Measuring and Modifying the Readability of English Texts with Large Language Models

Anonymous ACL submission

Abstract

The success of Large Language Models (LLMs) in other domains has raised the question of whether LLMs can reliably assess and manipulate the *readability* of text. We approach this question empirically. First, using a published corpus of 4,724 English text excerpts, we find that readability estimates produced "zero-shot" from GPT-4 Turbo exhibit relatively high correlation with human judgments ($r = 0.76$), out-performing estimates derived from traditional readability formulas. Then, in a pre-registered human experiment ($N = 59$), we ask whether Turbo can reliably make text easier or harder to read. We find evidence to support this hypothesis, though considerable variance in human judgments remains unexplained. We conclude by discussing the limitations of this approach, including concerns about data contamination, as well as the validity of the "readability" construct and its dependence on context, audience, and goal.

1 Introduction

The ease with which a text can be read or understood is called *readability*. Measuring and modifying readability has been a topic of interest for decades (Lively and Pressey, 1923; Flesch, 1948; Crossley et al., 2023b), with potential applications ranging from selecting and curating educational materials (Solnyshkina et al., 2017; Creutz, 2024; Liu and Lee, 2023) to making legal, medical, or other technical documents more accessible (Ghosh et al., 2022; Rosati, 2023; Chen et al., 2023). Methods for *assessing* readability, in turn, include: tests of reading comprehension, formulas incorporating basic text features (Lively and Pressey, 1923; Flesch, 1948) or psycholinguistic variables (Kyle and Crossley, 2015), and approaches using supervised learning to estimate readability from labeled text data (Schwarm and Ostendorf, 2005; Martinc et al., 2021).

Recent advances in Large Language Models (LLMs) (Brown et al., 2020) has led to interest in exploring the capacities and applications of these systems—including measuring and modifying the readability of text (Ribeiro et al., 2023; Li et al., 2023; Crossley et al., 2023a; Patel et al., 2023; Farajidizaji et al., 2023). In the current work, we approach this question empirically.

In Section 2, we describe in more detail past work on measuring and modifying readability of text automatically. We then empirically assess the ability of a state-of-the-art LLM (GPT-4 Turbo) to measure (Section 3) and modify (Section 4) the readability of text. Finally, we conclude by discussing the implications of the current work (Section 5), as well as its limitations (Section 6)—including the construct of "readability" itself.

2 Related work

As described in Section 1, efforts to quantify the readability of text date back at least a century (Lively and Pressey, 1923). For many decades, approaches relied on hand-crafted features thought to correlate with (or be causally implicated in) text readability, such as the average length of words or sentences (Flesch, 1948). As Vajjala (2022) describe, dominant approaches have gradually shifted towards treating readability assessment as a supervised machine learning problem, i.e., training a system to produce representations that facilitate the prediction of "gold standard "human readability judgments—though researchers continue to test the viability of hand-crafted features as an alternative or complementary approach (Deutsch et al., 2020; Wilkens et al., 2024). Pre-trained language models seem potentially well-suited to this task, and indeed, past work (Crossley et al., 2023b) suggests that fine-tuning these models can produce estimates that align closely with human judgments of readability.

Modifying readability has also been a topic of considerable interest, with most research focusing on making text easier to read, e.g., for journal abstracts (Li et al., 2023) or math assessments (Patel et al., 2023). Cardon and Bibal (2023) provide a useful overview of the distinct *operations* used in Automatic Text Simplification (ATS), including splitting up long sentences (Nomoto, 2023) and deleting or inserting individual words. As with work on measuring readability, this research has gradually shifted from explicit, rule-based approaches to systems that "learn" appropriate transformations using an annotated corpus (Cardon and Bibal, 2023), sometimes tailored with psycholinguistic features (Qiao et al., 2022).

Most relevantly, recent research has used *prompt engineering* approaches to ask whether Large Language Models (LLMs) can modify (Farajidizaji et al., 2023; Ribeiro et al., 2023; Liu et al., 2023; Creutz, 2024), with some studies even asking whether text cacn be modified to some *target readability level*, e.g., a target Flesch score (Flesch, 1948). Even with "zero-shot" prompting (i.e., no examples provided), LLMs appear to be surprisingly successful at modifying text readability in the desired direction—though not necessarily to the desired text level (Liu et al., 2023). In some cases, a residual correlation is found between the readability of the original text and the modified text (Farajidizaji et al., 2023).

3 Study 1: Measuring Readability

In Study 1, we asked whether a state-of-the-art LLM could be used to estimate the readability of text excerpts. We adopted an empirical approach to this question: given a corpus of human readability estimates (Crossley et al., 2023b), how well can an LLM equipped solely with instructions and a definition of readability produce outputs that correlate reliably with human judgments? We focus on the quality of LLM outputs generated "zero-shot" (i.e., without any labeled examples in the prompt). This study this mirrors other recent work (Dillion et al., 2023; Trott, 2024a; Aher et al., 2023; Gilardi et al., 2023) using LLMs for zero-shot annotation of text data.

3.1 Methods

3.1.1 CLEAR Dataset

We used the CommonLit Ease of Readability (CLEAR) Corpus (Crossley et al., 2023b), which contains human estimates of readability for 4,724 text excerpts. The CLEAR Corpus was produced by sampling text excerpts (between 140-200 words) from various databases (e.g., Project Gutenberg). It includes fiction and non-fiction, and spans a range from 1875 to 2020. Excerpts were normed by asking a sample of teachers to rate pairs of items for their relative readability. These pairwise judgments were then aggregated to create a readability index for each individual passage.

3.1.2 Model

Our primary goal was assessing the reliability of using a state-of-the-art LLM in estimating readability. To this end, we used GPT-4 Turbo, a proprietary LLM produced by OpenAI. We accessed Turbo using the OpenAI Python API (model name = *gpt-4-1106-preview*). Because Turbo is a closed-source model, it is unclear how many parameters the model has or how much data it was trained on.

3.1.3 Procedure

Turbo was provided with a system prompt ("You are an experienced teacher, skilled at identifying the readability of different texts."). Then, each text excerpt was presented to Turbo in a separate prompt (i.e., rather than in succession), along with instructions explaining that the goal was to rate the excerpt for how easy it was to read and understand, on a scale from 1 (very challenging to understand) to 100 (very easy to understand); the exact instructions provided to Turbo can be found in Appendix A.1. Turbo's responses were produced using a temperature of 0, with a maximum number of tokens of 3. Response strings were then converted to numeric values in Python.

3.2 Results

We first asked how well Turbo's ratings predicted human readability scores from the CLEAR dataset (Crossley et al., 2023b). A linear regression model predicting Human Readability from GPT-4 Turbo Ratings exhibited good fit ($R^2 = 0.58$). Turbo's ratings were positively correlated with Human Readability ($r = 0.76$) see also Figure 1. For comparison, the correlation between two random splits within the CLEAR corpus was only $r = 0.63$.

We then compared the predictive success of Turbo's ratings to several psycholinguistic variables that past work (Kyle et al., 2018) has found to be correlated with judgments about readability: log word frequency (Brysbaert and New, 2009),

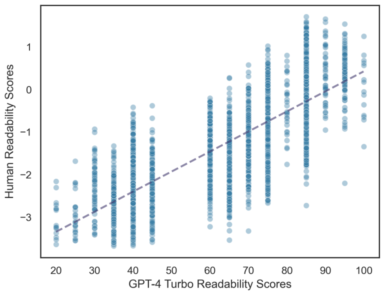

Figure 1: Relationship between ratings elicited by GPT-4 Turbo and average human readability judgments ($R^2 = 0.58$).

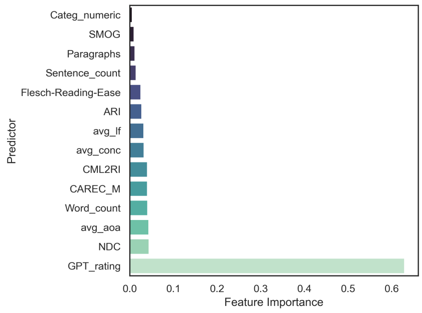

Figure 2: Feature importance scores for each predictor, as determined using a random forest regression.

word concreteness (Brysbaert et al., 2014), and word age of acquisition (Kuperman et al., 2012). For each variable, we calculated the *average* across all words in a given passage that occurred in the relevant dataset. A linear model including all three psycholinguistic predictors explained approximately 36% of the variance in human readability judgments ($R^2 = 0.36$). Each variable was significantly related: frequency $[\beta = 0.82, SE = 0.13, p < .001]$, concreteness $[\beta = 1.76, SE = 0.11, p < .001]$, and age of acquisition $[\beta = -0.56, SE = 0.06, p < .001]$. Thus, psycholinguistic properties of words in a passage are relevant for predicting readability judgments, but underperform ratings elicited from GPT-4 Turbo.[1]

We also considered several other potential correlates of readability included in the CLEAR corpus for each excerpt (see Figures 2 and 4 for a summary). Across all measures, Turbo's ratings were the most correlated with human judgments ($r = 0.76, p < .001$). We also compared the relative predictive power of each measure by entering them all as predictors in a random forest regression and visualizing the *feature importance scores* assigned to each predictor.[2] All measures were z-scored before fitting the model. As depicted in Figure 2, Turbo's ratings were assigned the highest feature importance (see A.2 for an analogous result using LASSO regression).

4 Study 2: Modifying Readability

In Study 2, we asked whether a state-of-the-art LLM could successfully *modify* (as opposed to simply *measure*) the readability of texts. We approached this question in the following way: given instructions to make a text excerpt *easier* or *harder*, does an LLM produce a modified version that an independent pool of human judges rate as easier or harder than the original? We also asked whether *automated measures* of readability (including ratings elicited from Turbo) co-varied with the experimental manipulation. This study was pre-registered on the Open Science Framework (OSF).[3]

4.1 Methods

4.1.1 Materials

To make this question empirically tractable, we selected a random sample of 100 excerpts from the original CLEAR corpus. Each excerpt was then presented to GPT-4 Turbo twice, with two different sets of instructions asking Turbo to make the excerpt easier or harder to read (exact prompting and instructions found in Appendix A.1). As in Study 1, Turbo was first provided with a system prompt ("You are an experienced writer, skilled at rewriting texts."); a temperature of 0 was used, and the maximum number of tokens was set to the number of tokens in the original excerpt, plus a "buffer" of 5 tokens. Additionally, we specified that the modified version should be of approximately the same length as the original.

This resulted in 300 items altogether. For the human study, these items were assigned to 6 lists using a Latin Square design, where each list had approximately 50 items. Note that in some cases, the modified version produced by Turbo cut-off in mid-sentence; we further modified these excerpts by removing the final sentence fragment. The experiment was designed on the Gorilla experimental

[1]Of course, taking the average of these variables across an entire passage is a relatively coarse measure and likely represents a *lower-bound* on their predictive efficacy.

[2]No maximum depth was used, and the random state was set to 0.

[3]A link to the pre-registration, as well as all code and data required to reproduce the analyses, will be provided after the anonymity period is over.

design platform (Anwyl-Irvine et al., 2018).

4.1.2 Participants

Our target N was 60 participants (10 per list). We anticipated a non-zero exclusion rate, so we intended to recruit 70 participants via Prolific; due to an error in the recruiting platform, we recruited only 69. As per our pre-registration, we excluded participants whose readability ratings for the *original* text excerpts exhibited a correlation with the gold standard was $r < .1$; this resulted in the removal of 10 participants. Participants were paid \$6.00 and the median completion time was 34 minutes and 21 seconds (an average rate of \$10.48 per hour). In the final pool of participants, 34 participants identified as female (22 male, 2 non-binary, and 1 preferred not to answer); the average self-reported age was 40.77 (SD = 14).

4.1.3 Procedure

Each participant rated the readability of a series of 50 text excerpts on a scale from 1 (very challenging to understand) to 5 (very easy to understand). Participants were instructed to consider factors such as "sentence structure, vocabulary complexity, and overall clarity"; they were also reminded to try to focus on the readability of the passage itself, as opposed to the complexity of the topic.

4.2 Results

We carried out three pre-registered analyses in R using the *lme4* package (Bates, 2011); see Appendix A.3 for more details. Human readability judgments were predicted by the contrast between *Easy* and *Hard* $[\chi^2(1) = 97.58, p < .001]$, between *Easy* and *Original* $[\chi^2(1) = 32.4, p < .001]$, and between *Hard* and *Original* $[\chi^2(1) = 74.75, p < .001]$. As depicted in Figure 3, excerpts in the *Easier* condition were rated as the most readable ($M = 4.48, SD = 0.8$), excerpts in the *Harder* condition were rated as the least readable ($M = 2.5, SD = 1.25$), with excerpts in the *Original* condition between the two ($M = 3.97, SD = 1.13$).

5 Discussion

Our primary question was whether state-of-the-art LLMs could be used to *measure* and *modify* the readability of a text excerpt. The first question was operationalized by assessing the ability of GPT-4 Turbo to produce readability ratings that correlated with a gold standard corpus (Crossley et al., 2023b). Turbo's ratings exhibited a strong correlation with

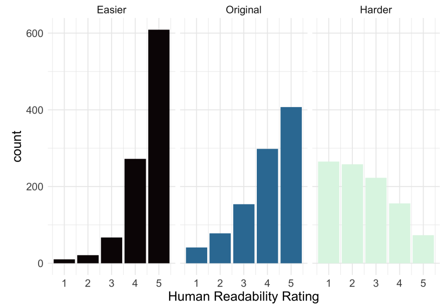

Figure 3: Distribution of human readability judgments for each text condition.

the gold standard ($r = 0.76$); consistent with other recent work using LLMs for text annotation (Trott, 2024b), this correlation was higher than the correlation between random splits of human ratings (Cross et al., 2023). Further, Turbo's ratings were the best predictor of human readability judgments of all the variables tested (see Study 3). The second question was operationalized by asking Turbo to produce easier or harder versions of 100 sample excerpts from the same corpus (Crossley et al., 2023b). In a pre-registered human study, participants consistently rated the *easier* versions as easier to read, and the *harder* versions as harder to read.

As with other recent work (Farajidizaji et al., 2023; Liu et al., 2023; Ribeiro et al., 2023), these results provide a proof-of-concept that LLMs may be useful for measuring and modifying text readability, at least as operationalized here. Unlike past work (Ribeiro et al., 2023; Farajidizaji et al., 2023), we do not investigate the question of modification to *target readability levels*, though we do collect novel human judgments to validate the success of GPT-4 Turbo's modifications (Study 4). Of course, considerable open questions about the viability of this approach remain. These include: uncertainty about the *quality* of the modified texts (Liu et al., 2023), which we did not assess here; the efficacy of further prompt engineering; and the *construct validity* of readability as a target measure. These questions are all explored in more detail in the Limitations section below.

6 Limitations

One limitation, particularly of Study 2, is scope: because we planned to collect human annotations for each excerpt, we considered only 100 text excerpts, and compared the performance of only one model (GPT-4 Turbo). The results of this study can be seen as a proof-of-concept, which future work can build on with larger samples and more sophisticated prompt engineering techniques.

A further limitation of Study 2 is that we did not assess the quality of the modified excerpts. In principle, then, some of the modified versions may not adequately summarize the target text. Evaluating the quality of summaries is notoriously difficult (Wang et al., 2019), though recent work (Liu et al., 2023) has made use of automated metrics like BERTScore (Zhang et al., 2020). Future work would benefit from another human study that asks directly about the *quality* of the modified texts.

A final limitation is the question of what the *construct* of readability means in the first place, and how best to measure it. Construct validity—whether a test measures what it was designed to measure—is by no means a new challenge for work in NLP generally (Raji et al., 2021) or readability specifically (Crossley et al., 2008). "Readability" may not be a unitary construct; different stakeholders likely construe readability in different ways depending on their goal (e.g., making a product manual accessible vs. curating educational materials) and audience (e.g., school-aged children vs. professionals). Further, different formulas or automated metrics emphasize different properties of a text, making implicit or explicit assumptions about the underlying construct. The current work relied on human judgments of readability as a "gold standard", using both existing corpora (Crossley et al., 2023b) and novel data (Study 2). By these metrics, using Turbo to measure and modify readability was modestly successful. Yet the ambiguity of the construct itself makes it challenging to determine whether these results generalize to other texts, contexts, goals, or audiences. Thus, future work could benefit from additional research on "benchmarking" readability itself and whether different benchmarks are needed for different construals of readability.

7 Ethical Considerations

All data collected from human participants has been fully anonymized before analysis or publication.

One potential risk with research on automatic text simplification is that tools will be deployed in various applied settings (e.g., education) before they are ready. As we discussed in the Limitations section (Section 6), we believe there are a number of open questions remaining with this kind of research and do not intend for these results to signal that LLMs could and should be used for measuring and modifying readability in an applied domain at this time.

Acknowledgments

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Alexander Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo Evershed. 2018. Gorillas in our midst: Gorilla. sc. *Behavior Research Methods*.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Douglas Bates. 2011. Mixed models in r using the lme4 package part 5: Generalized linear mixed models. *University of Wisconsin: Madison, WI, USA*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Chao-Yi Chen, Jen-Hao Yang, and Lung-Hao Lee. 2023. NCUEE-NLP at BioLaySumm task 2: Readability-controlled summarization of biomedical articles using the PRIMERA models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 586–591, Toronto, Canada. Association for Computational Linguistics.

Mathias Creutz. 2024. Correcting challenging Finnish learner texts with claude, GPT-3.5 and GPT-4 large language models. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 1–10, San Ġiljan, Malta. Association for Computational Linguistics.

Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai, and Miikka Silfverberg. 2023. Glossy bytes: Neural glossing using subword encoding. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 222–229, Toronto, Canada. Association for Computational Linguistics.

Scott Crossley, Joon Suh Choi, Yanisa Scherber, and Mathis Lucka. 2023a. Using large language models to develop readability formulas for educational settings. In *International Conference on Artificial Intelligence in Education*, pages 422–427. Springer.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023b. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2023. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. *arXiv preprint arXiv:2309.12551*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Sohom Ghosh, Shovon Sengupta, Sudip Naskar, and Sunny Kumar Singh. 2022. FinRAD: Financial readability assessment dataset - 13,000+ definitions of financial terms for measuring readability. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 1–9, Marseille, France. European Language Resources Association.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior research methods*, 50:1030–1046.

Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2023. Large language models and control mechanisms improve text readability of biomedical abstracts. *arXiv preprint arXiv:2309.13202*.

Fengkai Liu and John Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv:2311.09184*.

Bertha A Lively and SL Pressey. 1923. A method for measuring the" vocabulary burden" of textbooks: Educational administration and supervision,". *A method for measuring the" vocabulary burden" of textbooks: Educational Administration and Supervision*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Hiroki Nomoto. 2023. Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 76–82, Nusa Dua, Bali. Association for Computational Linguistics.

Nirmal Patel, Pooja Nagpal, Tirth Shah, Aditya Sharma, Shrey Malvi, and Derek Lomas. 2023. Improving mathematics assessment readability: Do large language models help? *Journal of Computer Assisted Learning*, 39(3):804–822.

Yu Qiao, Xiaofei Li, Daniel Wiechmann, and Elma Kerz. 2022. (psycho-)linguistic features meet transformer models for improved explainable and controllable text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 125–146, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

Domenic Rosati. 2023. GRASUM at BioLaySumm task 1: Background knowledge grounding for readable, relevant, and factual biomedical lay summaries. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 483–490, Toronto, Canada. Association for Computational Linguistics.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. *Journal of social studies education research*, 8(3):238–248.

Sean Trott. 2024a. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, pages 1–19.

Sean Trott. 2024b. Large language models and the wisdom of small crowds. *Open Mind*, 8:723–738.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2019. A task in a suit and a tie: paraphrase generation with semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7176–7183.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Example Appendix

A.1 Instructions for Study 1 and Study 2

In this section, we report the exact prompts used to elicit readability judgments from GPT-4 Turbo. Note that symbols like "EXCERPT" indicate that the text of the excerpt was inserted in this section of the prompt.

Study 1 Instructions:

> Read the text below. Then, indicate the readability of the text, on a scale from 1 (extremely challenging to understand) to 100 (very easy to read and understand). In your assessment, consider factors such as sentence structure, vocabulary complexity, and overall clarity.
>
> <Text>:EXCERPT</Text>
>
> On a scale from 1 (extremely challenging to understand) to 100 (very easy to read and understand), how readable is this text?. Please answer with a single number.

Study 2 Instructions:

> Read the passage below. Then, rewrite the passage so that it is easier/harder to read.
>
> When making the passage more/less readable, consider factors such as sentence structure, vocabulary complexity, and overall clarity. However, make sure that the passage conveys the same content.
>
> Finally, try to make the new version approximately the same length as the original version.
>
> <Text>:EXCERPT</Text>
>
> As described in the instructions, please make this passage easier/harder to read, while keeping the length the same.

A.2 Additional Statistical Analyses for Study 1

In this section, we report on the results of additional statistical analyses conducted on the Study 1 dataset. First, we include a correlation matrix (Figure 4) representing the relationship between the predictors considered; note that ratings elicited

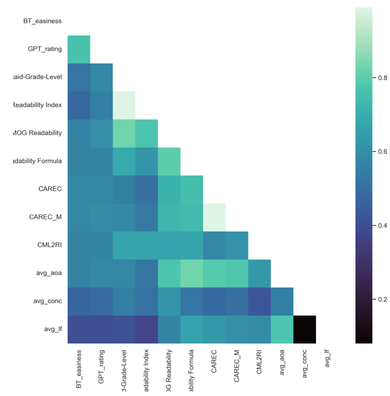

Figure 4: Correlation matrix between all the variables considered in Study 1. Correlation coefficients have all been transformed to absolute values for easier comparison.

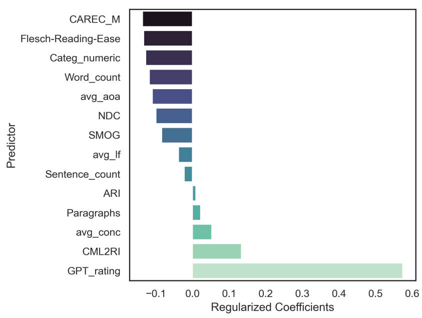

Figure 5: Regularized coefficients using Lasso regression.

from Turbo were the most correlated with human judgments.

Additionally, to expand on the random forest regression analysis conducted in the primary manuscript, we fit a Lasso regression model using the z-scored predictors. We first identified the optimal α parameter using cross-validation, then refit the model on the entire dataset.[4] The regression coefficients are depicted in Figure 5; as with the results of the random forest regression, Turbo's ratings have the largest absolute magnitude.

A.3 Additional Analysis Details for Study 2

In the case of fitting mixed effects models, we began with maximal random effects structure and reduced as needed for model convergence (Barr et al., 2013). Nested model comparisons were conducted by comparing a full model to a reduced

[4]Because our primary interest was in comparing the relative magnitude of coefficients, rather than analyzing model fit, we did not use cross-validation to analyze overall model fit.

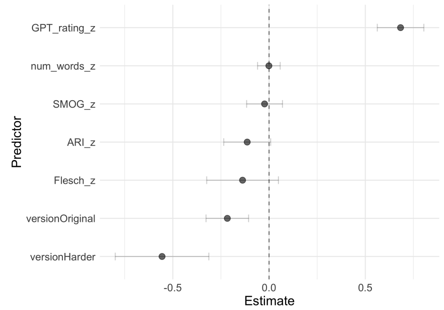

Figure 6: Coefficients in a mixed model predicting human readability judgments. Both text condition and Turbo's ratings exhibit independent effects.

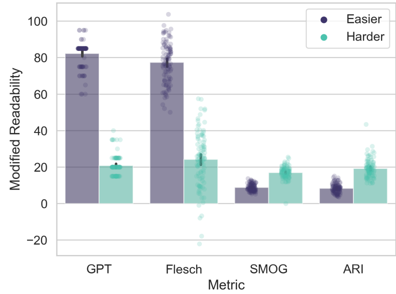

Figure 7: Comparison of automated readability scores for the modified text excerpts.

model omitting only the variable of interest, using a log-likelihood ratio test (LRT).

In an exploratory analysis, we asked whether ratings elicited by GPT-4 Turbo were also predictive of human judgments. A mixed model predicting human readability from both *Condition* and *Turbo rating* (along with control variables for other readability metrics) revealed significant effects of each variable, suggesting they explained independent variance. The coefficients for this exploratory analysis are depicted in Figure 6).

We also calculated the readability of the modified texts using automated readability formulas, e.g., the Flesch Reading Score (Flesch, 1948). We then asked whether the modified versions varied in the expected direction along each metric in question, according to whether Turbo was instructed to make the text easier or harder to read. We found that the modified versions varied in the expected direction according to automated readability metrics as well (see Figure 7).

Finally, consistent with (Farajidizaji et al., 2023), we found a consistent correlation between the readability of an *original* text excerpt and the *modified* version. That is, Turbo successfully modified texts to be easier or harder to read, depending on the

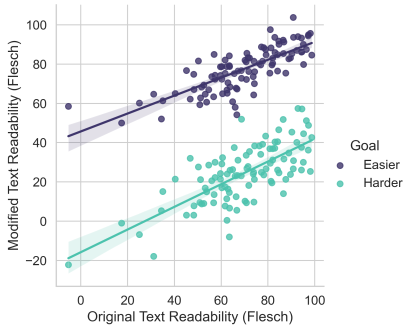

Figure 8: Comparison of Flesch readability for the original version and modified version, according to Turbo's instructions.

instructions, but the readability of the modified exhibited a residual correlation with the original text's readability (see Figure 8).