Can Llama 3 Accurately Assess Readability? A Comparative Study Using Lead Sections from Wikipedia

José Frederico Rodrigues[1], Henrique Lopes Cardoso[2], and Carla Teixeira Lopes[1]

[1] INESC TEC, Faculdade de Engenharia, Universidade do Porto
[2] LIACC, Faculdade de Engenharia, Universidade do Porto

Abstract. Text readability is vital for effective communication and learning, especially for those with lower information literacy. This research aims to assess Llama 3's ability to grade readability and compare its alignment with established metrics. For that purpose, we create a new dataset of article lead sections from English and Simple English Wikipedia, covering nine categories. The model is prompted to rate the readability of the texts on a grade-level scale, and an in-depth analysis of the results is conducted. While Llama 3 correlates strongly with most metrics, it may underestimate text grade levels.

Keywords: Readability Assessment · Large Language Models · Llama

1 Introduction

Text clarity is crucial for effective communication, understanding, and learning, particularly for those with lower information literacy. Readability affects how well readers engage with written content, whether in academic, medical, or everyday contexts. Gunning [9] stresses the importance of evaluating readability to ensure students are provided with materials at an appropriate difficulty level, while Manning [12] highlights writing strategies in healthcare to create clear, accessible messages. Complex terminology in fields like law and engineering poses similar challenges. Moreover, readability is core to user experience, especially as generative models are increasingly integrated into systems. Accurately assessing it is important to ensure systems are accessible to readers of varying abilities [16].

Conventional readability metrics generate scores based on elements like sentence length or word syllables but overlook factors such as content relevance or semantics, as shown in Table 1. Despite these limitations, they remain a simple way to estimate text readability. Large language models, however, are emerging as powerful tools in natural language processing, with the potential to accurately assess readability and overcome the conventional metrics' shortcomings.

Llama 3[3], announced on April 18, 2024, is a free model that can be run locally, making it ideal for this investigation due to the high volume of requests

[3] https://llama.meta.com/llama3/

Rodrigues, J.F., Lopes Cardoso, H., Teixeira Lopes, C. (2025). Can Llama 3 Accurately Assess Readability? A Comparative Study Using Lead Sections from Wikipedia. In: Grabis, J., Vos, T.E.J., Escalona, M.J., Pastor, O. (eds) Research Challenges in Information Science. RCIS 2025. Lecture Notes in Business Information Processing, vol 548. Springer, Cham. https://doi.org/10.1007/978-3-031-92471-2_6

Table 1. Traditional readability metrics

Metric	Features considered
FK [10]	Words per sentence, syllables per word
GF [4]	Words per sentence, complex words ($\geq$ 3 syllables)
SMOG [11]	Number of polysyllables per sentence
ARI [17]	Characters per word, words per sentence
DC [7]	Percentage of difficult words based on a list. Words per sentence
CL [5]	Characters per word, sentences per 100 words
LW [6]	Easy ($\leq$ 2 syllables) and difficult ($\geq$ 3 syllables) words per sentence

FK = Flesch-Kincaid Grade Level; GF = Gunning Fog Index; ARI = Automated Readability Index; DC = Dale-Chall; CL = Coleman–Liau; LW = Linsear Write

involved. We explore Llama 3's performance across multiple domains, comparing it to existing readability metrics. A new dataset is created using lead sections from English Wikipedia (EW) and Simple English Wikipedia (SEW), covering nine categories. We prompt the model to rate the readability on a grade-level scale, and we analyze its correlation with Table 1's readability metrics, which estimate the years of education required to understand a text.

2 Related Work

In studies by Naous et al. [13], Blaneck et al. [3], and Golan et al. [8] LLMs are directly applied for readability assessment. Naous et al. employed both supervised and unsupervised approaches with BERT, mBERT, and XLM-RoBERTa for English and multilingual readability tasks, fine-tuning them on the README++ dataset [13] annotated using Common European Framework of Reference for Languages (CEFR) standards. In other languages, language-specific models like AraBERT(Arabic) and RuBERT(Russian) were applied, and few-shot prompting was explored with GPT-4 and Llama 2. Blaneck et al. investigated German language readability using GBERT and GPT-2-Wechsel in ensemble approaches to enhance performance, while Golan et al. tested ChatGPT's ability to apply traditional readability formulas without relying on annotated datasets. Performance evaluation methods rely on metrics like Pearson Correlation and Root Mean Squared Error (RMSE), which were used to assess the accuracy of LLM predictions against human-annotated readability levels.

3 Dataset Creation and Experimental Setup

Our dataset [14], which is also suitable for the evaluation of text simplification tasks [15], includes lead section pairs from both EW and SEW, covering nine categories. Below, we present an example: the first excerpt is a lead section from English Wikipedia, while the second excerpt is its simplified counterpart from Simple English Wikipedia.

```
Tuition payments, usually known as tuition in American English and as tuition fees in
Commonwealth English, are fees charged by education institutions for instruction or
other services. Besides public spending (by governments and other public bodies),
private spending via tuition payments are the largest revenue sources for education
institutions in some countries. In most developed countries, especially countries in
Scandinavia and Continental Europe, there are no or only nominal tuition fees for
all forms of education, including university and other higher education.
```

```
Tuition payments, usually known as tuition in American English and as tuition fees in
Commonwealth English, are fees charged for students looking for a higher education.
Tuition payments are charged by colleges and universities include costs for lab
equipment, computer systems, libraries, facility upkeep and to provide a comfortable
student learning experience.
```

Despite the existence of several datasets suitable for readability assessment tasks, such as the README++ dataset [13], Newsela [19], and the PLABA dataset [1], none span multiple domains while maintaining consistency for their size and text sources. So, creating a new dataset was deemed necessary to ensure consistency, drawing all texts from the same source across different domains. EW and SEW were selected as the source of the texts for multiple reasons: EW articles are typically written for a general audience but tend to contain complex language, while SEW specifically aims to be more accessible, resulting in a wider range in readability levels across both encyclopedias. Wikipedia covers many topics, allowing the dataset to include many articles from various categories. Lastly, both EW and SEW are freely accessible, making it easy to source many lead sections without licensing issues.

To decide which categories the text samples would be extracted from, we leveraged SEW's category tree and determined the number of article pages of each sub-category directly under the "Everyday Life" and "Knowledge" categories. After analyzing the number of articles per category, we included categories with more than 100,000 articles. These categories were: "Culture", "Education", "Employment", "Entertainment", "Health", "Leisure", "Objects", "Science" and "Time". We traverse a given category and its subcategories to collect page titles. The page title acts as the article's unique identifier across EW and SEW. 10,000 lead section pairs were collected for each of the 9 selected categories, and there are no duplicate titles for each category. Overall, the dataset contains 133,240 unique lead sections and is publicly available in a research data repository[4].

The 8B parameter, instruction fine-tuned Llama 3 model, was chosen because of its smaller size and ability to run on consumer hardware. Inference is run on a local NVIDIA RTX 3090 GPU, using the Transformers library [18]. The following system prompt was defined to provide the model a general guideline of its task: "Your role is to rate the readability of texts that are provided to you.". To facilitate processing its responses, the model's temperature was set to a low value, 0.01, to make its replies follow the same format as much as possible. Lowering the temperature minimizes response variability, but due to the nature of readability assessment, this presented itself as an adequate alternative to requesting a specific format through prompt engineering, which proved ineffective,

[4] https://rdm.inesctec.pt/dataset/cs-2024-008

as the model's grading often mismatched its justification for its rating, compromising the validity of the results. The text to assess was provided with the prompt: "Consider the following text: {text}" followed by two new lines and the instruction: "Based on your own assessment, rate its readability on a grade-level scale."

4 Results

The readability assessment task was framed as a classification problem where the readability of texts was categorized into discrete grade levels. With this approach, we can directly map the readability of a text to an educational grade level, as the model was prompted to evaluate the readability of the lead sections by rating them on this scale. In this section, we present the findings of this investigation, organized into four subsections. We pre-processed the model's responses and the scores given by 7 traditional readability metrics, shown in Table 1 and calculated using textstat[5], so as to establish correlations and facilitate comparisons between grade levels. The traditional readability metric scores are floored, meaning a score of 6.7, for example, will correspond to grade level 6.

4.1 Grade Level Distributions

In its response, for both EW and SEW, the model attributes either a grade level to a text or a range spanning up to 3 levels, such as "10^{th}-grade to 12^{th}-grade". Llama provided readability ratings predominantly as ranges rather than single values. Specifically, 80.9% of Llama's responses were in the form of grade level intervals, while the remaining 19.1% were single values. Ranges of values output by Llama never spanned more than three grade levels, and the model's lowest and highest ratings attributed to a lead section were 2^{nd} to 3^{rd} grade and 12^{th} to 14^{th}, respectively. Overall, 53% of all ratings output by Llama were in the interval format of 9^{th} to 10^{th} grade. Higher readability ratings, indicating more complex text, were more common in the Science and Education categories. In contrast, the Leisure category rarely received higher ratings, suggesting that the lead sections in this category are deemed to be written at a relatively lower grade level. On the other hand, lower readability ratings were distributed across all categories more evenly, pointing toward a balanced presence of simpler texts.

While Llama tends to cluster its ratings within narrower intervals, traditional metrics seem to capture more variations in text complexity. The distribution of these scores post-processing is displayed for each metric in Figure 1. DC's scores, however, escape the trend by tightly clustering around grades 9 and 10, similar to the model.

4.2 Deviation Analysis

Llama's ratings in the form of ranges required conversion to single values for meaningful comparison with the processed traditional readability scores. For

[5] https://textstat.org/

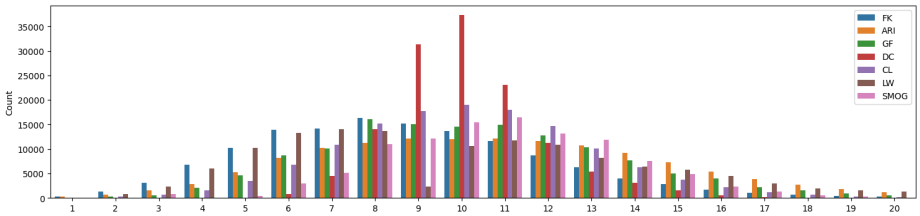

Fig. 1. Overall distribution of readability metric values (1 to 20).

each LLM rating provided as an interval, we selected the value within the interval that was closest to the grade level given by the metric we were comparing to.

The deviation between a metric's score and the model's rating refers to the difference between the rating provided by the Llama3 model and the readability score given by the traditional metric. To gain insights into how the model's assessment criteria align with established readability measures, we analyzed these deviations. In general, deviations between -3 and 3, displayed in Table 2, account for over 83% of all ratings. Dale-Chall stands out with 93.6% of its scores within this interval. The Dale-Chall metric shows the most significant alignment with the model, followed by Coleman-Liau. Overall, except for the Flesch-Kincaid, with 47% of positive deviations, results indicate that deviations are mostly negative, suggesting that the model is rating lead sections as simpler than the readability metrics convey.

Table 2. Percentage of deviations less than 0, greater than 0, and between -3 and 3.

LLM-Metric	FK	GF	SMOG	ARI	DC	CL	LW
= 0	27.1	27.3	29.4	21.0	**44.0**	32.8	15.3
<0	25.9	**46.6**	**55.9**	**54.1**	37.2	**42.1**	**44.6**
>0	**47.0**	26.1	14.7	24.8	18.8	25.0	40.0
Between -3 and 3	83.2	83.3	88.7	73.2	93.6	88.7	70.5

Bold highlights the highest deviation value/tendency (=0; <0; >0) for each metric.

Due to the imbalanced nature of the readability ratings, we report the macro-averaged Mean Absolute Error [2] between the traditional readability metric scores and the model's ratings across all categories, as displayed in Figure 2. This metric averages the MAE computed for each rating, giving them an equal weight. Notably, The Dale-Chall metric consistently shows one of the lowest errors across most categories, which aligns with the earlier observation of its narrower range and higher alignment with the model's ratings. The SMOG metric also shows relatively low errors in most categories, all of them practically identical, except for the Science and Education categories, where the error is slightly higher at 5.3 and 4.7, respectively. The Objects category seems to be where the error is lowest across most metrics, whereas the category where the error is highest varies.

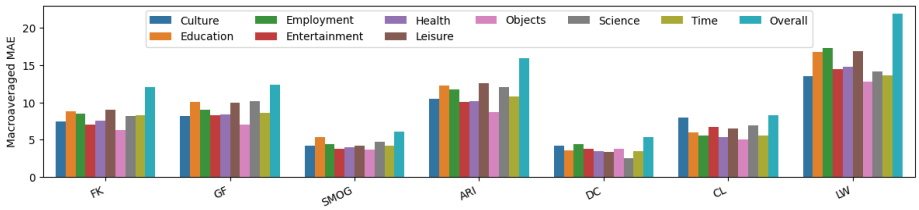

Fig. 2. Macro-averaged Mean Absolute Error between readability metrics and LLM ratings for all categories.

4.3 Correlation Analysis

To complement the analysis, Spearman's Rank Correlation, displayed in the left part of Table 3, is also reported. Overall, the model's predicted readability ratings have a strong positive correlation with all metrics except for DC, which shows a moderate positive association, suggesting a high degree of agreement between Llama and traditional metrics. Out of every metric, Llama's ratings have the strongest positive association with FK, peaking in the Education category. In contrast, the model's correlation is weakest with the DC metric in every category, exhibiting a moderate positive association. DC, however, achieved the lowest macro-averaged MAE among all the metrics. Furthermore, DC is the metric with the highest number of ties with the model's ratings, resulting in a much larger number of tied rank situations when calculating Spearman's rank correlation coefficient, which could impact the measure's accuracy.

Table 3. Left: Spearman's Rank Correlation between LLM ratings and readability metrics for all categories. **Right**: Percentage of cases where readability ratings for SEW were equal or lower than ratings for EW.

	Spearman's Correlation							Percentage of cases							
Category	**FK**	**GF**	**SM**	**AR**	**DC**	**CL**	**LW**	**LLM**	**FK**	**GF**	**SM**	**AR**	**DC**	**CL**	**LW**
Culture	0.81	0.76	0.76	0.77	0.51	0.71	0.78	96.8	87.7	87.0	93.0	86.9	74.3	83.7	86.1
Education	**0.82**	0.78	0.77	0.79	0.60	0.74	0.78	96.6	91.4	89.8	94.2	90.9	85.3	88.7	88.9
Employment	0.79	0.79	0.77	0.79	0.47	0.76	0.79	94.1	91.6	90.7	93.3	91.2	80.9	88.2	**90.4**
Entertainment	0.76	0.72	0.74	0.73	0.40	0.66	0.75	95.1	81.1	79.4	91.0	78.4	59.0	72.7	79.9
Health	0.79	0.78	0.76	0.78	0.61	0.75	0.77	96.0	**92.3**	**90.9**	94.0	**91.9**	84.7	**90.1**	89.9
Leisure	0.73	0.71	0.68	0.72	0.37	0.65	0.72	95.3	87.2	85.8	92.1	86.4	73.7	82.0	85.3
Objects	0.73	0.76	0.78	0.76	0.55	0.69	0.73	94.7	82.6	83.4	93.4	84.0	75.1	86.3	79.7
Science	0.81	**0.80**	**0.81**	**0.82**	**0.65**	**0.77**	0.78	**97.7**	92.1	90.8	**94.8**	90.9	**87.3**	89.5	89.2
Time	0.80	**0.80**	0.79	0.81	0.48	0.74	**0.80**	96.2	87.9	88.7	94.4	88.7	75.0	87.4	86.8
Overall	0.79	0.78	0.78	0.78	0.51	0.74	0.76	95.7	88.5	87.8	93.4	88.2	78.3	86.2	86.5

SM= SMOG; AR = ARI. Bold highlights the highest value per column.

Very strong positive correlations are observed across most metrics for the Science and Time categories, and for each metric except Linsear-Write, the strongest positive correlation is observed in the Science category. In contrast, Leisure is the category where each metric displays its weakest correlation, closely followed by the Entertainment category.

4.4 English Wikipedia vs Simple English Wikipedia

We compare readability assessments between standard and simplified Wikipedia lead sections since these simplified versions are human-generated and could include aspects that aren't considered by traditional metrics. We determine the percentage of cases where their ratings for SEW sections were equal to or lower than their ratings for EW. Results are shown in the right part of Table 3. Llama's ratings show the highest percentages across most categories, suggesting it rarely deems SEW lead sections as more complex. These results begin to showcase how an LLM could be a better choice over traditional metrics to assess readability. Metrics such as FK or SMOG rely on surface-level features like word length, sentence length, or syllable count, while a large language model can leverage a contextual understanding, which should align more closely with human judgments. Employing an LLM incurs a greater cost than computing these metrics, but it could be justified in situations where determining if a text's content is easier to understand is given more importance than determining if it is easier to read. Most metrics also display high percentages across all domains, validating the model's assessment. There is, however, a considerable difference in the DC percentages between the Entertainment and Science categories.

5 Conclusions

To study Llama 3's performance in the task of readability assessment, we create a new dataset spanning multiple categories. We explore the distribution and characteristics of the model's output and compare it with scores from traditional readability metrics. Llama tends to grade texts with a level interval instead of a single grade level. It correlates most strongly with FK and weakest with DC. It tends to grade texts as more readable than traditional metrics, except for FK. Lastly, Llama rarely determines SEW sections as more complex than their EW counterpart, surpassing traditional readability metrics when it comes to distinguishing simple from complex texts, indicating that it does not only rely on surface-level features. Overall, results suggest that Llama 3 can accurately assess readability while overcoming weaknesses inherent to traditional metrics.

References

1. Attal, K., Ondov, B., Demner-Fushman, D.: A dataset for plain language adaptation of biomedical abstracts. Scientific Data **10**, 8 (1 2023). https://doi.org/10.1038/s41597-022-01920-3
2. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. pp. 283–287 (01 2009). https://doi.org/10.1109/ISDA.2009.230
3. Blaneck, P.G., Bornheim, T., Grieger, N., Bialonski, S.: Automatic readability assessment of german sentences with transformer ensembles. arXiv preprint arXiv:2209.04299 (2022)
4. Brucker, C.: The gunning's fog index (or fog) readability formula. https://readabilityformulas.com/the-gunnings-fog-index-or-fog-readability-formula/, accessed: May 31, 2024

5. Coleman, M., Liau, L.: A computer readability formula designed for machine scoring (1975)
6. CSUN: Readability helps the level (11 2006), http://www.csun.edu/ vcecn006/read1.html, accessed: May 31, 2024
7. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational Research Bulletin **27**(2), 37–54 (1948), http://www.jstor.org/stable/1473669
8. Golan, R., Ripps, S.J., Reddy, R., Loloi, J., Bernstein, A.P., Connelly, Z.M., Golan, N.S., Ramasamy, R.: Chatgpt's ability to assess quality and readability of online medical information: Evidence from a cross-sectional study. Cureus (7 2023). https://doi.org/10.7759/cureus.42214
9. Gunning, T.G.: The role of readability in todays classrooms. Topics in Language Disorders **23**, 175–189 (7 2003). https://doi.org/10.1097/00011363-200307000-00005
10. Kincaid, J.P.P., Jr, R.F., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, http://library.ucf.edu
11. Laughlin, G.H.M.: Smog grading-a new readability formula. Journal of Reading **12**(8), 639–646 (1969)
12. Manning, D.T.: Writing readable health messages. Public health reports **96 5**, 464–5 (1981), https://api.semanticscholar.org/CorpusID:39039337
13. Naous, T., Ryan, M.J., Lavrouk, A., Chandra, M., Xu, W.: Readme++: Benchmarking multilingual language models for multi-domain readability assessment (5 2023)
14. Rodrigues, J.F., Teixeira Lopes, C., Lopes Cardoso, H.: Wikipedia and simple wikipedia lead section pairs for nine categories (2024), [Data set]. INESC TEC.
15. Rodrigues, J.F., Teixeira Lopes, C., Lopes Cardoso, H.: Evaluating llama 3 for text simplification: A study on wikipedia lead sections. In: Companion Proceedings of the ACM Web Conference 2024. WWW '25, Association for Computing Machinery (2025)
16. Roegiest, A., Pinkosova, Z.: Generative information systems are great if you can read. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. p. 165–177. CHIIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3627508.3638345, https://doi.org/10.1145/3627508.3638345
17. Smith, E.A., Senter, R.J.: Automated readability index. AMRL-TR. Aerospace Medical Research Laboratories (U.S.) pp. 1–14 (5 1967)
18. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
19. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics **3**, 283–297 (2015). https://doi.org/10.1162/tacl_a_00139, https://aclanthology.org/Q15-1021