STANDARDIZE: Aligning Language Models with Expert-Defined Standards for Content Generation

Joseph Marvin Imperial[Ω,Λ] **Gail Forey**[Λ] **Harish Tayyar Madabushi**[Λ]
[Λ]University of Bath, UK
[Ω]National University, Philippines
jmri20@bath.ac.uk

Abstract

Domain experts across engineering, healthcare, and education follow strict standards for producing quality content such as technical manuals, medication instructions, and children's reading materials. However, current works in controllable text generation have yet to explore using these standards as references for control. Towards this end, we introduce STANDARDIZE, a retrieval-style in-context learning-based framework to guide large language models to align with expert-defined standards. Focusing on English language standards in the education domain as a use case, we consider the Common European Framework of Reference for Languages (CEFR) and Common Core Standards (CCS) for the task of open-ended content generation. Our findings show that models can gain 40% to 100% increase in precise accuracy for Llama2 and GPT-4, respectively, demonstrating that the use of knowledge artifacts extracted from standards and integrating them in the generation process can effectively guide models to produce better standard-aligned content[1].

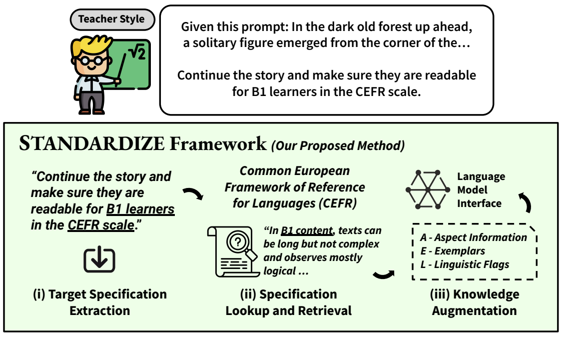

Figure 1: In contrast to the simple prompting method used by teachers, the proposed STANDARDIZE framework aims to improve the performance of generative models for content generation by using the fine-grained information found in expert-defined standards. The framework involves a three-part process starting with (i) the **extraction** of target specifications from the prompt, (ii) **lookup and retrieval** of information that matches the target specifications from the specified standard, and (iii) **knowledge augmentation** to produce artifacts that represent the standard itself for integration into the generation process with generative models.

1 Introduction

One of the most realized benefits of large language model (LLM) research is how it became widely adopted by the public. In particular, the rise of chat-style model interfaces, such as ChatGPT and Perplexity, has allowed non-technical users to fully utilize these tools in accomplishing day-to-day tasks and activities, such as getting help with writing, documenting code, and providing recommendations. A key technological advancement behind this is the use of reward-based methods such as Reinforcement Learning for Human Feedback (RLHF, Ouyang et al. (2022)), which allows embedding human preferences to generative models for better-aligned outputs with respect to the task at hand.

Despite the growing literature proposing complex algorithms and architectures for enriching the instruction-following capabilities of generative models, the missing puzzle piece that seems to have not garnered equal attention from the community is the integration of actual **standards** or **guidelines** crafted by domain experts as a reference for control. For example, in healthcare and engineering, well-documented standards are strictly followed in order to ensure the quality of processes. This includes the UK National Health Service (NHS) Injectable Medicines Guide (IMG) which contains instructions on how medical injectables should be mixed (Keeling et al., 2010) as well as the Simplified Technical English (STE)[2] which is a documented controlled language specification for writing technical manuals that are simple to read. Fol-

[1]Our code and data will be released upon publication.

[2]https://www.asd-ste100.org/

lowing these standards, even for domain experts, can be tedious, challenging, and even consequential in serious cases due to its complexity (Jones et al., 2021; Cousins et al., 2005). Thus, this research gap is an opportunity where the complex instruction-following capabilities of language models can provide assistance, particularly for tasks requiring the generation of text content since this is one of the areas where these models objectively perform well (Chung et al., 2022; Wei et al., 2021; Gatt and Krahmer, 2018).

Towards this end, we tackle the main research question: **How can we align large language models for content generation tasks using expert-defined standards?** We list our major contributions from this study as follows:

1. We introduce STANDARD-CTG, a new task formalizing the challenge of generating text using generative language models with expert-defined standards as an additional resource for control.

2. We propose STANDARDIZE, a retrieval-based framework using in-context learning that extracts knowledge artifacts from standards such as aspect information, exemplars, and manually crafted linguistic variables to improve the performances of generative language models for content generation.

3. We introduce high-performing baseline Llama2 and GPT-4 models for the task of STANDARD-CTG using two of the most widely recognized academic standards, CEFR and CCS.

2 Expert-Defined Standards

2.1 Background

According to the International Organization for Standardization (ISO)[3], **standards** are documented guidelines often containing rich detail in describing requirements, specifications, and criteria. These guidelines are defined and continuously improved by experts or interest groups in various domains, such as education, healthcare, and accounting, to name a few. Using standards ensures an institution's products and processes are consistent and reproducible (Sadler, 2017).

In the context of education and language assessment, standards are usually in the form of either (a) **content standards** such as documentations of a common language for ease of communication, writing, and content production, and (b) **performance standards** such as state-administered tests for reading and mathematical problem-solving competencies. This study focuses on content-based standards used in education and language assessment to be integrated into a generative model's text generation process. The alignment with existing standards for any generated text material is crucial to ensure quality and consistency before being used in classroom settings (La Marca et al., 2000).

2.2 Standards in Education and Language Assessment

We discuss the two selected English standards we consider as test cases for this study.

The Common European Framework of Reference for Languages (CEFR) is one of the well-known standard language framework[4] developed by The Council of Europe and used for assessing general language competencies such as reading, writing, and listening. The CEFR uses a six-point level scale of A1, A2, B1, B2, C1, and C2, which denotes increasing complexities in instructional content development. We use the level descriptors compiled by Natova (2021), which cover three aspects, namely (1) Meaning/Purpose, (2) Structure, and (3) Grammatical Complexity, describing the characteristics of desired content per level as shown in Table 1. We omit a fourth aspect of Reader's Knowledge Demands from the standard as this heavily depends on the reader's background knowledge and is entirely subjective (Forey, 2020; Forey and Cheung, 2019).

The Common Core Standards (CCS) is an academic standard[5] developed by the US National Governors Association and the Council of Chief State School Officers (CCSSO) which has been widely adopted by schools across the United States for its K-12 curriculum. In this study, we adapt the recommended model of CCS for assessing text complexity, which includes two main variables: (1) Qualitative Dimensions and (2) Quantitative Dimensions. However, similar to the CEFR standard, we do not include the last variable, which is Reader

[3] https://www.iso.org/standards.html

[4] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

[5] https://corestandards.org/

Level	Meaning / Purpose	Organisation / Stucture	Grammatical Complexity
A2	The text is clear and concrete, aiming to describe appearance, places, routines, preferences, or tell a simple story.	The text is often short and observes chronological and predictable structure.	The text contains comparison of adjectives, relative clauses, quantifiers, past simple of to be and full verbs, passive voice of present and past simple.
B1	The text is clear and concrete, aiming to describe appearance, places, routines, preferences, or tell a simple story. The text may also provide opinions and instructions or explanations, easy to understand and visualise, excluding ambiguity and diverse interpretations.	The text is can be long but not complex, and observes mostly chronological with unexpected changes of direction, digressions or flashbacks.	The text contains future forms, future in the past, 'used to' about repeated actions, present perfect simple, clauses for purpose and contrast, reporting statements, tag questions.
Linguistic Flags	Automatic Readability Formula, Type Token Ratio (2)	Total and average sentence and word lengths, Subordinating and coordinating conjunctions (3)	Age-of-Acquisition densities, entity density per sentence (2)

(a) The specifications provided by the Common European Framework of Reference for Languages (CEFR) cover aspects of meaning, organization, and grammatical complexity for two sample levels, A2 and B1.

Aspects	Qualitative (Meaning)	Qualitative (Syntax)	Quantitative
Description	The text can range from containing a single level of meaning to multiple levels of meaning based on complexity.	A text with low complexity tends to have simple, well-marked, and conventional structures, whereas a text of high complexity tends to have complex, implicit, and unconventional structures.	A text with many long words and/or sentences is harder to read than a text with many short words and/or sentences would be.
Linguistic Flags	Entity densities per sentence, Total proper noun density (2)	Type Token Ratio, Subordinating and coordinating conjunctions (2)	Total and average sentence and word lengths (2)

(b) The specifications of the Common Core Standards (CCS) cover qualitative and quantitative aspects. Unlike the CEFR, the CCS's model does not require categorization per level.

Table 1: From the CEFR and CCS academic standards, we manually identify linguistic flags as a form of knowledge artifact required for the proposed STANDARDIZE framework to improve the quality of language models for content generation. The full content of the standards can be found in the Appendix.

Considerations, as this requires professional judgment or a teacher's intervention. The description of each aspect of CCS is detailed in Table 1.

2.3 Standard-Aligned Content Generation (STANDARD-CTG)

Given the importance of adhering to expert-defined standards in the context of language assessment, we introduce the task of **standard-aligned content generation**. The overarching goal of STANDARD-CTG is to pave the way for new approaches that aim to integrate the conventional methodologies of controllable text generation in NLP with actual constraints provided by experts across interdisciplinary fields such as education, engineering, and medicine through documented standards. To align with terminologies used in education and other non-computing literature, in this work, we use the term *content generation* instead of *text generation* as usually seen in technical NLP literature.

We represent the task of STANDARD-CTG using the following formulation:

$$\mathrm{A} = C_{\mathrm{Stndrd}}(M(p, a, k_a), E) \tag{1}$$

where A quantifies the content alignment score of using a general evaluator C_{Stndrd} that tests the quality of a language model's M generated content against a collection of gold-standard examples E using inputs such as (a) a natural language prompt p, (b) information of some aspect a, and (c) transformed representation of an aspect k_a defined or extracted from the chosen standard. We pattern our major experiments in the succeeding sections based on this formulation.

3 The STANDARDIZE Framework

Our main hypothesis in this study is motivated by the fact that expert-defined standards are often very informative, lengthy, and complex. More specifically, we posit that in order for a generative model to produce content that is *aligned* with the specifications provided by a standard, the actual information found in the standard itself must be considered in the actual generation process. The challenge then is redirected towards *how* any information extracted can be represented as something that the generative model will find useful.

Towards addressing STANDARD-CTG, we propose STANDARDIZE, a retrieval-style in-context learning-based framework that exploits the rich information found in standards and transforms this into knowledge artifacts to improve the quality of content produced by generative models. Figure 1 encapsulates this framework in a visual manner. In the succeeding sections, we discuss the proposed STANDARDIZE framework more thoroughly.

3.1 Process

Target Specification Extraction is performed first to obtain informative tags in the prompt and to correctly match this information within the standards. For academic standards in language assessment, these specifications should provide information about *who* will be content delivered to (target audience) and using *what* specific standard out of many (CEFR or CCS). Thus, these two information tags are the basic required input for the process. As an example shown in Figure 1, the extracted specifications provided in the prompt are "A2 readers", which points to a particular group of learners requiring low-leveled reading materials, and "CEFR scale" which denotes the selected standard where properties of A2-level texts are described.

Specification Lookup and Retrieval is then performed next upon extracting the target specifications. A lookup process is done to find a match with the selected standard, usually in the form of a database or an external machine-readable file. The information from the standard in the form of *aspects* (or characteristics) that match the target specifications is then retrieved. The length and complexity of a standard's level of information regarding its specifications may vary. As shown in Figure 1 for the CEFR standard, the retrieved information that matches the desired level of complexity for the target audience (A2 readers) can be checked at Table 1.

Knowledge Augmentation is done last but is the most important process of the pipeline. We propose a further technical augmentation of information found in standards to obtain **knowledge artifacts** in the prompts. These knowledge artifacts can range from simple additional information already present in the standard to complex representations, such as incorporating actual linguistic features to control the granularity of the generation process. Recent works surveying the performance of open and closed models have shown that non-informative style of prompting language models, such as the teacher style shown in Figure 1, is effective only to

Aspect Criteria

Given this prompt: *In the dark old forest up ahead, a solitary figure emerged from the corner of the...*

Continue the story and make sure they are readable for B1 learners in the CEFR scale and observes the following specifications:

1. Meaning or Purpose: The text is clear and concrete, and tells a simple story.
2. Structure: The text is can be long but not complex, and observes mostly chronological with possible flashbacks.
3. Grammatical Complexity: The text may contain future forms, future in the past, repeated actions, present perfect simple forms.

Figure 2: A standard contains recommended characteristics of content across one or more **domain-specific aspects** or criteria. This figure shows an example of the CEFR standard where the set of criteria includes depth of meaning, structure, and grammatical complexity.

a certain extent and may be biased towards content generation in lower levels, such as A2 or B1 in the CEFR standards (Imperial and Madabushi, 2023b; Ribeiro et al., 2023).

3.2 Knowledge Artifacts for Content Generation

In this section, we discuss the knowledge artifacts used by the STANDARDIZE framework and how they are integrated into the generation setup via prompting.

Aspect Information (STANDARDIZE-*A*) is the most evident form of knowledge artifact as this pertains to the descriptive information provided in the standard. In the context of standards for content generation, aspect information is generally attributed to linguistic criteria of content with respect to its target audience. Figure 2 shows how aspect information from a standard (e.g., CEFR) can be integrated into the actual prompt. The addition of aspect criteria information ensures that the generative model will have access to *explicit characteristics* of the desired generated content in different dimensions.

Linguistic Signals (STANDARDIZE-*L*) represent the controllable variables of a standard that a generative model can use to steer the direction of content generation. In the STANDARDIZE framework, this process serves as a **rewrite function** where a generative model is asked to produce an initial content first using another method prompting (e.g., aspect information in Figure 2), and rewrites this

Exemplars

Given this prompt: *In the dark old forest up ahead, a solitary figure emerged from the corner of the...*

Continue the story and make sure they are readable for B1 learners in the CEFR scale.

Example books in the same level of complexity include Frankenstein by Mary Shelley, Wuthering Heights by Emily Bronte, and Midsummer Night's Dream by Shakespeare.

Figure 3: A standard contains recommended **exemplars** that serve as gold-standard reference. This figure shows an example of the CEFR standard where three well-known pieces of literature are provided as examples of content that conforms to the target level specified (B1).

by comparing linguistic flag values of the initially generated content against the mean value of a gold standard dataset of the target level. An example is illustrated in Figure 4 where the mean type-token ratio of a collection of gold-standard B1-level text (12.50) is added to the prompt while being compared to the current type-token value of the story, which is 4.22. A *verbalizer* is used to transform the computed linguistic flags into natural language prompts. The keywords *increase* and *decrease* are used in constructing the prompts to provide a sense of direction for the generative model.

In this work, we select 2 to 3 linguistic signals for both CEFR and CCS as reported in Table 1. The selection of what linguistic signal to use can be as simple as referring to what the definitions of aspects provide and need not be exhaustively many. For example, in CEFR, the Organization aspect is defined through different levels as *"text is often short and observes chronological and predictable structure"* for A2 and *"text is can be long but not complex"* for B1. Thus, we select *average sentence and word lengths* as a linguistic signal to capture this aspect.

Exemplars (STANDARDIZE-*E*) pertain to recommended examples by experts or developers of standards for reference of users. The addition of exemplars or any artifact found in the standard that showcases gold-standard output allows the generative model to have a sense of *implicit knowledge* during the content generation process. For example, in Figure 3, the exemplars for a B1-level content include *Frankenstein* by Mary Shelley, a well-known piece of gothic fiction. Although indirectly, any large language model trained using internet data (e.g., Wikipedia dumps) may have already formed a

Linguistic Flags

Given this story: *In the dark old forest up ahead, a solitary figure emerged from the corner of the...*

Rewrite the story and make sure they are readable for B1 learners in the CEFR scale. Use the following linguistic features to reach the target level of the story:

1. The type token ratio of the current story is 4.22 while the mean value in the target level is close to 12.50. Increase the complexity by aiming for higher type token ratio.

2. The average number of words of the current story is 510 while the mean value in the target level is close to 420. Decrease the complexity by aiming for lower average number of words.

Figure 4: A standard contains aspect definition which can be represented by flags such as **linguistic variables**. Given the mean values from gold-standard data in the target level, the generative model can then be steered to push the property of its generated content using **directional instructions** such as *increase* or *decrease*.

sense of knowledge of how this literature looks like (Karamolegkou et al., 2023; Petroni et al., 2019). We use the actual recommended exemplars from the CCS while we collected exemplars from the Penguin Readers publishing platform[6] which provides expert-curated literature for CEFR. The full list of exemplars for both standards can be found in the Appendix.

4 Experimental Setup

4.1 Tasks and Datasets

For this study, we specifically center our experimentation on the general task of **story** or **narrative generation**. We consider the subfield's rich literature and active research community in NLP (Alhussain and Azmi, 2021), as well as being one of the most common examples demonstrated across the education community regarding the use of generative text interfaces for content generation (Kasneci et al., 2023; Whalen et al., 2023). Further, we differentiate two tasks used in our work for story generation as listed below.

Task 1: Context Assisted Story Generation. For this setup, we provide preliminary context in the form of 50 to 70 words (or approximately 3 to 5 sentences) in the prompt to guide the generative language model in producing the story continuation. We select the CEFR as the standard of choice to evaluate this approach

[6] https://www.penguinreaders.co.uk/

and use the European Language Grid (ELG) corpus[7][8] compiled by Breuker (2022) to construct the prompts. The balanced corpus contains 300 CEFR-aligned English texts produced by experts and distributed across five levels A2, B1, B2, C1, C2 with 60 instances each. A1 is omitted due to lack of resources (n < 20).

Task 2: Theme Word Story Generation. In contrast to the previous setup, this method introduces only a single theme word for the generative language to produce a narrative from scratch, which allows for increased diversity in the content (Daza et al., 2016; Peng et al., 2018). To compile a theme words list, we select 50 random English noun words in plural form (e.g., *dragons, mysteries, voyages*) from the Corpus of Contemporary American English (COCA) (Davies, 2009) and prompt the generative model iteratively for each level in the standard. We investigate the application of CCS as the standard of choice in this setup.

4.2 Models

We cover a diverse set of generative language models for content generation, each with its own advantage, as discussed below. For the open models, we use a number of well-known models in the 2B-7B range, including **Llama2-Chat-7B** (Touvron et al., 2023a), **OpenChat-7B** (Wang et al., 2023), and **Longform-2.7B** (Köksal et al., 2023). For the closed model, we use **GPT-4-Turbo** (OpenAI, 2023). More information on the models can be found in Appendix A.

4.3 Evaluation Setup

Task Evaluation. For the context-assisted story generation task using CEFR standards, a Random Forest classifier is trained from a separate collection of Cambridge Exams dataset with CEFR labels following the works of Xia et al. (2016) and Imperial and Madabushi (2023a). This corpus is relatively balanced, with 331 reading passages distributed over labels A2 to C2 (roughly 60-70 each). We obtain the best accuracy with 0.912 using 79 length-normalized[9] linguistic features via the LFTK tool (Lee and Lee, 2023). For the theme word story generation using CCS standards, a binary XGBoost classifier is trained from the only CCS-aligned data found online and compiled by Flor et al. (2013) with an accuracy of 0.750 using the same feature length-normalized feature set as described above. Due to its limited size of 168, we grouped the dataset into binary categories, easy (grades $4-8$) and intermediate (grades $9-12$), with 48 and 73 documents per class, respectively. We also use the Cambridge Exams (Xia et al., 2016) and the official CCS-aligned data (Flor et al., 2013) for computing gold-standard means mentioned in Section 3.2. These datasets are exclusively used for evaluation. Therefore, we do not see any data contamination risks with this setup.

Alignment Metrics. To evaluate the models' content alignment or ability to generate content based on the target reading level, we use **precise accuracy** and **adjacent accuracy**. The former is a common metric for classification tasks, while the latter preserves the ordinality by counting labels off by one class as correct. We do not apply adjacent accuracy for CCS since the labels are binary.

5 Results

We discuss the results of our experiments procedures with the methods from the STANDARDIZE framework.

5.1 Standard Alignment via Overall Performance

The overall performance of models for CEFR and CCS are reported in Tables 2 and 3. For CEFR, the top-performing setup across the four models belongs to the STANDARDIZE framework. We report a 100% increase in performance with GPT-4 in precise accuracy (from $0.227 \rightarrow 0.480$) and a 43% increase for adjacent accuracy (from $0.630 \rightarrow 0.906$) compared to the teacher style method. The open models also gained substantial boosts in performance, such as Longform up by 23%, OpenChat up by 14%, and Llama2 by 58%. In terms of adjacent accuracies, GPT-4 remained the best model for preserving the ordinality of the labels with 0.906. With CCS, the general scores obtained in this setup

[7] Can be accessed by filling up the form: https://live.european-language-grid.eu/catalogue/corpus/9477

[8] We note that the ELG corpus is not included in any of the pretraining data reported from the documentation of the selected generative models for experimentation, which makes it a practical option to be used in this study.

[9] This pertains to using average-based features (e.g., average count of sentences) in order for the classifier to avoid being confounded by total-based features (e.g., total count of sentences).

Model	Precise Accuracy	Adjacent Accuracy	Fluency *(perplexity)*	Diversity *(distinct-n)*
Llama2 7B				
- Teacher Style	0.203	0.636	**13.189 ±4.88**	0.156 ±0.03
- STANDARDIZE-A	0.270	0.626	13.694 ±7.74	0.155 ±0.02
- STANDARDIZE-E	**0.320**	**0.683**	15.576 ±3.31	**0.188 ±0.01**
- STANDARDIZE-L	0.273	0.606	20.175 ±4.47	0.186 ±0.01
OpenChat 7B				
- Teacher Style	0.237	0.626	22.039 ±7.70	0.170 ±0.02
- STANDARDIZE-A	0.243	**0.630**	21.195 ±7.66	0.171 ±0.02
- STANDARDIZE-E	0.253	0.600	**13.931 ±2.97**	0.178 ±0.01
- STANDARDIZE-L	**0.270**	0.546	18.182 ±8.52	**0.179 ±0.02**
Longform 3B				
- Teacher Style	0.230	0.606	18.209 ±6.01	0.159 ±0.02
- STANDARDIZE-A	0.223	**0.610**	17.982 ±9.21	0.157 ±0.02
- STANDARDIZE-E	0.257	0.496	25.075 ±8.80	**0.192 ±0.11**
- STANDARDIZE-L	**0.283**	0.586	**16.926 ±6.91**	0.161 ±0.03
GPT-4				
- Teacher Style	0.227	0.630	27.357 ±6.30	0.187 ±0.08
- STANDARDIZE-A	0.397	0.846	29.729 ±9.58	0.174 ±0.01
- STANDARDIZE-E	0.307	0.703	30.357 ±9.79	0.182 ±0.01
- STANDARDIZE-L	**0.480**	**0.906**	**24.115 ±7.04**	**0.194 ±0.03**

Table 2: Experiment results comparing the conventional teacher style prompting with the STANDARDIZE framework for the Common European Framework of Reference for Languages (CEFR) standards.

Model	Precise Accuracy	Fluency *(perplexity)*	Diversity *(distinct-n)*
Llama2 7B			
- Teacher Style	0.473	17.936 ±4.32	0.184 ±0.01
- STANDARDIZE-A	0.400	22.070 ±1.75	0.171 ±0.01
- STANDARDIZE-E	0.590	**13.484 ±2.50**	**0.193 ±0.01**
- STANDARDIZE-L	**0.637**	15.066 ±2.47	0.191 ±0.01
OpenChat 7B			
- Teacher Style	0.427	16.116 ±12.39	0.166 ±0.05
- STANDARDIZE-A	0.567	19.444 ±2.57	0.172 ±0.01
- STANDARDIZE-E	0.550	**12.438 ±1.85**	0.178 ±0.01
- STANDARDIZE-L	**0.595**	13.734 ±2.53	**0.180 ±0.01**
Longform 3B			
- Teacher Style	0.400	13.657 ±5.39	**0.154 ±0.04**
- STANDARDIZE-A	0.573	17.918 ±4.74	0.148 ±0.01
- STANDARDIZE-E	0.490	14.277 ±2.79	0.151 ±0.02
- STANDARDIZE-L	**0.580**	**13.398 ±3.93**	0.148 ±0.04
GPT-4			
- Teacher Style	0.590	32.447 ±7.46	**0.195 ±0.01**
- STANDARDIZE-A	0.550	31.765 ±11.30	0.169 ±0.01
- STANDARDIZE-E	0.520	29.912 ±6.81	0.184 ±0.01
- STANDARDIZE-L	**0.610**	**26.912 ±6.11**	0.155 ±0.01

Table 3: Experiment results comparing the conventional teacher style prompting with the STANDARDIZE framework for the Common Core Standards (CCS).

are higher compared to CEFR with five classes. We see a similar pattern where all open and closed models obtained the best performance, with boosts ranging from 3% to 45% using linguistic signals (STANDARDIZE-L) to refine the generated content toward the target level. These findings provide concrete evidence that using the actual content of the standards and supplying better technical representations may be crucial when prompting language models to produce aligned content for classroom use.

5.2 Standard Alignment via Distributional Densities and Closeness

Using model-based classifiers as evaluators of standard alignment may be direct but may still impose a few weaknesses, such as dependence on accuracy performance as well as quantity of training data. With this, we explore the influence of using explicit controls via linguistic signals (STANDARDIZE-L) on the **distributional density** and **closeness** of the generated content against the gold-standard datasets for CEFR (ELG) and CCS (COCA) as mentioned in Section 4. For the distributional densities, we visualize the distributions in Figures 5 to 7 and also provide a comparison to a non-explicit method of control such as the teacher style prompting. For the distributional closeness reported in Table 4, we calculate the mean Euclidean distance of all the linguistic flags used for both standards and their levels listed in Table 1.

From the results, we observe that the general trend of using the best models with linguistic signals produces a more stable distribution across the variables it is explicitly controlling for (e.g., *average sentence length* or *type token diversity* as listed in Table 1), particularly with the CCS standards. We also notice that the distributions using STANDARDIZE-L also produce distributions closer to the mean (represented as a yellow star) from their corresponding gold-standard data. Moreover, in terms of closeness, using linguistic signals makes the quality of model generations more similar to the linguistic characteristics of the gold standard datasets in CEFR and CCS. Overall, these findings further strengthen the evidence of standard alignment by incorporating specific linguistic variables in the content generation process through the STANDARDIZE framework.

5.3 Fluency and Diversity of Generated Content

While our priority metrics are precise and adjacent accuracy, we also look at the level of **fluency** (measured as LM perplexity with GPT-2) and **content diversity** (measured as distinct n-grams) of using the STANDARDIZE framework.

In the case of fluency for models generating CEFR and CCS content, we don't see an obvious tradeoff and report relatively consistent performances with the STANDARDIZE-L setup. The best-performing model is still GPT-4 for both standards and Longform in terms of the open models. On the other hand, with the diversity metric, the most diverse batch of generated content comes from the teacher style method for CCS. But this may be on a case-to-case basis and task-dependent since we do not see the same tradeoff in performance with the CEFR standards in context-assisted story generation. Ultimately, our experiment procedure is focused on generating text content that aligns with the specified target level with respect to a standard. The standards that we applied in this study, CEFR and CCS, did not explicitly provide information on content creativity and how to measure this. Thus, we posit that creativity may be an interesting angle to explore in future works.

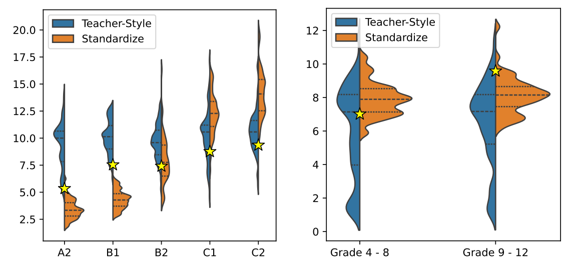

Figure 5: Distribution of **average sentence length** between CEFR using (left) and CCS (right) using their best performing models, GPT-4 and Llama2, with STANDARDIZE-L.

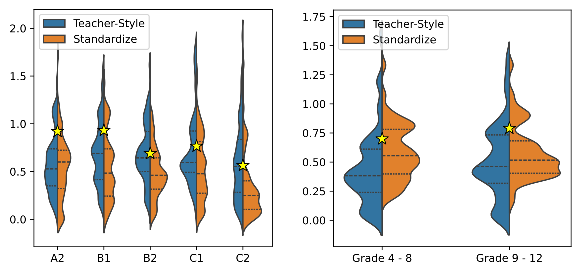

Figure 6: Distribution of **average entity density** between CEFR using (left) and CCS (right) using their best performing models, GPT-4 and Llama2, with STANDARDIZE-L.

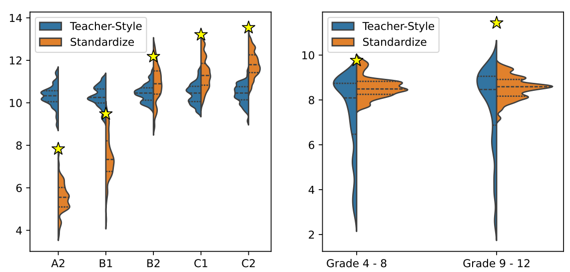

Figure 7: Distribution of **type token ratio** between CEFR using (left) and CCS (right) using their best performing models, GPT-4 and Llama2, with STANDARDIZE-L.

Setup	A2	B1	B2	C1	C2
Teacher Style	136.7	**96.7**	169.9	307.3	291.6
STANDARDIZE-L	**71.7**	124.6	**102.7**	**210.9**	**187.1**

Setup	Grade 4-8	Grade 9-12
Teacher Style	76.1	157.9
STANDARDIZE-L	**66.3**	**149.3**

Table 4: Mean Euclidean distances of generated content using simple teacher style prompting vs. STANDARDIZE-L for CEFR (top) and CCS (bottom).

6 Discussion

We discuss important points highlighting the real-world implications of our study within and beyond language model experimentations.

Validity on Global Education Context. Our main contribution, the STANDARDIZE framework, leverages the idea of a more holistic method for capturing the intricacies and complexities of educational standards for content generation. Our experiments with the CEFR and CCS standards showcase an opportunity for the generated texts of language model interfaces such as GPT-4, which are commonly used by educators and teachers, to be aligned with international language proficiency levels. Moreover, showing the effectiveness of STANDARDIZE on the aforementioned internationally recognized academic standards used in European and Northern American schools signifies the framework's strong potential for cross-curricula application. Thus, we invite future researchers to explore, validate, and propose derivations of our base framework for their own languages and language-specific standards for content generation.

Towards More Personalized Content Generation. Investigating the potential of generative models for personalized learning, such as providing adaptive feedback aligned with students' needs, is an active area in AI for education (Kasneci et al., 2023; Meyer et al., 2023; Sailer et al., 2023; Tack and Piech, 2022). This work contributes toward the goal of helping educators craft more personalized content for learners using the capabilities of large language models based on an assigned language proficiency level described by a standard. While we present a new task specifically targeted for the NLP community to encourage research in this direction (STANDARD-CTG as covered in Section 2.3), our results may already be useful for educators by providing context on better methods for generating level or target audience-specific texts by prompting language models using information found in educational standards.

7 Related Works

Complexity Controlled NLG. Research in complexity-controlled generation has been explored in the past, covering diverse facets in terms of text format, level granularity, and task variation. The work of Agrawal and Carpuat (2019) introduced controlling for specific complexity in the machine translation task. The following works of Agrawal and Carpuat (2023) and Ribeiro et al. (2023) explored grade-specific text simplification and summarization using control tokens and reinforcement learning, respectively. Currently, only two works have investigated incorporating CEFR for language learning content generation. Stowe et al. (2022) and Imperial and Madabushi (2023a) both made use of CEFR-aligned text for NLG but limited their studies to two levels, A1 and C2. However, none of them made use of the actual guideline information found in CEFR during the generation process.

Novelty. The STANDARDIZE framework is parallel to the work of Zhou et al. (2023), where a verbalizer is used to transform quantitative constraints into natural language for prompting, as well as the work of Ram et al. (2023) in the lookup and retrieval phase where aspect information is added in the prompt to influence model controllability. In comparison to all the works mentioned, our study's main novelty is capturing the *wholeness* of expert-defined standards, prioritizing fine granularity and not just one or two levels, as well as including information that can be represented as artifacts in the content generation process.

8 Conclusion

Standards are observed by interdisciplinary fields, particularly in the education domain, to enforce the quality and consistency of content. Our proposed STANDARDIZE framework using knowledge artifacts allows generative models such as Llama2 and GPT-4 to gain significant performance boosts (40% - 100%) in terms of content alignment guided by standards used in education such as CEFR and CCS. Moreover, we see a very promising potential for cross-domain and cross-standard generalization of our proposed method with the range of educational contexts around the world and invite future work to build on our baseline models.

Ethical Considerations

All datasets and corpora used in this study, such as the ELG (Breuker, 2022), Cambridge Exams (Xia et al., 2016), and CCS (Flor et al., 2013), are already established and accessible for research purposes. We observe a specific tone in the discussion of our experiments, emphasizing that the main motivation of the work is that language models such as GPT-4 can provide assistance in producing content that is more aligned or faithful with the constraints of standards such as CEFR or CCS *without* implying that they can replace experts in the field or produce better quality than the gold-standard data. Further, we also *do not* imply that any model enriched by any computational method to produce more standard-aligned content can replace the standard itself. Overall, we do not foresee any serious ethical issues in this study.

Limitations

Language Coverage of Standards. This work is mainly centered on the use of datasets and standards for the English language. While standards for language assessment, such as CEFR, have expanded through the years with versions to cover other languages, such as German, Czech, and Italian (Vajjala and Rama, 2018), we do not claim that our results will be able to generalize and have the same advantages with these languages. However, investigating this direction may be a good research opportunity for future work.

Dependence on Evaluation Methods. As observed in Section 5, we made sure to cover a variety of evaluation procedures for testing standard alignment instead of only using model-based methods such as a classifier. The limitation here is that trained classifiers are dependent on factors such as their accuracy, the quantity of data, the complexity of the training algorithm, and the quality of features. Thus, other means of evaluating alignment that is more direct, such as computed feature distances against a gold-standard dataset, is always recommended. Moreover, our model-based CEFR and CCS evaluators make use of artifacts such as datasets and tools for feature extraction from peer-reviewed papers (Xia et al., 2016; Flor et al., 2013). We are aware of paid third-party services online that promise more accurate classification of labels in CEFR, but they generally do not provide details on linguistic predictors used for prediction. Thus, this may not be a practical option for research.

Limited Data for Standards. Some standards for language assessment, such as the Common Core Standards in the US, have very limited data ($n < 200$) for training complex model-based evaluators. Our CCS classifier only reaches 0.750 in accuracy using the best held-out test set without risking overfitting. Researchers who want to extend this work to other standards and languages may face this data scarcity limitation. Thus, we recommend also adapting similarity-based evaluation using Euclidean distances as another form of evaluation method.

Human Evaluation. This study does not make use of human evaluation to judge the quality of the generated content of the models due to the substantial resources that this activity will require. Previous works in content generation only made use of human evaluation with non-domain experts for factors such as creativity, diversity, and relatedness (Ribeiro et al., 2023; DeLucia et al., 2021; See et al., 2019). However, these factors are not the top priority of this study. Doing human evaluation using standards requires multiple interactions with multiple domain experts for each aspect or constraint detailed in the standard, which can be very expensive. We posit that this may also be the reason why some existing standards, such as CCS, despite being developed a decade ago, still have very limited expert-produced resources online. Nonetheless, we follow the trend in recent works where automatic machine-driven methods are practical enough to provide evidence of alignment with standards that are correlated with humans (Arase et al., 2022; Settles et al., 2020; Wilkens et al., 2018; Tack et al., 2017).

Acknowledgements

This work made use of the Hex GPU cloud of the Department of Computer Science at the University of Bath. JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent AI [EP/S023437/1] of the University of Bath. We attribute the black icons used in Figure 1 to the collections of Design Circle and Victor Zukeran from the Noun Project and the colored teacher icon from Flaticon.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling Text Complexity in Neural Machine Translation. 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 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Sweta Agrawal and Marine Carpuat. 2023. Controlling Pre-trained Language Models for Grade-Specific Text Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic Story Generation: A Survey of Approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-Based Sentence Difficulty Annotation and Assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mark Breuker. 2022. CEFR Labelling and Assessment Services. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing Cham.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

D Cousins, B Sabatier, D Begue, C Schmitt, and T Hoppe-Tichy. 2005. Medication errors in intravenous drug preparation and administration: a multicentre audit in the UK, Germany and France. *Quality & Safety in Health Care*, 14(3):190.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Angel Daza, Hiram Calvo, and Jesús Figueroa-Nazuno. 2016. Automatic Text Generation by Learning from Literary Structures. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 9–19, San Diego, California, USA. Association for Computational Linguistics.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding Methods for Neural Narrative Generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Michael Flor, Beata Beigman Klebanov, and Kathleen M. Sheehan. 2013. Lexical Tightness and Text Complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 29–38, Atlanta, Georgia. Association for Computational Linguistics.

Gail Forey. 2020. A whole school approach to SFL metalanguage and the explicit teaching of language for curriculum learning. *Journal of English for Academic Purposes*, 44:100822.

Gail Forey and Lok Ming Eric Cheung. 2019. The benefits of explicit teaching of language for curriculum learning in the physical education classroom. *English for Specific Purposes*, 54:91–109.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023a. Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. *arXiv preprint arXiv:2309.05454*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023b. Uniform Complexity for Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12025–12046, Singapore. Association for Computational Linguistics.

Matthew D Jones, Anita McGrogan, DK Raynor, Margaret C Watson, and Bryony Dean Franklin. 2021. User-testing guidelines to improve the safety of intravenous medicines administration: a randomised in situ simulation study. *BMJ Quality & Safety*, 30(1):17–26.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright Violations and Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103:102274.

Susan Keeling, Robin Burfield, Christine Proudlove, and Katie Scales. 2010. The Injectable Medicines Guide Website. *British Journal of Nursing*, 19(19):S25–S28.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2023. LongForm: Optimizing Instruction Tuning for Long Text Generation with Corpus Extraction. *arXiv preprint arXiv:2304.08460*.

Paul M La Marca, Doris Redfield, and Phoebe C Winter. 2000. State Standards and State Assessment Systems: A Guide to Alignment. Series on Standards and Assessments.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted Features in Computational Linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20.

Ivanka Natova. 2021. Estimating CEFR Reading Comprehension Text Complexity. *The Language Learning Journal*, 49(6):699–710.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards Controllable Story Generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating Summaries with Controllable Readability Levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

D Royce Sadler. 2017. Academic achievement standards and quality assurance. *Quality in Higher Education*, 23(2):81–99.

Michael Sailer, Elisabeth Bauer, Riikka Hofmann, Jan Kiesewetter, Julia Glas, Iryna Gurevych, and Frank Fischer. 2023. Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, 83:101620.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do Massively Pretrained Language Models Make Better Storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. Controlled Language Generation for Language Learning Items. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–305, Abu Dhabi, UAE. Association for Computational Linguistics.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and Automated CEFR-based Grading of Short Answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 522.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with Universal CEFR Classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. OpenChat: Advancing Open-source Language Models with Mixed-Quality Dataa. *arXiv preprint arXiv:2309.11235*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Jeromie Whalen, Chrystalla Mouza, et al. 2023. ChatGPT: Challenges, Opportunities, and Implications for Teacher Education. *Contemporary Issues in Technology and Teacher Education*, 23(1):1–23.

Rodrigo Wilkens, Leonardo Zilio, and Cédrick Fairon. 2018. SW4ALL: a CEFR classified and aligned corpus for language learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

A Appendix

A.1 Corpus Statistics

We provide basic statistical information about the various corpora used in the study.

Level	Size	Average Word Count	Average Sentence Count
A2	60	186.55	18.91
B1	60	264.25	15.90
B2	60	517.71	31.71
C1	60	728.93	40.70
C2	60	749.73	37.55

Table 5: Statistics of the ELG corpus (Breuker, 2022) used for the CEFR context assisted story generation task.

Grade	Size	Average Word Count	Average Sentence Count
4 - 8	48	204.91	28.55
9 - 12	73	255.17	31.08

Table 6: Statistics of the official CCS-aligned corpus (Flor et al., 2013) used as gold-standard dataset for the STANDARDIZE-L artifact and for training the CCS classifier used in Section 5.

Level	Size	Average Word Count	Average Sentence Count
A2	64	60.87	11.53
B1	60	122.38	16.25
B2	71	265.35	37.03
C1	67	355.71	43.37
C2	69	333.86	38.41

Table 7: Statistics of the Cambridge Exams corpus (Xia et al., 2016) used as gold-standard dataset for the STANDARDIZE-L artifact and for training the CEFR classifier used in Section 5.

A.2 More Information on Selected Models

We set the minimum generated new tokens to 30 and the maximum to 300, as well as set the nucleus sampling decoding (top-p) to 0.95 as done with previous works on story generation (Imperial and Madabushi, 2023b; DeLucia et al., 2021; See et al., 2019). The actual sizes of the open models range from 5GB to 15 GB max.

Llama2-Chat (Touvron et al., 2023b) is one of the community-recognized open instruction-tuned models released by Meta and an improved version of Llama 1 (Touvron et al., 2023a). For this task, we use the 7B version[10] finetuned from over a million human preference data and optimized for chat and dialogue use cases. We prioritized the addition of this model in our study for its accessibility to the general NLP community.

Longform-OPT (Köksal et al., 2023) is a recent instruction-tuned model optimized for long text generation using the LongForm dataset. For this study, we use the OPT model variant[11] (Zhang et al., 2022) with 2.7B parameters as this version obtained the best performance for the short story generation task using the WRITINGPROMPTS dataset (Fan et al., 2018) against other instruction-tuned models such as Alpaca-LLaMA (Taori et al., 2023), FlanT5 (Chung et al., 2022), Tk-Instruct (Wang et al., 2022), and T0++ (Sanh et al., 2021).

OpenChat (Wang et al., 2023) is the most recent open model in our experiment setup, which currently is reported to be the best 7B model as of this writing and outperforms closed models such as ChatGPT (March) across a number of benchmark tasks such as GSM8K and TruthfulQA. In contrast to Llama and GPT models, which used RLHF (Ouyang et al., 2022), OpenChat is trained with mixed-quality data which is composed of high-quality expert data and sub-optimal web data with no preference labels. We use the 7B version[12] of this model variant released in January 2024.

[10] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[11] https://huggingface.co/akoksal/LongForm-OPT-2.7B

[12] https://huggingface.co/openchat/openchat-3.5-0106

GPT-4 (OpenAI, 2023) is the only closed model included in this study. We decide to add this model to our experiment for its global recognition through its easy-to-use interface among interdisciplinary fields, particularly in education (Kasneci et al., 2023). We use the version[13] finetuned with proprietary training data up to April 2023 with a 128K context window.

A.3 Exemplars List

We list the actual list of literary exemplars used for the STANDARDIZE framework. We manually selected at most three classical exemplars as reference for the language models.

Level	Exemplars
A2	*A Christmas Carol* by Charles Dickens *The Adventures Of Huckleberry Finn* by Mark Twain *The Little Prince* by Antoine de Saint-Exupery
B1	*Frankenstein* by Mary Shelley *Wuthering Heights* by Emily Bronte *Midsummer Night's Dream* by Shakespeare
B2	*Moby Dick* by Herman Melville *Jane Eyre* by Charlotte Bronte *Sense and Sensibility* by Jane Austen
C1	*Animal Farm* by George Orwell *Anna Karenina* by Leo Tolstoy *Great Expectations* by Charles Dickens
C2	*Oliver Twist* by Charles Dickens *Crime and Punishment* by Fyodor Dostoevsky *Les Miserables* by Victor Hugo

Table 8: The full exemplar list used for CEFR standards obtained from the Penguin Reader website (https://www.penguinreaders.co.uk/).

Grade	Exemplars
6-8	*Little Women* by Louisa May Alcott *The Adventures of Tom Sawyer* by Mark Twain *The Road Not Taken* by Robert Frost
9-12	*Jane Eyre* by Charlotte Brontë *The Great Gatsby* by F. Scott Fitzgerald *Fahrenheit 451* by Ray Bradbury

Table 9: The full exemplar list used for CCS standards obtained from the official website (https://www.thecorestandards.org/ELA-Literacy/).

A.4 Libraries and Dependencies

We have used the following dependencies and Python libraries for the study:

- Linguistic Feature Tool Kit (LFTK) (Lee and Lee, 2023)
- Spacy (https://spacy.io/)
- Scikit-Learn (https://scikit-learn.org/stable/)
- OpenAI API (https://openai.com/blog/openai-api)

[13] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Level	Meaning / Purpose	Organisation / Stucture	Grammatical Complexity
A2	The text is clear and concrete, aiming to describe appearance, places, routines, preferences, or tell a simple story.	The text is often short and observes chronological and predictable structure.	The text contains comparison of adjectives, relative clauses, quantifiers, past simple of to be and full verbs, passive voice of present and past simple.
B1	The text is clear and concrete, aiming to describe appearance, places, routines, preferences, or tell a simple story. The text may also provide opinions and instructions or explanations, easy to understand and visualise, excluding ambiguity and diverse interpretations.	The text is can be long but not complex, and observes mostly chronological with unexpected changes of direction, digressions or flashbacks.	The text contains future forms, future in the past, 'used to' about repeated actions, present perfect simple, clauses for purpose and contrast, reporting statements, tag questions.
B2	The text provides opinions and instructions/explanations, easy to understand and visualise, excluding ambiguity and diverse interpretations. The text also gives description, classification, argumentation or a combination of these, allowing greater ambiguity and various interpretations.	The text can be long but not complex, and observes chronological or spatial with possible statement of various aspects of a phenomenon.	The text contains past continuous, past perfect, passive voice of perfect and continuous, 'would' about habits, reporting questions, infinitives and -ing forms.
C1	The text may serve different purposes and may be combined with multiple levels of meaning. The descriptions and instructions in the text are detailed and may be hard to visualise.	The text is often lengthy, complex, and observes logical organisation, starting with a claim followed by reasons, proving it, or changing view-points.	The text contains compound adjectives, conditional sentences, inversion, future perfect, cleft and non-finite clauses, modals about the past.
C2	The text may serve different purposes and may be combined with multiple levels of meaning. The text may also show exploration of hypotheses, causes and effects, etc. The details of the text are complex to follow and visualise.	The text is often lengthy, complex, and observes presentation which may start with the ending/final result and go back to the possible causes.	The text contains combination of multiple adjectives, inversion with hardly and only when, comment clauses, non-finite perfect clauses, ellipsis, passive impersonal constructions.
Linguistic Flags	Automatic Readability Formula, Type Token Ratio (2)	Total and average sentence and word lengths, Subordinating and coordinating conjunctions (3)	Age-of-Acquisition densities, entity density per sentence (2)

(a) The specifications provided by the Common European Framework of Reference for Languages (CEFR) cover aspects of meaning, organization, and grammatical complexity for two sample levels, A2 and B1.

Aspects	Qualitative (Meaning)	Qualitative (Syntax)	Quantitative
Description	The text can range from containing a single level of meaning to multiple levels of meaning based on complexity.	A text with low complexity tends to have simple, well-marked, and conventional structures, whereas a text of high complexity tends to have complex, implicit, and unconventional structures. Simple texts tend to relate events in chronological order, while complex texts make more frequent use of flashbacks, flash-forwards, and other manipulations of time and sequence.	That text that has longer words and longer sentences are more difficult to read than shorter ones. A text with many long words and/or sentences is thus rated by these formulas as harder to read than a text with many short words and/or sentences would be.
Linguistic Flags	Entity densities per sentence, Total proper noun density (2)	Type Token Ratio, Subordinating and coordinating conjunctions (2)	Total and average sentence and word lengths (2)

(b) The specifications of the Common Core Standards (CCS) cover qualitative and quantitative aspects. Unlike the CEFR, the CCS's model does not require categorization per level.

Table 10: The full content of the CEFR and CCS standards.

TARGET LEVEL: A2 (Elementary Level)

CONTEXT: Corryvale Farm is in the west of England. Stephen and Jenny Blackler keep cows and chickens and sell milk and eggs. But most of their money comes from Jenny's sheep. She has 50 sheep and each one has enough wool for six sweaters. Every six months Stephen cuts this wool off the sheep so they won't get too warm.

TEACHER STYLE: In the lush, verdant lands of the west of England, nestled among rolling hills and meandering streams, lay Corryvale Farm. It was a picturesque farm, embodying the pastoral beauty that many dreamt of, yet few had the fortune to call home. This was the realm of Stephen and Jenny Blackler, a couple whose lives were as intertwined with the land as the ancient oak roots were with the earth. Stephen, a robust man with hands as rough as the bark of the trees he often climbed, tended to the cows and chickens with a steady patience. He rose with the sun, greeted by the gentle mooing of the cows and the cheerful chaos of the clucking chickens. The rhythm of milking the cows and collecting eggs was a comforting routine that marked the passage of days on the farm.

STANDARDIZE-A: At Corryvale Farm, Stephen would wake up early. Even if it was as cold as ice outside, he started the day with the same routine. First, he would feed the cows. There, in the fields, were 30 healthy cows. These cows were beautiful, larger than any other cows in the west of England. They were the pride of the farm. Then, he would go to the chickens. He had roughly 100 chickens that laid the best eggs every morning. The chickens were quite noisy and more excited than the quiet cows, making mornings lively on the farm. Jenny was not a big fan of waking up early, she preferred to sleep a bit more. However, her work started after breakfast. Her favourite part of the day was when she went to see her sheep. She loved them dearly. She always said that her 50 sheep were the cleanest and happiest sheep in the entire country. She was so proud of them.

STANDARDIZE-E: Stephen and Jenny loved these days on their Corryvale Farm, filled with the soft clucking of chickens and lowing of cows. They took pride in their daily duties of caring for all the animals, especially the sheep that were their main source of income. One cold morning, Jenny took her pail to milk the cows while Stephen checked the hens' house for fresh eggs. Their cheeks were red from the cool air and they could see their breath when they talk. But they didn't mind the chill for the promise of spring was around the corner. Meanwhile, back in the barn, Jenny's 50 sheep huddled together, their woolly coats growing thicker in the cool weather. This would mean only one thing. It was shearing time! Every six months, like clockwork, Stephen would sheer the sheep's wool. For each sheep he could fill a bag with enough wool to knit six sweaters. The wool was not just soft but warm too, a perfect shield against the bitter winter.

STANDARDIZE-L: Corryvale Farm is in the west of England. It is Stephen and Jenny Blackler's farm. They have cows and chickens. They sell milk and eggs. Jenny also has 50 sheep. The sheep have wool. Every six months, Stephen cuts the wool. This keeps the sheep cool. They make money from the wool.

Table 11: Sample generations with the teacher style method and variations of the STANDARDIZE framework using the best model (GPT-4) for the **context-assisted story generation** observing CEFR standards.

TARGET GRADE: 4 - 8

THEME WORD: Curses

TEACHER STYLE: In the small village of Ashwood, nestled in the heart of the Whispering Woods, a young apprentice named Eira stumbled upon an ancient tome hidden deep within the village library. The tome was bound in black leather and adorned with strange symbols that seemed to shift and writhe like living things. As soon as Eira opened the book, a puff of noxious smoke erupted from its pages, filling the air with a pungent stench. The smoke coalesced into a low, menacing whisper that seemed to seep into Eira's mind. "Curse you, Eira," the voice hissed. "You have disturbed the rest of the ancient ones. Now you will pay the price." Eira felt a cold dread creeping up her spine as the voice continued to taunt her. "You have unleashed a power beyond your control. Prepare to face the consequences."

STANDARDIZE-A: Once upon a time, in a small town nestled in the heart of a dense forest, there stood a grand mansion rumored to be haunted. The mansion had been abandoned for decades, its grandeur and beauty slowly being consumed by the encroaching trees and vines. The locals avoided the mansion, whispering tales of curses and ghostly apparitions that roamed its halls. One summer evening, a group of curious children stumbled upon the mansion while exploring the woods. They cautiously approached the mansion, their hearts racing with every creak and groan of the old wooden doors.

STANDARDIZE-E: Once upon a time, in a small town nestled in the heart of a dense forest, there stood a grand mansion that had been abandoned for decades. The mansion was rumored to be haunted by the ghosts of its former owners, who had died under mysterious circumstances. The townspeople whispered stories of strange occurrences that had taken place within the mansion's walls. Some claimed to have seen ghostly apparitions flitting about in the windows at night, while others swore they had heard eerie whispers emanating from the abandoned halls. One stormy evening, a group of brave adventurers decided to explore the cursed mansion, determined to uncover its secrets. As they approached the entrance, the wind howled and the trees creaked ominously, as if warning them away.

STANDARDIZE-L: In the small town of Willow Creek, there stood a grand mansion that had been abandoned for decades. The once-grand estate was now a relic of the past, its grandeur faded and its windows shattered. The locals whispered stories of the mansion's dark history, how it was cursed by a witch who had lived there many years ago. According to the legend, the witch had been wronged by the mansion's owner, who had refused to grant her a simple wish. Enraged, the witch placed a curse on the mansion, ensuring that no one would ever be able to live there peacefully. The curse was said to cause strange occurrences and unexplained noises to echo through the empty halls, making it impossible for anyone to enter or leave the mansion without experiencing terrifying events. One stormy night, a group of brave adventurers decided to investigate the rumors and put an end to the curse. They entered the mansion, their flashlights casting eerie shadows on the walls. As they explored the dusty rooms, they began to hear strange noises, like whispers in the darkness.

Table 12: Sample generations with the teacher style method and variations of the STANDARDIZE framework using the best model (Llama2) for the **theme word story generation** observing CCS standards.