation methods. In Adam Pawłowski, Ján Mačutek, Sheila Embleton, and George Mikros, editors, *Language*

and Text: Data, Models, Information and Applications, volume 356 of *Current Issues in Linguistic Theory*, pages 163–176. John Benjamins Publishing Company.

Jio Oh, Steven Euijong Whang, James Evans, and Jindong Wang. 2026. Classroom AI: Large language models as grade-specific teachers. *npj Artificial Intelligence*, 2:28.

Panagiota Papadimitriou and Chrysoula Savvaki. 2026. *Istoria G' Dimotikou: Vivlio Mathiti [History, Grade 3: Student's Book]*. 24grammata. Digital Library of School Textbooks, Textbook Registry Code 39, Identifier 20.500.14040/43632.

Sofia Stamou, Eleftherios Kalogeros, Matthew Damigos, and Manolis Gergatsoulis. 2027. Leveraging LLMs to build text narratives from CIDOC CRM. In *Metadata and Semantic Research: 19th Metadata and Semantics Research Conference, MTSR 2025, Thessaloniki, Greece, December 15–19, 2025, Revised Selected Papers*, volume 3057 of *Communications in Computer and Information Science*. Springer.

Aris R. Terzopoulos, Lynne G. Duncan, Mark A. J. Wilson, Georgia Z. Niolaki, and Jackie Masterson. 2017. HelexKids: A word frequency database for Greek and Cypriot primary school children. *Behavior Research Methods*, 49(1):83–96.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598. Association for Computational Linguistics.

Micah Watanabe. 2025. ChatGPT, can you make this text easier?: Investigating the utility of LLMs in easing text readability. *International Journal of Artificial Intelligence in Education*, 35(6):3922–3942.