Evaluating the Evaluators: Are readability metrics good measures of readability?

Isabel Cachola, Daniel Khashabi* and **Mark Dredze***
Johns Hopkins University, Baltimore, MD 21211
{icachola, danielk, mdredze}@cs.jhu.edu

Abstract

Plain Language Summarization (PLS) aims to distill complex documents into accessible summaries for non-expert audiences. In this paper, we conduct a thorough survey of PLS literature, and identify that the current standard practice for readability evaluation is to use traditional readability metrics, such as Flesch-Kincaid Grade Level (FKGL). However, despite proven utility in other fields, these metrics have not been compared to human readability judgments in PLS. We evaluate 8 readability metrics and show that most correlate poorly with human judgments, including the most popular metric, FKGL. We then show that Language Models (LMs) are better judges of readability, with the best-performing model achieving a Pearson correlation of 0.56 with human judgments. Extending our analysis to PLS datasets, which contain summaries aimed at non-expert audiences, we find that LMs better capture deeper measures of readability, such as required background knowledge, and lead to different conclusions than the traditional metrics. Based on these findings, we offer recommendations for best practices in the evaluation of plain language summaries. We release our analysis code and survey data.

JHU-CLSP/eval-the-eval-readability

1 Introduction

In the field of Natural Language Processing (NLP), plain language summarization (PLS) distills complex documents, such as scientific articles, into accessible summaries for non-expert audiences while preserving essential meaning (Chandrasekaran et al., 2020). The COVID-19 pandemic highlighted the critical need to make scientific knowledge accessible to the general public (Wang et al., 2020). By enhancing public engagement with research, PLS can help bridge the gap between expert knowledge and general understanding.

Although human evaluation remains the gold standard for assessing summary quality and readability, the high cost and slow turnaround (Liu et al., 2022) have led many researchers to rely on automatic evaluation metrics for evaluating PLS summaries (Goldsack et al., 2022; Guo et al., 2021). Although these metrics have been validated in fields such as education and law (Thorndike, 1936; Han et al., 2024), their effectiveness in reflecting readability in the context of PLS remains unproven.

Are automated readability metrics appropriate evaluators for the task of PLS? We explore whether the definition of readability as implemented by automated measures matches the definition used by the PLS research community. Additionally, given that Language Models (LMs) can reason over complex language tasks (Brown et al., 2020; Wei et al., 2022; Yang et al., 2024), we explore whether LMs can judge the readability of a summary. Given these motivations, we ask the following research questions (RQs).

RQ1 What is the current standard of evaluation in PLS literature? We review PLS literature by collecting relevant papers published in *ACL venues from 2013 to 2025 and note the readability evaluation method used in the study. We find that the majority of papers focus on a small number of traditional readability metrics, such as Flesch-Kincaid grade Level (FKGL) (Flesch, 1952). This finding motivates our analysis of the suitability of traditional readability metrics for PLS evaluation.

RQ2 How well do traditional readability metrics correlate with human readability judgments? Since the PLS research community employs these traditional metrics (RQ1), we assess their suitability by measuring their correlation with human readability judgments. A low correlation would suggest that a metric is inadequate for eval-

*Equal advising.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24023–24039
November 4-9, 2025

uating PLS readability, and would necessitate the PLS research community identify and move to better metrics. To the best of our knowledge, this work is the first to compare traditional readability metrics to human readability judgments for PLS.

RQ3 How well do LM-based evaluators correlate with human readability judgments? Traditional readability metrics primarily use lexical features, such as the number of syllables in a word, to measure readability. In contrast, LMs may capture more complex attributes of readability than traditional metrics, such as the inclusion of necessary context and explanation of key concepts. The findings of this research question have important implications for both the best practices in evaluation of PLS and the broader NLP community's understanding of LM capabilities.

RQ4 What do LM-based evaluators reveal about the readability of popular summarization datasets? Researchers often rely on traditional readability metrics when assessing summaries in new methods or datasets. However, if these metrics correlate poorly with human judgments, the resulting conclusions may be flawed. Similarly, existing datasets, which often arise from data of convenience, may be poorly suited to PLS research. This RQ explores what LM-based evaluators reveal about the readability of popular summarization datasets and how LM-based conclusions differ from those based on traditional readability metrics.

We answer these questions through the following contributions. First, we survey PLS papers published in *ACL venues and find that the most popular metric for readability evaluation is Flesch-Kincaid Grade Level (FKGL) (Flesch, 1952). Motivated by these findings, we then compare 8 traditional readability metrics to human judgments. We show that 6 of the 8 metrics have a poor correlation (less than 0.3 Pearson correlation) with human judgments, including FKGL, indicating that these metrics are poor measures of readability for PLS. Additionally, we compare the judgments of 5 LMs to human judgments and show that LMs outperform the traditional metrics. We demonstrate that LMs have promising potential as evaluators by reasoning over more complex attributes of readability. We use LM evaluators to re-evaluate 10 summarization datasets and show that some summarization datasets intended for PLS achieve similar readability scores to those aimed at expert audiences, calling into question the utility of these data. Finally, based on a thorough analysis of current readability evaluation practices, we offer recommendations for best practices in PLS evaluation and identify opportunities for future work.

2 Related Works

Summarization evaluation. PLS research often introduces either datasets (Goldsack et al., 2022; Crossley et al., 2021; Liu et al., 2024; Manor and Li, 2019), methods (Guo et al., 2022; August et al., 2022; Luo et al., 2022; Ji et al., 2024; Flores et al., 2023), or both (Guo et al., 2021; Zaman et al., 2020; Chandrasekaran et al., 2020). The majority of prior work use a combination of readability metrics, such as Flesch Reading Ease (Flesch, 1943) or the Gunning-Fog Index (Gunning, 1952) to validate the readability of their dataset or generations. Readability metrics are typically reported in conjunction with more general summarization metrics, such as ROUGE (Lin, 2004) or BertScore (Zhang* et al., 2020). General summarization evaluation is a well-studied area, with ongoing work analyzing both the efficacy of summarization metrics (Fabbri et al., 2020; Khashabi et al., 2022; Goyal et al., 2022) and designing metrics that better align with human judgments (Liu et al., 2023c, 2022). Guo et al. (2023) analyzed how perturbations in plain language summaries affect results of general summarization metrics. In this work, we focus on readability metrics, rather than general summarization metrics, with the goal of understanding how well readability metrics measure readability for PLS.

Readability Metrics. While readability metrics are well studied in fields such as education (Thorndike, 1936; DuBay, 2004; Sibeko and van Zaanen, 2022) and linguistics (Carla Pires and Vigário, 2017), there is little work studying how well these metrics perform for the task of PLS. Most traditional metrics were not designed specifically for PLS, or even for evaluation in Computer Science. The most common origin of traditional metrics is the need to assess the readability of K-12 school texts (Dale and Chall, 1948; Coleman and Liau, 1975). Linsear Write was introduced in the book, *Gobbledygook has gotta go*, published by the US Department of the Interior for the purposes of measuring the complexity of government communications (O'hayre, 1966). As readability metrics rely primarily on lexical features (Rush, 1985), prior work has offered criticism of readability metrics, showing that they can be easily

manipulated to provide better scores with changes that do not substantially improve the readability of summaries (Tanprasert and Kauchak, 2021). Other work has looked at which linguistic attributes are correlated with readability metrics (Štajner et al., 2012). To the best of our knowledge, our work is the first to measure the correlation of readability metrics with human readability judgments.

LMs as Evaluators. Recent advances in LMs have shown that they are capable of reasoning over complex language (Brown et al., 2020; Wei et al., 2022; Yang et al., 2024). LMs have been shown to be effective evaluators in other natural language tasks (Li et al., 2025; Zhang et al., 2024; Nedelchev et al., 2020; Liu et al., 2023a), including related summarization tasks (Song et al., 2024). Given this success in prior work, we hypothesize that LMs are capable of evaluating the readability of plain language summaries. In particular, we hypothesize that LMs can reason over more complex attributes of readability, such as the background required or whether technical concepts are explained.

3 Experimental Setup

3.1 Current PLS evaluation standards RQ1

We aim to conduct a thorough literature survey of the standard practices in readability evaluation for PLS. We collect papers[1] from the ACL Anthology[2] that mention one of the following key phrases: "plain language summarization," "readable summaries," or "lay summarization." We exclude papers published for a shared task from annotation and assume the participants use the metrics designated by the shared task organizers. Our goal is to understand the decisions made by researchers, and including shared task papers in this survey would over-represent the decisions made by the task organizers. We report the evaluation methods used by the shared tasks and the number of participants to represent the impact of the evaluation choices. We identify 55 papers that match our criteria. We annotate the papers for relevance to PLS, the type of publication (Main conference, Findings, or Workshop), and which readability evaluation metrics are used. We exclude papers from the survey not relevant to PLS, resulting in 18 relevant papers from the years 2013 to 2025. The most common reasons for relevance exclusion include using "readable" in a different word sense (e.g. "human readable" vs "machine readable") or just citing a PLS paper. We report the number of papers that use each metric.

3.2 Comparing traditional readability metrics to human judgments RQ2

Human Annotated Data. To measure the correlation between readability metrics with human judgments, we use the dataset collected by August et al. (2024). This dataset contains 60 summaries of 10 scientific papers in a variety of domains. Each paper has both expert written and machine written summaries (generated using GPT-3.) The summaries are annotated on a scale of 1 to 5 for the annotator's reading ease of the article. 1 indicates a very poor reading ease, while 5 indicates a very high reading ease. For each summary, we take the average of the annotators' scores to calculate the correlations with readability evaluations as described below. August et al. (2024) originally collected this dataset to better understand human preferences in scientific summarization. In this paper, we extend their work by applying their findings to summarization evaluation metrics. To the best of our knowledge, this is the only available dataset of human judgments for PLS. Appendix A contains additional dataset details.

Traditional readability metrics. We consider "traditional" readability metrics to be those most commonly used in PLS literature. These metrics are well-established, and have been used in past work as judges of readability (Chandrasekaran et al., 2020). This term excludes LM-based evaluations, discussed in § 3.3. We consider 8 readability metrics: Flesh-Kincaid Grade Level (FKGL) (Flesch, 1952), Flesch Reading Ease (FRE) (Flesch, 1943), Dale Chall Readability Score (DCRS) (Dale and Chall, 1948), Automated Readability Index (ARI) (Smith and Senter, 1967), Coleman Liau Index (CLI) (Coleman and Liau, 1975), Gunning Fog Index (GFI) (Isnaeni, 2017), Spache (Spache, 1953) and Linsear Write (LW) (O'hayre, 1966). All of the metrics, except for DCRS and Spache, use lexical features such as number of syllables or length of sentences to measure readability. DCRS and Spache use word familiarity to measure readability, assuming that more common words are easier to read (Dale and Chall, 1948; O'hayre, 1966).[3]

[1] On May 7th, 2025

[2] https://aclanthology.org/

[3] We use the py-readability-metrics package to calculate the readability scores.

Quantifying alignment between traditional metrics and humans. We report the Pearson and Kendall-Tau correlation of each metric listed above with the human judgments collected by August et al. (2024). Except for LW and FRE, all metrics provide a lower score for higher readability, while the human judgments provide a higher score for higher readability. To calculate correlations, we multiply the scores by -1 (except for LW and FRE), so that text rated as more readable by traditional metrics will be positively correlated with human judgments.

3.3 LMs as evaluators of readability RQ3

We experiment with the following 5 LMs as evaluators of readability: Mistral 7B (Jiang et al., 2023), Mixtral 7B (Jiang et al., 2024), Gemma 7B (Team, 2024), Llama 3.1 8B, and Llama 3.3 70B (Dubey et al., 2024). We experiment with 3 prompts and report the prompts in appendix B. We report the Pearson and Kendall-Tau correlations of the scores provided by each LM with the human judgments.

3.4 Analysis of summarization datasets RQ4

To test the ability of our results to generalize to datasets outside of the one collected by August et al. (2024), we include datasets with intended audiences more specific than "general" - experts and kids. We expect expert-targeted datasets to be given low readability scores and kid-targeted datasets to have high readability scores.

Expert targeted datasets. We include 3 expert-targeted datasets: arXiv, PubMed (Cohan et al., 2018) and SciTLDR (Cachola et al., 2020). arXiv and PubMed are collections of abstracts in the Computer Science and Biomedical domains, respectively (Cohan et al., 2018). SciTLDR is a collection of short, expert-targeted, one sentence summaries of Computer Science papers. We expect our methods to provide low readability scores. Additionally, the comparison of SciTLDR to arXiv and PubMed allows us to test if the scores are length dependent.

Kid-targeted dataset. The Science Journal for Kids (SJK) dataset is a collection of summaries of scientific papers in a variety of domains, intended for kids (Stefanou et al., 2024). Given that this dataset is targeted to kids, we expect it would receive high readability scores.

General audience datasets. In addition to the datasets listed above, we evaluate 6 popular datasets intended for PLS: CDSR (Guo et al., 2021), PLOS (Goldsack et al., 2022), eLife (Goldsack et al., 2022), Eureka (Zaman et al., 2020), CELLS (Guo et al., 2022), and SciNews (Liu et al., 2024). These datasets are intended for a general audience. CDSR, PLOS, and CELLS are written by journal editors or experts. eLife Sciences gives paper authors the option to write "eLife digests," with the goal of "cutting jargon and putting research in context." [4] The Eureka dataset was collected from EurekaAlert, which hosts press releases about research for scientific journalists. Finally, SciNews is a collection of scientific news reports, written by science reporters.

Dataset	Audience	Domain	# Docs	# Tokens
arXiv	Experts	CS	6440	163
PubMed	Experts	Medicine	6658	205
SciTLDR	Experts	CS	618	19
SJK	Kids	Varied	284	142
CDSR	General	Healthcare	284	221
PLOS	General	Biomed	1376	195
eLife	General	Biomed	241	383
Eureka	Journalists	Varied	1010	662
CELLS	General	Biomed	6311	162
SciNews	General	Varied	4188	615

Table 1: Comparison of the datasets analyzed in this paper. The first 4 are datasets in with a specific target audience. The following 6 datasets are commonly used in PLS literature. We report the number of documents (# Docs) in the test set as well as the average number of tokens (# Tokens).

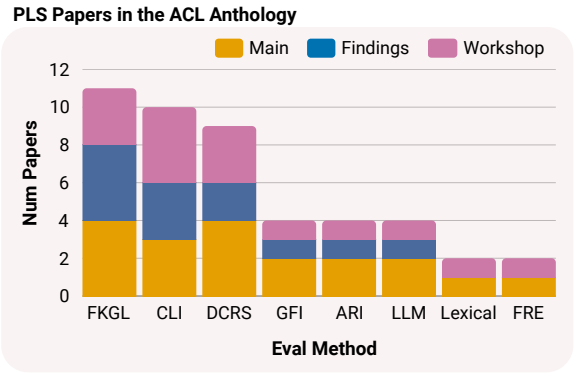
Table 1 contains a comparison of the summarization datasets analyzed in this paper. We use the test split of each dataset for our analysis and we report the intended audience, domain, number of documents in the test set, and average number of white-space delineated tokens. In total, we analyze 10 popular scientific summarization datasets.

4 Results

4.1 Current PLS evaluation standards RQ1

We found 18 ACL Anthology papers on the task of PLS and 3 shared tasks, representing 81 additional papers. Figure 1 shows the literature survey results, excluding metrics used by a single paper. FKGL is the most popular metric, followed by CLI and DCRS. LM-based evaluations are uncommon (4 of the 18 papers). The shared task BioLaySumm used FKGL and DCRS for both years, adding CLI in 2024. BioLaySumm 2025 is ongoing at the time of writing; the organizers plan to use FKGL, DCRS, and CLI. Our survey shows that PLS is an increasingly popular topic of study, as the number

[4] https://elifesciences.org/digests



Shared Tasks

Task Name	# Participants	Metrics Used
BioLaySumm @ BioNLP 2025	TBD	FKGL, DCRS, CLI
BioLaySumm @ BioNLP 2024	53	FKGL, DCRS, CLI
BioLaySumm @ BioNLP 2023	20	FKGL, DCRS
CL-LaySumm @ SDP 2020	8	Human Eval, Lexical

Figure 1: Evaluation metrics used by papers published in the ACL Anthology. We report the count of papers using each method out of a total of 18 papers. We additionally report the evaluation strategies used by PLS shared tasks and the number of participants.

of participants in shared tasks increased from 8 in 2020 to 53 in 2024, emphasizing the importance of PLS evaluation. Less popular metrics include GFI, ARI, lexical proxies (e.g., number of sentences in a document), and FRE. In § 4.2, we place the highest importance on the results of the most commonly used evaluation metrics: FKGL, CLI, and DCRS.

4.2 Comparing traditional readability metrics to human judgments RQ2

In Table 2a, we report the Pearson and Kendall-Tau correlation of 8 traditional readability metrics with human judgments. We find that 6 of the 8 metrics have less than 0.3 Pearson correlation with human judgments. DCRS and CLI have the highest correlation, achieving 0.2 Pearson points higher correlation than the most popular metric, FKGL (§ 4.1). FKGL receives only 0.16 Pearson and 0.08 Kendall-Tau correlation, indicating little to no correlation with human judgment.

Table 3 shows an example summary and readability scores, along with its human judgment. The human annotators gave the example summary an average rating of 4.05/5; they found the text fairly readable. However, the majority of traditional metrics give the summary poor readability scores: college level or higher. This is likely because the text includes domain-specific vocabulary, such as "acute respiratory distress syndrome (ARDS)," which is penalized by traditional metrics. Traditional readability metrics do not account for elements of the summary that make it more readable, such as defining ARDS as "a very serious lung disease" and explaining the scientists' motivation to "test a new method of lung damage diagnosis."

Metric	Pearson	Kendall Tau
FKGL	0.16	0.08
CLI	0.36	0.20
DCRS	**0.37**	**0.24**
GFI	0.21	0.11
ARI	0.10	0.02
FRE	0.29	0.15
Spache	0.13	0.04
LW	-0.06	-0.03

(a) Traditional metric scores correlation with human judgment.

Model	Pearson	Kendall Tau
Mistral 7B	0.52	0.44
Mixtral 7B	0.54	0.41
Gemma 7B	0.54	0.43
Llama 3.1 8B	0.45	0.34
Llama 3.3 70B	**0.56**	**0.35**

(b) LM scores correlation with human judgment.

Table 2: We report the Pearson and Kendall-Tau correlation of each metric with human judgment. Tab.2a contains the correlation of traditional readability metrics with human judgment. DCRS and CLI have the highest correlation with human judgment. Notably, the most popular metric, FKGL, as shown in §4.1, has low correlation with human judgment. Tab.2b contains the correlation of LM models as evaluators with human judgment. All 5 models achieve higher correlation than all of the traditional metrics.

```
On a scale of  1 to 5, what is the reading ease of the following text?
1 indicates the text requires expert background knowledge and 5 indicates the
text is readable to the general population. \n Assume the reader is an adult. Do
not use Flesch-Kincaid or other readability formulas. Use your own judgment to
rate the text. \n\n Format the output as follows: \n
Score: <score> \n Reason: <reasoning> \n\n  Text: {SUMMARY}
```

Figure 2: The best performing prompt of the 3 we tested. We report the results of this prompt in Table 2b and the results of the remaining prompts in Appendix B.

4.3 LMs as evaluators of readability RQ3

Traditional readability metrics rely on lexical proxies and do not measure other elements of a summary that could make it more readable, such as definitions of technical terms, explanations of important concepts, or descriptions of impact and motivation. LMs have been shown to perform well on many language understanding tasks (Brown et al., 2020; Srivastava et al., 2023), indicating that they have some understanding of language. We hypothesize that this knowledge will translate well to the task of PLS, and the LMs will be able to reason about more complex features of a summary that impact the readability.

We experiment with 3 prompts. We report best performing prompt in Figure 2 and its results in Table 2b; the other prompts and their results are in Appendix B. All of the LMs outperform the traditional

Scientists create a device which can detect the onset of acute respiratory distress syndrome (ARDS), a very serious lung disease, by measuring chemicals in patients' exhaled breath The researchers wanted to test a new method of lung damage diagnosis by analyzing patient breath samples. In particular, the researchers were looking for better ways to detect acute respiratory distress syndrome (ARDS), a form of lung injury that causes inflammation and severe damage. [...] a much larger group of test subjects is necessary to further validate their method. This new method of breath analysis could be a noninvasive, cost effective way to diagnose and track ARDS, and could potentially be modified to screen for other serious conditions as well.

(a) Excerpt of an example summary. This summary is written by an expert and is labeled as a low complexity summary.

Metric	Score	S-12	US Grade Level
FKGL ↓	13.9	12	College
CLI ↓	12.7	12	College
DCRS ↓	11.3	8.9	College graduate
GFI ↓	18.6	12	Above college graduate
ARI ↓	16.7	13	College graduate
FRE ↑	50.2	50	12th grade
Spache ↓	8.7	12	9th grade
LW ↑	19.5	60	College graduate

(b) Scores given be each metric for the example summary. ↓ indicates a lower score is more readable while ↑ indicates a higher score is more readable. We provide "S-12", the score each metric would assign US grade 12, to help contextualize the scores. We additionally translate each score to the US grade level.

Model	Score
Mistral 7B	4
Mixtral 7B	4.5
Gemma 7B	4
Llama 3.1 8B	4
Llama 3.3 70B	4

(c) Scores given be each model for the example summary The scores are on a scale of 1-5, with 5 being the most readable.

Table 3: 3b contains an example summary from August et al. (2024)'s dataset. 3b contains each metric's score for the example summary. 3c contains each model's readability scoring for the example summary. On average, the human annotators rated this summary a 4.05/5, indicating they found the summary fairly readable. All the LM evaluators rate the summary a 4 or 4.5 out of 5, agreeing with the human annotators. In contrast, 6 out of 8 of the traditional metrics rate the summary at a college reading level or higher, which is considered low readability.

metrics in correlation with human judgments. The best performing model, `Llama 3.3 70B`, outperforms the best traditional metric, DCRS, by nearly 0.2 Pearson points. We conduct significance testing and report the p-values comparing the LM results to the traditional metrics in Appendix C.

Performance in this task is not solely a factor of model size, as we see that smaller models perform similarly to the larger models. The difference in performance between the LMs is small, indicating that most generally well-performing models can be good judges of readability.

Table 3 contains an example summary and its associated scores from each LM. The human annotators rated the example summary a 4.05 out of 5 on reading ease. All models gave the summary a rating of 4 or 4.5 out of 5. The reasoning provided by `Llama 3.3 70B` states that the "concepts discussed, such as analyzing breath samples and identifying chemical compounds, are also explained in a way that is easy to understand." The model notes that the summary "may require some effort and attention," contributing to the model's reasoning for assigning the summary a 4/5 rather than a 5/5. This output indicates that the model is using its language reasoning abilities to rate the summary on attributes deeper than lexical features.

Dataset	Mean	Median	Var
arXiv	1.31	1	0.23
PubMed	1.99	2	0.19
SciTLDR	1.86	2	0.32
SKJ	4.40	4	0.24
CDSR	3.49	4	0.52
PLOS	2.06	2	0.26
eLife	3.18	3	0.65
Eureka	3.21	3	0.67
CELLS	2.23	2	0.50
SciNews	3.37	4	0.64

Table 4: Readability scores on a scale of 1 to 5, as judged by `Llama-3.3-70B`, 5 being the most readable. We report the mean, median, and variance of each score.

4.4 Analysis of summarization datasets RQ4

We analyze scientific summarization datasets using the LM evaluators. We use `Llama 3.3 70B`, the best performing model from § 4.2. In Figure 3, we include histograms of the readability scores for all 10 tested datasets, to visualize the distributions. In Table 4, we report the mean, median, and variance of the readability scores for each dataset.

We've shown that most LM judgments of readability correlate higher with human judgments than traditional metrics. In order to further validate our findings, we begin our analysis with 4 datasets with specific target audiences - experts or kids.

Expert-targeted datasets. We experiment with 3 datasets intended for expert readers: arXiv, PubMed, and SciTLDR. ArXiv, PubMed, and Sc-

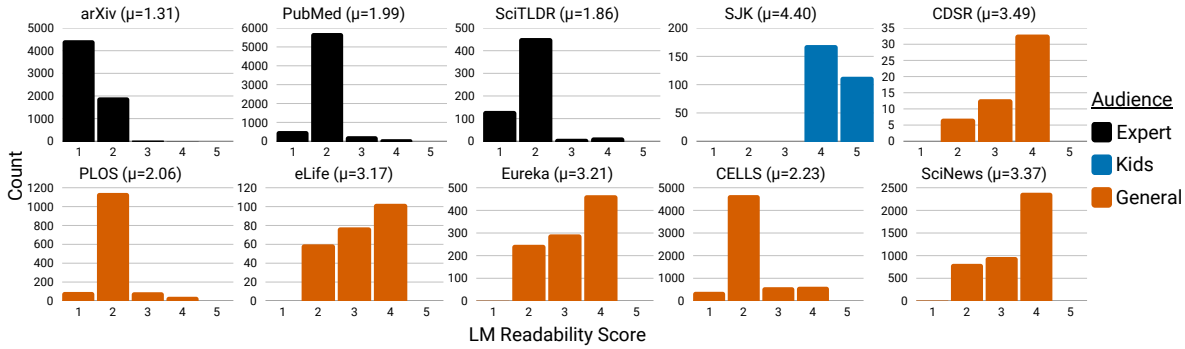


Figure 3: Histogram of LM readability scores and the mean scores (μ) for each dataset, as judged by `Llama 3.3 70B`. As we can see from the results, PLOS and CELLS are judged to be similarly readable to the expert targets datasets (arXiv, PubMed, and SciTLDR). The most readable PLS datasets are CDSR and SciNews.

iTLDR receive low readability scores, averaging less than 2/5. This matches our expectations since summaries intended for an expert audience typically have low readability for non-experts. We also note that SciTLDR receives similarly low readability scores, despite containing significantly shorter summaries than the arXiv and PubMed datasets. This shows that the LM evaluator is not simply favoring shorter summaries as more readable.

Kid-targeted dataset. SJK receives high readability scores, with an average readability of 4.40. The results of the expert and kid targeted datasets match our expectations of readability scores, and serve to support the analysis of the remaining 6 general-audience datasets below.

General audience datasets. We analyze 6 popular PLS datasets: CDSR, PLOS, eLife, Eureka, CELLS, and SciNews. PLOS and CELLS receive mean readability scores of 2.06 and 2.23, respectively. These scores are similar to the expert-targeted datasets described above, indicating that these two datasets may not be well-suited for PLS. SciNews and CDSR receive the highest readability scores, with average scores of 3.49 and 3.37, respectively, indicating that they are the well suited for the task of PLS.

Keyword analysis. To understand the LM's reasoning for assigning scores, we use the `YAKE!` algorithm to extract keywords from the reasoning provided by the LM evaluator for why each summary was provided with a specific score (Campos et al., 2020). Figure 4a contains the keywords stratified by score and Figure 4b contains the keywords stratified by dataset. When stratified by score, the model mentions issues such as requiring "expert background knowledge" and "using specialized terms" for summaries with readability scores of 1 or 2. For summaries with scores of 4 or 5, the model references how the summaries include "explanations for the non-expert" and explains "concepts in an accessible" manner. When stratified by dataset, for datasets with generally low readability scores, such the model mentions issues such as requiring "specialized knowledge" or that the text

Score	Keywords
1	expert background knowledge, text also assumes, text requires expert, background knowledge, highly technical, text assumes
2	using technical terms, text discusses complex, strong background knowledge, require specialized knowledge, using specialized terms
3	understandable for a general, adults with some medical, making it a challenging, readable with some basic, understandable for the general, audience than just medical, vocabulary of the text
4	explanation for the non-expert, context in a clear, explanations for the non-expert, terms for a non-expert, understandable for some readers
5	concepts in an accessible, language that is easy, text to be accessible, easy for a general, concepts that are easy, text uses simple language

(a) Keywords stratified by score.

Expert-Targeted Datasets	
arXiv	familiar with the specific, expertise in this area, research in a specialized, suggests that a significant, likely for an academic
PubMed	knowledge about the disease, background or some familiarity, structure of a scientific, professionals with a strong
SciTLDR	context for these terms, specific to these fields, audience with some technical, networks and the concept, fields such as artificial
Kid-Targeted Dataset	
SJK	easy for most adults, straightforward and the concepts, easy for an adult, readers with a basic, concepts in a clear
General Audience Datasets	
CDSR	text assumes some basic, assumes some basic knowledge, basic knowledge of medical, general adult audience, medical
PLOS	using technical terms, require specialized knowledge, strong background knowledge, discusses complex concepts
eLife	explanation of the concepts, understand for a general, explanation of these concepts, understanding of the concepts, without such a background
Eureka	context for a non-expert, unfamiliar to some adult, non-expert with some basic, context of the research, terms are not overly
CELLS	audience with no science, topic is a specialized, foundation in these fields, readers with some scientific
SciNews	understandable by those without, understand all the details, includes a few specialized, make it more readable, understanding of these fields

(b) Keywords stratified by dataset.

Figure 4: Keywords mentioned in the reasoning of the LM evaluator for why a summary was given a certain readability score. Figure 4a contains the keywords stratified by score and Figure 4b contains keywords stratified by dataset.

is "likely for an academic." The model also mentions the domain specific knowledge required such as Pubmed's focus on "disease[s]." For datasets with generally high readability scores, such as SJK and SciNews, the model mentions how the summaries are "easy for most adults" and how the text is "understandable by those without" background knowledge. This keyword analysis indicates LMs are attributing their judgements to deeper attributes that contribute to readability compared to traditional metrics.

LM evaluators vs. traditional metrics. Finally, we compare the results of the analysis using traditional metrics to LM evaluators of readability. In this analysis, we focus on `Llama 3.3 70B`, the best performing LM, and FKGL, the most popular readability metric. Table 5 compares the average LM readability and FKGL score for each dataset, and how each metric would rank the datasets. All but 1 dataset changed their ranking depending on the metric used. arXiv has the largest delta, ranking 10th in readability according to the LM evaluator and 2nd according to FKGL. FKGL ranking arXiv as the 2nd most readable is particularly concerning, as this dataset is a collection of scientific abstracts, intended for an expert audience. To measure disagreement, we convert each metric into binary scores of "high readability" and "low readability." For FKGL, we consider any summary given a score of under 12 points to have high readability. FKGL considers any text above 12 to be college reading level. For the LM evaluator, we consider any summary given a score of 3 or higher to have high readability. By converting the scores to binary labels, we calculate the Cohen's Kappa score (McHugh, 2012) for agreement as 0.17, indicating the two metrics have fair but not substantial agreement. We provide examples of this disagreement in Table 6. This analysis shows how the evaluation metrics we use can greatly influence the conclusions we draw.

5 Discussion

We found PLS an increasingly popular area of study, but researchers primarily rely on a handful of traditional metrics for evaluation. However, we found that traditional metrics are imperfect measures of readability and LM evaluators can draw significantly different, and more accurate, conclusions about PLS datasets than when using FKGL, the most common metric.

	LM Eval		FKGL		
Dataset	S	R	S	R	ΔR
arXiv	1.31	10	11.53	2	+8
PubMed	1.99	8	14.14	5	+3
SciTLDR	1.86	9	15.66	10	−1
SKJ	4.40	1	8.41	1	0
CDSR	3.49	2	14.08	4	−2
PLOS	2.06	7	15.44	9	−2
eLife	3.18	5	11.87	3	+2
Eureka	3.21	4	14.87	6	−2
CELLS	2.23	6	15.35	8	−2
SciNews	3.37	3	14.98	7	−4

Table 5: The mean score (S) and rank (R) for each dataset, as judged by an LM evaluator and FKGL. ΔR represents the change in rank from the LM evaluator to the FKGL scores.

5.1 Why traditional readability metrics are insufficient measures of readability

We consider 2 explanations for the poor correlation of readability metrics with human judgments: definitional inconsistency or measurement error. Definitional inconsistency means that the definition of "readable," as measured by the metrics, differs from the definition of "readable," as considered by human judges. Measurement error means that, even if we have the correct definition, we are not measuring readability properly. We argue that there is evidence for both problems.

On definitional inconsistency, the majority of readability metrics originated in the education domain. Traditional readability metrics typically define a "readable" text as one with an appropriate text complexity for the number of years of education (i.e., a text has a US 9th grade reading level) (Gunning, 1952; Coleman and Liau, 1975; Flesch, 1952). In contrast, the field of PLS typically defines a "readable" text as one that gives a non-expert, adult reader an overall understanding of the source article. These different definitions have different implications for the resulting text. If optimizing for education-appropriate text complexity, we can measure the complexity of the vocabulary or sentences. However, using the PLS definition of readability, we should measure features such as whether the text includes explanations of technical terms or how much background is required to understand the concepts.

Traditional readability metrics also suffer from measurement error. Even if we assume a consistent definition, traditional metrics do not properly measure readability. They measure lexical properties, such as number of syllables in a word, which penalizes summaries for using clearly defined technical terms. Traditional metrics also do not measure deeper features that contribute to readability, such as how much background is required to understand

Wind power is an important source of renewable energy, but some people are concerned that conventional wind turbines are too loud and too hazardous for birds and bats. We wanted to create a new kind of wind energy harvesting machine based on the jiggling motion of cottonwood tree leaves in the wind, which would be quieter and safer for wildlife. After building and testing artificial cottonwood leaves that moved and created electricity in the wind, we found that they didn't produce enough energy to feasibly use for electricity production. We also tried building a cattail-like device to generate electricity when it swayed in the wind, [...]	Introduction. Accumulation of glycochenodeoxycholic acid (GCDC) in serum has a clinical significance as an inductor of pathological hepatocyte apoptosis, which impairs liver function. Inhibition of GCDC accumulation can be used as a marker in therapy. This study was aimed to quantify the serum level of GCDC in obstructive jaundice patients. Methodology. GCDC acid level in the serum was quantified using high performance liquid chromatography (HPLC) technique according to Muraca and Ghoos modified method. It was performed before and after decompression at day 7 and day 14. The sample was extracted with solid phase extraction (SPE) technique on SPE column. The results were analyzed using SPSS V 16.0 ($P < 0.05$) [...]
(a) FKGL = 16.47 (College-graduate), LM score = 4/5.	(b) FKGL = 10.0 (10th grade), LM score = 1/5.

Table 6: Examples of disagreement between FKGL and the LM evaluator. 6a contains an example from the SJK dataset that the LM rated high readability and FKGL rated low readability. 6b contains a summary from the Pubmed dataset that the LM rated low readability while FKGL rated high readability.

the text. In § 4.3, we show that LMs are better able to reason over these more complex attributes.

Table 6 shows examples in which the LM evaluator and FKGL disagree on the readability. FKGL rates a summary from the SJK dataset as having a graduate-college reading level, while the LM rates it as highly readable (Table 6a). Although the summary explains the concepts well, long words such as "harvesting" and "electricity" likely cause FKGL to rate the summary as less readable. Table 6b has a Pubmed example, which the LM rates as having low readability, while FKGL assigns the summary a 10th grade reading level. This example contains many short words, such as "GCDC" and "SPE", which are favored by FKGL. Although short, these technical words that are not well defined. For example, the "GCDC" is defined as "glycochenodeoxycholic acid," but is not otherwise explained. In general, we notice that FKGL favors acronyms, which are often present in technical text.

5.2 Recommendations and Future Directions

We find that many traditional readability metrics have poor correlation with human judgments and that LMs provide better judgments. However, LM-evaluators are an imperfect solution since they are subject to bias and a lack of interpretability (Liu et al., 2023b; Wang et al., 2023; Shen et al., 2023; Stureborg et al., 2024). Therefore, we recommend a multi-faceted evaluation of PLS that uses a combination of traditional readability metrics and LM evaluators. Specifically, we recommend using DCRS and CLI, which have the highest correlation with human judgments. We recommend discontinuing use of FKGL for PLS, the current most popular metric, due to low correlation with human judgment. We recommend using LMs as additional metrics, especially for more qualitative evaluations, such as the keyword analysis conducted in § 4.3. These types of analyses give a more holistic view of the benefits and downsides of datasets and methods. Finally, we recommend that PLS research use datasets with higher readability scores (§ 4.3), such as CDSR and SciNews. We recommend that PLOS and CELLS be considered general scientific summarization datasets and not plain language datasets. This recommendation is particularly impactful as PLOS has been used in every year the shared task BioLaySumm has occurred (Goldsack et al., 2024, 2023).

Future work should focus on constructing metrics that better align with human judgments of readability in both definition and measurement (§ 5.1). We show that LMs are promising and worthy of future work that can decrease bias and improve interpretability. Dataset collection should focus on collecting highly readable summaries and consider deeper attributes of readable summaries, such as explanations of technical concepts.

Limitations

The conclusions of this paper are limited to the task of plain language summarization, and are not intended to apply to other applications of readability metrics, such as judging the age-level appropriateness of educational material. Additionally, our human judgments and experiments focused on the summarization of scientific articles, and may not generalize to PLS in other domains, such as law or clinical notes. Finally, our experiments are limited to the English language, and our findings may not apply to other languages. We leave the exploration of readability evaluation in other domains and languages to future work.

Ethical Considerations

This paper involves the use of LMs for generation and evaluation. LMs have been shown to generate factually incorrect information and are subject to bias (Venkit et al., 2024; Stureborg et al., 2024). Additionally, the use of language models contributes to the environmental footprint of our field (Schwartz et al., 2020). However, this paper focuses on the evaluation of plain language summarization, which has the potential to make scientific knowledge more accessible to the general population. Therefore, we believe that the benefits of this work outweigh the potential harms.

Acknowledgments

We'd like to thank the authors of August et al. (2024) for sharing their human-annotated dataset with us. We'd additionally like to thank Tal August for his insightful guidance in creating this paper. DK was supported by ONR (N00014-24-1-2089).

References

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know your audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Annual Meeting of the Association for Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Afonso Cavaco Carla Pires and Marina Vigário. 2017. Towards the definition of linguistic metrics for evaluating text readability. *Journal of Quantitative Linguistics*, 24(4):319–349.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, W. Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *North American Chapter of the Association for Computational Linguistics*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Scott Andrew Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2021. The commonlit ease of readability (clear) corpus. In *Educational Data Mining*.

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

William H DuBay. 2004. The principles of readability. *Impact Information*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar,

Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li,

Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Rudolf Flesch. 1952. "simplification of flesch reading ease formula". *Journal of Applied Psychology*.

Rudolf Franz Flesch. 1943. Marks of readable style: a study in adult education. In *Teachers College Contributions to Education*.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873, Singapore. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *ArXiv*, abs/2209.12356.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Yue Guo, Tal August, Gondy Leroy, Trevor A. Cohen, and Lucy Lu Wang. 2023. APPLS: Evaluating evaluation metrics for plain language summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor A. Cohen. 2022. Retrieval augmentation of large language models for lay language generation. *Journal of biomedical informatics*, page 104580.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Yu Han, Aaron Ceross, and Jeroen Bergmann. 2024. The use of readability metrics in legal text: A systematic literature review. *ArXiv*, abs/2411.09497.

Nur Rachma Isnaeni. 2017. Readability of english written materials. *Elite : English and Literature Journal*, 1(1):179–191.

Yuelyu Ji, Zhuochun Li, Rui Meng, Sonish Sivarajkumar, Yanshan Wang, Zeshui Yu, Hui Ji, Yushui Han, Hanyu Zeng, and Daqing He. 2024. RAG-RLRC-LaySum at BioLaySumm: Integrating retrieval-augmented generation and readability control for layman summarization of biomedical texts. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 810–817, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2022. GENIE: Toward Reproducible and Standardized Human Evaluation for Text Generation. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2025. Exploring the reliability of large language models as customized evaluators for diverse NLP tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10325–10344, Abu Dhabi, UAE. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023b. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Annual Meeting of the Association for Computational Linguistics*.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq R. Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir R. Radev. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Annual Meeting of the Association for Computational Linguistics*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. *ArXiv*, abs/2210.04705.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Rostislav Nedelchev, Jens Lehmann, and Ricardo Usbeck. 2020. Language model transformers as evaluators for open-domain dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6797–6808, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John O'hayre. 1966. *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.

R. Timothy Rush. 1985. Assessing readability: Formulas and alternatives. *The Reading Teacher*, 39(3):274–283.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Johannes Sibeko and Menno van Zaanen. 2022. An analysis of readability metrics on english exam. *Journal of the Digital Humanities Association of Southern Africa*, 3(01).

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan

Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (TMLR).

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.

Loukritia Stefanou, Tatiana Passali, and Grigorios Tsoumakas. 2024. Auth at biolaysumm 2024: Bringing scientific content to kids. In *Proceedings of the ACL 2024 BioNLP Workshop*, Bangkok, Thailand. A paper presented at the BioLaySumm 2024 shared task on lay summarization of biomedical research articles.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *ArXiv*, abs/2405.01724.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*.

Gemma Team. 2024. Gemma.

Edward L. Thorndike. 1936. *The Elementary School Journal*, 36(6):470–472.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An audit on the perspectives and challenges of hallucinations in nlp. In *Conference on Empirical Methods in Natural Language Processing*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Annual Meeting of the Association for Computational Linguistics*.

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif R. Aljohani, and Raheel Nawaz. 2020. HTSS: A novel hybrid text summarisation and simplification architecture. *Inf. Process. Manag.*, 57:102351.

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*.

Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 33–42, New York, NY, USA. Association for Computing Machinery.

Appendix

A Human annotated dataset details

We use the human annotated data collected by August et al. (2024). The dataset includes 60 summaries over 10 papers, 6 summaries per paper. Of the 6 summaries, 2 are written by experts and 4 are machine written by GPT3. The 10 papers were sampled from the top 10% of papers from *r/science*, a subreddit dedicated to public discussions of scientific papers. These papers were chosen as a proxy for scientific topics the general public is most interested in. The dataset was annotated by 593 Mechanical Turk workers in total across the three tasks in the original study. Table 7 contains the distributions of scores assigned by the human annotators.

Score	5	4	3	2	1
%	37	29	17	10	7

Table 7: Percentage of scores assigned in human annotated dataset for reading ease.

In order to measure inter-annotator agreement, we bin the scores into a binary "high-readability" and "low readability." Summaries given scores of 3 or higher are considered highly readable while summaries assigned scores less than 3 are considered to have low readability. We use Cohen's Kappa to calculate an inter-annotator agreement of 0.6. This is a moderate agreement for a somewhat subjective task, indicating that there is some general notion of readability. We also note that this is significantly higher than the agreement between traditional metrics and LMs (0.17 as shown in § 4.4).

B LM readability evaluation prompts

We experiment with 3 prompts, shown in Table 10. The `Simple Prompt` simply asks the LM to rate the text for reading ease on a scale of 1 to 5. The American Society for Cell Biology (ASCB) provides guidelines for best practices in scientific communication.[5] In the `ASCB Prompt`, we provide these guidelines to the LM as guidance for rating the readability. Finally, the `Own Reasoning Prompt` is similar to the `Simple Prompt`, but with the additional instruction for the LM to use it's own judgment to rate the text, rather than traditional readability formulas, such as FKGL.

We report the Pearson and Kendall-Tau correlation of each prompt with human judgment in Table 8. The `Own Reasoning Prompt` performs the best when averaged across all models. We found that the models tended to over-rely on the guidance provided in the `ASCB Prompt`, providing lower scores if the conditions are not met. For the `Simple Prompt`, the models would occasionally try to calculate FKGL or another readability metric, rather than using its own reasoning. This is likely because FKGL is strongly associated with readability in the models' training data. We found that the `Own Reasoning Prompt` struck the right balance between providing enough instructions that the model is able to understand the task without providing too much information for the model to over-rely on. However, it is notable that the `ASCB Prompt`, the worst performing prompt, still achieves higher correlation with human judgment than FKGL, the most popular traditional metric.

[5] ASCB Best Practices in Science Communication

Model	Simple	ASCB	Own
Mistral 7B	0.46	0.54	0.52
Mixtral 7B	0.46	0.47	0.54
Gemma 1.1 7B	0.55	0.33	0.54
Llama 3.1 8B	0.54	0.56	0.45
Llama 3.3 70B	0.59	0.58	0.56
Mean Corr.	0.52	0.50	0.52

(a) Pearson Correlation.

Model	Simple	ASCB	Own
Mistral 7B	0.32	0.40	0.44
Mixtral 7B	0.36	0.41	0.41
Gemma 1.1 7B	0.42	0.24	0.43
Llama 3.1 8B	0.38	0.35	0.34
Llama 3.3 70B	0.36	0.38	0.35
Mean Corr.	0.37	0.36	0.39

(b) Kendall-Tau Correlation.

Table 8: Pearson and Kendall-Tau Correlation with human judgment for each prompt listed in Table 10. `Own Reasoning` prompt performs the best averaged across all models.

C Statistical Significance

We use the William's test to calculate statistical significance of the difference in performance between each LM evaluator and traditional metric (Graham and Baldwin, 2014). We report the p-values in Table 9. The difference in Pearson correlation between `Llama 3.3 70B`, the best performing model, the traditional metrics is statistically significant, except for DCRS and CLI. The Pearson correlation difference between the LM evaluators and FKGL, the most popular metric, is statistically significant, except `Llama 3.1 8B`. The Kendall-Tau values show that the Mistral, Mixtral, and Gemma models are statistically significant over most of the traditional metrics. This supports our suggestions from § 5.2, in which we recommend using a combination of the best performing traditional metrics (DCRS and CLI) with LM evaluators, while discontinuing the use of FKGL.

	LW	Spache	FRE	ARI	GFI	DCRS	CLI	FKGL
Mistral 7B	6.51E-04	0.02	0.05	0.01	0.05	0.19	0.18	0.02
Mixtral 7B	6.90E-04	0.01	0.03	0.01	0.04	0.16	0.15	0.02
Gemma 7B	9.65E-04	0.02	0.03	0.01	0.04	0.17	0.15	0.02
Llama 3.1 8B	2.00E-03	0.04	0.14	0.03	0.10	0.34	0.31	0.06
Llama 3.1 70B	3.66E-04	0.01	0.02	0.01	0.03	0.14	0.13	0.02

(a) Pearson correlation p-values.

	LW	Spache	FRE	ARI	GFI	DCRS	CLI	FKGL
Mistral 7B	0.01	0.01	0.03	0.01	0.04	0.13	0.10	0.03
Mixtral 7B	0.01	0.02	0.05	0.02	0.06	0.19	0.14	0.04
Gemma 7B	0.01	0.02	0.03	0.02	0.05	0.16	0.10	0.03
Llama 3.1 8B	0.02	0.05	0.12	0.05	0.12	0.31	0.25	0.09
Llama 3.1 70B	0.03	0.06	0.09	0.05	0.12	0.30	0.24	0.09

(b) Kendall-Tau p-values.

Table 9: William's test p-values comparing the difference in performance between each LM and each traditional metric. Values that are statistically significant (p-value < 0.05), are highlighted in green.

Simple Prompt

On a scale of 1 to 5, what is the reading ease of the following text? 1 indicates the text requires expert background knowledge and 5 indicates the text is readable to the general population. Assume the reader is an adult. \n \n
Format the output as follows: \n
Score: <score> \n
Reason: <reasoning> \n
Text: {SUMMARY}

ASCB Guidelines Prompt

On a scale of 1 to 5, what is the reading ease of the following text? 1 indicates the text requires expert background knowledge and 5 indicates the text is readable to the general population. Characteristics of a highly readable text include: \n
- Know your audience, and focus and organize your information for that particular audience. \n
- Focus on the big picture. What larger problem is your work a part of? What major ideas or issues does your work address? How will your work help global understanding of some issue? \n
- Avoid jargon. If you must use a technical term, make sure to explain it, but simplify the language. \n
- Try to use metaphors or analogies to everyday experiences that people can relate to. \n
- Underscore the importance of public support for exploratory research and scientific information, and the role of this information in providing the context for effective policy making. \n \n
Assume the reader is an adult. Do not use Flesch-Kincaid or other readability formulas. Use your own judgment to rate the text. \n \n
Format the output as follows: \n
Score: <score> \n
Reason: <reasoning> \n \n
Text: {SUMMARY}

Own Reasoning Prompt

On a scale of 1 to 5, what is the reading ease of the following text? 1 indicates the text requires expert background knowledge and 5 indicates the text is readable to the general population. \n
Assume the reader is an adult. Do not use Flesch-Kincaid or other readability formulas. Use your own judgment to rate the text. \n \n
Format the output as follows: \n
Score: <score> \n
Reason: <reasoning> \n \n
Text: {SUMMARY}

Table 10: Prompts we tested. `Own Reasoning` is the best performing prompt, as reported in Table 8.