Zero-shot Large Language Models for Automatic Readability Assessment

Riley Grossman
New Jersey Institute of Technology
Newark, NJ
rag24@njit.edu

Yi Chen
New Jersey Institute of Technology
Newark, NJ
yi.chen@njit.edu

Abstract

Unsupervised automatic readability assessment (ARA) methods have important practical and research applications (e.g., ensuring medical or educational materials are suitable for their target audiences). In this paper, we propose a new zero-shot prompting methodology for ARA and present the first comprehensive evaluation of using large language models (LLMs) as an unsupervised ARA method by testing 10 diverse open-source LLMs (e.g., different sizes and developers) on 14 diverse datasets (e.g., different text lengths and languages). Our findings show that our proposed prompting methodology outperforms prior methods on 13 of the 14 datasets. Furthermore, we propose LAURAE, which combines LLM and readability formula scores to improve robustness by capturing both contextual and shallow (e.g., sentence length) features of readability. Our evaluation demonstrates that LAURAE robustly outperforms prior methods across languages, text lengths, and amounts of technical language.

1 Introduction

Measuring the readability of text has many practical applications. For example, it is important when selecting or creating resources for educating school children (Karaca, 2024) and second language learners (Xia et al., 2016). It is also critical in ensuring that patients who are diagnosed with a new disease or preparing to undergo treatment can understand, process, and act upon health-related information (Bhatt et al., 2024; Lin et al., 2024). Healthcare institutions such as the American Medical Association and the National Institutes of Health provide recommendations for the readability level of patient materials to best improve patient outcomes (Rooney et al., 2021).

Readability measurement has long been an important component of academic research as well. For example, prior work has used readability measures to ensure that resources are understandable for target audiences (Lang et al., 2025), to validate text simplification or summarization methods (Picton et al., 2025; Wu et al., 2025), and to examine document readability as a predictor of future outcomes (e.g., financial report readability and subsequent stock performance) (Zhang et al., 2025).

The time and costs required to accurately label the readability level of a large text corpus inspired the development of automatic readability assessment (ARA) tools. The earliest ARA tools were formulas that measured readability as a linear combination of shallow text features, such as average sentence length and syllables per word (Flesch, 1948). The development of machine learning (ML), and specifically attention-based transformer models, led to the development of supervised ARA tools that combined text embeddings and linguistic features (Xia et al., 2016; Lee et al., 2021). While supervised methods substantially improved accuracy over traditional formulas, they require annotated corpora, technical expertise, and additional computational resources, which has limited their adoption. Most recently, researchers have proposed unsupervised language model-based ARA tools to balance the accuracy of deep learning methodologies with the usability of readability formulas (Trott and Rivière, 2024; Martinc et al., 2021).

However, we find that current research still heavily relies on readability formulas. We search the *Scopus* database for peer-reviewed papers published in 2025 with "readability" and at least one of the words "formula", "NLP", "LLM", "ML", "AI", or "BERT" in the titles, keywords, or abstract. We manually identify the papers that apply ARA tools (excluding proposals for new methods). We find two papers using supervised BERT models, two papers with unsupervised measures based on pretrained language model's (PLM) surprisal, and five papers with zero-shot prompted LLMs. In contrast, we find 337 papers using readability formulas. Readability formulas are applied across

a diverse set of medical (Uysal, 2025), linguistics (Bao et al., 2025), business (Chabot, 2025), AI (Hossen Rujeedawa et al., 2025), and governance (Raman et al., 2025) research papers, and in many reputable conferences and journals, including the *AAAI Conference on Artificial Intelligence* (Wu et al., 2025), *Scientific Reports* (Uysal, 2025), *PLoS ONE* (Ozdemir Kacer, 2025), *Finance Research Letters* (Pathak, 2025), and *Journal of Medical Internet Research* (Lang et al., 2025).

Although some delay may be expected, we believe that there are legitimate reasons why unsupervised language model-based ARA methods have not yet replaced readability formulas. Trott and Rivière (2024) first showed that zero-shot prompting of GPT-4 models outperformed traditional readability formulas on one English dataset. While this limited evaluation shows the potential of the method, it may also inhibit adoption as researchers do not know if the performance generalizes to technical texts (e.g., medical resources), non-English languages, or when using free open source models.

In this paper, we make four contributions. First, we propose two methodological advances for using LLMs for unsupervised ARA: 1) we compute readability scores as an expected value over the output token probabilities, and 2) we prompt the model on the same scale used by manual annotators and include detailed definitions of each readability level when available. Experimental results show that prompting LLMs with our two proposed techniques leads to improved performance in comparison to Trott and Rivière (2024) on 13 of 14 datasets, with particularly strong gains on non-English datasets.

Second, we propose a novel readability assessment method called **L**LM-based **A**utomatic **U**nsupervised **R**eadability **A**ssessment **E**nsemble (LAURAE), which combines zero-shot prompted LLM scores and readability formula scores, based on weights derived from the LLM's stated confidence in its rating. By considering both high-level contextual understanding from LLMs and shallow features (e.g., sentence length and number of polysyllabic words) from readability formulas to produce a holistic readability assessment, LAURAE outperforms readability formulas on all 14 datasets and standalone LLMs on 11 datasets.

Third, we present a more comprehensive evaluation of LLM-based methodologies for unsupervised ARA. Specifically, our evaluation is performed on 10 open-source LLMs with varying sizes, developers, and multilingual capabilities, and 14 datasets that vary in language, text length, and ground truth type. The evaluation shows that LAURAE is generalizable and an effective unsupervised ARA tool regardless of text length, language, and amount of technical content. Our finding that using the zero-shot methodology proposed in Trott and Rivière (2024) for open-source LLMs only outperforms the best readability formula on 6 of 14 datasets, highlights the importance of our more thorough evaluation.

Fourth, all of the code and datasets, including two Greek textbook datasets we curated, are publicly available[1], providing resources to practitioners who need readability assessment tools and researchers for future work.

In conclusion, our evaluation supports the adoption of LAURAE for unsupervised ARA in practice or research. In particular, LAURAE is useful whenever high-quality readability assessment for a corpus of texts is desired and manual annotation is not plausible or too costly. Moreover, our findings show that combining contextual features from LLMs with shallow features from readability formulas enhances robustness in unsupervised ARA. Future research should explore whether combining zero-shot LLM ratings with shallow unsupervised NLP tools can further increase performance and robustness in other tasks such as sentiment analysis or toxicity detection.

2 Related Work

In this section, we review automatic readability assessment (ARA) methods. Early work in this area developed traditional readability formulas that rely on shallow text characteristics such as average sentence or word length, the number of technical words, and the number of polysyllabic words. While many of these formulas were created decades ago (Flesch, 1948; Smith and Kincaid, 1970), new research continues in this area for languages that were previously unsupported (Laurs, 2024; Zhu et al., 2024). Traditional readability formulas are widely used since they are easy to implement and require no training data. However, as demonstrated in Section 6, their effectiveness is limited.

The next phase of ARA extracted linguistic features (e.g., number of noun phrases) from a text, as inputs to supervised machine learning (ML) models (Xia et al., 2016; Chatzipanagiotidis et al.,

[1] https://github.com/rag24/LAURAE

2021; Vajjala and Meurers, 2012). The availability of attention-based pretrained language models (PLMs) led to new methods that finetuned BERT-based models for ARA (Martinc et al., 2021), including those which augmented BERT representations with extracted linguistic features (Li et al., 2022; Imperial, 2021; Deutsch et al., 2020; Hou et al., 2022; Lee et al., 2021). Although these supervised approaches improved performance, they are often not applicable due to the costs of obtaining a large manually labeled training dataset.

Towards solving this issue, Lee and Vajjala (2022a) proposed a neural pairwise ranking model, which can be applied to unseen datasets through transfer learning when the target and training datasets are very similar. Martinc et al. (2021) proposed the ranked sentence readability score (RSRS), an unsupervised method that utilizes PLMs to quantify sentence-level readability by calculating the unexpectedness of each word. However, as shown in Section 6, RSRS does not consistently outperform readability formulas.

The most related work to ours is Trott and Rivière (2024), which proposes zero-shot prompting GPT-4 Turbo and GPT-4o for ARA. However, the evaluation is limited to a single prompting technique and English dataset. Another recent work zero-shot prompts closed source GPT models and the ChatGLM2-6B and Meta-Llama-3-8B models to assess readability on English and Chinese sentences. However, the lack of experimentation with prompts and larger open-source models results in poor performance (Liu et al., 2025).

In this work, we first address the limitations in prior works by proposing a new methodology for zero-shot prompting LLMs for unsupervised ARA, and evaluating this methodology when applied to 10 open-source LLMs and 14 diverse benchmark datasets. Furthermore, we propose a new method, LAURAE, that combines zero-shot prompting of LLMs with readability formula scores, and demonstrates robust and generalizable performance.

3 Methodology

We first propose a new method for obtaining unsupervised ARA scores from large language models (LLMs) in Section 3.1. Then, in Section 3.2, we introduce our proposed method, LAURAE, which ensembles the LLM readability scores from Section 3.1 with readability formulas scores, using weights based on the LLM's verbalized confidence in its rating.

3.1 Prompting Zero-shot LLM

We propose two changes to the methodology for prompting LLMs in an unsupervised ARA task. First, we investigate whether or not including a definition of the readability scale in the prompt (when available) improves LLMs' zero-shot performance. Second, we calculate expected readability scores as an expected value over the models' output token probabilities, following prior research on using LLMs as raters (Liu et al., 2023; Lai et al., 2025). This is in contrast with the original proposal and current state-of-the-art (Trott and Rivière, 2024), which does not experiment with different prompts and considers only the number in the generated output text when producing a readability score.

3.1.1 Readability Scale Definitions

Human annotators either explicitly, or implicitly (e.g., school textbook datasets), rated the readability of the texts in each benchmark dataset. These ratings are used as the ground-truth readability scores, and for 7 of the 14 datasets, human raters relied on the Common European Framework of Reference for Languages (CEFR) to produce ratings. CEFR provides six levels of language proficiency, defined by the capabilities of learners at each level. For the datasets where manual raters used the CEFR scale and definitions, we prompt the LLM to generate scores on the same scale and include the definitions in the prompt. The six levels of CEFR proficiency (A1-C2) are converted to integers (1-6).

For the remaining datasets, we use a similar prompting approach to the existing literature (Trott and Rivière, 2024). We prompt the model for a readability score on an arbitrary scale (i.e., whole number value between 1 and 9) based on several key factors (e.g., grammar and clarity) as well as the models' own definition of readability. For comparison, we also test this prompting approach for the CEFR datasets. We display the actual prompts for each dataset in Appendix A.

3.1.2 Expected Value of Output Tokens

Instead of only considering the generated output token as in Trott and Rivière (2024), we calculate readability scores as an expected value over the models' output token probabilities.

Formally, for the i-th generated token, the LLM produces a logit vector $z_i \in \mathbb{R}^{|V|}$, where V is the

LLM vocabulary. Applying the softmax function to z_i yields probability distribution p_i, such that each element p_{ij} corresponds to the probability of generating the j-th vocabulary token V_j, as the i-th token in the sequence. Let n be the position in the generated text where the readability score is located. We create a new array r, with two rows and $|V|$ columns such that r_{1k} equals the k-th highest probability in p_n, and r_{2k} is the corresponding token index j, so that $p_{nj} = r_{1k}$.

In a zero-temperature setting, the model will generate the highest ranked-token, $V_{r_{21}}$. This is referred to as the vanilla score (Lai et al., 2025). To calculate the expected value score, we first check whether $V_{r_{21}}$ is numeric and proceed to the next highest-ranked token until $V_{r_{2k}}$ is no longer numeric. The expected value score, s_{LLM}, is then computed as:

$$s_{LLM} = \sum_{m=1}^{k-1} r_{1m}(V_{r_{2m}}) \quad (1)$$

3.2 LAURAE

We propose a new unsupervised automatic readability assessment (ARA) method that ensembles ratings from shallow unsupervised ARA techniques and zero-short prompting of instruction-tuned LLMs. In many cases, especially with longer or more technical texts that are heavily context-dependent, we expect zero-shot LLMs to outperform shallow unsupervised ARA methods. However, LLMs may struggle to rate the readability of very short texts that lack contextual information, or types of text that they are not well-suited for, such as children's stories (Bhandari and Brennan, 2023; Valentini et al., 2023). In these cases, unsupervised ARA methods such as readability formulas that focus on shallow features (e.g., sentence length or the number of polysyllabic words), may perform well. Thus, combining the scores of a shallow ARA method with the scores of a zero-shot LLM may improve overall performance by increasing robustness and generalizability.

Since we are operating in an unsupervised setting, we can not use hyperparameter tuning to select the optimal weights for the two readability scores in the ensemble. Instead, we propose using the LLM's self-reported confidence to determine the weights. Prior work has shown that, compared to entropic uncertainty measures, LLMs produce more reliable and accurate confidence scores when prompted in natural language (Tian et al., 2023). Thus, we prompt each model to state its confidence, on a 1-9 scale, that the generated readability score will align with human raters' scores (full prompts in Appendix A). We again apply the expected value scoring technique from Section 3.1.2 to the LLM's confidence score, and divide the score by 10, to obtain our confidence weight c. The 1-9 scale ensures that each ARA method receives at least 0.1 weight in the final score. Given the LLM's readability score, s_{LLM}, and the readability formula's score, s_{rf}, we compute LAURAE's readability score as

$$c(\frac{s_{LLM} - \mu_{LLM}}{\sigma_{LLM}}) + (1 - c)(\frac{s_{rf} - \mu_{rf}}{\sigma_{rf}}), \quad (2)$$

where μ_{LLM}, σ_{LLM}, μ_{rf}, and σ_{rf} are the dataset-level means and standard deviations of the respective readability scores.

Dataset	Language	N	Avg. Length	Ground Truth
ReadMe	English	296	22	CEFR Rating
	French	185	25	
	Hindi	163	23	
	Arabic	206	24	
	Russian	178	23	
MedReadMe	English	1140	25	
Cambridge	English	300	579	
CLEAR	English	1890	201	non-CEFR Rating
Greek Language	Greek	393	161	
Greek History	Greek	804	209	
OneStop	English	567	782	
Asset	English	485	21	Comparison
Vikidia	English	150	596	
	French	150	509	

Table 1: Datasets Used in Evaluation

4 Experimental Setup

We start by introducing the experimental setup used to evaluate the zero-shot prompting ARA capabilities of LLMs (Section 5) as well as our proposed ensemble method, LAURAE (Section 6). Prior work on unsupervised ARA has yet to evaluate LLMs' ARA capabilities on diverse texts (e.g., differing text lengths and languages). Furthermore, the capabilities of leading open-source LLMs have not been tested, which may limit adoption. Our experimental setup aims to more comprehensively investigate whether LLMs can replace traditional readability formulas as the dominant unsupervised ARA method.

Datasets We utilize the publicly available ReadMe (Naous et al., 2024), MedReadMe (Jiang

and Xu, 2024), Cambridge (Xia et al., 2016), CommonLit Ease of Readability (CLEAR) (Crossley et al., 2023), OneStop (Vajjala and Lučić, 2018), Asset (Alva-Manchego et al., 2020), and Vikidia (Lee and Vajjala, 2022b) datasets. The 14 selected datasets and their key characteristics are shown in Table 1. Due to the limited availability of datasets with non-English texts longer than a sentence, we curated two additional datasets from publicly available Greek textbooks (see Appendix B). The datasets vary across five key dimensions: 1) there are six different languages considered; 2) there are seven datasets with sentence-length texts (i.e., $\leq$25 words on average) and seven datasets with paragraph- or article-length texts; 3) there are seven datasets with readability scores and definitions based on the Common European Framework of Reference for Languages (CEFR) scale, 4) the MedReadMe dataset contains healthcare texts, which allows for evaluation of ARA on texts with technical language, and 5) there are eleven datasets with ground truth readability scores for each text and three datasets where the ground truth indicates which of two comparable texts is more readable.

Evaluation Metrics. For the 11 datasets that provide ground truth ratings, we report the Pearson correlation between each method's readability scores and the ground truth ratings, following the existing literature (Naous et al., 2024; Jiang and Xu, 2024; Martinc et al., 2021; Trott and Rivière, 2024). For the three datasets that provide ground truth comparisons, we report how accurately each method can identify the more readable of the two texts. To test for statistically significant differences, we use Steiger's modification of Williams's test (Steiger, 1980) to compare correlations and McNemar's test (McNemar, 1947) to compare accuracies.

Open Source LLMs. We evaluate the ARA capabilities of a diverse set of open-source, instruction-tuned LLMs: *Llama 3.1 8B/70B* and *Llama 3.2 3B* (AI@Meta, 2024); the *Aya Expanse* series (Dang et al., 2024); the *Gemma 2* series (Gemma Team, 2024); *Mixtral 8x7B* (Jiang et al., 2024) and *Phi-4* (Abdin et al., 2024). Models vary from being monolingual (e.g., Phi-4 and Gemma 2 series) to specifically designed for multilingual capabilities (e.g., Aya Expanse series). Model size also varies from 2 billion (i.e., Gemma 2B) to 70 billion (i.e., Llama 70B) parameters.

4.1 Baseline Methods

Readability Formulas. For English texts, we evaluate *FKGL* (Kincaid et al., 1975) and *ARI* (Smith and Kincaid, 1970) due to their popularity and use in related works (Naous et al., 2024; Jiang and Xu, 2024). We also evaluate *OSMAN* (El-Haj and Rayson, 2016) for Arabic text, *Lix* (Anderson, 1983) for Hindi and Greek texts, and the adapted versions of *Flesch Reading Ease (FRE)* for French and Russian texts using the *textstat*[2] Python package.

RSRS. Martinc et al. (2021) proposed the Ranked Sentence Readability Score (RSRS) as a pretrained language model-based (PLM) unsupervised ARA method that uses a PLM's quantification of word unexpectedness and traditional shallow readability indicators such as sentence length (details in Appendix C). We tested several variations (see Appendix D) and present results for RSRS implemented with a multilingual BERT model (*mBERT*) (Devlin et al., 2019).

LLM-v-ns. A methodology for zero-shot prompting LLMs for ARA that uses vanilla readability scores (instead of expected value readability scores discussed in Section 3.1.2) and only prompts LLMs for readability scores on an arbitrary scale. This is most similar to the method proposed in Trott and Rivière (2024).

5 Zero-shot Performance of Open Source LLMs

In this section, we investigate the performance of three groups of methods: (1) the shallow unsupervised ARA methods (i.e., RSRS and readability formulas), (2) the zero-shot ARA performance of 10 popular open-source LLMs with the LLM-v-ns prompting methodology, and (3) the performance of the same 10 LLMs with our proposed zero-shot prompting methodology that includes the CEFR scale in prompts (for the seven relevant datasets) and computes readability scores as the expected value over the model's output token probabilities. The performance of these methods on 14 diverse readability datasets is shown in Figure 1.

5.1 Comparison among existing LLMs-based methods

All included LLMs have English language capabilities, and thus, we can compare all of the models' performances on the English datasets (i.e., the first

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

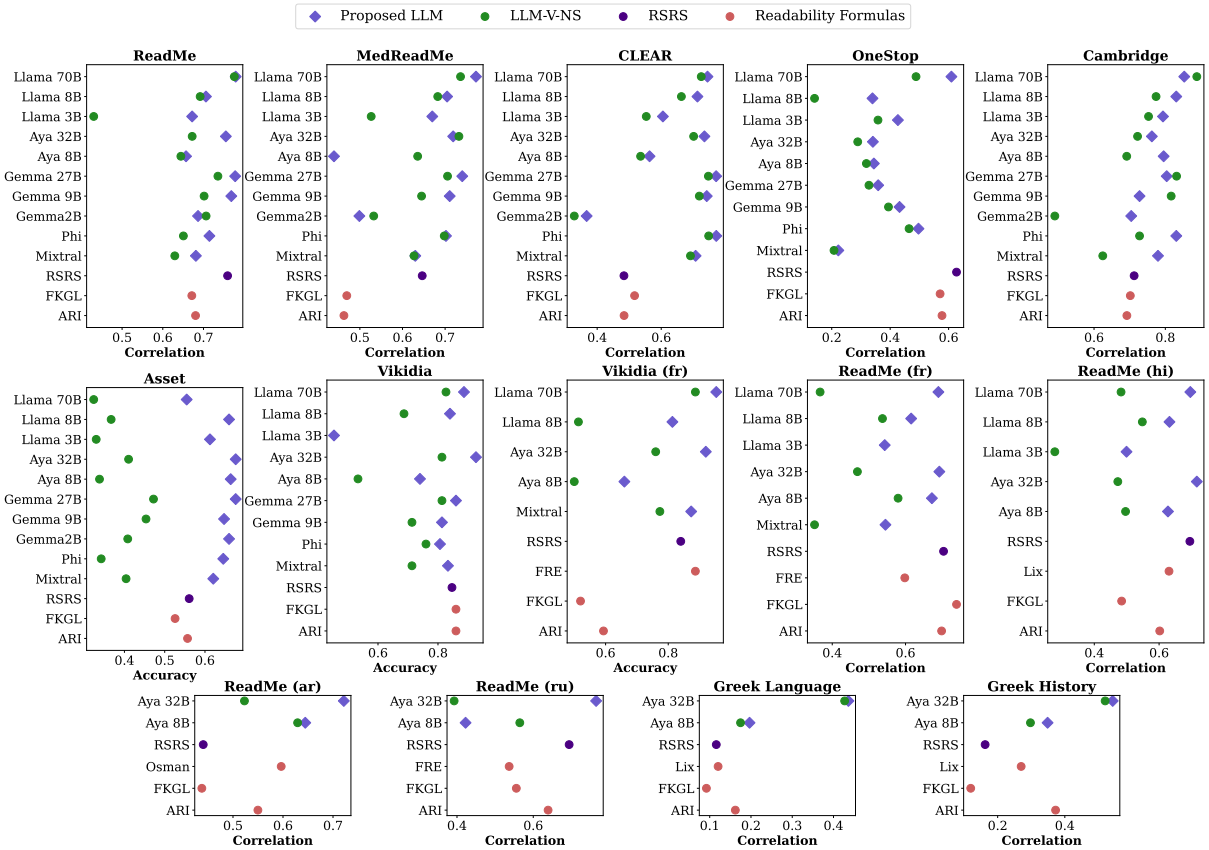

Figure 1: Proposed Zero-shot LLM Method versus Unsupervised ARA Baselines on Benchmark Datasets (Note: observations performing worse than 0.5 points behind the best method on each dataset are removed)

seven datasets in Figure 1). The Llama 70B model, the largest LLM tested, is the top-ranked LLM four times, second-ranked once, and third-ranked once. Phi, Gemma 27B, and Aya 32B each perform as the top-ranked LLM on one English dataset. We conclude that the Llama 70B model has the best overall performance of the tested LLMs on English datasets.

The only dataset that Llama 70B performs poorly on is Asset, which was uniquely created by human raters selecting the simpler text between two short sentences output by text simplification tools. Llama 70B rates the two sentences as equally readable 13.61% of the time, which does not count as a correct prediction. No other LLM does this. Upon inspection, we found that many of the sentence pairs differed by just one or two words, which suggests that rating the two sentences as equally readable is reasonable.

For the seven non-English datasets, we only compare the performance of LLMs that were trained on texts in the target language. For the two French datasets and one Hindi dataset, we compare the Llama and Aya Expanse models. Surprisingly, Aya 32B outperforms Llama 70B on two of the three datasets, despite being smaller in size, highlighting the benefits of a focused approach to developing multilingual capabilities (Dang et al., 2024). The four datasets shown in the last row of Figure 1 have languages that are only supported by the Aya Expanse series, and Aya 32B always outperforms Aya 8B. In summary, of the models tested, Aya 32B has the best performance on non-English datasets.

We use Llama 70B for English datasets and Aya 32B for non-English datasets in the rest of the paper to replicate a truly unsupervised setting where the LLM can not be chosen by testing performance on a validation dataset.

5.2 Evaluation of Our Proposed LLM-based Method

Now we present the evaluation of our proposed improvements to ARA with zero-shot prompted LLMs. Across 14 datasets, our proposed method is the best performer on 11 datasets. The exceptions are the OneStop dataset where RSRS outperforms our proposed method by 0.018 points in correlation, the Cambridge dataset where the LLM-v-ns outper-

forms our method by 0.035 points in correlation, and the French ReadMe dataset where the FKGL and ARI formulas, as well as RSRS, outperform our method by 0.006-0.047 points in correlation. Even in a fully unsupervised setting (i.e., Llama 70B for English, and Aya 32B for non-English, texts), our method still performs the best on 10 datasets.

We are especially interested in comparing our proposed zero-shot prompting methodology to LLM-v-ns. Regardless of LLM, our method generally outperforms the baseline LLM-v-ns. Even on the Cambridge dataset, where the best performance is achieved by Llama 70B with the LLM-v-ns method, our proposed method performs better for 7 of the 10 evaluated LLMs. Focusing only on the recommended LLMs from Section 5.1, our method outperforms LLM-v-ns on 13 datasets (with statistically significant differences for all but the English ReadMe and Greek Language datasets).

5.2.1 Ablation Studies

Recall that our proposed method has two improved components: 1) prompting the LLM for scores using the same scale as manual labelers, and 2) conducting expected value scoring. In Table 2, we isolate the effect of these two components for the recommended LLM (see Section 5.1) on each dataset. The "Expected Value" column shows the performance differences that resulted from expected value scoring, as opposed to vanilla. The use of expected value scores increases model performance on all 14 datasets, with 12 being significant increases. Improvements are particularly large for the datasets with ground truth comparisons instead of ratings (i.e., Vikidia and Asset). This is likely due to the expected value scoring technique reducing ties when comparable texts receive the same vanilla readability score.

The "Scale Included" column shows the isolated performance differences due to prompting the model to rate readability on the CEFR scale. This evaluation only applies to the seven datasets with ground truth values on the CEFR scale. Inclusion of the CEFR scale increases performance significantly for five of the seven datasets, and leads to particularly large increases for the non-English datasets. This suggests that LLMs may have less stored information about readability in non-English languages, and therefore benefit more from a well-defined readability scale.

Dataset	Expected Value	Scale Included
Greek Lang.	0.009	-
Greek History	0.022^{**}	-
Vikidia (fr)	0.160^{***}	-
Vikidia	0.060^{**}	-
Asset	0.231^{***}	-
CLEAR	0.020^{***}	-
OneStop	0.121^{***}	-
MedReadMe	0.014^{***}	0.026^{***}
Cambridge	0.032^{***}	-0.059^{***}
ReadMe	0.018^{**}	-0.026^{*}
ReadMe (fr)	0.016	0.214^{***}
ReadMe (hi)	0.058^{***}	0.204^{***}
ReadMe (ar)	0.029^{**}	0.177^{***}
ReadMe (ru)	0.030^{**}	0.339^{***}
Average	**+ 0.059**	**+ 0.125**

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Performance Differences when Including CEFR Scale in Prompt and Using Expected Value Scoring by Dataset

6 LAURAE

In this section we evaluate the performance of our proposed ensemble method LAURAE. Section 6.1 shows that by combining LLM readability scores with readability formula scores, LAURAE improves the performance and generalizability of the LLM-based approach presented in Section 5. We perform two additional analyses to verify LAURAE's effectiveness of using the LLM's verbal confidence scores as ensemble weights in Sections 6.2 and 6.3. Finally, we investigate whether LAURAE can benefit from dataset-level ensemble weights in Section 6.4.

6.1 LAURAE Evaluation

As shown in Table 3, LAURAE, using readability formula scores, outperformed each of the three baselines. Although slightly worse in terms of overall performance, LAURAE still outperforms the baseline methods when RSRS replaces readability formulas as the shallow feature readability score (see Appendix E). Consistent with Section 5.1, we use Llama 70B for English datasets and Aya 32B for non-English datasets. For simplicity, we select the best readability formula on each dataset.

The main takeaway from Table 3 is that LAURAE outperforms all baselines on 13 of 14 datasets. Only the LLM-v-ns baseline method outperforms LAURAE on the Cambridge dataset. This may be attributed to its better performance than our proposed zero-shot prompting methodology on this dataset (see Section 5.2). Notably, LAURAE re-

Dataset	LAURAE	LLM-v-ns	Formula	RSRS
Greek Lang.	**0.43**	0.427	0.162^{***}	0.116^{***}
Greek Hist.	**0.572**	0.52^{***}	0.373^{***}	0.163^{***}
Vikidia (fr)	**0.953**	0.76^{***}	0.887^{*}	0.84^{***}
Vikidia	**0.9**	0.827^{***}	0.86	0.847^{*}
Asset	**0.629**	0.324^{***}	0.557^{***}	0.561^{**}
CLEAR	**0.735**	0.725^{*}	0.517^{***}	0.484^{***}
OneStop	**0.654**	0.488^{***}	0.577^{**}	0.627
MedReadMe	**0.77**	0.736^{***}	0.469^{***}	0.646^{***}
Cambridge	0.86	$\mathbf{0.888}^{*}$	0.702^{***}	0.713^{***}
ReadMe	**0.798**	0.776	0.68^{***}	0.759
ReadMe (fr)	**0.75**	0.469^{***}	0.739	0.704
ReadMe (hi)	**0.754**	0.473^{***}	0.631^{**}	0.695
ReadMe (ar)	**0.757**	0.523^{***}	0.596^{***}	0.441^{***}
ReadMe (ru)	**0.803**	0.393^{***}	0.639^{***}	0.694^{***}
Average	***0.74***	*0.595*	*0.599*	*0.592*

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Performance Evaluation of proposed LAURAE (Note: best result bolded and significance testing indicates difference from LAURAE result)

duces the performance gap between our proposed prompting methodology and LLM-v-ns on that dataset (see Figure 2). LAURAE's improved performance is significantly different from the performance of RSRS, LLM-v-ns, and readability formulas on 10, 11, and 12 of the datasets, respectively.

A second takeaway is that across the 14 datasets, all three baselines have a similar average performance. This is a surprising finding because prior work had shown that the LLM-v-ns methodology applied to GPT-4 outperformed traditional readability formulas (Trott and Rivière, 2024). However, they only studied the CLEAR dataset, for which our findings are consistent (Table 3). This new insight shows the importance of comprehensive evaluation, as we did with 14 diverse datasets.

6.2 Verbal Confidence Score Calibration

Now we study the effectiveness of using the LLM's verbal confidence scores as ensemble weights in LAURAE. Prior research has shown that LLMs are often overconfident in their predictions (Tian et al., 2023; Zhou et al., 2023). To verify that verbal confidence scores are a useful proxy for LLM accuracy, we show that the correlation between LLM and ground-truth readability scores is stronger among ratings associated with higher LLM confidence scores in each dataset. For each of the 11 datasets with ground truth readability ratings (see Table 1), we compute the verbal confidence score associated with the 25th and 75th percentiles. We then separate out the texts corresponding to verbal confidence scores in the top and bottom quartiles of verbal confidence scores respectively. In all but one dataset (Greek History), the highest-confidence quartile showed substantially stronger correlations with the ground truth than the lowest-confidence quartile. On average, the difference between the top and bottom quartiles was 16.7 correlation points. The Greek History dataset had low performance in both quartiles (correlations of 0.1096 and 0.4108, respectively), indicating that the issue is with the overall model performance on this dataset, and not the verbal confidence scores.

We only evaluate the usefulness of verbal confidence scores in the datasets with a ground truth rating because for those datasets with ground truth comparisons (i.e., Text 1 is simpler than Text 2), our method rates each text independently. In other words, the confidence scores only indicate how confident the LLM is in rating the readability of a single text, not in the comparison between two texts.

6.3 LAURAE Ablation Studies

We compare the performance between LAURAE and the proposed standalone zero-shot LLM method in Figure 2. We also compare with three other ensemble variations to show the effectiveness of verbal confidence scores as weights:

1) LAURAE-naive, a method that combines the standardized LLM and readability formula scores with equal weights;

2) LAURAE-entropy, a method that derives an ensemble weight based on the Shannon entropy of the LLM's output distribution for the readability score token[3]; and

3) LAURAE-minmax, a method that applies minmax normalization at the dataset-level to ensure confidence scores are distributed across the entire interval $[0, 1]$ based on relative confidence.

Our findings show that combining the contextual LLM and shallow readability formula scores brings further improvements in comparison with our proposed zero-shot prompting methodology. Furthermore, we show that, on average, verbal confidence scores are the best technique for determining ensemble weights.

[3]Specifically, the LLM and readability formula weights in the ensemble are $1 - H$, and H, respectively, where H is entropy.

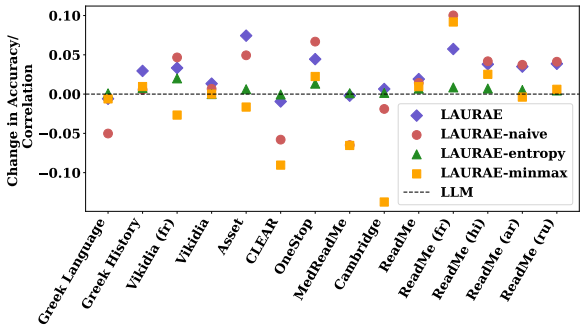

Figure 2: LAURAE Ablation Study Results

In Figure 2, the standalone LLM is considered as the baseline performance, and we plot differences in performance for LAURAE and its variants. Only the LAURAE-minmax variant performs worse than a standalone LLM, by -0.013 points on average. LAURAE-entropy, LAURAE-naive, and LAURAE all outperform a standalone LLM, by 0.006, 0.015, 0.027 points on average, respectively.

In the datasets where a standalone LLM is the best performing method (i.e. Greek Language and CLEAR datasets), LAURAE and LAURAE-entropy display robustness by keeping the differences small (i.e., less than 0.01 points in terms of accuracy/correlation). In contrast, LAURAE-naive performs at least 0.05 points worse than a standalone LLM on three datasets.

In summary, empirical evaluation shows that the proposed LAURAE is the best performing variant due to its higher average performance, and demonstrated robustness.

6.4 Dataset-Level Weights

We test whether a variation on LAURAE that aggregates verbal confidence scores to the dataset-level to reduce noise can improve effectiveness.

It is clear from Table 3 that LLMs are better suited to evaluate readability on some datasets (e.g., Cambridge and ReadMe) than others (e.g., Asset and Greek textbook datasets). One possible way to determine how useful LLM readability scores are for a dataset is to utilize the average verbal confidence score from all texts in the dataset (distribution of verbal confidence scores within each dataset is reported in Appendix F). Given potential noises in individual text-level confidence scores, we test the effectiveness of a new LAURAE variant that replaces the individual text confidence score, c, in Eq. 2 with the average confidence score from all texts in the same dataset. We call this variant LAURAE-agg. On average, LAURAE-agg improves performance by 0.0007 points in terms of accuracy/correlation across the 14 datasets, as reported in Appendix F.

7 Conclusion

In comparison to prior work, we more thoroughly investigate whether zero-shot prompting LLMs is an effective method for unsupervised ARA, by using 14 diverse benchmark datasets. This thorough evaluation yields outcomes that differ from those reported in narrow-scope studies in the literature Trott and Rivière (2024). It also reveals the need for language models specifically designed to have multilingual capabilities.

We also made methodological contributions. We propose two changes to the prompting methodology: 1) prompt the LLM for readability scores on the same scale used by manual labelers, 2) and compute readability as an expected value over the output token probabilities. We further propose LAURAE, a novel method that combines readability scores from LLMs and traditional formulas using weights determined by the LLM's confidence. This combination harnesses the strengths of standalone LLMs (e.g., ability to rate readability with respect to a specific definition), while also avoiding some of their pitfalls (e.g., low performance on unseen writing styles) for increased robustness.

LAURAE consistently outperforms existing unsupervised ARA methods by substantial margins. Besides supporting widespread adoption of LAURAE, these findings suggest that future research should investigate if combining zero-shot LLMs with shallow feature methods can benefit other unsupervised NLP tasks.

Acknowledgments

Research reported in this publication was supported in part by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UM1TR004789, the NSF under Grant No. CNS2237328 as well as by the Martin Tuchman'62 Chair Endowment and the Leir Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. We thank the Computer Technology Institute and Press "Diophantus" for providing open access to Greek textbooks.

8 Limitations

There are two additional requirements for using our proposed method compared to using the traditional readability formulas: 1) Python literacy, and 2) access to computational resources.

First, while minimal coding is required for readability assessment using traditional formulas, as Python packages like Textstat generate scores with a single line of code, Python literacy is required to use our proposed method. This may seem like a barrier for some users (e.g. healthcare professionals, and educators). We will try to alleviate this issue by providing detailed instructions along with all the code to aid users. We hope users will ultimately weigh the performance gains of our method against its requirements when selecting an unsupervised readability assessment method.

Second, computational resources are required for running LLMs. In this research, for any LLM with 14 billion or fewer parameters, we use one Nvidia A100 GPU for model inference. We use two Nvidia A100 GPUs for inference with Aya Expanse 32B, and three Nvidia A100 GPUs for inference with the Mixtral and Llama 70B models. All variations of RSRS are performed with a single Nvidia A100 GPU. One possibility that can be explored in future research is whether performance degrades significantly with quantized versions of the LLM (e.g. 8-bit floating point precision instead of 16-bit like we use). If performance is comparable, this would reduce the demand for computational resources.

One limitation of LAURAE in comparison to readability formulas is that LAURAE is not interpretable. The performance gap between the two methods is substantial enough that this is not a reason to use readability formulas; however, it should encourage future research to investigate the generated natural language explanations from LLMs prompted to perform readability assessment. As shown in Appendix A, we do prompt the model to explain its output after providing readability and confidence scores. In this current work, we did not evaluate the plausibility or faithfulness of the explanations, because the focus was on improving the performance of unsupervised ARA methods. However, future research should explore whether these explanations are useful. If the explanations are useful (or can be made useful through changes to the prompting methodology), it would allow for an interpretable version of LAURAE because both components would be interpretable.

Similarly, the evaluation is based on correlation with ground truth readability scores rather than accuracy of absolute readability scores (e.g., 5th grade-level or A2 CEFR level). Though we do not utilize them in this way, LLMs can simply choose a grade or CEFR level readability score. Future research should LLMs absolute readability scoring accuracy and whether ensemble approaches offer improvements in this area.

An additional minor limitation of our paper is that the evaluation is limited to the datasets we could obtain from past papers or curate ourselves. Thus, we are only able to test our paper against one medical dataset (i.e., MedReadMe). Similarly, we were unable to find other open-access textbook repositories to test whether the poor performance of our method on the Greek textbooks was a byproduct of the limited training resources available for Greek, or if it is a limitation of our method (e.g., RSRS struggles to perform well on textbooks targeted to children).

9 Ethical Considerations

We do not condone the use of our method for evaluating individuals' writing. Although the method outperforms traditional readability formulas, it does not have perfect accuracy and could even display biases against certain writing styles. Our method is intended for use in applications where a collection of texts need to be automatically rated for readability. For example, it can be used to compare medical resources for patients and determine which texts are more readable. It may also be used to test the effectiveness of text simplification methods.

Second, LLMs are associated with high energy costs. Typically papers report the energy costs and environmental harm of training LLMs. Our proposed method does not require additional training, but uses model inferences (Samsi et al., 2023), which still have some energy costs. We encourage users to take these costs into consideration when choosing an unsupervised ARA method.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli

Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics.

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. Camembert 2.0: A smarter french language model aged to perfection. *Preprint*, arXiv:2411.08868.

Tong Bao, Yi Zhao, Jin Mao, and Chengzhi Zhang. 2025. Examining linguistic shifts in academic writing before and after the launch of chatgpt: a study on preprint papers. *Scientometrics*, 130(7):3597 – 3627.

Prabin Bhandari and Hannah Brennan. 2023. Trustworthiness of children stories generated by large language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 352–361, Prague, Czechia. Association for Computational Linguistics.

Chaitanya Bhatt, Ethan Lin, Laura E. Ferreira-Legere, Cynthia A. Jackevicius, Dennis T. Ko, Douglas S. Lee, Kathryn Schade, Sharon Johnston, Todd J. Anderson, and Jacob A. Udell. 2024. Evaluating readability, understandability, and actionability of online printable patient education materials for cholesterol management: A systematic review. 13(8). Publisher: Wiley.

Miia Chabot. 2025. Textual and ai-based analysis of climate disclosures: Evidence from the european energy sector. *British Accounting Review*.

Savvas Chatzipanagiotidis, Maria Giagkou, and Detmar Meurers. 2021. Broad linguistic complexity analysis for Greek readability classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–58. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN — a novel Arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255, Portorož, Slovenia. European Language Resources Association (ELRA).

Rudolph Flesch. 1948. A new readability yardstick. 32(3).

Gemma Team. 2024. Gemma.

Spyridoula Georgatou. 2016. Approaching readability features in greek school books. Master's thesis, Eberhard Karls Universität Tübingen.

Muhammad Irfaan Hossen Rujeedawa, Sameerchand Pudaruth, and Vusumuzi Malele. 2025. Unmasking ai-generated texts using linguistic and stylistic features. *International Journal of Advanced Computer Science & Applications*, 16(3).

Shudi Hou, Simin Rao, Yu Xia, and Sujian Li. 2022. Promoting pre-trained LM with linguistic features on automatic readability assessment. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 430–436. Association for Computational Linguistics.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Chao Jiang and Wei Xu. 2024. MedReadMe: A systematic study for fine-grained sentence readability in medical domain. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17293–17319, Miami, Florida, USA. Association for Computational Linguistics.

Mehmet F. Karaca. 2024. Is artificial intelligence able to produce content appropriate for education level? a review on ChatGPT and gemini. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference*, AICCONF '24, pages 208–213. Association for Computing Machinery.

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

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Peng Lai, Jianjie Zheng, Sijie Cheng, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. 2025. Beyond the surface: Enhancing llm-as-a-judge alignment with human via internal representations. *arXiv preprint arXiv:2508.03550*.

Iain A. Lang, Angela King, Kate Boddy, Ken Stein, Lauren Asare, Jo Day, and Kristin Liabo. 2025. Jargon and readability in plain language summaries of health research: Cross-sectional observational study. 27.

Thomas Laurs. 2024. Towards a readability formula for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 170–175, Torino, Italia. ELRA and ICCL.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686. Association for Computational Linguistics.

Justin Lee and Sowmya Vajjala. 2022a. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813. Association for Computational Linguistics.

Justin Lee and Sowmya Vajjala. 2022b. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Wenbiao Li, Wang Ziyang, and Yunfang Wu. 2022. A unified neural network model for readability assessment with feature projection and length-balanced loss. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7446–7457. Association for Computational Linguistics.

Michael X. Lin, Gavin Li, David Cui, Priya M. Mathews, and Esen K. Akpek. 2024. Usability of patient education–oriented cataract surgery websites. 131(4):499–506.

Fengkai Liu, Tan Jin, and John S. Y. Lee. 2025. Automatic readability assessment for sentences: neural, hybrid and large language models. 59(3):2265–2296.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. 47(1):141–179. Place: Cambridge, MA, Publisher: MIT Press.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

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.

Emine Ozdemir Kacer. 2025. Evaluating ai-based breastfeeding chatbots: quality, readability, and reliability analysis. *PLoS ONE*, 20(3 March).

Jalaj Pathak. 2025. Impact of judgment readability on financial crimes. *Finance Research Letters*, 75.

Bryce Picton, Saman Andalib, Aidin Spina, Brandon Camp, Sean S. Solomon, Jason Liang, Patrick M. Chen, Jefferson W. Chen, Frank P. Hsu, and Michael Y. Oh. 2025. Assessing ai simplification of medical texts: Readability and content fidelity. *International Journal of Medical Informatics*, 195:105743.

Raghu Raman, Vinith Kumar Nair, Sofi Dinesh, and Ramana Acharyulu. 2025. Comparative analysis of chatgpt and bard in digital governance: Accuracy, adaptability, and readability insights. *Digital Government: Research and Practice*, 6(2).

Michael K Rooney, Gaia Santiago, Subha Perni, David P Horowitz, Anne R McCall, Andrew J Einstein, Reshma Jagsi, and Daniel W Golden. 2021. Readability of patient education materials from high-impact medical journals: A 20-year analysis. 8:2374373521998847.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. *Preprint*, arXiv:2310.03003.

Edgar A. Smith and J. Peter Kincaid. 1970. Derivation and validation of the automated readability index for use with technical materials. *Human Factors*, 12(5):457–564.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

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.

Emre Uysal. 2025. Evaluation of the readability level of the package inserts of topical antifungal drugs. *Scientific Reports*, 15(1).

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. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598, Singapore. Association for Computational Linguistics.

Jiaxing Wu, Lin Ning, Luyang Liu, Harrison Lee, Neo Wu, Chao Wang, Sushant Prakash, Shawn O'Banion, Bradley Green, and Jun Xie. 2025. Rlpf: Reinforcement learning from prediction feedback for user summarization with llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25488–25496.

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.

Xiaoxu Zhang, Bo Wang, and Guangze Liu. 2025. Readability of financial reports and stock price crash risk. *Finance Research Letters*, 86:108489.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Shuqin Zhu, Man Zhang, and Dongdong Guo. 2024. Automatic prediction of text readability for international chinese language education. In *Proceedings of the 2024 International Conference on Innovation in Artificial Intelligence*, ICIAI '24, pages 65–71. Association for Computing Machinery.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian. *Preprint*, arXiv:2309.10931.

A Prompts

We report the exact prompts used for each of the datasets in our study. There are two variations of the prompts. For the seven datasets (Cambridge, MedReadMe and five ReadMe datasets) where ground truth ratings are based on the CEFR scale, we prompt the model to output a score on the CEFR scale and provide definitions of each level. For the remaining seven datasets, we prompt the model to produce readability scores on an arbitrary 1-9 scale, considering a few specific factors of readability (e.g., sentence structure and grammar), and the model's own definitions of readability. For comparison, we do prompt the models to provide a readability score on the arbitrary scale for the seven CEFR datasets too.

In the early stages of this project, we revised the prompts several times to ensure that the models consistently output numbers on the requested scale (e.g., 1-9). Ultimately, we settled on requesting a specific format for the model output to ensure that we could easily extract readability scores from each model output. We also evaluated the open-source Falcon-7B-Instruct model in our initial experiments; however, it was excluded from the final experiments because we were unable to find a prompt that encouraged the model to reliably return numeric readability scores.

A.1 CEFR prompts

All of the five ReadMe datasets and the MedReadMe dataset have the same prompt because the ground-truth ratings are on a 6-point scale based on CEFR levels. The CEFR level definitions are from Naous et al. (2024). For all LLMs, the prompt is:

Rate the readability of the text between 1 (very easy) and 6 (very challenging) using the following scale: 1 = Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required. 2 = Can understand short, simple texts on familiar matters of a concrete type 3 = Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension. 4 = Can read with a large degree of independence, adapting style and speed of reading to different texts and purpose 5 = Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of specialty, provided he/she can reread difficult sections. 6 = Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. You may use both the provided scale and your own understanding of readability to determine the most appropriate score. Where on the scale of readability does this text rate: [INSERT TEXT]. Additionally, state how confident you are that your rating will align with human raters, with a whole number value between 1 and 9. Answer with this format: Answer: [SCORE] Confidence: [Confidence Score] Explanation: [EXPLANATION]

A.1.1 Cambridge

The Cambridge dataset is also based on the CEFR scale, but there are no texts at the easiest level (i.e., A1), so we adjust the prompt slightly:

Rate the readability of the text between 1 (easy) and 5 (very challenging) using the following scale: 1 = Can understand short, simple texts on familiar matters of a concrete type 2 = Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension. 3 = Can read with a large degree of independence, adapting style and speed of reading to different texts and purpose 4 = Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of specialty, provided he/she can reread difficult sections. 5 = Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. You may use both the provided scale and your own understanding of readability to determine the most appropriate score. Where on the scale of readability does this text rate: [INSERT TEXT]. Additionally, state how confident you are that your rating will align with human raters, with a whole number value between 1 and

9. Answer with this format: Answer: [SCORE] Confidence: [Confidence Score] Explanation: [EXPLANATION].

A.2 Arbitrary Readability Scale Prompts

All datasets are prompted to generate a readability score based on an arbitrary scale where 1 indicates the text is very easy to understand and 9 indicates the text is very difficult to understand. Several of the benchmark datasets have texts that were not explicitly rated by manual annotators, but were implicitly rated based on comparisons to another text (e.g., Asset or Vikidia datasets) or the text being written to a specific audience (e.g., Greek textbook datasets). For these datasets, we append an additional short description about where the texts are sourced from. This description is added to the beginning of the prompt. We include these descriptions after the base prompt.

A.2.1 Base Prompt

Rate the readability of the text with a whole number value between 1 (very easy to understand) and 9 (very difficult to understand). Consider factors such as sentence structure, vocabulary or grammar complexity, and overall clarity, as well as your own understanding of readability. Where on the scale of readability does this text rate: 's'. Additionally, state how confident you are that your rating will align with human raters, with a whole number value between 1 and 9. Answer with this format: Answer: [SCORE] Confidence: [Confidence Score] Explanation: [EXPLANATION]

A.2.2 Greek Language Textbooks

The Greek Language textbooks dataset is based on textbooks that are designed for Greek schoolchildren between second and sixth grade, so we adjust the prompt to include this context:

Rate the readability of excerpts from Greek language textbooks targeted for students between second and sixth grades.

A.2.3 Greek History Textbooks

The Greek History textbooks dataset is based on textbooks that are suitable for Greek schoolchildren between fourth and twelfth grade, so we adjust the prompt to include this context:

Rate the readability of excerpts from Greek History textbooks targeted for students between fourth and twelfth grades.

A.2.4 Vikidia Datasets

Both Vikidia datasets are sourced from Wikipedia articles that were manually rewritten to be suitable for children audiences. do not have any readability scale. Thus, we adjust the prompt to include this context:

These Wikipedia articles are either intended for adult audiences or manually rewritten for children audiences.

A.2.5 Asset Dataset

The Asset dataset was created by a manual evaluation of text simplification methodologies applied to short sentences. Thus, we adjust the prompt to include this context:

All of these sentences were rewritten to compare different text simplification methodologies.

B Greek Textbook Datasets

We aim to evaluate the proposed method on datasets that vary in text length, text content, and language. However, one type of dataset we could not obtain[4] was a non-English dataset with texts that were longer than a single sentence and a ground truth rating instead of a ground-truth comparison between two similar texts.

One paper (Georgatou, 2016) created a readability dataset from the open-access repository of Greek textbooks available at `http://ebooks.edu.gr/ebooks/`. Although we did not find this exact dataset, we were able to create a similar dataset from the repository.

We first collected three language textbook PDFs for each grade of 2nd, 4th, and 6th graders and the history textbook PDFs for 4th, 6th, 10th and 12th grade. To extract a meaningfully long passage, we only considered passages which were at least ten lines long. We manually removed any passages which were copyright information, author information, or lists of reading materials. We were left with 393 passages from the language textbooks and 804 passages from the history textbooks. The ground truth readability score of each passage is the grade level of the textbook the excerpt is from.

C RSRS

Proposed by Martinc et al. (2021), the Ranked Sentence Readability Score (RSRS) is calculated at

[4] Although these datasets exist in prior research, we could not find a publicly available one that suited our needs and did not hear back from the authors we contacted.

Dataset	mBERT	XLM-R	Monolingual
Greek Lang.	0.116	0.112	**0.159**
Greek Hist.	0.163	0.138	**0.191**
Vikidia (fr)	0.84	0.833	**0.853**
Vikidia	0.847	**0.867**	0.847
Asset	0.561	**0.563**	0.532
CLEAR	0.484	0.428	**0.54**
OneStop	**0.627**	0.618	0.597
MedReadMe	0.646	0.567	**0.68**
Cambridge	**0.713**	0.656	0.65
ReadMe	0.759	0.746	**0.782**
ReadMe (fr)	0.704	**0.707**	0.688
ReadMe (hi)	**0.695**	0.619	0.605
ReadMe (ar)	0.441	**0.6**	0.466
ReadMe (ru)	**0.694**	0.621	0.522
Average	***0.592***	*0.577*	*0.579*

Table 4: Performance Evaluation of Pretrained Language Models for RSRS. Note: Best performance bolded.

the sentence level, as defined below. Note that the RSRS of a document is the average RSRS of all its sentences.

$$\text{RSRS} = \frac{\sum_{i=1}^{s} \sqrt{i} \times WNLL(i)}{s} \quad (3)$$

where s is the sentence length (i.e. the number of tokens in the sentence), and i is the i^{th} ranked token in terms of the smallest word-negative log likelihood (WNLL).

To calculate the WNLL for a token, we first mask the target token and pass the masked sentence to a pre-trained language model (e.g., BERT). The language model makes a prediction for the masked token. Specifically, the output is a vector, y_p, whose length equals the vocabulary size. The j^{th} value of y_p represents the model's predicted probability of the masked token being the j^{th} token in the vocabulary. The WNLL of a target token is then computed as:

$$\text{WNLL} = -(y_t \log y_p + (1 - y_t) \log (1 - y_p)) \quad (4)$$

where y_t is a vector whose length equals the vocabulary size, and corresponds to the ground truth value for the target token (i.e., it has a 1 in the index position for the ground truth token and 0s for all other positions). If the model expects the ground truth value with high probability, the WNLL will be small, otherwise (i.e. the token is unexpected), the WNLL will be large.

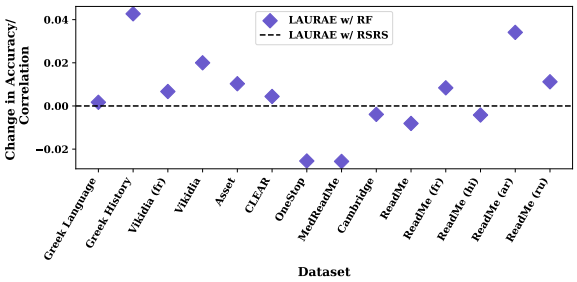

Figure 3: LAURAE Ablation Study for Shallow Unsupervised ARA Method

D Selecting PLM for RSRS

For the RSRS method Martinc et al. (2021), we follow test three variants of RSRS[5] We implement RSRS with two multilingual BERT-based models: *mBERT* (Devlin et al., 2019) and *XLM-R* (Conneau et al., 2020). We also choose one monolingual BERT-based model for each dataset depending on the language: *BERT* (Devlin et al., 2019) for English, *IndicBERTv2* (Doddapaneni et al., 2023) for Hindi, *ARBERTv2* (Abdul-Mageed et al., 2021) for Arabic, *ruBert* (Zmitrovich et al., 2023) for Russian, *CamemBERT(a)v2* (Antoun et al., 2024) for French, and *GreekBERT* (Koutsikakis et al., 2020) for Greek.

We present the results for RSRS implemented with mBERT, XLM-R, and a monolingual BERT-based model in Table 4. Although RSRS with the monolingual BERT-based models is the highest performing method on 6 of the 14 datasets, the RSRS variant with mBERT has the highest average performance. Our findings algin with Naous et al. (2024), who found that RSRS implemented with multilingual BERT models had better average performance than RSRS implemented with monolingual BERT-based models. Thus, we report RSRS with mBERT as the underlying PLM in the main results. However, even if we were to replace the results in Table 3 with another RSRS variant, it would not change our main findings. LAURAE would still be the highest performing method on 13 of 14 datasets.

E LAURAE with RSRS

In Figure 3, we plot the performance difference between using readability formulas as the shallow unsupervised method in LAURAE and using the RSRS as the shallow unsupervised method. It

[5]Our implementation of RSRS is based on code that was made available at: https://github.com/kinimod23/GRANT.

Dataset	LAURAE	LAURAE-agg	Mean	Std Dev
Greek Lang.	0.430	0.436	7.170	0.486
Greek Hist.	0.572	0.568	7.736	0.484
Vikidia (fr)	0.953	0.960	7.563	0.439
Vikidia	0.900	0.900	8.025	0.447
Asset	0.629	0.627	8.032	0.283
CLEAR	0.735	0.736	7.962	0.187
OneStop	0.654	0.654	7.934	0.076
MedReadMe	0.770	0.772	8.223	0.355
Cambridge	0.860	0.860	8.183	0.373
ReadMe	0.798	0.796	8.096	0.465
ReadMe (fr)	0.750	0.748	7.975	0.288
ReadMe (hi)	0.754	0.752	7.978	0.365
ReadMe (ar)	0.757	0.760	7.905	0.433
ReadMe (ru)	0.803	0.806	7.886	0.321
Total	*0.740*	*0.741*	*7.955*	*0.422*

Table 5: Verbal Confidence Score Distribution and Performance Evaluation of proposed LAURAE-agg

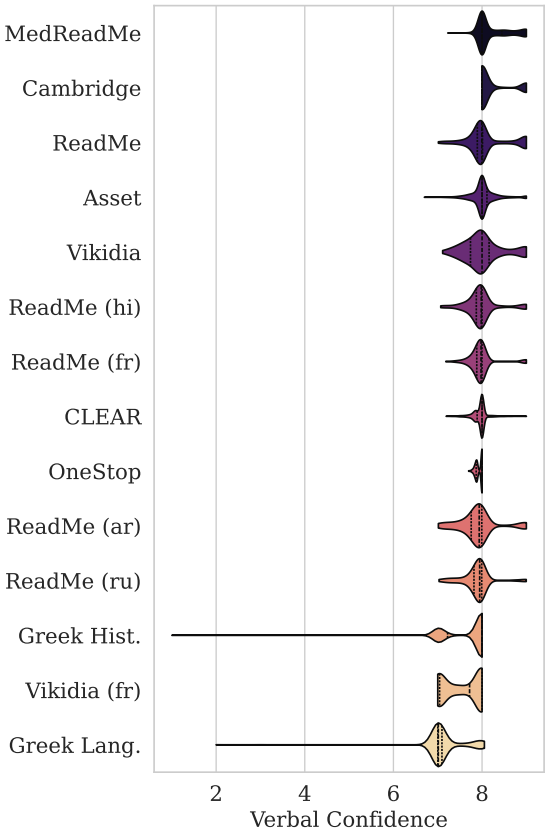

Figure 4: Distribution of Verbal Confidence Scores by Dataset

shows that LAURAE with readability formulas outperforms LAURAE with RSRS on 9 of 14 datasets. The average performance increase is 0.005 points. Although a relatively small average difference, we combine this with the fact that, in comparison to RSRS, readability formulas are less computationally expensive, more interpretable, and have more user-friendly platforms (e.g., Readable.com) that allow less technical users to implement them. Thus, we recommend implementing LAURAE with readability formula scores for the shallow feature score.

F Verbal Confidence Score Distribution and LAURAE-agg performance

The distribution of confidence scores (means/standard deviations reported in Table 5, and distributions visualized in Figure 4) raises several areas for future research.

First, both within and across datasets, verbal confidence scores are clustered near the high end of the range (i.e., 1-9) that LLMs were prompted to provide confidence scores on. Even for datasets where performance is poor (e.g., Greek textbook datasets), the confidence scores are still rarely under 7. Thus, it would be an important area of future research to determine if there are techniques to improve confidence score calibration. This may include adaptations from prior works that prompt LLMs to explicitly consider other answers in their explanations (Tian et al., 2023).

Second, we found that LLM confidence scores are systematically lower when texts are non-English. On average, confidence in non-English datasets is just 0.7678, compared to 0.8071 for English datasets. These differences are not fully explained by lower performance on non-English datasets, nor by the use of Aya 32B model for non-English texts. For example, we compared the verbal confidence and actual performance using the Llama 70B model for the English Vikidia and French Vikidia datasets. While accuracy is higher on the French Vikidia dataset (0.947 versus 0.887), the verbal confidence scores are higher for the English Vikidia dataset (8.025 versus vs. 7.571). This suggests that LLMs may be relatively underconfident when rating non-English texts (likely due to a smaller amount of training data compared to English texts), and this encourages future research to assess and improve LLM confidence calibration across languages.