



Citation for published version:

Imperial, JM & Tayyar Madabushi, H 2023 'Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models' arXiv.

Publication date:
2023

[Link to publication](#)

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models

Joseph Marvin Imperial^{Ω,Λ} Harish Tayyar Madabushi^Λ

^ΛBathNLP Lab, University of Bath, UK

^ΩNU Human Language Technology Lab, National University, Philippines

jmri20@bath.ac.uk htm43@bath.ac.uk

Abstract

Readability metrics and standards such as Flesch Kincaid Grade Level (FKGL) and the Common European Framework of Reference for Languages (CEFR) exist to guide teachers and educators to properly assess the complexity of educational materials before administering them for classroom use. In this study, we select a diverse set of open and closed-source instruction-tuned language models and investigate their performances in writing story completions and simplifying narratives—tasks that teachers perform—using standard-guided prompts controlling text readability. Our extensive findings provide empirical proof of how globally recognized models like ChatGPT may be considered less effective and may require more refined prompts for these generative tasks compared to other open-sourced models such as BLOOMZ and FlanT5—which have shown promising results¹.

1 Introduction

The introduction of public-facing text generative models with easy-to-use interfaces, such as ChatGPT by OpenAI², Perplexity Ask by Perplexity AI³, and Bard by Google⁴, has catalyzed the research progress of large language models (LLMs) that can follow and execute complex instructions in human language. This particular advantage over regular language models has seen a rapid growth of appreciation and utilization across a number of disciplines and sectors, such as medicine and healthcare (Thirunavukarasu et al., 2023; Singhal et al., 2023), teaching and assessment in education (Tack and Piech, 2022; Kasneci et al., 2023; Wang and Demszky, 2023), business and e-commerce

(Paul et al., 2023), and software development (Chen et al., 2021; Rozière et al., 2023; Muennighoff et al., 2023) to name a few.

One of the primary drivers of this advancement in LLMs is *instruction tuning*. This process involves finetuning an LLM on a diverse collection of multi-task corpora transformed in an instruction-answer pair format, which in turn allows the model to learn and improve upon tasks it was not trained on (Wei et al., 2021; Wang et al., 2022a). In the same vein, other advancements explored the involvement of human raters where a reward-driven language model learns from the aggregated preferences and is incentivized through reinforcement learning if its generated content from a series of executed instructions is acceptable (Ziegler et al., 2019; Ouyang et al., 2022). These training methodologies, in essence, allow LLMs to have some form of knowledge in relation to what aligns with humans and bridge the gap between the LLM-oriented goal of next token prediction and a user-oriented objective. Likewise, specifications from various instruction-answer corpora act as signals of constraint to control a model’s output (Zhang et al., 2023b).

However, one of the main research gaps that these powerful instruction-following models may need to be rigorously tested with is the *ability to capture human standards*. Standards or domain-specific frameworks are expert-defined set of rules that humans follow in various interdisciplinary fields. For example, a teacher must be properly knowledgeable of assessment standards such as the Common European Framework of Reference for Languages (CEFR) for evaluating the quality of text-based educational content before they can use it in a classroom setting (Jones and Saville, 2009). Therefore, if LLMs such as ChatGPT are to be utilized to generate educational content for the teacher, then it would be ideal for these models to be evaluated or trained based on how they accept

¹We will release the code and model generations upon publication.

²<https://openai.com/blog/chatgpt>

³<https://www.perplexity.ai/>

⁴<https://blog.google/technology/ai/bard-google-ai-search-updates/>

inputs, such as prompting or finetuning, to acquire some form of knowledge of how CEFR works and how it is used to assess the quality of texts.

In this work, we tackle the main research question: **To what extent can instruction-tuned large language models capture readability level specifications from prompts and reflect it to the generated content?** Towards this end, our primary contributions are as follows:

1. To the best of our knowledge, our work is the first to explore the readability-alignment capabilities of a diverse set of open and close-sourced instruction-tuned large language models.
2. Our findings provide empirical and quantitative evidence of the true performances of models such as ChatGPT and LLaMA for the tasks of story completion and simplification often performed by non-technical users such as teachers to produce classroom-ready content.
3. Our work is the first to explore the evaluation of instruction-tuned LLM generations using actual text assessment standards in practice, including the Common European Framework of Reference for Languages (CEFR).

2 Readability Knowledge Alignment of Large Language Models

2.1 Background

Instruction-tuned language models are developed to be used by the wider non-technical and interdisciplinary audiences of the general public. As such, users may impose or desire to have current domain-specific and expert-outlined standards in their respective fields integrated into these models for seamless use. For example, simple text prompts with grade-level specifications such as *"Write a story for second-grade readers."* are often used and suggested by academic groups for teachers and educators who want to produce classroom-ready materials using commercial generative tools such as ChatGPT (Staake, 2023; Herft, 2023). This notion, however, assumes that these models already have some knowledge of how text readability assessment metrics, such as Flesch Kincaid Grade Level, work and also assumes that they can generate any text conforming to any readability level specification on the fly. In this study, we put this assumption to stringent tests and formally frame the

task as evaluating for *readability knowledge alignment*. We discuss our experimental procedures in this section concerning the choice of instruction-tuned models to be investigated, metrics for evaluation, and corpora for prompting generations from models.

2.2 Selected Models

We explore a diverse set of open and closed-source instruction-tuned large language models to assess their capability to follow readability specifications from the prompts and reflect it to their generated content. We consider a model's *standard* size with respect to the selection that will be included in our main experiments. For example, if LLaMA 2 has multiple models ranging from 7B, 13B, and 70B, we select the one with 7B parameters as this is considered the base model that is accessible by most. To further clarify, we did not perform any finetuning method as these models are already finetuned towards maximizing their instruction-following capabilities.

LLaMA 2 (Touvron et al., 2023b) is an improved version of the original LLaMA 1 model (Touvron et al., 2023a) with an added mix of publicly available online data and pretrained with over 2T tokens with a context length of 4096. Specifically, we use the 7B model⁵ finetuned for chat with over 1M human annotations using the Reinforcement Learning from Human Feedback (RLHF) method (Ziegler et al., 2019).

FlanT5 (Chung et al., 2022) is another enhanced instruction-tuned language model built on top of the T5 model (Raffel et al., 2020) with 11B parameters. For this study, we use the FlanT5-Base model⁶ hosted in Huggingface with 250M parameters and trained with over 14M examples from instruction datasets including Muffin (Wei et al., 2021), T0-SF (Sanh et al., 2021), and Natural Instructions V2 (Wang et al., 2022b).

BLOOMZ (Muennighoff et al., 2022) by BigScience⁷ is an enhanced version of the multilingual language model BLOOM (Scao et al., 2022) through finetuning on xP3 which is a compilation of multilingual multitask learning datasets in 46

⁵<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

⁶<https://huggingface.co/google/flan-t5-base>

⁷<https://huggingface.co/bigscience>

languages with English prompts. We use the standard 3B model⁸ hosted on Huggingface for our experiments. We included this multilingual language model in our study to diversify the models being investigated and see if finetuning on multilingual instruction-tuned datasets can affect the performances for our complexity-specific prompting tasks.

Longform-T5 (Köksal et al., 2023) is a recent model finetuned using the Longform dataset on top of the various architectures such as T5-XL, OPT, and LLaMA 1. The Longform dataset contains over 27,739 LLM-generated instructions and long text pairs from parsed structured corpora and reformulated NLG tasks derived from existing corpora such as C4 (Raffel et al., 2020), WikiHow (Koupae and Wang, 2018), BigBench (Srivastava et al., 2023), and StackExchange (Longpre et al., 2019). We use the standard 3B T5-XL model⁹ hosted on Huggingface for this study.

Dolly is one of the earlier instruction-tuned models released subsequently after ChatGPT. The model is finetuned with a publicly accessible dataset containing 15K human-generated prompt-response pairs collated by Databricks conforming to tasks such as classification, closed and open QA, summarization, and trained on top of EleutherAI’s 3B Pythia model (Biderman et al., 2023). We use the standard 3B model¹⁰ for this study available on Huggingface.

ChatGPT (GPT-3.5-Turbo) is the only closed-source model we consider within our computing budget. We include this model in our experimentation since ChatGPT is globally recognized and one of the few models with a publicly accessible interface. For this study, we use the latest regular-sized GPT-3.5-Turbo context model covering up to 2021 in its training data through the OpenAI API¹¹. We label this model as *close-sourced* since there are no publicly available reports about its data and training procedures.

2.3 Assessment Standards as Evaluation Metrics

We select two standard metrics used by educators in assessing the quality and complexity of texts in a classroom setting described below:

Flesch Kincaid Grade Level (FKGL) (Kincaid et al., 1975) is a simple but long-standing readability formula used in all aspects of text quality assessment both in globally recognized text editing software such as Microsoft Word as well as in text complexity and simplification research (Wubben et al., 2012; Shardlow, 2014; Scarton and Specia, 2018; Alva-Manchego et al., 2020; Maddela et al., 2021; Alva-Manchego et al., 2021; Tanprasert and Kauchak, 2021). Derived from the original Flesch Reading Ease formula (Flesch, 1948), FKGL considers surface-level variables such as the total number of words TW , sentences TS , and syllables TSL . In terms of output, FKGL provides a score x within the range $[0, 18]$, where lower values indicate easier readability (e.g. short stories) and higher values denote increased complexity (e.g. academic papers). We show the formula of FKGL below:

$$FKGL = 0.39\left(\frac{TW}{TS}\right) + 11.8\left(\frac{TSL}{TW}\right) - 15.59 \quad (1)$$

Common European Framework of Reference for Languages (CEFR)¹² is one of the most well-known language learning assessment metrics globally developed by the Council of Europe and is often used as a basis to grade complexity levels of reading materials and educational content for foreign language learners. CEFR uses a six-point reference scale (A1, A2, B1, B2, C1, C2), which denotes increasing levels of complexity when used to grade texts for various learners. In order to identify the CEFR levels of the generated texts of the instruction-following LLMs used in the study, we use the separate SVM classifier model from the work of Xia et al. (2016) trained with the Cambridge Exams dataset composed of CEFR-ready data from A2 to C2. The SVM model was developed by extracting over 150+ linguistic features ranging from traditional, lexico-semantic, parse tree, and discourse-based features and performs at an accuracy of 0.803, as reported in the paper. We tried training the feature set using an optimized

⁸<https://huggingface.co/bigscience/bloomz-3b>

⁹<https://huggingface.co/akoksal/LongForm-T5-XL>

¹⁰<https://huggingface.co/databricks/dolly-v2-3b>

¹¹<https://platform.openai.com/docs/guides/gpt>

¹²<https://www.coe.int/en/web/common-european-framework-reference-languages>

Random Forest, which obtained a higher accuracy of 0.836, and used this model instead for this work.

2.4 The European Language Grid (ELG) Data

For this study, we requested the CEFR corpus from the **European Language Grid (ELG)**¹³ which contains over 1,200 text passages from a diverse range of genres such as fiction, science, and history distributed over the six CEFR scales (A1 to C2). From the data, we selected only those text passages that strictly belong to one scale (ex. C2) and disregarded the A1 level due to having only 24 documents and to also conform to the CEFR classifier by [Xia et al. \(2016\)](#) used for generation analysis. We balanced the number of entries for each level (60) in order to have a uniform distribution and even comparison for later discussion of results.

We describe in Table 1 an overview and some basic statistics of the collected ELG dataset. From the Table, a linear relationship can be observed where as the CEFR complexity level increases from A2 to C2, the variables of average word count, sentence count, and corresponding FKGL levels also accumulate.

Levels	Size	Ave WC	Ave SC	Ave FKGL
A2	60	186.55	18.91	3.32
B1	60	264.25	15.90	6.83
B2	60	517.71	31.71	6.91
C1	60	728.93	40.70	8.61
C2	60	749.73	37.55	9.88

Table 1: Statistics of ELG dataset for used prompting instruction-following LLMs. Size denotes the number of document instances per level, Ave WC is the average word count, Ave SC is the average sentence count, and Ave FKGL is the average Flesch Kincaid Grade Level score.

3 Prompt-Based Story Completion

Our first choice of generation task to measure the generation quality of instruction-following language models is the open-ended story completion. We selected this task as it aligns with the natural task of teachers prompting language model-driven interfaces such as ChatGPT for educational content generation such as stories or short narratives ([Kasneji et al., 2023](#); [Whalen et al., 2023](#)).

¹³<https://live.european-language-grid.eu/catalogue/corpus/9477>

3.1 Procedure

For the prompt-based story completion setup, we split each narrative entry from the ELG corpus into prompt-continuation pairs. Each prompt is composed of 50-70 words to provide enough context for the language models, and we set the specifications for each model to generate text with a minimum of 30 and a maximum of 300 new tokens, respectively. In terms of decoding, we set the nucleus sampling hyperparameter $top-p$ to 0.95 following the recommendation of [DeLucia et al. \(2021\)](#) stating a value of 0.9 or higher is the best for narrative generation.

As reported in Table 2, we use four styles of instructional prompting where specific grade levels, the name of the assessment framework, and its description are added iteratively to find out if the increasing information on readability specification will be captured and have a substantial effect on the complexities of instruction-following models’ generation quality. We customized the different levels of instructional prompts for both the FKGL and CEFR assessment standards. We replace the {text} token with the prompts from the ELG corpus before sending the entire instruction to each model for generation.

3.2 Results and Insights

Figures 1 and 2 report the performances of the six instruction-tuned models for the story completion task evaluated using the FKGL and CEFR. Actual values from the formula are used for FKGL, while accuracy scores are used to report a model’s performance for CEFR. We include additional tables for the mean and standard deviations of FKGL scores in Appendix A.

Formula-based readability specifications in prompts struggle to control model generations.

Using the FKGL as guiding information for generating story completions for Grade 2, none of the models in any of the prompt iterations with increasing readability information specification achieved acceptable performance that is within the range of $1 < FKGL(x) < 3$. This finding may indicate that formula-based text complexity metrics aside from FKGL, such as SMOG ([Mc Laughlin, 1969](#)), Dale-Chall ([Dale and Chall, 1948](#)), and Coleman-Liau Index ([Coleman and Liau, 1975](#)) that use other forms of predictors beyond total word, sentence, and syllable counts may also not be captured well by instruction-tuned language models unless an ex-

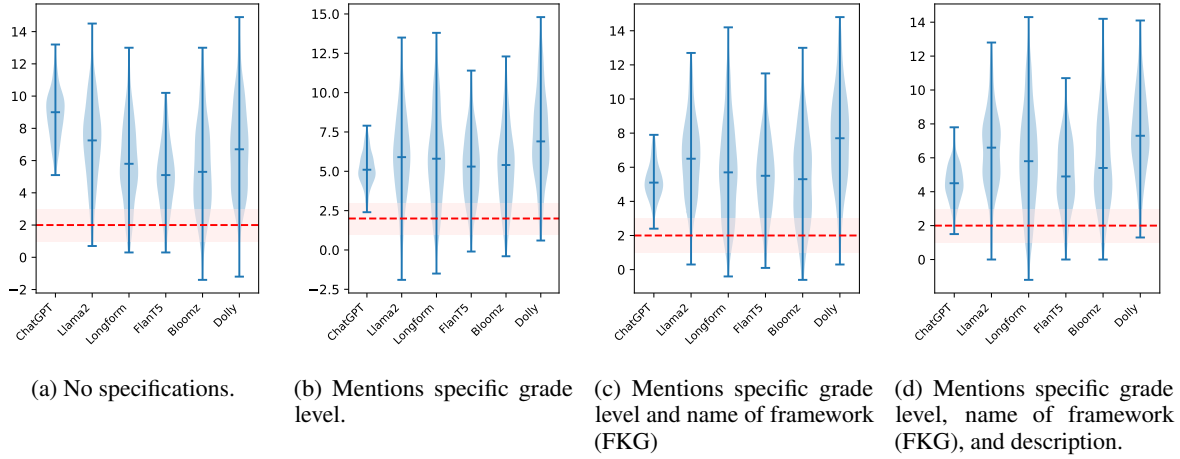


Figure 1: Performance via mean Flesch Kincaid Grade Level (FKGL) scores of each instruction-tuned language model for each prompt specification style for the **story completion subtask**. The red line and shading indicate the center and the region of acceptable values that are within the target complexity level of the generated text, which is Grade 2.

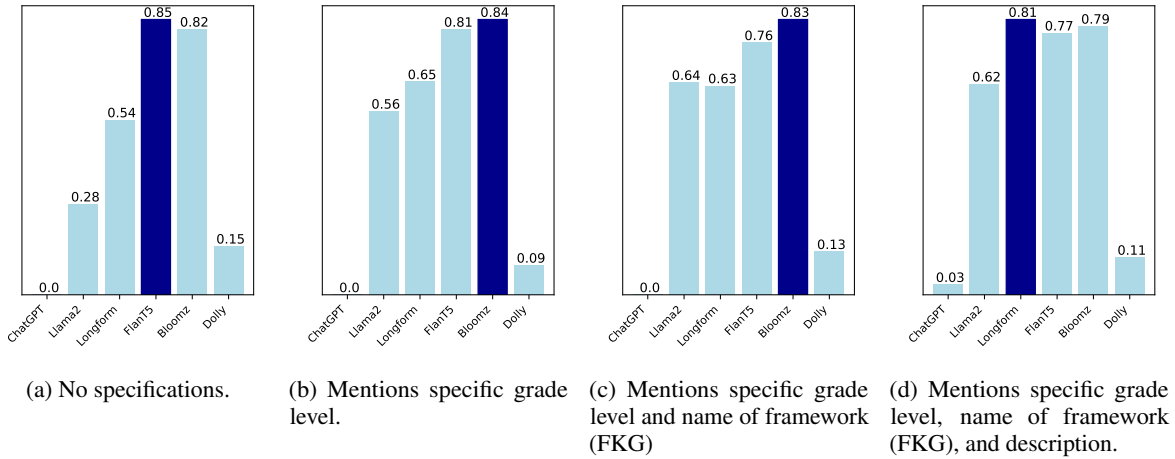


Figure 2: Performance via accuracy scores of each instruction-tuned language model for each prompt specification style for the **story completion subtask** on the Common European Framework of Reference for Languages (CEFR) standard. The top performing model is highlighted in dark blue.

PLICIT series of computation is provided within the prompts. This limitation may prove to be counter-intuitive as the desired goal is to have the models approximate the readability levels internally to guide its generations instead of the use, but nonetheless, it is still an interesting research challenge.

Going deeper into the analysis, we look at the mean and standard deviations of each model for each iteration style. Without any specifications of grade level, metric, and description, ChatGPT (GPT-3.5-Turbo) achieved the worst performance with a mean of 8.832 ($SD = 1.549$) for its FKGL scores from its generations while FlanT5 obtained the closest to the desired range $1 < FKGL(x) < 3$ with 5.133 ($SD = 2.063$). Interestingly, while none of the models were able to provide

generations within the acceptable boundary for FKGL, we observe that only one model, ChatGPT (GPT-3.5-Turbo), showed stable *improving* scores with the increasing detailedness of the readability information specification in the prompts with a mean trend of $8.832 \rightarrow 5.155 \rightarrow 5.224 \rightarrow 4.567$. We attribute the performance of this model to its implementation of RLHF to improve alignment to human preferences across a range of tasks (Ouyang et al., 2022). Moreover, since this model is the only one in the set to have a public-facing interface that teachers and educators use, this finding provides empirical support to the various published recommendations by the education community (Staake, 2023; Herft, 2023) to further *specify* the readability level and assessment framework

of choice when using these models for content generation, especially ChatGPT.

Prompt Style	Prompt Content
No grade level specifications.	<p>(Write a story using the following prompt)</p> <p>[Simplify the following narrative] {text}</p>
Mentions specific grade level (Grade 2 or A2).	<p>(Write a story that is readable by Grade 2 learners using the following prompt)</p> <p>[Simplify the following narrative for Grade 2 learners] {text}</p> <p>(Write a story that is readable by A2 learners in the using the following prompt)</p> <p>[Simplify the following narrative for A2 learners] {text}</p>
Mentions specific grade level and name of the framework (FKG or CEFR).	<p>(Write a story that is readable by Grade 2 learners in the Flesch-Kincaid Grade Level scale using the following prompt)</p> <p>[Simplify the following narrative for Grade 2 learners in the Flesch Kincaid Grade scale] {text}</p> <p>(Write a story that is readable by A2 learners in the CEFR scale using the following prompt)</p> <p>[Simplify the following narrative for A2 learners in the CEFR scale] {text}</p>
Mentions specific grade level, name of framework (FKG or CEFR), and description.	<p>(Write a story that is readable by A2 learners in the CEFR scale using the following prompt. Text assessed as A2 level in CEFR uses basic sentence patterns, explicit information and a limited number of information points)</p> <p>[Simplify the following narrative for Grade 2 readers in the Flesch-Kincaid Grade scale. The Flesch-Kincaid Grade scale looks at total words, total sentences, and total syllables in a text] {text}</p> <p>(Write a story that is readable by A2 learners in the CEFR scale using the following prompt. Text assessed as A2 level in CEFR uses basic sentence patterns, explicit information and a limited number of information points)</p> <p>[Simplify the following narrative for A2 learners in the CEFR scale. Text assessed as A2 level uses basic sentence patterns, explicit information, and limited number of information points] {text}</p>

Table 2: The various iterations of instructional prompts used for the generation setup of the **(story completion)** and **[narrative simplification]** tasks with respect to information of grade level, framework, and description specifications.

Models finetuned on diverse multi-task instruction datasets are more aligned with readability. Using CEFR as the guiding standard for readability level specification, we see favorable results from open-sourced models such as BLOOMZ, FlanT5, LLaMA 2, and Longform, which all include extremely diverse instruction-tuned datasets for their finetuning phase. FlanT5 obtained the

best performance for no specification prompts with 0.85 accuracy while BLOOMZ performs the best of all models for prompts that specify target grade level and assessment metric name with 0.84 and 0.83 accuracies, respectively. Longform and LLaMA 2, on the other hand, have the most observable improvements across the board, where the accuracies for generating aligned story completions with respect to the prompts increases linearly as the information on readability is expanded: 0.54 \rightarrow 0.65 \rightarrow 0.63 \rightarrow 0.81 for Longform and 0.28 \rightarrow 0.56 \rightarrow 0.64 \rightarrow 0.62 for LLaMA 2.

In terms of poorly performing models, ChatGPT and Dolly obtained 0 – 13% accuracies across all prompts. Upon manual inspection of the generated outputs of these two models, we see a misclassification rate of over 90% from these models due to the tendency that they produced outputs are one level higher than the target level, which is B1 instead of A2 in the CEFR scale. This finding means that these models generate texts but are not as precise with the readability specifications as the other open-sourced ones like BLOOMZ and LLaMA 2 for the CEFR scale. While we do not know what datasets were used for training ChatGPT as it is closed-source, we attribute the poor performance of Dolly to the very limited variety of instruction datasets with a size of only 15K used for its finetuning compared to the diverse multi-task data used in FlanT5, Longform, LLaMA 2, and BLOOMZ (Muennighoff et al., 2022; Chung et al., 2022; Köksal et al., 2023; Touvron et al., 2023a)

4 Prompt-Based Narrative Simplification

Our second choice of generation task is to measure the capability of instruction-following language models to simply short text passages and narratives into a target readability level. Similar to story completion, this task is also aligned with how teachers can use these models to simplify a piece of educational content if it is too complex for a target learner audience (Kasneci et al., 2023; Whalen et al., 2023; Pu and Demberg, 2023).

4.1 Procedure

For narrative simplification, we select only the advanced levels on the CEFR scale, which are C1 and C2, from the ELG dataset. The justification for this is that since the task is simplification, we want the initial text to come from a higher level. A total of 120 advanced-level entries were obtained, and

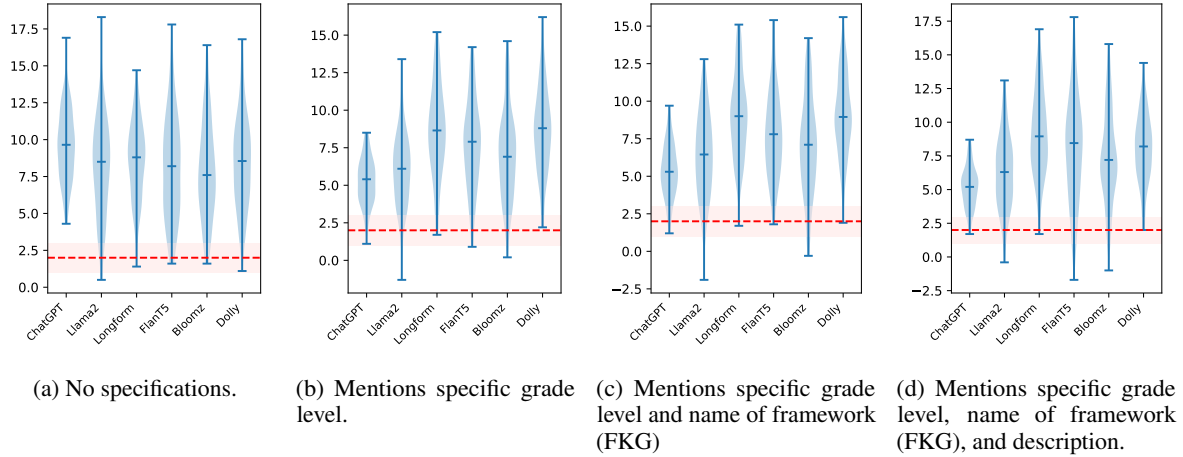


Figure 3: Performance via mean Flesch Kincaid Grade Level (FKGL) scores of each instruction-tuned language model for each prompt specification style for the **narrative simplification subtask**. The **red line and shading** indicate the center and the region of acceptable values that is within the target complexity level of the generated text, which is Grade 2.

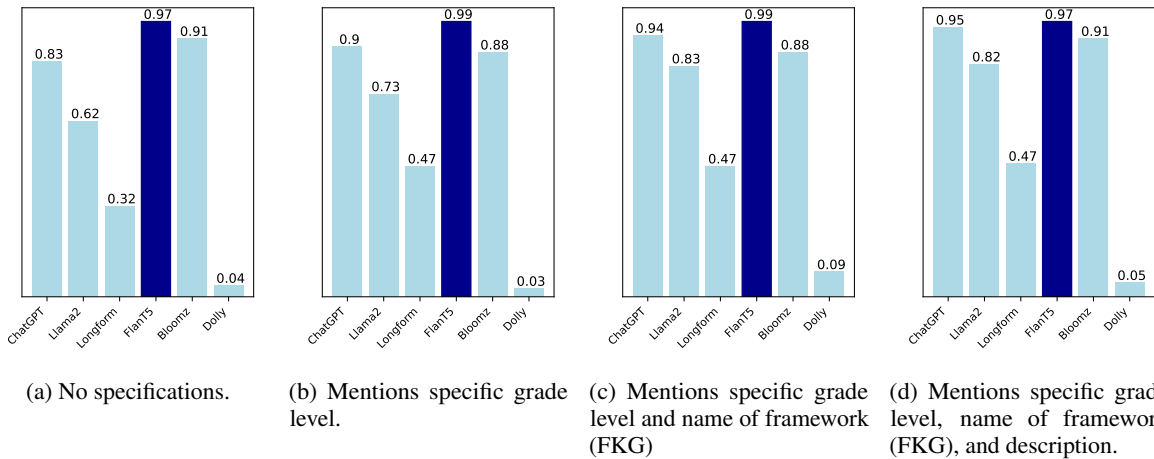


Figure 4: Performance via accuracy scores of each instruction-tuned language model for each prompt specification style for the **narrative simplification subtask** on the Common European Framework of Reference for Languages (CEFR) standard. The top performing model is highlighted in **dark blue**.

we split each one to get the first 100-150 words to be appended with the instructional prompts for simplification. We specified the models to generate at least a minimum of 30 and a maximum of 300 new tokens. A nucleus sampling hyperparameter $top-p$ to 0.95 is also used. Similar to story completion, we use four styles of instructional prompting where specific grade levels, the name of the assessment framework, and its descriptions as reported in Table 2.

4.2 Results and Insights

Instruction-tuned models also struggle in simplification tasks using FKGL specifications. Referring back to the average FKGL scores per CEFR level presented in Table 1, the advanced C1 and

C2 levels have a mean of 8.91 and 9.88, respectively, while the target level for this narrative simplification task is A2 with 3.32. Looking at the performances of models illustrated in Figure 3, similar to the story completion subtask, we see that controlling for the readability level, regardless of how informative the prompt is proves to be challenging for all instruction-tuned models evaluated in the study. Models including BLOOMZ, Longform, FlanT5, and Dolly all show similar patterns of inconsistencies across all four prompt styles with various levels of readability specifications. While none of the models were able to produce generations that are within the acceptable range of $1 < FKGL(x) < 3$ for narrative simplification, the ChatGPT and LLaMA 2 models

show improvement of scores as the readability information provided with the prompt is enhanced with $9.570 \rightarrow 5.285 \rightarrow 5.390 \rightarrow 5.210$ and $8.221 \rightarrow 6.137 \rightarrow 6.471 \rightarrow 6.339$ for each model respectively. We also report a difference of 4.36 and 1.882 from the prompt with no specification of target readability level vs. the prompt with the readability level, metric name, and description for ChatGPT and LLaMA 2, respectively.

From this finding, we echo the same inference from the story completion task, where the reason why these models were not able to fully capture the desired reading level from the generations can be attributed to the need for actual computation information present in the prompt. We also attribute the improvement shown by ChatGPT and LLaMA 2 to the efficacy of the RLHF algorithm and rejection sampling (Ouyang et al., 2022; Touvron et al., 2023a,b) used for optimizing these models, which may have helped in the refinement of generation quality as the prompt becomes more informative. Still, we encourage specifying necessary information about the target audience’s reading level and the type of assessment used when prompting models in order to minimize the generation of overly complex texts.

Less diverse corpora for instruction-following models results in consistently poor simplification ability. Using the CEFR framework to guide instruction-tuned models for narrative simplification obtained better results in general compared to using FKGL. We report the accuracies of models in simplifying advanced-level passages from the C1 and C2 scale of the ELG corpus down to the desired readability level of A2 in Figure 4. From the results, FlanT5 is the best model with consistent performances across all prompts with an average accuracy of 98%—even the ones without specification of target reading level. We cross-examined existing literature and came across several works that support T5-based models’ general performance for sentence and narrative-level simplification for English (Sun et al., 2023; Maddela et al., 2023). The second best-performing models are taken by ChatGPT, BLOOMZ, Longform, and LLaMA 2, which all showed consistent minor improvements as the prompts became more detailed by adding the specific name of the framework and the characteristic of the target readability level. Lastly, the Dolly model performed the worst for the task without

an accuracy not going beyond 10%. Upon manual reviewing of the outputs of this model, we see that most of its generations are classified under one level higher, B1, than the target reading level, A2. We attribute this poor performance to the low diversity of instruction dataset used for Dolly compared to the collection of multitask corpora used for finetuning FlanT5 models (Chung et al., 2022).

5 Related Work

The majority of literature on evaluating instruction-tuned models has spotlighted ChatGPT due to its global recognition amongst interdisciplinary fields. Specifically, these evaluation works have focused on aspects such as multilinguality (Bang et al., 2023; Gowriraj et al., 2023; Zhang et al., 2023a), reasoning (Qin et al., 2023; Laskar et al., 2023), truthfulness (Laskar et al., 2023), toxicity (Guo et al., 2023; Ouyang et al., 2022) to name a few. In terms of incorporating forms of control to guide generations, related works have explored style (Keskar et al., 2019), tone (Sennrich et al., 2016), topic coherence (Tang et al., 2019; Chang et al., 2021; Krishna et al., 2022), sentiment and emotion (Dathathri et al., 2019; Khalifa et al., 2020), and text complexity (Imperial, 2022; Pu and Demberg, 2023; Murgia et al., 2023). The main gap in literature that our study fills is the evaluation of LLMs and their alignment with real-world text assessment standards used by teachers, such as the CEFR framework.

6 Conclusion

In this work, we tackled a unique perspective of evaluating the capabilities of instruction-tuned language models by integrating readability-specific information anchored on realistic assessment standards such as the CEFR framework used by teachers and educators. Our findings expose the advantages and weaknesses of open and closed-source generative models such as LLaMA, FlanT5, and ChatGPT for the story completion and narrative simplification tasks, in which we trace back each model’s performance to the quality of instruction datasets used for finetuning them. We hope this study sheds light on both the technical and non-technical audiences, especially the members of the education community, regarding the true capabilities of these generative models in producing educational content.

Limitations

On use of FKGL for measuring simplification systems. We are well aware of the limitations of FKGL for evaluating the performances of simplification systems as highlighted in [Tanprasert and Kauchak \(2021\)](#). However, our choice of metrics and assessment standards, FKGL and CEFR, is made through the selection of those that are often used by teachers and educators in assessing the complexities of texts. Metrics such as SARI ([Xu et al., 2016](#)) and BLEU ([Papineni et al., 2002](#)), on the other hand, are researcher-facing technical metrics used for engineering and evaluating simplification systems. Nonetheless, combining all of these technical and non-technical metrics and their interactions may be a good future study for this work.

On experiments exclusively with English data. All experiments, findings, and insights in this work only apply to English, as evidenced by the language of the datasets used. Thus, our findings may not generalize if similar research derived from this work is to be done with other languages using other models, such as those trained with multilingual data.

On the use of base versions of instruction-tuned models. As mentioned in Section 2, we used the standard sizes of generative models since we did not have the required hardware to use the largest versions of a model family (ex. 70B version of LLaMA 2). The analysis of the effects of scale for these models in terms of capturing readability standards may be pursued as future work of this study.

Ethics Statement

The ELG corpus is publicly accessible through a request form provided by the website. We use the six open and closed-source instruction models only for the tasks of story completion and narrative simplification in this study. We believe the model generations to be free of harmful content to an average reader.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification:](#)

[Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. [The \(un\)suitability of automatic evaluation metrics for text simplification](#). *Computational Linguistics*, 47(4):861–889.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Haw-Shiuan Chang, Jiaming Yuan, Mohit Iyyer, and Andrew McCallum. 2021. [Changing the mind of transformers for topically-controllable language generation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2601–2611, Online. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

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.

- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Srinivas Gowriraj, Soham Dinesh Tiwari, Mitali Potnis, Srijan Bansal, Teruko Mitamura, and Eric Nyberg. 2023. [Language-agnostic transformers and assessing ChatGPT-based query rewriting for multilingual document-grounded QA](#). In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 101–108, Toronto, Canada. Association for Computational Linguistics.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
- Andrew Herft. 2023. [A Teacher’s Prompt Guide to Chatgpt](#). *Herft Educator*.
- Joseph Marvin Imperial. 2022. Uniform complexity for text generation. *arXiv preprint arXiv:2204.05185*.
- Neil Jones and Nick Saville. 2009. European language policy: Assessment, learning, and the cefr. *Annual Review of Applied Linguistics*, 29:51–63.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2020. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
- 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*.
- Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.
- Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. [RankGen: Improving text generation with large ranking models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. [A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.
- Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. [An exploration of data augmentation and sampling techniques for domain-agnostic question answering](#). In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 220–227, Hong Kong, China. Association for Computational Linguistics.
- Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. [Controllable text simplification with explicit paraphrasing](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. [LENS: A learnable evaluation metric for text simplification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.
- G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.
- Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. 2023. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.
- Emiliana Murgia, Maria Soledad Pera, Monica Landoni, and Theo Huibers. 2023. Children on chatgpt readability in an educational context: Myth or opportunity? In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 311–316.
- 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.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Justin Paul, Akiko Ueno, and Charles Dennis. 2023. Chatgpt and consumers: Benefits, pitfalls and future research agenda.
- Dongqi Pu and Vera Demberg. 2023. [ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Carolina Scarton and Lucia Specia. 2018. [Learning simplifications for specific target audiences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.
- Jill Staake. 2023. [20 Ways Teachers Can Use Chatgpt To Make Their Lives Easier](#). *We Are Teachers*.
- Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. [Teaching the pre-trained model to generate simple texts for text simplification](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355, Toronto, Canada. 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.
- Hongyin Tang, Miao Li, and Beihong Jin. 2019. [A topic augmented text generation model: Joint learning of semantics and structural features](#). 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 5090–5099, Hong Kong, China. Association for Computational Linguistics.
- Teerapaun Tanprasert and David Kauchak. 2021. [Flesch-kincaid is not a text simplification evaluation metric](#). In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, pages 1–11.
- 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*.
- Rose Wang and Dorottya Demszky. 2023. [Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv e-prints*, pages arXiv–2204.
- 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.
- Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. [Sentence simplification by monolingual machine translation](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.
- 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.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, and Alham Fikri Aji. 2023a. Multilingual large language models are not (yet) code-switchers. *arXiv preprint arXiv:2305.14235*.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Mean and standard deviations of FKGL scores from model generations.

A.2 Sample generations from different prompt styles.

Models	Prompt Style #1	Prompt Style #2	Prompt Style #3	Prompt Style #4
ChatGPT	8.832 (1.549)	5.155 (1.087)	5.224 (1.060)	4.567 (1.128)
BLOOMZ	5.618 (2.840)	5.379 (2.579)	5.343 (2.713)	5.949 (2.854)
Longform	5.935 (2.622)	5.907 (2.952)	5.882 (2.871)	5.950 (3.028)
FlanT5	5.133 (2.063)	5.343 (2.234)	5.555 (2.204)	5.051 (2.036)
Dolly	6.777 (2.753)	7.182 (2.853)	7.659 (2.818)	7.443 (2.478)
Llama 2	7.165 (2.597)	5.970 (2.804)	6.614 (2.346)	6.487 (2.305)

Table 3: Mean and (standard deviation) of FKGL scores of each model for each prompt iteration of the story completion subtask.

Models	Prompt Style #1	Prompt Style #2	Prompt Style #3	Prompt Style #4
ChatGPT	9.570 (2.368)	5.285 (1.526)	5.390 (1.762)	5.210 (1.516)
BLOOMZ	7.975 (3.312)	7.086 (2.974)	7.049 (3.248)	7.468 (3.146)
Longform	8.585 (2.660)	8.616 (3.104)	9.070 (2.772)	9.136 (3.111)
FlanT5	8.158 (3.486)	7.728 (2.979)	7.876 (2.911)	8.018 (3.920)
Dolly	8.453 (3.092)	8.616 (2.864)	9.081 (2.671)	8.033 (2.541)
Llama 2	8.221 (3.257)	6.137 (2.603)	6.471 (2.824)	6.339 (2.698)

Table 4: Mean and (standard deviation) of FKGL scores of each model for each prompt iteration of the narrative simplification subtask.

PROMPT: Write a story using the following prompt:

People have always made buildings. We need houses to keep us warm and dry and we build stadiums so we can watch football matches and pop concerts. We use other buildings like museums to keep beautiful things in. Many people work together on a building...

CHATGPT: Once upon a time, in a bustling city called Archwood, people had always believed in the power of buildings. Each structure stood tall, carrying an essential purpose in the lives of its inhabitants. From cozy homes that sheltered them from the harsh elements to grand stadiums that ignited their passion for sports and entertainment, buildings were the fabric that held their society together.

LLAMA 2: How many bricks does a building need? That’s a great question. It depends on the size of the building. A small house might need 100 bricks, while a skyscraper could need 10,000 bricks. But no matter how many bricks a building needs, it’s the people who make it happen. They work together, using their different skills and talents, to create something amazing. And when they’re finished, they stand back and admire their creation, proud of what they’ve achieved.

LONGFORM: The architecture is the building that the architects made. This architect worked really hard and produced lots of drawings and sketches.

FLANT5: We are building a new school at the old school. This year our team is building a home. We have many people working together on the building. We need buildings to keep us warm and dry, stadiums to keep us warm and dry, and other things to keep beautiful things in. The architect decides how the building is going to look and draws pictures that show people his or her ideas.

BLOOMZ: Buildings are very strong and last a long time. Some of our buildings can last forever. Many of the people building buildings are old. They’ve worked in the building for many years.

DOLLY: Other people work on the plans, doing different jobs like laying bricks, painting the outside, and installing the electricity and plumbing. Then someone lets the people decide if they like the building. If they don’t like it they can change it. If they like it, it’s built. People even come up with new ideas for buildings as new jobs are needed.

Table 5: Sample generations using prompt with *no readability specifications* for the **story completion subtask**.

PROMPT: Write a story that is readable by Grade 2 learners in the Flesch-Kincaid Grade Level scale using the following prompt. The Flesch-Kincaid Grade scale considers the total words, total sentences, and total syllables in a text:

People have always made buildings. We need houses to keep us warm and dry and we build stadiums so we can watch football matches and pop concerts. We use other buildings like museums to keep beautiful things in. Many people work together on a building...

CHATGPT: Once upon a time, there were lots of buildings in a big city. People loved to make buildings because they were really important. Buildings keep us safe and dry, just like our cozy houses. One kind of building that people love is a stadium. A stadium is a massive place where we can watch exciting football matches and concerts with our favorite pop stars. We cheer loudly when our team scores a goal or when the singer sings our favorite song. It's so much fun!

LLAMA 2: The builder is in charge of building the building. He buys the materials and uses machines to put them together. The electricians and plumbers make sure the building has electricity and running water. The engineers make sure the building can withstand earthquakes and hurricanes. Buildings are important because they protect us from bad weather and keep us warm and dry.

LONGFORM: People have always made buildings. We need houses to keep us warm and dry and we build stadiums so we can watch football matches and pop concerts. We use other buildings like museums to keep beautiful things in. Many people work together on a building. The architect decides how the building is going to look and draws pictures that show people his or her ideas.

FLAN T5: People have always made houses. Many people have always made houses. One person likes to build a house more than other people. Another likes to make a house more beautiful than others. Then one person brings his or her ideas and teaches them the following things.

BLOOMZ: Then the builder starts work building the building. People use the building to do the things that the architect has decided. The building has people inside who live in it. People also work on a building. It has people inside it. People often ask, "Who are you?" Question 1: How do people keep warm and dry?

DOLLY: Sometimes people hire people to do jobs that make buildings work. The bricklayer cuts stones from a quarry and builds the walls of the stadium and the house. A painter puts the walls white so they will catch the light. In the end, the bricks and stones make the buildings that people use to be happy.

Table 6: Sample generation using prompt with *full readability specifications* (grade level, standard name, and description) using FKGL for the **story completion subtask**.

PROMPT: Write a story that is readable by A2 learners in the CEFR scale using the following prompt. Text assessed as A2 level uses basic sentence patterns with memorised phrases, explicit information, and limited number of information points:

People have always made buildings. We need houses to keep us warm and dry and we build stadiums so we can watch football matches and pop concerts. We use other buildings like museums to keep beautiful things in. Many people work together on a building...

CHATGPT: Once upon a time, in a small town, there were people who loved to build buildings. They knew that buildings were very important for us because they help us stay warm and dry. One day, the people in the town decided that they needed a big stadium. They wanted a place where they could enjoy watching exciting football matches and exciting pop concerts. So, they all worked together to build it.

LLAMA 2: They call the people who build the building contractors. Contractors work with architects to build the building.

LONGFORM: People have always made buildings. We need houses to keep us warm and dry and we build stadiums so we can watch football matches and pop concerts. We use other buildings like museums to keep beautiful things in. Many people work together on a building. The architect decides how the building is going to look and draws pictures that show people his or her ideas.

FLANTS5: At first people didn't want buildings. They wanted houses. At first they looked at houses. Then they looked at soccer matches. They looked at stadiums. They looked at concerts. They looked at baseball games. People have always built buildings. People are building houses to keep us warm and dry.

BLOOMZ: The builder builds the building. They start to build a city and they also build villages. People come to cities to look for something to do. A building is made of lots of things. When a building is made, they are put on a truck and delivered to the town, village, or village to village.

DOLLY: The builders and masons put the bricks together to make the walls, and the electricians and plumbers do the wiring and the plumbing, so the building can be used. We use brick, stone and wood. Brick and stone are both made from rocks that are found in the ground and are therefore the building materials that people have been using for thousands of years. Brick is the oldest and most common building material. Stone is also a hard material, although a bit softer than brick. Wood is not a building material, because people do not like the idea of using wood for buildings.

Table 7: Sample generation using prompt with *full readability specifications* (grade level, standard name, and description) using the CEFR scale for the **story completion subtask**.

PROMPT: Simplify the following narrative:

We changed lives for a day! Have you ever dreamed about swapping lives with someone else for a day? Perhaps you think it would be more fun to do something quite different for a change? We fixed it for four people - read how it went. Amanda (23) a young mother with two daughters swapped lives with her friend Cindy (30) who works as a waitress at a glitzy London restaurant and entertainment complex. A. Amanda's story I'd never been inside anywhere remotely like it before. I couldn't believe how dark and noisy it was - there were so many people and such a brilliant atmosphere...

CHATGPT: We made dreams come true for a day. Four people swapped lives and had an unforgettable experience. Amanda, a young mother with two daughters, traded places with her friend Cindy, a waitress at a glamorous London restaurant. Amanda was amazed by the dark and noisy atmosphere and the lively crowd.

LLAMA 2: I had to wear a ridiculous uniform and serve drinks to all these strange people. I was so nervous and felt like I was going to be sick. Cindy was great, though. She took me under her wing and showed me how to do everything. We chatted and laughed and had a great time. I was amazed at how much I enjoyed myself! Amanda was a complete nightmare! She was so nervous and clumsy that she knocked over a tray of drinks and spilled them all over the floor.

LONGFORM: A beautiful day for a perfect swap. As Amanda writes: I spent the afternoon at home with my two little daughters. It was my first day of work since I was a waitress! I loved being in the sun with my friends and enjoying the company of my work colleagues! It was one of the best days I've had in my life. I love the fact that I get to give my daughters a chance to learn about another culture while also working part-time.

FLAN T5: Amanda and Cindy had changed their lives for a day. Amanda wanted to do something differently and to try something new.

BLOOMZ: After five minutes, I was so excited it was too hard to take my eyes off of my phone. After we got inside, my eyes never closed. I felt overwhelmed with emotion, but more excited than nervous - my friend could not believe how good it was. I just knew we were going to love it. B. Cindy's story when we walked up to the restaurant, I was really nervous, especially because I was nervous and a bit shy. The staff were so nice and friendly, the music was so much fun, and we were on the balcony overlooking the city - it was great!

DOLLY: Everyone was so busy, the noise was overwhelming and I felt intimidated and out of place. It was a bit scary too. Cindy seemed so calm though, it was like she could multitask and juggle all of her work and friends and social engagements so easily. Her energy was infectious - I felt like I could easily do what she was doing. So we set a day to swap and the morning of the day we swapped our lives, we went to the restaurant and enjoyed a nice breakfast together - chatting and laughing. Then we went to the apartment building where Cindy works and she introduced me to the complex.

Table 8: Sample generations using prompt with *no readability specifications* for the **narrative simplification subtask**.

PROMPT: Simplify the following narrative for Grade 2 readers in the Flesch-Kincaid Grade scale. The Flesch-Kincaid Grade scale looks at total words, total sentences, and total syllables in a text:

We changed lives for a day! Have you ever dreamed about swapping lives with someone else for a day? Perhaps you think it would be more fun to do something quite different for a change? We fixed it for four people - read how it went. Amanda (23) a young mother with two daughters swapped lives with her friend Cindy (30) who works as a waitress at a glitzy London restaurant and entertainment complex. A. Amanda's story: I'd never been inside anywhere remotely like it before. I couldn't believe how dark and noisy it was - there were so many people and such a brilliant atmosphere...

CHATGPT: We switched lives for a day! Do you ever imagine what it would be like to be someone else for a day? We made it happen for four people. Amanda, who is a young mother with two daughters, switched lives with her friend Cindy. Cindy works at a fancy restaurant and entertainment place in London. Amanda's story: I had never been inside a place like that before. It was so dark and loud, with lots of people and a great feeling in the air.

LLAMA 2: I was amazed at how Cindy coped with it all. I was scared of everything! I loved the food, though, especially the fish and chips. I took a photo of everything I ate and drank. I couldn't wait to get home and tell my husband and daughters about my adventure. B. Cindy's story It was so strange to be at home with Amanda's two little girls. They were so well-behaved and quiet, unlike my own kids! I was amazed at how quickly Amanda got everything done - she was like a whirlwind of activity. I was exhausted just watching her! I loved the quietness of the house and the way the sunlight streamed in through the windows. It was so peaceful.

LONGFORM: I have just started work in the restaurant. It is very loud and busy. I like my job, I like my colleagues and the atmosphere is brilliant. It was such a change to be living with Cindy for the day. I could really see how she thinks.

FLAN T5: I can't believe how dark and noisy it was - there were so many people and such a brilliant atmosphere.

BLOOMZ: I was surprised that it was a restaurant and not some kind of bar. Amanda and Cindy swapped lives with each other at the West London restaurant for a whole day for a change.

DOLLY: This is why I suggested we go for a drive instead. I'm so looking forward to not being here any more, I'll be able to get to know Cindy all over again! Cindy's story My day began with a bit of a shock. I was expecting Amanda to be someone I'd never met, but when my turn came I found out she was the old me. I was surprised and confused but decided to act as if nothing was wrong. Things were pretty chaotic when we arrived so we split up to cover more ground. As I explored the place I started to wonder what I'd done wrong.

Table 9: Sample generations using prompt with *full readability specifications* using FKGL for the **narrative simplification subtask**.

PROMPT: Simplify the following narrative for A2 learners in the CEFR Scale. Text assessed as A2 level uses basic sentence patterns with memorised phrases, uses explicit information and limited number of information points:

We changed lives for a day! Have you ever dreamed about swapping lives with someone else for a day? Perhaps you think it would be more fun to do something quite different for a change? We fixed it for four people - read how it went. Amanda (23) a young mother with two daughters swapped lives with her friend Cindy (30) who works as a waitress at a glitzy London restaurant and entertainment complex. A. Amanda's story: I'd never been inside anywhere remotely like it before. I couldn't believe how dark and noisy it was - there were so many people and such a brilliant atmosphere...

CHATGPT: We changed lives for a day! Have you ever thought about trading lives with someone else for a day? Maybe you think it would be more fun to do something completely different for a change? We made it happen for four people. Read about Amanda and Cindy's swap. Amanda (23) is a young mother with two daughters. She switched lives with her friend Cindy (30), who works as a waitress at a fancy restaurant and entertainment complex in London. A. Amanda's story: I had never been to a place like it before. It was really dark and loud. There were so many people and the atmosphere was amazing.

LLAMA 2: Cindy showed me around and introduced me to everyone, and I felt like I was part of something special. We chatted for hours and I learned so much about her life and the people she works with. I couldn't believe how busy she was all the time! B. Cindy's story Amanda was amazing! I was really nervous at first, but she was so friendly and easy to talk to. I loved showing her around and introducing her to everyone. I was so glad she could come and see what I do every day. We had a great time together and I felt like I'd known her for ages.

LONGFORM: They were all young mothers with two daughters. Cindy was the host mother, Amanda was the waitress. We arranged that both women would work at the same restaurant and entertainment complex in London for an entire day. Cindy was a waitress at a popular restaurant in London, while Amanda was a busy mother with two daughters. They both got to experience life in the opposite direction and swapped lives on the day. We asked them to enact a story they would tell of the day and they did it. 1. They took their daughters to a lovely day out and then drove back to their normal home in England. 2. They got to spend a whole day with their partner or partner's family, then they returned to their normal lives. 4. They were all treated to a delicious dinner, and then they had a bath.

FLAN15: I was thrilled to be seated in an establishment of such grandeur. There was such a large number of people in an environment that was so intimate and unique.

BLOOMZ: The glitzy interior really surprised me - I felt very happy and content after that day. Amanda, 23, changed lives for a day with a friend.

DOLLY: It was overwhelming but such a relief once we were in and I knew Cindy wouldn't get too drunk. The waiter took our order, I ordered a glass of white wine and a pint of ale. We started with appetizers, I asked if she had any of her own. She said there were some chips in the kitchen and that she'd fetch them. She returned with about eight pieces of chips, which I wolfed down in a few big chunks, before I realized I'd left the kitchen. I returned with about 10 rashers of bacon and three rashers of streaky bacon, which she cut up and added to the chips.

Table 10: Sample generations using prompt with *full readability specifications* using CEFR for the **narrative simplification subtask**.