Readability Reconsidered: A Cross-Dataset Analysis of Reference-Free Metrics

Catarina G. Belem[1], Parker Glenn[2], Alfy Samuel[2], Anoop Kumar[2], Daben Liu[2],

[1]University of California Irvine, [2]Capital One

Correspondence: cbelem@uci.edu

Abstract

Automatic readability assessment plays a key role in ensuring effective and accessible written communication. Despite significant progress, the field is hindered by inconsistent definitions of readability and measurements that rely on surface-level text properties. In this work, we investigate the factors shaping human perceptions of readability through the analysis of 897 judgments, finding that, beyond surface-level cues, information content and topic strongly shape text comprehensibility. Furthermore, we evaluate 15 popular readability metrics across five English datasets, contrasting them with six more nuanced, model-based metrics. Our results show that four model-based metrics consistently place among the top four in rank correlations with human judgments, while the best performing traditional metric achieves an average rank of 8.6. These findings highlight a mismatch between current readability metrics and human perceptions, pointing to model-based approaches as a more promising direction.

1 Introduction

Readability assessment can be used to determine the level of comprehension of a piece of text (DuBay, 2004; Collins-Thompson, 2014). In domains such as science communication (Kerwer et al., 2021; August et al., 2023), health (Friedman and Hoffman-Goetz, 2006; Hershenhouse et al., 2024), law (Curtotti et al., 2015; Cheong et al., 2024), and education (Vajjala and Lučić, 2018), readability assessment plays a key role in making information accessible to individuals regardless of their background or cognitive needs (Collins-Thompson, 2014). It is important for highly-specialized fields characterized by dense jargon and complex language (Friedman and Hoffman-Goetz, 2006; Han et al., 2024), as well as for applications engaging with users of varied familiarity with the domain (Joshi et al., 2025; Puech et al., 2025).

One challenge in advancing automatic readability assessment is that *readability* is an overloaded term, measured in different ways by prior work. Some studies treat readability as *text difficulty*, using surface-level properties such as word length, word frequency, and various word type counts (Flesch, 1948; Kincaid et al., 1975; Leroy et al., 2008). Others broaden the definition of readability to consider syntactic and discourse-level organization, including cohesion and coherence properties (Graesser et al., 2004; Petersen, 2007; Pitler and Nenkova, 2008; Feng et al., 2010; Eslami, 2014; Zhuang et al., 2025). A third line of work views readability as a combination of text characteristics and information content (Xia et al., 2016; August et al., 2024).

Taken together, the diversity of interpretations highlight the difficulty of pinning down readability, and have led to the continued use of proxy metrics that may not fit the task, domain, or are misaligned with human comprehension judgments (Ahmed, 2023; Liu and Lee, 2023; Han et al., 2024).

2 Related Work

Readability Datasets. Despite growing interest in readability assessment, high-quality datasets remain scarce (Xia et al., 2016). Existing document-level datasets can be subdivided into *parallel* corpora (Vajjala and Lučić, 2018; August et al., 2024; Joshi et al., 2025) and non-parallel corpora (Lu et al., 2022; Crossley et al., 2024) and span various tasks and content type, including literary and informational (Crossley et al., 2024), academic (August et al., 2024), or information-seeking content (Lu et al., 2022; Joshi et al., 2025). Recently, sentence-level datasets have also been introduced (Arase

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 47–69

et al., 2022; Naous et al., 2024).

Readability Metrics. While human judgments remain the gold standard for readability evaluation, their collection is often time-consuming and expensive (Rooein et al., 2024). Automated metrics have emerged as a cheaper and quicker alternative. Examples include metrics relying on basic linguistic features, including sentences, words, and syllables counts, average reading time (Demberg and Keller, 2008), language model perplexity (Collins-Thompson, 2014; Pitler and Nenkova, 2008), and fraction of functional (Leroy et al., 2008, 2010) or uncommon words (August et al., 2024). Surface-form features have been further combined to form *readability tests*, such as the Automatic Reading Index (Senter and Smith, 1967), Dale-Chall Readability Score (Dale and Chall, 1948), Flesch-Kincaid Reading Ease (Flesch, 1948), and Linsear Write Formula (Klare, 1974). Despite critiques of brittleness (Rooein et al., 2024; Collins-Thompson, 2014) and limited domain suitability (Leroy et al., 2010), these formulas continue to be used. Recently, both fine-tuning (Arase et al., 2022; Naous et al., 2024) and LLM-as-a-judge approaches (Rooein et al., 2024; Trott and Rivière, 2024) have been proposed to capture more abstract and nuanced aspects of readability. However, since these methods rely on implicitly learned representations, they are regarded as less interpretable than those grounded in surface-level textual features.

3 How Do Humans Perceive Readability?

Given the divergent definitions of readability and continued reliance on surface-form metrics, we take a human-centric perspective, asking: *What guides human perceptions of readability?* To address this question, we analyze a subset of the ELI-WHY (GPT-4) (Joshi et al., 2025) dataset, designed to study whether LLMs can generate explanations tailored to various readability levels. The dataset comprises GPT-4–generated explanations for 299 "Why" questions, each annotated by humans into three readability levels— Elementary , High School , and Graduate —along with accompanying rationales justifying their judgments. Each question–explanation pair was independently rated by three annotators, and final labels were determined via majority vote. For additional details, see the original paper. Table 11 (in Appendix) shows randomly selected examples of human rationales for each readability level.

Exploring Human Rationales. Although Joshi et al. (2025) collected human rationales supporting readability judgments, their analysis primarily focuses on the labels themselves, offering limited insight into the factors shaping human perceptions. We complement their study by providing a quantitative perspective on the key factors driving human text comprehension through the analysis of human rationales. Two authors of this paper annotated the human-provided readability rationales for 90 ELI-Why question–answer pairs, balanced evenly across classes. Building on the original human annotation instructions, each rationale was labeled with one or more of the following categories:

- *Wording/Terminology*: presence of scientific words, abbreviations, or complex synonyms;
- *Sentence Structure*: comments on sentence length or the number of concepts;
- *Examples/Analogies*: mentions of examples or analogies as key factors;
- *Details and Depth*: mentions of the presence or absence of details;
- *Curriculum-based*: links the information content or topic to a specific education level.

Figure 1 shows the consensus vote across readability classes. The average sample-level Jaccard index for the obtained annotations is 0.91, indicating high agreement between the two annotators. *Wording/Terminology* emerges as the predominant rationale for readability judgments, with annotators distinctions in lexical complexity (*e.g.*, "Words like adherence are too advanced for elementary school") or simplicity ("uses basic words"). The *Curriculum-based* category is invoked far more often to justify High School and Graduate judgments than Elementary , with annotators noting that "The scientific terms... require an introductory background or some foundational knowledge" or that "a concept that will be brought up in chemistry classes in undergrad." Conversely, *Examples/Analogies* is disproportionately used to support Elementary judgments, with comments such

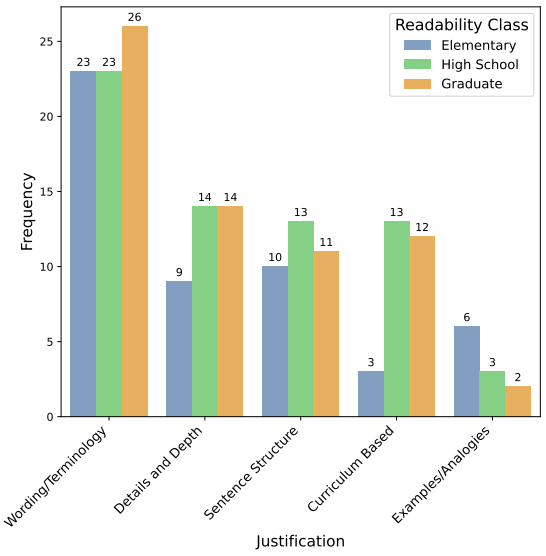


Figure 1: **Distribution of justification reasons across 90 examples in ELI-WHY (GPT-4)**. Counts are based on the consensus over 2-way annotations.

as "Examples are... what you'd say to a toddler" or "The analogies used... make it more accessible to elementary level.". Notably, both categories rely on comprehension and common-sense reasoning that go beyond surface-level textual properties.

4 Re-evaluating Readability Metrics

Motivated by the gap between surface-form textual cues and human perceptions of readability, we investigate how well existing readability metrics correlate with human judgments across five diverse datasets (see statistics in Table 1).[1]

4.1 Datasets

SCIENTIFIC PAPERS (August et al., 2024) consists of 180 total query-focused summaries about 10 different academic papers (*e.g.*, What did the paper do?) and cover topics from public policy to nanotechnology. Summaries were carefully curated by an expert science writer to reflect three levels of complexity: Low, Medium, and High.

CLEAR (Crossley et al., 2024) contains 4.7k text excerpts sourced from open digital libraries including Project Gutenberg and Wikipedia. The texts are self-contained and cover both cover literary and informational content. Approximately 111k pairwise readability judgments from 1.1k annotators were aggregated under a Bradley-Terry model to obtain continuous readability scores.

ELI-WHY (GPT-4) (Joshi et al., 2025) includes 897 GPT-4–generated explanations tailored to three readability levels— Elementary, High School, and Graduate —each annotated with human-assigned labels and rationales. Likewise, **ELI-WHY (HUMAN)** is a smaller dataset with 123 answers that were manually curated.

SCIENCEQA (Lu et al., 2022) is a multi-modal science reasoning dataset consisting of 21k multiple choice questions sourced from K-12 curriculum, covering various subjects (*e.g.*, natural science, language science, and social science). Each example is associated with a reference solution (or *explanation*) and reference knowledge (or *lecture*), both of which are written at the readability level of the intended student audience. We randomly sample 200 text-only examples per grade for our analysis.

4.2 Metrics [2]

Surface-form metrics consist of direct counts of properties of the text, such as characters, syllables, monosyllables, polysyllables, words, and sentences. These also include other specialized variants such as estimated reading time in seconds, number of difficult words, and functional words.

Psycholinguistic metrics, known as *readability tests*, are typically formulated as weighted sums of ratios involving surface-form properties. For instance, *Automatic Readability Index* is based on characters-to-words and words-to-sentences ratio (Senter and Smith, 1967), the *Flesch Kincaid Reading Ease* on words-to-sentences and syllables-to-words (Flesch, 1948), and *Dale-Chall Readability* on the fraction of difficult words and words-to-sentences ratio (Dale and Chall, 1948). An exception is the *Linsear Write Formula*, which distinguishes easy from hard words using syllable counts and computes their frequencies in a text sample (Klare, 1974). We additionally report values for other popular metrics (Coleman and Liau, 1975; Gunning, 1952; Harry and Laughlin, 1969).

Model-based metrics are categorized into two main classes: *fine-tuned metrics* (Zhuang et al., 2025) and *LLM-as-a-judge metrics* (Zheng et al., 2023). In this work, we use two fine-tuned metrics based on ModernBERT (Warner et al., 2024) –

[1] Additional dataset details are available in Appendix A.

[2] We refer readers to Appendix B for additional details.

Dataset	Size	Label Type	Labels	Avg. #WORDS	Avg. #SENTS
SCIENTIFIC PAPERS (August et al., 2024)	180	categorical	Low ≺ Medium ≺ High	65.93	2.22
CLEAR (Crossley et al., 2024)	1000	continuous	N/A	199.23	9.45
ELI-WHY (GPT-4) (Joshi et al., 2025)	897	categorical	Elementary ≺ High School ≺ Graduate	144.21	6.97
ELI-WHY (HUMAN) (Joshi et al., 2025)	117	categorical	Elementary ≺ High School ≺ Graduate	99.03	4.22
SCIENCEQA (Lu et al., 2022)	2295	categorical	Grade 1 ≺ Grade 2 ≺ ... ≺ Grade 12	183.08	13.26

Table 1: Dataset statistics, including dataset size, readability label type (continuous vs categorical), average number of words and sentences across examples.

META RATER (READABILITY) and META RATER (PROFESSIONALISM), which were recently introduced to evaluate texts along readability and professionalism dimensions, respectively (Zhuang et al., 2025). The former considers factors such as clarity, coherence, vocabulary complexity, and sentence structure with the goal of assessing whether a reader can understand a written text, whereas the latter relies on the depth and content accessibility to determine the degree of expertise or knowledge required to comprehend a text. Additionally, we include a complementary BERT-based metric—README++ (Naous et al., 2024)—which predicts readability in terms of language learning capabilities through the use of the 6-point Common European Framework of Reference for Languages scale.

We test three different LLM-as-a-judge approaches, including the zero-shot continuous score approach by Trott and Rivière (2024) (dubbed LLM-AS-A-JUDGE CONTINUOUS 0-100). We also test a categorical setting, in which a model is tasked with predicting one of three readability labels - Elementary, High School or Graduate. We prompt the model with the same instructions provided to human annotators in Joshi et al. (2025) and, in the 5-shot setting, include the five example annotations (two Elementary, two Graduate, one High School). All LLM-as-a-judge approaches are performed using `Llama-3.3-70B-Instruct` with greedy decoding (`temperature=0`).

4.3 Results & Discussion

An ideal metric should correlate strongly with human judgments of readability. To operationalize this, and given that readability labels are ordinal, we map the discrete labels to monotonically increasing numeric values ranging from 0 to $k-1$. We apply a similar transformation to the outputs of model-based metrics to obtain numerical values and then compute the correlation between metric outputs and human annotations using the Kendall Tau-b coefficient (Kendall, 1938).[3][4] To assess overall performance, we report the average rank order across all datasets (`Avg. Rank`).

Table 2 shows that ***model-based metrics systematically achieve stronger correlations with human judgments***, surpassing surface-form and psycholinguistic metrics by up to 0.24 absolute points. Notably, all three LLM-as-a-judge metrics consistently rank in the top three (average ranks 2.4–3.2), followed closely by the fine-tuned META RATER (PROFESSIONALISM) and README++ models. Looking at the disagreements between metrics, we find LLM-as-a-judge metrics to be more sensitive to specialized terminology and sentence structure, whereas fine-tuned models like README++ are more sensitive to information density and presence of connectors and cohesive devices. Comparing META RATER (PROFESSIONALISM) with META RATER (READABILITY), the latter shows an average correlation rank of 21.0, falling below psycholinguistic and surface-form metrics, where the best traditional metric achieves 8.6. This may be because examples are generally clear, grammatically correct, and coherent, leading the model to systematically assign the same readability class. Conversely, because META RATER (PROFESSIONALISM) reflects the depth and expertise demanded

[3] We use the implementation available in `scipy.stats`.

[4] See Appendix B for details on the categorical-to-numerical mappings used for each metric.

Type	Metric	SCIENTIFIC PAPERS (August et al., 2024)	CLEAR (Crossley et al., 2024)	ELI-WHY (GPT-4) (Joshi et al., 2025)	ELI-WHY (HUMAN) (Joshi et al., 2025)	SCIENCEQA (Lu et al., 2022)	Avg. Rank
Surface-form	# Words	0.16*	-0.06*	0.46*	0.15	0.28*	17.0
	# Sentences	0.25*	0.23*	0.38*	-0.07	0.09*	17.0
	Avg. Sentence Length	-0.15	-0.25*	0.21*	0.40*	0.39*	16.4
	Avg. Reading Time (s)	0.20*	-0.23*	0.47*	0.25*	0.32*	14.8
	# Syllables	0.22*	-0.28*	0.47*	0.28*	0.33*	13.2
	# Monosyllables	0.08	0.16*	0.39*	0.01	0.22*	18.8
	# Polysyllables	0.31*	-0.33*	0.46*	0.47*	0.41*	9.6
	# Difficult Words	0.26*	-0.40*	0.45*	0.46*	**0.48***	8.6
	TE Score	0.35*	-0.18*	0.34*	0.34*	0.06*	17.2
Psycholinguistics	Automatic Readability Index	0.07	-0.33*	0.36*	0.56*	0.40*	11.0
	Coleman Liau Index	0.30*	-0.32*	0.31*	0.54*	0.35*	16.8
	Dalle Chall Readability Score	0.37*	-0.37*	0.37*	0.52*	0.22*	12.4
	Flesch Reading Grade	0.15	-0.36*	0.37*	0.58*	0.40*	11.6
	Flesch-Kincaid Reading Ease	-0.32*	0.37*	-0.35*	-0.58*	-0.36*	11.8
	Gunning Fog	0.15*	-0.37*	0.39*	0.57*	0.37*	14.0
	Linsear Write Formula	-0.06	-0.31*	0.24*	0.45*	0.40*	14.2
	SMOG Index	0.14	-0.38*	0.37*	0.59*	0.37*	12.2
Model-based	README++	0.40*	**-0.45***	**0.50***	0.50*	0.44*	6.2
	Meta Rater (readability)	-0.17	0.14*	0.00	0.00	0.09*	21.0
	Meta Rater (professionalism)	**0.49***	-0.40*	**0.51***	**0.67***	0.44*	**4.2**
	LLM-as-a-judge (0-shot)	**0.57***	**-0.50***	**0.49***	**0.73***	**0.60***	**2.4**
	LLM-as-a-judge (5-shot)	**0.61***	**-0.55***	0.43*	**0.71***	**0.61***	**3.2**
	LLM-as-a-judge (continuous 0-100)	**-0.56***	**0.59***	**-0.53***	**-0.68***	**-0.52***	**2.4**

Table 2: **Rank correlations between readability metrics and human judgments of correctness across 5 datasets.** We report the Kendall Tau coefficient and boldface the four metrics exhibiting strongest correlations with human judgments. *indicates correlation coefficients with p-value < 0.01.

by each input, we hypothesize it better aligns with human perceptions of readability which go beyond lexical and syntactic cues (see Section 3).

Together these results demonstrate the strong performance of LLM-as-a-judge metrics. However, we highlight the trade-off with inference cost, as LLM-based evaluations typically require generating text for each instance, making them slower and more resource-intensive approaches than fine-tuned models. We also note that despite achieving the strongest correlations with human judgments (up to 0.73), ***model-based metrics remain far from perfect alignment***, suggesting room for improvement.

Overall, **no single model-based metric consistently dominates**: while the continuous LLM-as-a-judge metric achieves the highest correlations on three datasets, it underperforms relative to LLM-AS-A-JUDGE (0-SHOT) on ELI-WHY (HUMAN) and SCIENCEQA. The two metrics differ considerably: the continuous variant penalizes texts containing numbers and named entities (*e.g.*, "The Barber of Seville"), whereas the discriminative one is more sensitive to scientific terminology (*e.g.*, "hydrophobic effect", "endergonicity"), complex sentence structures, and equations. Despite its finer granularity, the continuous approach shows marked score saturation in SCIENCEQA (Li et al., 2025), with 81.30% of scores confined to three values.

Surface-form metrics outperform psycholinguistic metrics on 4 (out of 5) datasets. With the exception of ELI-WHY (HUMAN) dataset, Table 2 shows that there is always a simpler surface-level metric (*e.g.*, # DIFFICULT WORDS, or #SYLLABLES) that is on par or outperforms popular metrics, such as the Automatic Readability Index or the Flesch Kincaid Reading Ease. Upon further analysis, we find that the stronger correlation observed for average sentence length in the ELI-WHY (GPT-4) can be attributed to length bias in the generations, where perceived readability is linked to the explanation's length (see Figure 3).

5 Conclusion

This work tackles the inconsistency of readability definitions (and metrics) in the literature by showing that human perceptions of readability go beyond lexical and syntactic features, also considering topic and information content. Furthermore, we benchmark 20+ reference-less metrics–including LLM-as-a-judge and fine-tuned models–across five datasets. Our results show that model-based metrics correlate more strongly with human judgments than popular readability metrics, suggesting they capture more nuanced features. Together, these findings call for clearer definitions of readability and more rigorous validation of metrics, paving the way for assessments that better reflect how humans understand text.

Limitations

The analysis conducted in this paper is limited to the available datasets in the English language, therefore providing limited generalization to other languages. While we are partially motivated by the lack of high quality labeled data in other languages, a few exceptions exist namely in the French language (François and Fairon, 2012). Future work may consider expanding on this work through the creation of additional readability datasets in other languages or by expanding our analysis to other languages.

Section 3 concerns the investigation of the main factors shaping human readability judgments. While our findings are intuitive and generally aligned with prior discussion in the literature (August et al., 2024; Klare, 1974), they are based on information extracted from a single dataset in QA, potentially leading to concerns about their generalizability. However, reasoning judgments are not widely available in readability datasets, making it non-trivial to extend this analysis to other datasets. Future work could include building additional datasets, therefore, facilitating the expansion of this analysis to other domains and tasks.

Lay Summary

Readability assessment helps ensure that information can be understood by people with different backgrounds and abilities. A key goal is to automate this process and reduce the need for human evaluation.

Many datasets and methods have been developed for automatic readability assessment, but they often rely on different definitions of what makes text readable. Even today, most approaches still use basic measures, like the number of words, syllables, or sentences, to estimate readability.

In this work, we show that people's perceptions of readability depend on more than simple text features—they are strongly influenced by the content and topic of the text. We compare traditional readability measures with more advanced model-based metrics across five datasets and find that conventional measures often fail to capture what humans consider readable. Our results emphasize the need for clearer, standardized definitions of readability and for moving beyond simple, surface-level metrics.

Acknowledgments

We thank the anonymous reviewers, the members of the Capital One research team for their helpful feedback.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2024. GPT4 Technical Report. *Preprint*, arXiv:2303.08774.

Arif Ahmed. 2023. Beyond vocabulary: Capturing readability from children's difficulty. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 134–141, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans. Comput.-Hum. Interact.*, 30(5).

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2454–2469, New York, NY, USA. Association for Computing Machinery.

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

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research running title: Computational assessment of text readability.

S.A. Crossley, Y. Tian, P. Baffour, A.Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Michael Curtotti, Eric McCreath, Tom Bruce, Sara Frug, Wayne Weibel, and Nicolas Ceynowa. 2015. Machine learning for readability of legislative sentences. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, page 53–62, New York, NY, USA. Association for Computing Machinery.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

William H. DuBay. 2004. The principles of readability.

Hedayat Eslami. 2014. The effect of syntactic simplicity and complexity on the readability of the text. *Journal of Language Teaching and Research*, 5(5).

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.

Daniela B. Friedman and Laurie Hoffman-Goetz. 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education amp; Behavior*, 33(3):352–373.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, amp; Computers*, 36(2):193–202.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Robert Gunning. 1952. The technique of clear writing. *(No Title)*.

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

G Harry and Mc Laughlin. 1969. Smog grading - a new readability formula. *The Journal of Reading*.

Jacob S. Hershenhouse, Daniel Mokhtar, Michael B. Eppler, Severin Rodler, Lorenzo Storino Ramacciotti, Conner Ganjavi, Brian Hom, Ryan J. Davis, John Tran, Giorgio Ivan Russo, Andrea Cocci, Andre Abreu, Inderbir Gill, Mihir Desai, and Giovanni E. Cacciamani. 2024. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer and Prostatic Diseases*, 28(2):394–399.

Brihi Joshi, Keyu He, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, and Xiang Ren. 2025. Eli-why: Evaluating the pedagogical utility of language model explanations. *Preprint*, arXiv:2506.14200.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Martin Kerwer, Anita Chasiotis, Johannes Stricker, Armin Günther, and Tom Rosman. 2021. Straight from the scientist's mouth—plain language summaries promote laypeople's comprehension and knowledge acquisition when reading about individual research findings in psychology. *Collabra: Psychology*, 7(1).

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

George R. Klare. 1974. Assessing readability. *Reading Research Quarterly*, 10(1):62.

Gondy Leroy, Stephen Helmreich, and James R. Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International Journal of Medical Informatics*, 79(6):438–449. Special Issue: Information Technology in Health Care: Socio-technical Approaches.

Gondy Leroy, Stephen Helmreich, {James R.} Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: beyond readability formulas. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 394–398.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025. Evaluating scoring bias in llm-as-a-judge. *Preprint*, arXiv:2506.22316.

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.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Glenda M. McClure. 1987. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, PC-30(1):12–15.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Sarah E. Petersen. 2007. *Natural language processing tools for reading level assessment and text simplification for bilingual education*. Ph.D. thesis, USA. AAI3275902.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311, Vienna, Austria. Association for Computational Linguistics.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report.

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of English texts with GPT-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. 2025. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10856–10896, Vienna, Austria. Association for Computational Linguistics.

A Additional Details: Datasets

In this section, we provide additional details about the datasets. Table 1 summarizes the general statistics about the five datasets considered in this study, including the readability label type, the size of the dataset, but also the average example length in terms of word count and sentence count.

A.1 SCIENCEQA (Lu et al., 2022)

SCIENCEQA is collected from elementary and high school science curricula sourced from IXL learning[5] and with topics ranging from natural, social,

[5]https://www.ixl.com

SCIENCEQA readability example

```
Lecture: {{lecture}}
Explanation: {{explanation}}
```

Figure 2: Formatting of each SCIENCEQA example. Whenever examples miss the corresponding `{{lecture}}` or `{{explanation}}` fields, we we omit them from the template above.

and natural sciences. To ensure coverage across grades 1–12, we sample from the full dataset. We draw 200 examples per grade, except for 1st grade where only 95 are available, yielding 2295 examples overall. Although primarily a multiple-choice QA dataset, it also includes a *lecture* covering the knowledge needed to answer each question and a solution outlining how to use it to derive the answer. For every question, we compute the readability by concatenating the two fields as demonstrated in Figure 2. For some qualitative examples, see Table 3. To compute the correlation with human judgments, we use grades 1-12 as the readability judgments (12-way classification), where a higher grade implies added difficulty in comprehending a text.

A.2 CLEAR (Crossley et al., 2024)

CLEAR consists of 4.7k text excerpts sampled from online digital libraries. Each example is curated to ensure the text is self-contained and composed of full sentences. Unlike the other datasets, the readability score in CLEAR is continuous and represents the easiness of comprehension of a given text (*BT_easiness*). We refer to the original paper for additional details regarding the dataset. Table 5 illustrates a few examples from this dataset and corresponding readability score. To balance efficiency with generalization, we randomly sample 1k examples without replacement from the original dataset and use them for our correlation analysis. Table 6

A.3 SCIENTIFIC PAPERS (August et al., 2023)

SCIENTIFIC PAPERS dataset is a parallel corpus for readability, comprising 3 human-edited variants of the same summary for each example. Table 7 shows three human-curated versions of the question "What did the paper find?" at different complexity levels. The correlation analysis considers we all examples and map the ordinal classes— Low $\prec$ Medium $\prec$ High —onto a 0–2 scale.

A.4 ELI-WHY (GPT-4) (Joshi et al., 2025)

Our analyses reveal the presence of length bias, where there seems to exist a correlation between the length of GPT4-generated explanations and human perceived readability (see Figure 3). In fact, we observe a propensity for responses deemed higher readability to be longer, which can be explained by the added detail and specificity often emphasized by human experts. Future work could explore ways of mitigating this bias by enforcing strict generation lengths or, if a reference document with relevant information is available by controlling the information content within each generation (August et al., 2024).

A.5 ELI-WHY (HUMAN) (Joshi et al., 2025)

Table 8 illustrates a few randomly selected examples for the ELI-WHY (HUMAN) datasets. These explanations were manually curated by two authors of the paper.

B Automated Metrics

This section discusses the implementation details of the metrics evaluated in the main paper. All experiments are implemented in Python: `textstat`[6] is used to compute surface-form and psycholinguistic metrics; `transformers` is used to implement the model-based metrics, including fine-tuned and LLM-as-a-judge approaches.

B.1 Psycholinguistics Metrics

The metrics listed below are commonly referred to as *readability tests* and commonly used to gauge the difficulty that human readers may have in understanding a given text.

Automatic Readability Index (ARI) (Senter and Smith, 1967) estimates the US grade level needed to comprehend a text. To do so, it uses the ratio of characters-to-words and words-to-sentences. Intuitively, these ratios capture the idea that longer words and longer sentences are more difficult to grasp. The character counts include both numbers and letters. A score of 1 and 14 would match that of a Kindergarten and a College student,

[6] https://pypi.org/project/textstat/

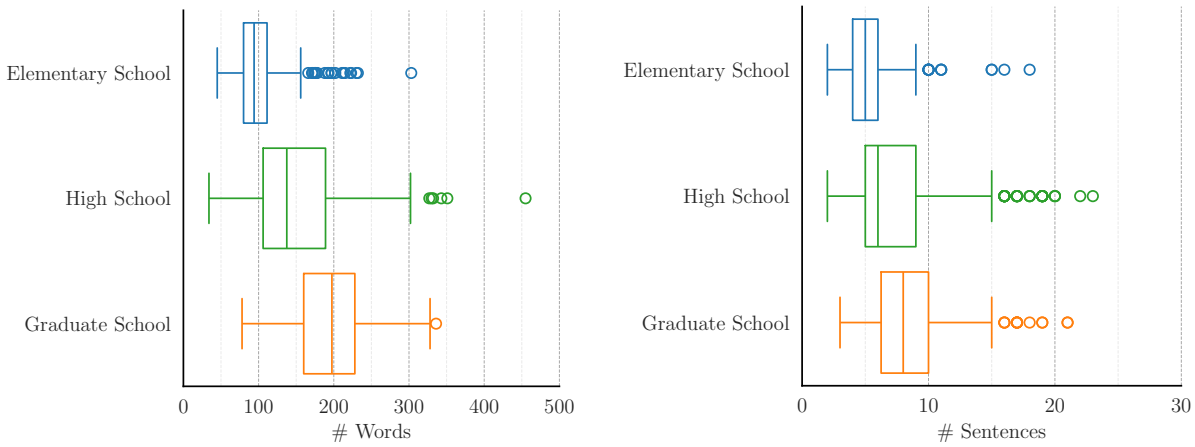


Figure 3: Distribution of number of words (# WORDS) and sentences (# SENTENCES) per readability label in the ELI-WHY (GPT-4) dataset.

respectively.

$$\lceil 4.71 \left(\frac{\texttt{\#chars}}{\texttt{\#words}}\right) + 0.5 \left(\frac{\texttt{\#words}}{\texttt{\#sentences}}\right) - 21.43 \rceil$$

Coleman Liau Index (CLI) (Coleman and Liau, 1975) similarly to ARI, it also yields an estimate of the minimum US grade level necessary to understand a piece of text. It is defined in terms of the average counts of letters and sentences per 100 words in a text sample.

$$0.0588 \cdot \texttt{\#letters} - 0.296 \cdot \texttt{\#sentences} - 15.8$$

Dale Chall Readability (Dale and Chall, 1948) leverages the fraction of difficult words in the document, as well as the average word-to-sentence count ratio to gauge the difficulty of a given text. By design, the metric relies on a pre-defined subset of 3k words that is empirically expected to be familiar to the majority of 4th graders. The formula is designed such that scores ≤ 4.9 match grade 4 and below, and scores ≥ 10 match grades 16 and above. Below we write the new Dale-Chall Formula:

$$\lfloor 64 - 0.95 \left(\frac{\texttt{\#difficult\_words}}{\texttt{\#words}}\right) - 0.69 \left(\frac{\texttt{\#words}}{\texttt{\#sentences}}\right) \rfloor$$

Flesch-Kincaid Reading Ease (FKRE) and **Flesch-Kincaid Grade Level (FKGL)** (Flesch, 1948) rely on the same core properties of language, such as average word length and average sentence length, differing only in the coefficients. The formulas were defined by the US Navy to gauge the readability of the technical material and later adopted by a few US states to impose readability requirements on various legal documents (*e.g.*, insurance policies) (McClure, 1987). The FKRE is defined in as follows:

$$206.835 - 1.015 \left(\frac{\texttt{\#words}}{\texttt{\#sentences}}\right) - 84.6 \left(\frac{\texttt{\#syllables}}{\texttt{\#words}}\right)$$

whereas the FKGL is defined as:

$$0.39 \left(\frac{\texttt{\#words}}{\texttt{\#sentences}}\right) + 11.8 \left(\frac{\texttt{\#syllables}}{\texttt{\#words}}\right) - 15.59$$

Gunning Fog Index (GFI) (Gunning, 1952) provides an estimate of the number of formal education required to understand the text on a first reading. It works by first computing the average sentence length, *i.e.*, word-to-sentence ratio of a passage and then computing the ratio of complex words in the passage. In this formula, complex words are defined as words with 3+ syllables that are not proper nouns, familiar words, or compound words. Conventionally, scores range between 6 and 17 which indicate that 6th grade and College Graduate are necessary to be able to understand a piece of text, respectively.

$$0.4 \left[\left(\frac{\texttt{\#words}}{\texttt{\#sentences}}\right) + 100 \left(\frac{\texttt{\#complex\_words}}{\texttt{\#words}}\right) \right]$$

Linsear Write Formula (LWF) (Klare, 1974) counts the number of easy and hard words in a 100-word sample. To distinguish easy from hard words, it utilizes the number of syllables in each word: polysyllable words are considered hard words, whilst words with less than 3 syllables are considered easy. It was originally designed to gauge

the readability of the technical manuals used in the US Air Force.

$$r = \frac{3 \cdot \#\texttt{hard\_words} + 1 \cdot \#\texttt{easy\_words}}{\#\texttt{words}}$$

where the final linsear write score is given by

$$\text{LWF} = \begin{cases} r/2 & \text{if } r > 20 \\ r/2 - 1 & \text{else} \end{cases}$$

SMOG grade (Harry and Laughlin, 1969) was proposed as a more accurate and easier to compute alternative to Gunning Fog Index. It is defined in terms of polysyllable counts (words with 3+ syllables) across three 10-sentence long texts.

$$1.043\sqrt{\#\texttt{polysyllables} \cdot \frac{30}{\#\texttt{sentences}}} + 3.1291$$

B.2 Model-based Metrics

META RATER (PROFESSIONALISM) and **META RATER (READABILITY)** (Zhuang et al., 2025) are two fine-tuned based metrics, both operationalized using a ModernBERT-base model. The models are designed to evaluate the *degree of required expertise* and *ease of understanding* in a 0-5 point scale, respectively. To obtain the metric score associated with a given text, each text is fed through the model and the class with maximum probability is selected (*i.e.*, greedy prediction). This score is then used to compute the correlation with human judgments.

README++ (Naous et al., 2024) is a model-based metric that grounds readability assessment in the capabilities of second-language learners. Specifically, we use `tareknaous/readabert-en`, a BERT-based model fine-tuned on the English portion of the README++ corpus—a sentence-level readability dataset spanning multiple domains (*e.g.*, finance, economics, poetry, agriculture). Readability scores are provided on a six-point scale aligned with the Common European Framework of Reference for Languages (CEFR), where higher values indicate greater language proficiency.

Since README++ was originally trained on single sentences, we hypothesize that it may not generalize well to multi-sentence inputs, such as those in SCIENCEQA or ELI-WHY (GPT-4). To address this limitation, we adopt a bottom-up approach: for each document, we first compute the README++ score for each sentence, then average them to obtain a document-level score (README++ (AVG)). We also evaluate another variant, README++ (MAX), which reflects the hypothesis that advanced readers can understand simpler texts, but not vice versa. Table 9 summarizes the results. While both README++ and README++ (MAX) exhibit the same average rank (1.8), we observe that README++ exhibits stronger correlations with human judgments in 3 (out of 5) evaluated datasets. Notably, README++ (AVG) exhibits a an average rank of 2.4, suggesting that this variant systematically under performs the other two variants in terms of correlating with human judgments. For brevity, and because of its superior performance, we restrict the analysis in the main paper to the original method—README++.

LLM-AS-A-JUDGE (0-SHOT) and **LLM-AS-A-JUDGE (5-SHOT)** are prompt-based strategies to extract the readability level from any text. Specifically, we use `Llama-3.3-70B-Instruct` [7], a popular open-source instruction-following model. To ensure that models' predictions strongly align with human readability judgments, we re-use instructions previously provided to humans (Joshi et al., 2025). The prompt is discriminative in nature, being designed to extract 3-way readability labels – Elementary, High School, Graduate. Originally, the prompt includes 5 examples of readability judgments spanning the three classes, which we refer to as LLM-AS-A-JUDGE (5-SHOT). Although these examples improve alignment with human judgments and help constrain output structure, they add runtime overhead. We therefore evaluate a 0-shot version that replaces examples with explicit format instructions. Since our goal is to compute correlations with human judgments, which can be expressed as either categorical or continuous, we map textual labels to numbers. Treating the labels as ordinal (Elementary $\prec$ High School $\prec$ Graduate), we assign them a 0–2 scale for correlation analysis.

LLM-AS-A-JUDGE (CONTINUOUS 0-100), first proposed in Trott and Rivière (2024), elicits continuous 0-100 readability scores from `GPT-4-Turbo` and `GPT-4o-mini` (Achiam et al., 2024), with higher values denoting easier texts to understand. We follow the same prompt as in the original paper (Figure 4), but replace the model

[7] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

LLM-AS-A-JUDGE (CONTINUOUS 0-100)

```
{
    "content": "You are an experienced teacher
    skilled at identifying the readability of
    different texts.",
    "role": "system"
}, {
    "content": "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.\n<Text>{{text}}</Text>\nOn 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.",
    "role": "user"
}
```

Figure 4: Prompt used to extract a 0-100 continuous score associated with the ease of readability of a given text. The placeholder `{{text}}` is either the explanation to a question or the text excerpts depending on the dataset being evaluated.

with `Llama-3.3-70B-Instruct` to ensure comparability among LLM-as-a-judge metrics.[8]

In the main paper, we ensure the reproducibility of LLM-as-a-judge evaluations by reporting correlations obtained from greedy generations (`temperature=0`).[9] This decoding strategy is not only deterministic but also commonly adopted in prior work (Trott and Rivière, 2024; Gu et al., 2025), being representative of the most likely (or modal) behavior of the LLM.

C Human Perceptions of Readability

In the main paper, we examine the reasons driving the human's annotations of various perceived readability levels. To this end, we employ various automatic pattern extraction techniques, including frequency-based analysis (represented in the form of wordclouds) and n-gram feature importance. The following sections provide additional details about each of these experiments.

[8] `Llama-3.3-70B-Instruct` consistently generates a number between 1–100.

[9] Continuous LLM-as-a-judge approaches (LLM-AS-A-JUDGE (CONTINUOUS 0-100)) are configured to generate at most 3 tokens, whereas the discriminative approaches (LLM-AS-A-JUDGE (0-SHOT) and LLM-AS-A-JUDGE (5-SHOT)) are configured to generate at most 20 tokens. We then extract the corresponding readability label through the use of regular expressions.

C.1 Frequency-based Analysis

As part of our analysis, we conduct a frequency-based analysis of the rationales behind the readability judgments provided by the human annotators in the ELI-WHY (GPT-4) dataset.

Methodology. We conduct our analysis by first separating the dataset into three subsets according to the perceived readability level of the GPT4-generated explanations. In doing so, we obtain a total of 324, 694, and 182 examples corresponding to the Elementary, High School, and Graduate, respectively. Subsequently, we merge the annotators justification field for each subset, remove the English stopwords (as provided by the NLTK library). To aggregate words with similar meanings, we further lemmatize each word using the WORDNETLEMMATIZER[10].

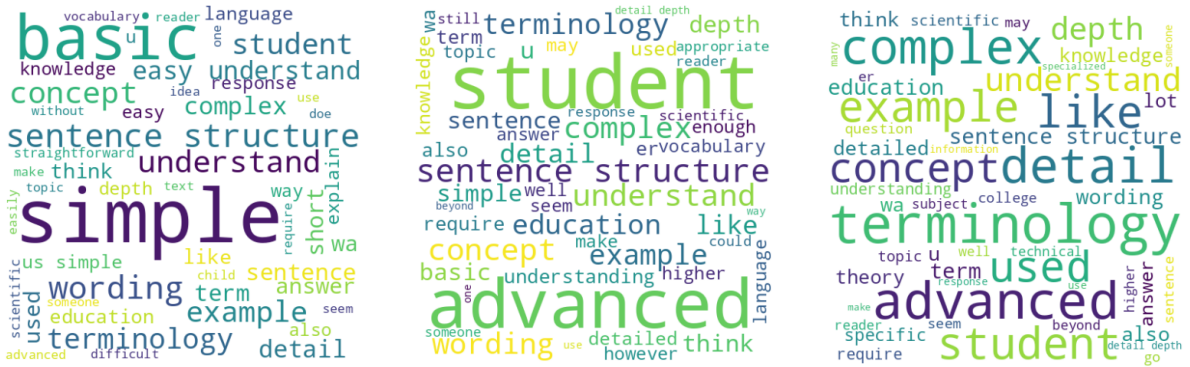
C.2 Predictive Analysis

We also conduct a model-based approach to determine the discriminative power of different phrases for each readability class. In this analysis, each annotator's justification is considered to be an individual document and both term and document frequencies are used to determine the readability class of a annotators' justifications.

Methodology. Similarly to the frequency-based analysis, we first decompose the ELI-WHY (GPT-4) dataset into three exclusive subsets based on the human perceived readability label. Additionally, we expand the justification field into individual documents, resulting in 707, 1665, and 416 total documents for Elementary, High School, and Graduate, respectively. As preprocessing steps, we remove the English stopwords using the NLTK default list, lemmatize the text using the WORDNETLEMMATIZER, and lowercase the text. Finally, we compute the term-to-document frequency matrix using SKLEARN's TFIDFVECTORIZER. To ensure that we capture complex phrases and not just individual words, we consider n-grams where $n \in \{1, 2, 3, 4\}$ and, to avoid overfitting to terms that appear in a single document, set MIN_DF=2.

Having the term-to-document frequency matrix, we adopt a one-vs-all approach, where we itera-

[10] https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html



(a) Elementary School (b) High School (c) Graduate School

Figure 5: Frequency-based analysis of the language expressions used by human annotators when judging the perceived readability of various GPT4-generated explanations in ELI-WHY (GPT-4). These word clouds are collected over 324, 694, and 182 examples annotated for Elementary, High School, and Graduate, respectively.

tively fit a linear model to discriminate justifications of one class (*e.g.*, Elementary) from justifications outside of this class (*e.g.*, High School and Graduate). While focusing on linear models such as logistic regression allow us to directly examine the predictive importance of different n-grams, it pre-assumes that the most is a strong predictor. With the intent of selecting a good predictive model, we perform hyperparameter optimization using 10-fold cross-validation while using predictive accuracy as the evaluation criteria. We consider the following hyperparameters and employ grid search:

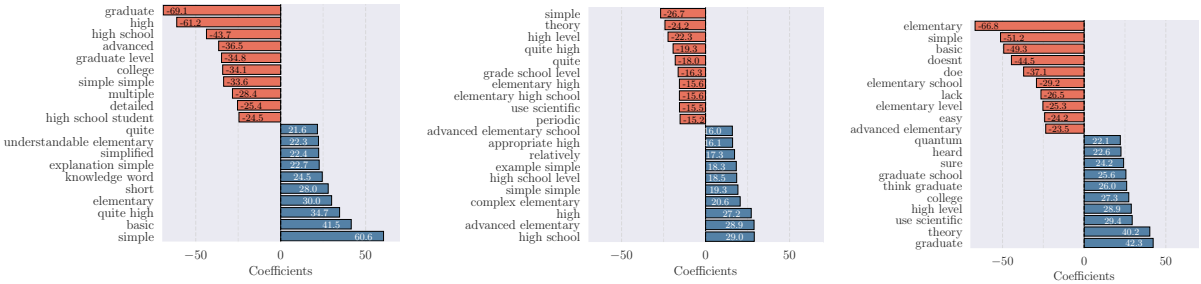
- `estimator = LogisticRegression()`
- `max_iter = {100, 300}`
- `C = {0.01, 0.1, 1, 10, 100, 500}`
- `penalty = {l1, l2, elasticnet}`
- `solver = {liblinear, saga}`

We list the best obtained models for each readability class in Table 10. Across all readability classes, we find that the fitted logistic regression outperforms a simple baseline that predicts the majority class (MAJORITY ACCURACY) by at least 3% and up to 15% absolute points.

D Related Work

In this section, we extend the discussion of readability metrics provided in the main paper. Specifically, we elaborate on the limitations of the previously proposed LLM-as-a-judge approaches and remaining challenges.

Readability Assessment using LLMs. Rooein et al. (2024) show that combining yes/no prompts with conventional metrics yields stronger correlations with human judgments than using either set of metrics alone. Trott and Rivière (2024) use 0-shot prompts to extract continuous readability scores which correlate strongly with human judgments. In spite of promising results, these approaches have seen little adoption in practice. Their reliance on repeated prompting introduces significant inference overhead, making them costly for large-scale evaluation or use as reward functions. They also require allocating part of the already limited readability data to calibrate combinations or thresholds, further limiting their practicality. Finally, although prior work has explored continuous readability assessments with LMs, to our knowledge their ability to distinguish coarse-grained readability classes remains unexplored.



(a) Elementary School (b) High School (c) Graduate School

Figure 6: Regression analysis of the language expressions used by human annotators when judging the perceived readability of various GPT4-generated explanations in ELI-WHY (GPT-4). These values clouds are collected over 324, 694, and 182 examples annotated for Elementary , High School , and Graduate , respectively.

Table 3: Randomly selected ScienceQA examples across 6 different readability classes (**grade**).

Grade	Subject (Category)	Formatted Example
1	language science (comprehension strategies)	Explanation: A book is made of paper.\nA book tells a story.\nA teacher may read a book out loud.
3	natural science (weather and climate)	Lecture: The atmosphere is the layer of air that surrounds Earth. Both weather and climate tell you about the atmosphere.\nWeather is what the atmosphere is like at a certain place and time. Weather can change quickly. For example, the temperature outside your house might get higher throughout the day.\nClimate is the pattern of weather in a certain place. For example, summer temperatures in New York are usually higher than winter temperatures.\n\nExplanation: Read the text carefully.\nWhere Sarah lives, winter is the rainiest season of the year.\nThis passage tells you about the usual precipitation where Sarah lives. It does not describe what the weather is like on a particular day. So, this passage describes the climate.
5	natural science (traits and heredity)	Lecture: Organisms, including people, have both inherited and acquired traits. Inherited and acquired traits are gained in different ways.\nInherited traits are passed down through families. Children gain these traits from their parents. Inherited traits do not need to be learned.\nAcquired traits are gained during a person's life. Some acquired traits, such as riding a bicycle, are gained by learning. Other acquired traits, such as scars, are caused by the environment. Children do not inherit their parents' acquired traits.\n\nExplanation: People are not born knowing how to cook. Instead, many people learn how to cook. So, cooking is an acquired trait.
7	natural science (designing experiments)	Lecture: Experiments can be designed to answer specific questions. When designing an experiment, you must identify the supplies that are necessary to answer your question. In order to do this, you need to figure out what will be tested and what will be measured during the experiment.\nImagine that you are wondering if plants grow to different heights when planted in different types of soil. How might you decide what supplies are necessary to conduct this experiment?\nFirst, you need to identify the part of the experiment that will be tested, which is the independent variable. This is usually the part of the experiment that is different or changed. In this case, you would like to know how plants grow in different types of soil. So, you must have different types of soil available.\nNext, you need to identify the part of the experiment that will be measured or observed, which is the dependent variable. In this experiment, you would like to know if some plants grow taller than others. So, you must be able to compare the plants' heights. To do this, you can observe which plants are taller by looking at them, or you can measure their exact heights with a meterstick.\nSo, if you have different types of soil and can observe or measure the heights of your plants, then you have the supplies you need to investigate your question with an experiment!
9	language science (literary devices)	Lecture: Figures of speech are words or phrases that use language in a nonliteral or unusual way. They can make writing more expressive.\nA euphemism is a polite or indirect expression that is used to de-emphasize an unpleasant topic.\nThe head of Human Resources would never refer to firing people, only to laying them off.\nHyperbole is an obvious exaggeration that is not meant to be taken literally.\nI ate so much that I think I might explode!\nAn oxymoron is a joining of two seemingly contradictory terms.\nSome reviewers are calling this book a new classic.\nA paradox is a statement that might at first appear to be contradictory, but that may in fact contain some truth.\nAlways expect the unexpected.\n\nExplanation: The text uses an oxymoron, a joining of two seemingly contradictory terms.\nOpen secret is a contradiction, because open describes something that is freely or publicly known, and a secret is hidden.

Table 3: Randomly selected ScienceQA examples across 6 different readability classes (**grade**). (continued)

Grade	Subject (Category)	Formatted Example
11	language science (word usage and nuance)	Lecture: Words change in meaning when speakers begin using them in new ways. For example, the word peruse once only meant to examine in detail, but it's now also commonly used to mean to look through in a casual manner.\nWhen a word changes in meaning, its correct usage is often debated. Although a newer sense of the word may be more commonly used, many people consider a word's traditional definition to be the correct usage. Being able to distinguish the different uses of a word can help you use it appropriately for different audiences.\nBritney perused her notes, carefully preparing for her exam.\nThe traditional usage above is considered more standard.\nDavid perused the magazine, absentmindedly flipping through the pages.\nThe nontraditional usage above is now commonly used, but traditional style guides generally advise against it.\n\nExplanation: The first text uses travesty in its traditional sense: a ridiculous imitation; a parody.\nDoug's ill-researched essay about the Space Race received a poor grade because it presented such a travesty of the actual historical events.\nThe second text uses travesty in its nontraditional sense: a disappointment or a tragedy.\nDoug realized that his essay about the Space Race was a bit inaccurate, but he still thought it a travesty that such an entertaining essay should receive a poor grade.\nMost style guides recommend to use the traditional sense of the word travesty because it is considered more standard.

Table 4: Randomly selected examples from the CLEAR dataset. In contrast to other datasets, each example is associated with a continuous readability score obtained by fitting a Bradley–Terry model on pairwise human judgments of reading ease.

Readability Score	**Category**	**Text**
-2.91	Info	It must not be supposed that, in setting forth the memories of this half-hour between the moment my uncle left my room till we met again at dinner, I am losing sight of "Almayer's Folly." Having confessed that my first novel was begun in idleness–a holiday task–I think I have also given the impression that it was a much-delayed book. It was never dismissed from my mind, even when the hope of ever finishing it was very faint. Many things came in its way: daily duties, new impressions, old memories. It was not the outcome of a need–the famous need of self-expression which artists find in their search for motives. The necessity which impelled me was a hidden, obscure necessity, a completely masked and unaccountable phenomenon. Or perhaps some idle and frivolous magician (there must be magicians in London) had cast a spell over me through his parlour window as I explored the maze of streets east and west in solitary leisurely walks without chart and compass. Till I began to write that novel I had written nothing but letters, and not very many of these.
-1.44	Info	In the second place, the Emperor is an exceedingly intelligent and highly cultivated man. His mental processes are swift, but they go also very deep. He is a searching inquirer, and questions and listens more than he talks. His fund of knowledge is immense and sometimes astonishing. He manifests interest in everything, even to the smallest detail, which can have any bearing upon human improvement. I remember a half hour's conversation with him once over a cupping glass, which he had gotten from an excavation in the Roman ruin called the Saalburg, near Homburg. He always appeared to me most deeply concerned with the arts of peace. I have never heard him speak much of war, and then always with abhorrence, nor much of military matters, but improved agriculture, invention, and manufacture, and especially commerce and education in all their ramifications, were the chief subjects of his thought and conversation. I have had the privilege of association with many highly intelligent and profoundly learned men, but I have never acquired as much knowledge, in the same time, from any man whom I have ever met, as from the German Emperor.
-1.21	Literary	Moreover Grandmother Grant always dressed in one fashion; she had a calico dress for morning and a black silk for the afternoon, made with an old-fashioned surplice waist, with a thick plaited ruff about her throat; she sometimes tied a large white apron on, but only when she went into the kitchen; and she wore a pocket as big as three of yours, Matilda, tied on underneath and reached through a slit in her gown. Therein she kept her keys, her smelling-bottle, her pocket-book, her handkerchief and her spectacles, a bit of flagroot and some liquorice stick. I mean when I say this, that all these things belonged in her pocket, and she meant to keep them there; but it was one peculiarity of the dear old lady, that she always lost her necessary conveniences, and lost them every day.\n"Maria!" she would call out to her daughter in the next room, "have you seen my spectacles?"\n"No, mother; when did you have them?"\n"Five minutes ago, darning Harryś stockings; but never mind, thereś another pair in the basket."
-0.37	Literary	The others were watching him closely. They guessed something of the nature of what must be passing through Ned's mind, for both Jack and Teddy followed his gaze up the uneven shore. Jimmy had the glasses again, and was busily engaged in scrutinizing the blur on the distant horizon, which all of them had agreed must be smoke hovering close to the water. Perhaps he half-believed the fanciful suggestion made by Teddy, with reference to Captain Kidd, and was wildly hoping to discover some positive sign that would stamp this fairy story with truth. All the previous adventures that had befallen himself and chums would sink into utter insignificance, could they go back home and show evidences of having made such a romantic discovery up there in the Hudson Bay country.\n"See the feather they say he always wore in his hat, Jimmy?" asked Frank.
0.06	Literary	The other day, as I was walking through a side street in one of our large cities, I heard these words ringing out from a room so crowded with people that I could but just see the auctioneer's face and uplifted hammer above the heads of the crowd.\n"Going! Going! Going! Gone!" and down came the hammer with a sharp rap.\nI do not know how or why it was, but the words struck me with a new force and significance. I had heard them hundreds of times before, with only a sense of amusement. This time they sounded solemn.\n"Going! Going! Gone!"\n"That is the way it is with life," I said to myself - "with time." This world is a sort of auction room; we do not know that we are buyers: we are, in fact, more like beggars; we have brought no money to exchange for precious minutes, hours, days, or years; they are given to us. There is no calling out of terms, no noisy auctioneer, no hammer; but nevertheless, the time is "going! going! gone!"

Table 5: Randomly selected examples from the CLEAR dataset. In contrast to other datasets, each example is associated with a continuous readability score obtained by fitting a Bradley–Terry model on pairwise human judgments of reading ease. (continued)

Readability Score	**Category**	**Text**
0.19	Info	There are various kinds of pitcher-plants. Some are shorter and broader than others; but they are all green like true leaves, and hold water as securely as a jug or glass. They grow in Borneo and Sumatra, hot islands in the East. The one shown in the drawing grows in Ceylon.\nSome grow in America; but they are altogether different from those in Borneo and Ceylon. One beautiful little pitcher-plant grows in Australia: but this is also very different from all the rest; for the pitchers, instead of being at the end of the leaves, are clustered round the bottom of the plant, close to the ground.\nAll these pitcher-plants, though very beautiful to look at, are very cruel enemies to insects: for the pitchers nearly always have water in them; and flies and small insects are constantly falling into them, and getting drowned.

Table 6: Comparison of readability scores between the original CLEAR (Original) and the 1k subsample used to conduct the correlation analysis (Subsample).

Statistic	**Original**	**Subsample**
Count	4724	1000
Mean	-0.96	-0.97
Std	1.03	1.06
Min	-3.68	-3.68
25%	-1.70	-1.74
50%	-0.91	-0.89
75%	-0.20	-0.20
Max	1.71	1.71

Table 7: Randomly selected examples from the SCIENTIFIC PAPERS dataset, spanning all three readability classes.

Complexity Level	Text
Low	The researchers found that women who lived in countries that received less US foreign aid during the policy used less contraceptives and had both more pregnancies and more abortions during the years that the policy was in place. They also noted that the effects of the policy reversed once it had been rescinded, further strengthening the researchers' hypothesis that the Mexico City Policy has an effect on a nation's observed patterns of reproductive behavior.
Medium	The researchers found that abortions and pregnancies increased when the Mexico City Policy was in effect, which they correlate to a decreased availability in contraception during those years. They also found that the effects varied by exposure to the policy, as women in high exposure countries were more likely to experience abortion when the policy was enacted and less likely when it wasn't in effect. The alternating patterns of reproductive behavior depending on whether the policy was enacted also strengthens the researchers' hypothesis that it has a not insubstantial effect on abortion rates in sub-Saharan Africa.
High	When US support for international family planning organizations was conditioned on the policy, coverage of modern contraception fell and the proportion of women reporting pregnancy and abortions increased, in relative terms, among women in countries more reliant on US funding. Although the degree to which abortions increase when contraceptive supply is curtailed is poorly characterized, one analysis estimated that, depending on the total fertility of the population, a 10% decline in contraceptive use would lead to a 20-90% increase in abortions. The researchers posit that the observed changes in abortion could be due to changing availability of modern contraception, and that a change in the use of modern\ncontraception would be expected to result in a change in pregnancy rates. Women in high-exposure countries experienced a relative increase in abortion (and decrease in modern contraceptive use) when the policy was enacted and a relative decrease in abortion (and increase in modern contraceptive use) when the policy was rescinded.
Low	Study looks at pushup capacity and heart health, finding that those who could do the most (over 40) push ups had the lowest risk of heart disease.
Medium	Study examines the relationship between a person's push up ability and their physical health, finding that push ups are a good indicator of a person's cardiovascular fitness.
High	Association Between Pushup Exercise Capacity and Future Cardiovascular Events Among Active Adult Men

Table 8: Examples of different explanations for the ELI-WHY (HUMAN) for the questions "Why do we enjoy horror movies or stories?" and "'Why does DNA have a double helix structure?". Each set of three examples refers to the same question.

Readability	Topic	Formatted Explanation
Elementary	Psychology	All the same reasons people like sad songs, Halloween, war documentaries, apocalyptic fiction, etc. etc. It's like any other film genre. Horror movies can be artistic; the performances can be entertaining; the movies can be well-constructed or conceived; they can be relatable or provide personal insight.
High School	Psychology	According to these researchers, stimulation is one of the driving forces behind the consumption of horror. Exposure to terrifying acts like stories of demonic possession or alien infestation can be stimulating both mentally and physically. These experiences can give rise to both negative feelings, such as fear or anxiety, and positive feelings, such as excitement or joy. And we tend to feel the most positive emotions when something makes us feel the most negative ones.
Graduate	Psychology	"The horror film occupies in popular culture roughly comparable to that of horror literature. That is to say, it is generally ignored, sometimes acknowledged with bemused tolerance, and viewed with alarm when it irritates authority - rather like a child too spirited to follow the rules that rendition has deemed acceptable" (p. ix), a view that is echoed elsewhere. For example, Tudor (1997) noted that "a taste of horror is a taste for something seemingly abnormal and is therefore deemed to require special attention" (p. 446). Part of the reason for the disdain, apart from the broad and base nature of the content, may be the relative cheapness of horror film: these are often much less expensive to create than are other genre films such as westerns, comedies, or science fiction.
Elementary	Biology	DNA is made up of small components called nucleotides. A nucleotide is made up of 3 parts: a phosphate group, a sugar, and a base. The base can be 1 of 4 varieties: A, T, C, or G. Simply put, because of their structures, A and T bond nicely with one another and C and G bond nicely as well. As a result, a single strand of DNA will bond to a strand with another strand with a "complimentary" sequence of bases. In other words, there will be two strands with opposite, for lack of a better word, sequences of bases. The production of strands that complement one another is a result of how DNA copies itself (known as DNA replication). The second strand's base sequence is ordered based on the first strandś sequence.\n\nSo that answers the question of why DNA is a double helix instead of a single helix. But why is it a helix at all? Why not a shaped like a ladder? In a cell, most of the material present is water. Water is shaped in such a way where positive and negative charges aren't spread evenly throughout the waterś molecules. This is known as being a polar molecule. Polar molecules "like" being around other polar molecules. Non-polar molecules don't "like" being around polar molecules like water. In DNA, the bases are non-polar, but the phosphate groups are polar. As a result, the preferred shape puts the phosphates in contact with the water and the bases are covered by being on the inside. The twisting shape of DNA reduces the extent to which the bases are exposed to the water in the outside environment.\n\nTL;DR: DNA is made of 3 components: bases, sugars, and phosphates. The sugars bond with one another explaining the double-configuration of DNA. To be stable, the bases need to stay in the inside of the molecule and phosphates need to make up the moleculeś exterior, and a helical shape allows this to happen.
High School	Biology	The three-dimensional structure of DNA—the double helix—arises from the chemical and structural features of its two polynucleotide chains. Because these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones wind around each other to form a double helix, with one complete turn every ten base pairs.
Graduate	Biology	The stability of double-stranded DNA (dsDNA) at physiological temperature is due to the hydrogen bonding between complementary bases and the stacking between neighboring bases. However, these base-stacking interactions are of the order of magnitude of a few k_B T thermal energy and the thermal fluctuations can lead (even at physiological temperature) to local and transient unzipping of the double helix.

Metric	SCIENTIFIC PAPERS (August et al., 2024)	CLEAR (Crossley et al., 2024)	ELI-WHY (GPT-4) (Joshi et al., 2025)	ELI-WHY (HUMAN) (Joshi et al., 2025)	SCIENCEQA (Lu et al., 2022)	Avg. Rank
README++	**0.40**	-0.45	**0.50**	0.50	**0.44**	1.8
README++ (AVG)	0.23	-0.49	0.26	**0.68**	0.38	2.4
README++ (MAX)	0.35	**-0.51**	0.43	0.57	0.42	1.8

Table 9: Rank correlations between variants of the README++ metric and human judgments of correctness across 5 datasets. We boldface the variant exhibiting strongest correlation with human judgments. We report the Kendall Tau coefficient. All correlation coefficients are statistically significant with p-value < 0.01.

Table 10: Hyperparameter configurations of the Logistic Regression models fit for each readability class. We use a grid search to find the optimal combination over the hyperparameters C, PENALTY, and SOLVER. The best configuration is defined as the best achieving accuracy determined using 10-fold cross-validation.

Readability Class	Hyperparameters	Majority Accuracy (%)	Best Accuracy (%)
Elementary	C = 100 max_iter = 300 penalty = l1 solver = saga	74.64	88.05
High School	C = 500 max_iter = 100 penalty = l1 solver = saga	59.72	75.11
Graduate	C = 100 max_iter = 300 penalty = l1 solver = saga	85.08	88.34

Table 11: Human rationales underlying readability judgments across 3 different readability classes: Elementary, High School, Graduate. Each row refers to the analysis of the same "Why" question but different GPT-4 explanation, being sourced from ELI-WHY (GPT-4) (Joshi et al., 2025).

Elementary	**High School**	**Graduate**
- It's probably too verbose for elementary levels, but I think people reading at that level could understand this explanation. The words are short enough. - The explanation uses basic English language to interpret why humans are inclined towards social interactions. There are not many technical or professional terminologies, making it easy to understand. The sentence structures are simple, making it easy to follow.	- Pretty easy and straight forward to understand. Not using complex words or scientific words. - The sentences are short in length and easy to digest. It uses terms like "elements" and "conductivity and ductility" which require deeper understanding of elements and reactions. - The explanation is written in a way that is easy to understand, but the details and some of the words used such as "corrosion" would make it difficult for an elementary reader to comprehend. However, the material is not so specialized that you would learn it on the graduate level, meaning this falls into the high school reading level.	- The terminology seems higher level and more complicated than elementary or high school; - This is borderline HS/GS to me. But the terms "parasocial" and "existential fears" are a bit much for a typical high school student. It should be simplified a bit for an HS student.
- The details are very surface level and it uses simple wording. - Simple sentence structure with simple and short explanations. Not detailed or in depth. - They used simple wording and examples to make their point. - It uses simple words like electricity, and can be easily understood - It gives clear examples like copper being easy to stretch and not rusting, the sentences are short and straightforward. It gives enough detail to understand why copper is used in wires.	- The wording/terminology, examples, and details suggest high school-level engagement. It lacks the technicality of graduate school while being too advanced for elementary school; - Using terminology like "ritual", "theological" and "philosophical" which requires basic knowledge of these terms. Depth and detail are also moderate levels but not quite a graduate level understanding; - Wording Terminology, Sentence Structure, Details and depth	- No way most high school students could follow this; - The details and depth show of a graduate school person answering this.
- Simple wording, a concept that most students of elementary school age should be able to grasp. Also not too many details. - The explanation uses simple and direct language without complex terminology, making it accessible to children or adults with basic education. - I think this text's wording, examples, sentence structure, and amount of detail are simple enough for an elementary-age student to comprehend.	- This response includes references to Alzheimer's, which I think would be outside the understanding of a typical 4th grader. It also references brain waves, which I think is covered in high school-level science courses. - It uses more elevated vocabulary than Elementary School, however the lack of citations and more complex concepts and narrative structure make it less than Graduate School.	- The language is more advanced and mentions more specific scientific theories. - The amount of detail and specific terminology make me think it is a graduate level.