Towards Empowering Consumers through Sentence-level Readability Scoring in German ESG Reports

Benjamin Josef Schüßler[♢], **Jakob Prange**[♠]
[♢]University of Augsburg, Germany
[♠]German Center for Addiction Research in Childhood and Adolescence (DZSKJ),
University Medical Center Hamburg-Eppendorf, Germany
benjamin.schuessler@uni-a.de, j.prange@uke.de

Abstract

With the ever-growing urgency of sustainability in the economy and society, and the massive stream of information that comes with it, consumers need reliable access to that information. To address this need, companies began publishing so called Environmental, Social, and Governance (ESG) reports, both voluntarily and forced by law. To serve the public, these reports must be addressed not only to financial experts but also to non-expert audiences. But are they written clearly enough? In this work, we extend an existing sentence-level dataset of German ESG reports with crowdsourced readability annotations. We find that, in general, native speakers perceive sentences in ESG reports as easy to read, but also that readability is subjective. We apply various readability scoring methods and evaluate them regarding their prediction error and correlation with human rankings. Our analysis shows that, while LLM prompting has potential for distinguishing clear from hard-to-read sentences, a small finetuned transformer predicts human readability with the lowest error. Averaging predictions of multiple models can slightly improve the performance at the cost of slower inference.[1]

Keywords: sentence-level readability, German ESG reports, crowdsourcing

1. Introduction

In order to make transparent how corporate economic goals align with, contribute to, or violate sustainability goals, policymakers demand written reporting on environmental, social, and governance topics, in short, ESG reports.[2] Next to *greenwashing* (the intentional or negligent misrepresentation of one's sustainability strategy to sound more positive and marketable than it really is, de Freitas Netto et al., 2020), another challenge is ensuring the reports' accessibility to their diverse audiences. This is even more important for layperson consumers than for other stakeholder groups such as economic auditors or financial analysts. The latter know exactly what they are looking for and, in case of unclear language, can consult with legal or public relations experts. This is usually not the case for consumers, who may be on their own and may read exploratorily, to gather information from scratch. Quoting EU Directive 2024/825, also known as the Empowering Consumers Directive,[3] *"[i]n order to contribute to the proper functioning of the internal market, based on a high level of consumer protection and environmental protection, and to make progress in the green transition, it is essential that consumers can make informed purchasing decisions and thus contribute to more sustainable consumption patterns. That implies that traders have a responsibility to provide* ***clear, relevant and reliable information****."*

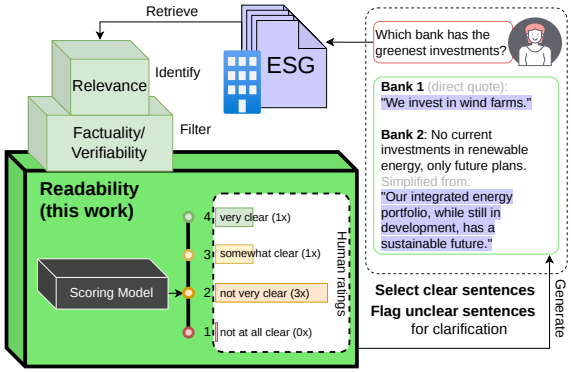

Figure 1: Readability as a foundation of consumer empowerment from ESG reports.

In this work, we focus on clarity as a fundamental requirement for consumer accessibility, and evaluate automatic readability scorers against the judgments of layperson readers (figure 1).

Automatic readability assessment (ARA) is the task of estimating how easy a text is to read and understand. Often, this is measured at the document level using broad sentence and word length

[1]Code and dataset extension available at: github.com/schuesslerbenjamin/Sentence-level-Readability-Scoring-in-German-ESG-Reports. Trained models available at: huggingface.co/schuesslerbenjamin/Sentence-level-Readability-Scoring-in-German-ESG-Reports

[2]EU Directive 2022/2464, also known as the Corporate Sustainability Reporting Directive (CSRD): eur-lex.europa.eu/eli/dir/2022/2464/oj/eng

[3]eur-lex.europa.eu/eli/dir/2024/825/oj/eng

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 26–41

statistics to assess generic readability. For example, the Flesch Reading Ease test (Flesch, 1948) is intended to rate English educational books on a scale ranging from school grades to professional scientist difficulty.

Here, on the other hand, we are interested in measuring readability in German ESG reports and in more fine-grained grammatical patterns than length. As our target group, we envision, for example, a young adult deciding on a sustainable bank to open their first account or investment plan with, or a family choosing an electricity provider for their home. Already overwhelmed with the multitude of companies to choose from and only able to skim very short excerpts of reports from each company, they rely on a retrieval-augmented generation (RAG; cf. Kleinle et al., 2024) or recommender system (cf. Hillebrand et al., 2023). We consider a previously unaddressed requirement for such a system, namely to maximize the clarity of the presented content: systems should (a) prefer easily readable sentences for direct extraction and (b) if a sentence is highly relevant but difficult to read, it should be simplified. Rather than trying to replicate coarse document-level scores like Flesch, we thus propose to model readability at the sentence level.

We aim to answer two questions: **RQ1:** How readable are German sustainability reports? And **RQ2:** How to model sentence-level readability? Concretely, we contribute:

- an in-depth data analysis of German ESG reports through a crowd-sourcing annotation study, finding largely clearly written sentences but also subjective variation;
- a comparison of different model types, including generative Large Language Models (LLMs), regressions, and a custom feature-based classifier, finding lower prediction error in small finetuned models and higher ranking correlation, albeit on a shifted scoring scale, in one of the LLMs;
- an ablation of syntactic features, highlighting their relative importance in sentence-level readability prediction;
- and a discussion of sentence-level readability in the context of other factors of consumer empowerment through ESG reporting.

2. Related Work

To find similar research, we systematically queried the typical research databases (see Appendix A).

Readability of German texts. While most research on readability has focused on English texts (Collins-Thompson, 2014), some approaches have also been adapted to the German language. Amstad (1978), for example, adapts the Flesch Reading Ease formula by Flesch (1948) by changing the factor for the word length to consider that German words tend to be longer. More recent research includes creating more sophisticated readability formulae (e.g, "Hohenheimer Komplexitätsindex für Politikersprache" (HKPS, German for Hohenheim Complexity Index for Political Language) by Kercher (2013)), improving the readability for people with learning difficulties (e.g., Jablotschkin et al. (2024)), or analyzing how difficult language learners perceive the readability of texts (e.g., Weiss and Meurers (2022)).

Furthermore, the GermEval 2022 shared task on text complexity assessment of German texts by Mohtaj et al. (2022) is based on sentences from articles in the areas of society, science, and history of the German Wikipedia. It motivated a wide range of approaches, the best of which was an ensemble of GBERT and GPT-2 submitted by Blaneck et al. (2022) and achieved a 0.195 MSE (0.442 RMSE) on a 7-point rating scale.

Our study, instead, focuses on the readability as perceived by *native speakers* who are laypersons in the *ESG domain*.

Readability of ESG reports. Smeuninx et al. (2020) compare the performance of traditional readability formulae with a few modern NLP methods when predicting the readability of English ESG reports. They find that the former lack in performance, especially when the syntax varies. In general, Smeuninx et al. (2020) identify that ESG reports can be difficult to read, in some instances even more complex than financial reporting.

Among other linguistic aspects, Huang et al. (2024) analyze the readability of Chinese ESG reports and their impact on the ESG scores over time. Bonn and Gaida-Albers (2024) investigate how report readability, among other parameters, correlates with the overall sustainability of German companies ("ESG-Score" assigned by auditors), but they do not *predict* readability and the reports they analyze are written in English.

Methodologically very similar to our work are Vajjala and Meurers (2012), who compare syntactic features against "traditional features" like word length and sentence length and achieve 0.023 MSE (0.15 RMSE) on a 5-point rating scale. But they, again, work with English texts in the educational domain rather than German ESG reports.

3. Data

For our experiments, we use the dataset[4] from the SustainEval GermEval shared task on understanding sustainability reports (Prange et al., 2025). It consists of short excerpts sampled from the German Sustainability Code (Deutscher Nachhaltigkeitskodex),[5] where companies can voluntarily publish ESG reports and receive feedback and resources to prepare for legally required and audited CSRD reporting. Specifically, each datapoint consists of four consecutive sentences in German, of which the last is the target sentence receiving annotation and the preceding ones are provided for context. Statistics are given at the top of table 1.

Readability Annotation. We extend the "verifiability" annotations used for the SustainEval shared task with layperson readability judgments via crowd-sourcing. Training and evaluation crowd annotators were recruited via Prolific and paid above German minimum wage. The actual annotation was carried out via SoSciSurvey on GDPR-compliant servers in Germany. Development crowd annotation was carried out on a different platform, also according to German minimum wage and GDPR standards. The change in annotator pools likely led to the difference in agreement and score distributions. In all cases, the only information disclosed by annotators was that they speak German as their primary language. Annotators were identified only by anonymous IDs, which enabled us to exclude annotators from future annotation rounds if they were too fast or always assigned the same category. While the three context sentences were shown, annotators were asked specifically to rate their understanding of only the target sentence on a forced-choice Likert scale (*How well do you understand the sentence?* 1: not at all, 2: rather not, 3: somewhat, 4: very clearly).

Agreement. Most sentences were rated by 5 annotators (some by 4 and very few by 6), and most of the time (72.3–87.7%), a majority of at least 3 annotators assign the same rating (middle part of table 1). Due to the lack of annotator identities in the data, we were not able to compute chance-corrected agreement metrics such as Cohen's κ or Krippendorff's α. To gain a more comprehensive measure of agreement, we introduce *Mode Agreement*, which cleanly handles anonymity and varying numbers of annotations per sentence. For each sentence, we count how many annotators agree on the most common rating (the mode), and divide by the number of annotations that sentence received (see equation 1). If all annotators agree, the Mode Agreement is one. If no annotators agree, the Mode Agreement is zero.

	Train	Dev	Eval
# Sentences	960	267	407
Ø Words / sentence	16.92	17.50	17.32
Ø Syllables / word	2.28	2.32	2.28
Inter-Annotator Agreement			
≥ 3 agree	86.8%	72.3%	87.7%
Mode agreement	70.3%	60.4%	70.1%
Readability Annotations [1.0; 4.0]			
Avg. mean	3.515	3.200	3.526
Avg. standard deviation	0.505	0.691	0.501
Avg. majority vote	3.695	3.431	3.709
# Actual majority votes			
1.0	5	0	0
1.5	1	0	0
2.0	21	11	7
2.5	19	14	4
3.0	167	90	81
3.5	76	38	35
4.0	671	114	280

Table 1: Dataset statistics.

$$\text{Mode Agreement} = \begin{cases} \frac{\text{Mode's Frequency}}{\text{\# Annotations}}, & \text{if Mode's Frequency} \geq 2 \\ 0, & \text{else.} \end{cases} \quad (1)$$

Weighing the majority agreement in this way, while still in an acceptable range of 60.4–70.3%, paints a somewhat less optimistic picture than simply counting how often a majority exists (see table 1). Based on this, we decided to account for outlier noise in crowd-sourcing by aggregating the annotations using the majority vote instead of the mean over all votes. Only in the case of ties, we take the mean of the tied votes.

Score distribution. The bottom part of table 1 shows that the annotations are skewed towards very easily readable texts. This is most substantial in the training and evaluation splits, where more than two thirds of all instances were assigned the highest readability score of 4. This can have several reasons: Firstly, the annotators are self-declared native laypeople. We expect that most laypeople generally perceive texts in their native language as at least somewhat readable. Secondly, crowd-workers might fear that if they rate a text as not understandable, they might not be allowed to answer the other questions and get paid less. To ensure models learn to predict scores across the full scale during training, we randomly oversample the underrepresented rating classes in the training split until they match the most common

[4] github.com/SustainEval/sustaineval2025_data/
[5] deutscher-nachhaltigkeitskodex.de

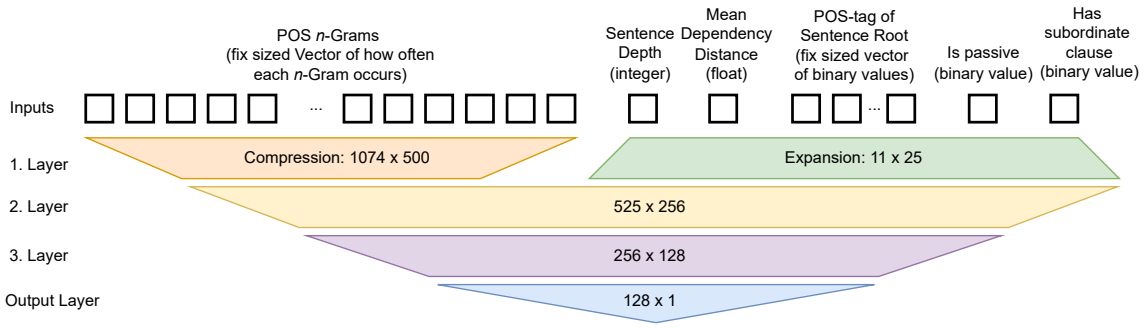

Figure 2: Structure of our syntax-based ARA model.

class. We do not manipulate the distribution in the development and evaluation splits. Scores are normalized to [0.0; 1.0] for model training, inference, and evaluation.

4. Whitebox Readability Models

In line with our goal to predict sentence-wise readability in ESG reports in a transparent and interpretable manner, we explore a simple but effective whitebox model, where input features can be controlled, parameter sizes are small, and runtimes are fast (section 4.1). We hypothesize that to capture actual human judgments in a specialized domain, established formulae with linear coefficients preset to a fixed educational setting are not sufficient and test this with a baseline model (section 4.2). On the other hand, state-of-the-art language models (section 5) may be more powerful than we need and their large blackbox architectures restrict the linguistic insight they may provide.

Given the use case outlined in the introduction, where a system needs to judge how presentable individual sentences are to a layperson user, we focus on the lower-bound setup of providing only the target sentence to the models, without context. Within-sentence grammatical patterns are complementary to across-sentence semantic and pragmatic aspects of readability like verifiability, coherence, and cohesion. These aspects likely also have a large impact on readability and are modeled to various extents by pretrained (L)LMs. By examining quantitative and qualitative differences in readability scores assigned by the various models, we can approximate which facets of readability can be determined from syntax alone and which stem from other linguistic properties.

4.1. Syntactic Features

Inspired by the work of Liu et al. (2025) and Smeuninx et al. (2020), we design a feed-forward neural network on top of syntactic features extracted from the input sentence (figure 2). Most notably, to limit the impact of the Part-of-Speech (POS)-tag n-Grams, their vector is compressed in the first layer before being concatenated with the other features. For more implementation details, see appendix B.

Our model uses the following features:

Part-of-Speech-tag n-grams. The first syntactical feature analyzes the grammatical structure of a sentence, as it can have an impact on the syntactical complexity and, thus, on the readability of a sentence (Razon and Barnden, 2015). This is based on the idea that if a sentence structure is observed more often, a reader is more likely to understand it easily (Kauchak et al., 2017). Using a sliding window, we represent a shallow view of a sentence's syntactic structure as count features of POS-tag bigrams and trigrams. We filter for punctuation symbols. Before training, we generate all bigrams and trigrams that appear in the training set and expand them into individual features, representing how often each bigram and trigram appears in the sentence. From the training data, we extract 158 unique bigrams and 916 unique trigrams resulting in 1,074 total n-gram feature dimensions.

Depth of the dependency tree. As proposed by Yngve (1960), we calculate the depth of the dependency tree to estimate the hierarchical complexity of a sentence. In a dependency tree, every word except for the root has exactly one head that it refers to. The deeper the dependency tree of a sentence, the higher its syntactic complexity, and thus—we hypothesize—the harder it is for the reader to understand the relations of the words within the sentence.

Mean dependency distance. We also calculate the mean dependency distance as proposed by Liu (2008). The dependency distance is thereby defined as the number of words between a word and its head. Using the mean dependency distance instead of summing up all distances within a

sentence prevents longer sentences from getting disproportionally higher scores (Liu, 2008).

Part-of-Speech-tag of the root. There is exactly one word in every sentence that has no head in the dependency tree, the root. It can have a major impact on a sentence's readability (Dell'Orletta et al., 2011). We extract all root POS-tags within the training data and expand them into binary variables. We find that verbs or auxiliary verbs are usually the root of sentences in our dataset. Thereby, our approach allows the model to find relations between all possible POS-tags of sentence roots and the readability of a sentence.

Passive voice. Sentences written in passive voice can also be harder to read. Thus, Smeuninx et al. (2020) analyze the readability of whole documents and calculate the proportion of sentences that are in passive voice. Since we are only working with single sentences, we create a binary variable indicating whether the text is in passive voice. We consider a sentence to be passive voice if it contains a participle that has a form of "werden" (the equivalent of passive *to be*) as its head, or if it includes a passivized subject.

Subordination. Finally, sentences consisting of multiple clauses can be more complicated than sentences with fewer clauses. Smeuninx et al. (2020) calculate the average number of subclause-introducing elements per sentence to represent the degree of subordination in a document. We adopt their idea to our sentence-level ARA task and create a binary variable that indicates whether there is at least one subordinate conjunction in the sentence.

4.2. Baselines

To test whether our selected linguistic features are more informative in our setting than established work on readability suggests, we compare with two baselines. Both baselines are trained on our German ESG-report data to account for domain effects.

Sentence Length. We train a simple linear regression model using only the number of words per sentence. This approach was used by Crossley et al. (2007) as a proxy for syntactic complexity.

Readability Formulae. Representing the traditional research on readability, we train an XGBoost model (Chen and Guestrin, 2016) over scores calculated using established readability formulae.[6] We select the following formulae due to their relevance and applicability to German sentences: the *Flesch-Reading-Ease* test introduced by Flesch (1948) for English texts and adapted to German texts by Amstad (1978); the *Hohenheim Complexity Index for Political Language* (HKPS, Kercher, 2013); the proportion of polysyllabic words, based on the idea of the *Gunning Fox* (Gunning, 1952) and *SMOG indices* (McLaughlin, 1969); the *Vienna Educational Text Formula* (Bamberger and Vanacek, 1984); and the Swedish readability index LIX (Björnsson, 1968). See appendix C for details on the formulae and their implementation.

5. Blackbox Readability Models

As reference, we also compare the syntactic features model with two types of modern language models: a finetuned classifier on top of a pretrained transformer encoder and instruction-tuned generative LLMs. This is to set a practical upper bound in terms of predictive power. If a whitebox model reaches or surpasses the blackbox models' prediction accuracy, the whitebox model should clearly be preferred. Otherwise, a tradeoff between accuracy, speed, and interpretability needs to be found.

5.1. XLM-RoBERTa Encoder-Classifier

For the first language model, we use a transformer encoder model and task-specifically finetune it to the ARA task. This approach follows Tseng et al. (2019) and can simultaneously consider several linguistic layers of a text, including semantic and syntactic aspects, making it more powerful in principle than the syntax-based model. Since we define ARA as a regression task, we train the model's final layer as a regression head (see appendix D for more details).

This allows the model to effectively predict readability scores on a scale from zero to one. We compare several BERT-like encoder models on the development set, and select the multilingual `XLM-RoBERTa-base`[7] and `XLM-RoBERTa-large`[8] models (Conneau et al., 2020; Liu et al., 2019) based on their performance.

5.2. Generative LLMs

For the second language model, we test instruction-tuned LLMs on the ARA task, focusing on instruction-tuned models pretrained on datasets that include German texts.

[6] We also experimented with aggregating the scores using Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net, but XGBoost led to the best results overall.

[7] huggingface.co/FacebookAI/xlm-roberta-base

[8] huggingface.co/FacebookAI/xlm-roberta-large

Type	Model	MSE (↓)	MAE (↓)	Kendall τ (↑)	Ø Time per Sentence (↓)	# Params
□ Whitebox	Sentence length baseline	0.1859	0.4017	–0.2290	0.0003s	1
	Readability formulae baseline	0.0394	0.1588	0.0863	**0.0018s**	5
	Syntactic features (ours)	0.0389	0.1502	0.0534	0.0261s	∼0.7M
■ XLM-RoBERTa	base	0.0295	**0.1114**	0.2461	0.0035s	∼278M
	large	0.1325	0.3373	–0.2198	NA	∼550M
■ LLMs	Qwen 3 4B Instruct 2507	0.1119	0.2469	0.2822	0.0503s	∼4,000M
	Gemma 3 4B it	0.2396	0.4402	0.0448	0.3110s	∼4,000M
	Llama 3 8B instruct	0.2347	0.4230	–0.1906	0.7100s	∼8,000M
Combinations	Syntax + XLM-base	**0.0264**	0.1141	0.1676	0.0296s	∼279M
	Syntax + Qwen	0.0485	0.1777	0.2304	0.0764s	∼4,001M
	XLM-base + Qwen	0.0358	0.1432	**0.2857**	0.0538s	∼4,278M
	Syntax + XLM-base + Qwen	0.0292	0.1355	0.2627	0.0799s	∼4,279M

Table 2: Results of the experiments. ↓ indicates that a lower value is better and ↑ indicates that a larger value is better. The best value per metric is bold and the best per model type is underlined. For details on the experimental setup see appendix F.

We compare the `Llama 3 8B instruct` model[9] by Dubey et al. (2024) with the `Qwen 3 4B Instruct 2507` model[10] by Yang et al. (2025) and the `Gemma 3 4B it` model[11] by Kamath et al. (2025). We analyze these LLMs, as they are highly relevant in the current research and have been extensively researched.

We prompt the models with similar instructions as the human annotators. While the dataset itself is in German, we prompt the models in English as previous research found that several LLMs are biased to internally pivot towards English due to imbalanced training data (Wendler et al., 2024). We instruct the models to classify the readability of a sentence into four classes, each coded with a number from one through four. Similar to the human annotations, these numbers are then scaled down to the same range from zero to one. We apply one-shot prompting (see appendix E), because early experiments on the development data split showed that zero-shot prompts lead to worse performance. For the shots, we randomly sample sentences and their readability score from the training data.

6. Experimental Results

All performance metrics can be found in table 2, while table 3 contains average predicted scores. Details on the experimental setup can be found in appendix F.

Metrics. We use Mean Squared Error (MSE) as our main metric for analysis and as the loss function during training. We also report the Mean Absolute Error (MAE), as it is more robust to outliers than the MSE, and more interpretable because it is true-to-scale. For both error metrics, a lower score is considered better, where 0 is the best possible error and 1 is the worst possible.

	Train	Dev	Eval
Sentence length baseline	2.523	2.535	2.531
Readability formulae baseline	3.527	3.551	3.528
Syntactic features	3.608	3.555	3.596
XLM-RoBERTa base	3.790	3.933	3.920
Qwen 3	3.077	3.075	3.037
Human Annotation	3.695	3.431	3.709

Table 3: Average model predictions.

Additionally, we report a rank order correlation score to analyze whether a model can correctly identify which sentences are easier to read than others. The sentences are sorted by the predicted scores, and then this order is compared to the sorted list of gold-standard annotations (Collins-Thompson, 2014). Rank correlation measures the extent to which the predicted order aligns with human annotations across entire datasets. A model that can distinguish between easy and hard-to-read sentences but has a systematic bias to too low or high scores, provides more value than a model that makes incorrect predictions in both directions. MSE does not capture this difference, which is why a rank order correlation score is needed. We use the Kendall τ coefficient, variant b, introduced by Kendall (1945) as it is often used in the research and accounts for ties. To calculate the scores, we

[9] huggingface.co/meta-Llama/Llama-3.1-8B-Instruct
[10] huggingface.co/Qwen/Qwen3-4B-Instruct-2507
[11] huggingface.co/google/gemma-3-4b-it

use the Python implementation by the library `scipy` by Jones et al. (2001)[12]. A Kendall τ score of $+1$ indicates perfect correlation or correct ranking, relative to the ground truth. A score of 0 indicates no correlation, and -1 shows that the rankings are inverted.

6.1. Individual Models

Sentence Length. This simple baseline yields mediocre error metrics and the Kendall τ score is worse than random, indicating that the model fails to distinguish easy from hard-to-read sentences. Further, this shows that sentence length by itself is not a good predictor of readability.

Readability Formulae. Aggregating the scores of several traditional readability formulae leads to better results than the simple sentence-length baseline, according to all metrics. This shows that the additional sentence parameters and weightings used in readability formulae allow for a better prediction of the readability than relying on the number of words alone.

Syntactic Features. With our proposed feature-based approach, we aim to predict sentence readability by having the model learn to analyze several syntactic patterns. This model is more complex than the two baselines, both in terms of its input features and degrees of freedom of its hidden layers. This added representational capacity leads to a similar error rate and ranking performance as the formulae.

To investigate the importance of individual features, we conduct an ablation study (table 4). A feature is more important to the model if its removal strongly negatively affects the performance, i.e. increases the error rate or lowers the Kendall τ score. Removing trigrams worsens the performance the most, according to all three metrics. Thus, it is the most influential feature. Passivization (by the error metrics) and the depth of the sentence (by the Kendall τ score) are, respectively, the second most important features.

Interestingly, removing the bigrams very slightly improves the performance of the model, according to the error metrics. A possible reason is that the information in the bigrams is already part of the trigrams and the model compresses the n-gram input vector to a fixed width. However, according to the Kendall τ scores, ablating any feature leads to worse performance, indicating their necessity to correctly distinguish easy from hard sentences.

[12]We also experiment with rounding the predictions to the next .5 before calculating the Kendall τ score, but as this leads to similar results, we keep the original calculation.

Ablated Feature	MSE	MAE	Kendall τ
Sentence Depth	+0.0043	+0.0282	–0.0904
Dependency Dist.	–0.0046	–0.0104	–0.0286
Sentence Root	–0.0045	–0.0321	–0.0123
Is Passive	+0.0116	+0.0329	–0.0690
Has Subordination	+0.0036	+0.0039	–0.0508
Bigrams	–0.0031	–0.0033	–0.0467
Trigrams	**+0.0227**	**+0.0617**	**–0.1103**
All Features	0.0369	0.1502	0.1203

Table 4: Ablation of the Syntax-based ARA model on the evaluation data split. We report the differences in metrics to the complete model (last row).

XLM-RoBERTa. We observe a strong improvement over the syntax model in all metrics. This can be argued with the transformer model's higher degrees of freedom to fit to the task. The XLM-RoBERTa model thereby predicts with the smallest errors of all individual models. We also find, rather surprisingly, that the large model variant performs much worse than the base model variant, only slightly beating the simple sentence length baseline. This may be due to insufficient training conditions that fail to saturate the many parameters of the large model.

Generative LLMs. Prompting different LLMs, we find that Llama and Gemma fail to correctly estimate the readability, as indicated by high error metrics and very low Kendall τ scores. Qwen outperforms the other LLMs in every metric. Thus, we use Qwen as the representative LLM going forward. The highest Kendall τ score out of all the individual models indicates that the LLM is better at distinguishing easy from hard-to-read sentences than XLM-RoBERTa and the whitebox models. However, as it has not been finetuned on our specific dataset and rating scale, it is much worse than our syntax model, the readability formula model, and XLM-RoBERTa at assigning scores that are numerically close to the human ratings. Specifically, it assigns lower scores on average than most other models and humans (table 3).

Error Analysis. Consider the examples in figure 3. The top sentence is lengthy and syntactically complex. There is some variation in human judgments but consensus is clearly "not very readable". This is reflected in all model predictions being less than the top score, though the syntax model and XLM-RoBERTa are still (too) optimistic, while Qwen matches the human vote. The bottom sentence is short and not complex, but understanding it requires access to the preceding context, which was provided to humans but not to models. This lead to a bimodal distribution in human judgments, as

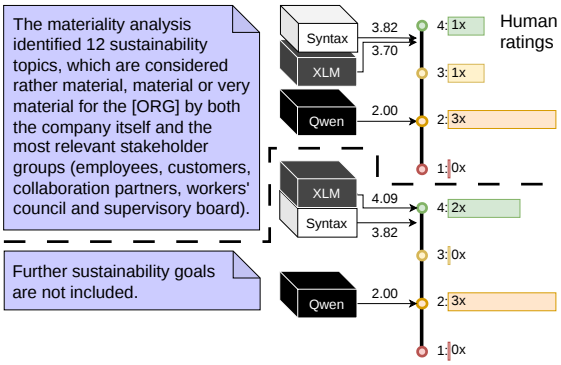

Figure 3: Two examples of different length and superficial complexity (translated from German).

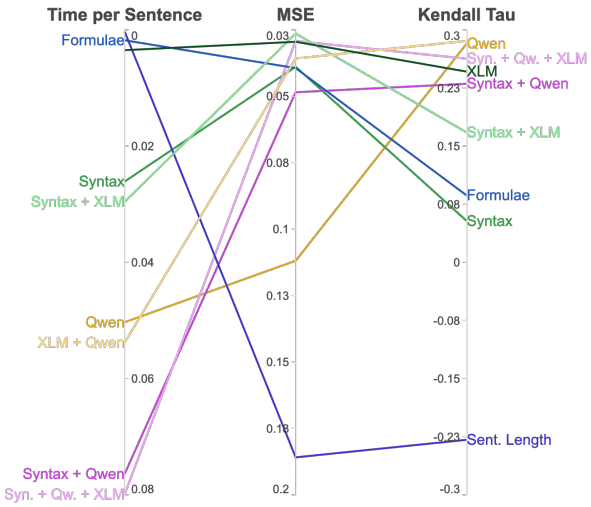

Figure 4: Time-performance trade-offs. All axes rank models top-down from best to worst.

some annotators likely focused on the low syntactic complexity (high readability) while others emphasized context-dependence (low understanding). Models diverge similarly, and Qwen again happens to match the majority vote which in this case is only narrowly decided. Note that Qwen and the Syntax model assign the same score to these two examples.

6.2. Model Combinations

To account for different aspects of readability influencing individual models differently, we also experiment with averaging the predictions of the three models.[13] The combination of the syntactic model and RoBERTa has the lowest MSE (but not MAE)[14] out of all experimental settings, but only by a small margin. Combining Qwen's and RoBERTa's predictions slightly outperforms Qwen's individual Kendall τ score. Overall, simple mean aggregation does improve predictions slightly, but not substantially.

6.3. Trading off Errors and Speed

Some models can rate the readability of a sentence faster while making larger prediction errors than others (figure 4). Three models appear viable for this tradeoff: The readability formulae baseline is very fast while making small errors. Combining syntax and RoBERTa is slower, but makes even smaller errors. The most viable option is simply using RoBERTa, which is almost as fast as the readability formulae approach and makes almost as little errors as the combination. Any combination with an LLM has higher computational cost, slower speed, and larger score deviations.

7. Discussion

We set out to analyze how we can automatically measure the readability of German ESG reports in a way that aligns with non-expert human judgments. In doing so, we gained insight into the following two research questions:

RQ1: How readable are German ESG reports? Our data analysis reveals that on average laypersons perceive German ESG reports as easy to read. However, some potentially crucial sentences are unclear and there is considerable variation in judgment between readers. While existing research characterizes ESG reports as generally hard to read (e.g., Pombinho et al., 2024), our sentence-level approach allows a more nuanced evaluation and, thus, enables more nuanced solutions to the problem.

In contrast to similar research that analyzes specific demographics like second language learners (e.g., Vajjala and Meurers (2012)), we have no clearly defined target audience. Our approach allows us to analyze how an average German speaker might perceive the readability of German texts. However, without a clearly defined audience, defining rules for what makes a text readable is difficult. We see this reflected in imperfect agreement among annotators. Therefore, future research could investigate personalized readability systems (Benjamin, 2012; Bailin and Grafstein, 2001) while considering that not only the grammatical structure of the target sentence, but also its

[13] Additionally, we tried other aggregation methods, including Linear Regression, Ridge Regression, and XGBoost. However, all models performed similarly, thus we chose mean aggregation for simplicity.

[14] MSE is more sensitive than MAE to individual datapoints with large errors. So the difference between the model with the lowest MSE and the model with the lowest MAE lies in deviating less from the ground truth on outliers versus getting the majority of the data closer to it.

dependence on local and external context might be important.

RQ2: How to model sentence readability? We find that more complex models (more parameters) take longer to rate the readability of a sentence and tend to outperform smaller models in terms of the Kendall τ score. The LLM outperforms the other individual models in this metric, showing that it is better at delineating easy from hard-to-read sentences. However, as the LLM was not specifically tuned to the task like the other models, it performs the worst according to the MSE. XLM-RoBERTa has the best tradeoff between low MSE and fast inference.

Intuitively and according to the literature (e.g., Collins-Thompson, 2014; Vajjala and Meurers, 2012), word choice and lexical complexity play an important role as well. However, we were not able to replicate this effect in our domain and audience in pilot experiments with word frequency features.

Readability depends on the audience. A central difficulty with estimating readability is that it depends on genre and domain, as well as the audience. Traditional formulae like Flesch Reading Ease involve coefficients finetuned to the educational domain, which we account for by training a new regressor on our German ESG data. And while ESG reports are likely authored by trained writers who ensure high readability standards for expert readers, expectations may be different for the average consumer.

Empowering Consumers. As shown in figure 1, readability is a crucial building block of true consumer empowerment from ESG reporting, next to other important factors like factuality (Diggelmann et al., 2020; Florstedt et al., 2025; Luo et al., 2025) and verifiability (Prange et al., 2025). There are likely inter-correlations between these different aspects of how a company's ESG report is written and that company's actual sustainability strategy (Bonn and Gaida-Albers, 2024). Although the predictions of even the best models analyzed in this work are far from perfect agreement with the annotators, their scores still provide an indication for the readability as perceived by laypeople and substantiate the complexity of the task.

8. Conclusion

In this work, we applied different readability scoring methods to German ESG reports. We evaluated these methods using error and rank correlation metrics, as well as their insight into what makes a sentence hard-to-read (whitebox versus blackbox). Our results show that prompting LLMs has the potential to distinguish clear from hard-to-read sentences. However, a small task-specifically finetuned transformer model predicts human readability with the smallest error. Averaging predictions of multiple models can slightly improve the performance at the cost of slower inference.

Feature-based models and other explainability methods, which we leave to future work, can identify individual linguistic patterns that impact readability. Thereby, future research could contribute to transparency and consumer empowerment, consumer protection and, through more sustainable consumption patterns, environmental protection and the green transition.

9. Limitations

Naturally, any model is an abstraction of reality. Thus, our models are also limited in several ways. Other hyperparameters, LMs, and more complex prompt engineering could lead to different results. Further behavioral and mechanistic explainability methods could allow more thorough investigations of the whitebox (e.g. gradient-based) and even blackbox models (e.g. discretization-based). Finally, analyzing German ESG reports on the document level could be interesting, especially regarding the coherence between adjacent sentences. Furthermore, we identify the following two major problems:

Difficulties in assessing readability. During our experiments, we find that complex context sentences can influence the perception of readability of consecutive sentences. Although the human annotators were tasked to only rate the target sentence, they were able to see the context sentences which might have impacted their ratings. However, our readability models were not able to see the context, leading to an information asymmetry. This poses a general problem to the task of sentence-level readability assessment.

Furthermore, we see a high level of subjectivity in the annotations as seen in the mediocre agreement on the readability ratings. To limit the influence of outliers, we use the majority vote to aggregate the individual annotations into a single gold truth. However, as the provided examples show, a strong disagreement can influence the majority vote drastically as well. To solve this problem, Benjamin (2012) proposes a personalized readability model trained on the user's browser behavior. However, privacy concerns arise when tracking such personal data. We assume that for only a few users the benefits would outweigh the risks.

Class imbalance. Our data shows a strong class imbalance towards easier-to-read sentences.

Given that we only analyze the readability as perceived by native speakers, this can be argued with their fundamentally good understanding of German sentences. Furthermore, crowd-workers might fear getting rejected from the task and thus paid less if they admit to not understanding the task, or may simply overestimate themselves. This bias towards very easy sentences is in contrast to existing research that describes ESG reports as ambiguous (Bingler et al., 2024) and more complex than financial reports (Smeuninx et al., 2020). This discrepancy might have to do with our focus on sentence-level rather than document-level readability, and may either be a true effect or an artifact of how judgments were collected. We invite future research to replicate and compare different methodologies.

Future work may also address the class imbalance not only at training time but also at test time. In a simple case, for example, performance can be broken down by gold rating, evaluating instances rated as perfectly clear by all annotators separately from all instances that at least one annotator had at least some trouble understanding.

10. Acknowledgments

The authors gratefully acknowledge the HPC resources used during early experiments that were provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the BayernKI project v110ee. BayernKI funding is provided by Bavarian state authorities. The majority of the work for this paper was done at the chair of Computational Linguistics of Prof. Dr. Annemarie Friedrich at the University of Augsburg and we are very grateful for their support. Further, we thank the anonymous reviewers for their constructive feedback. We also thank Nina Prange for her input on the public communication aspect of consumer empowerment.

11. Bibliographical References

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich.

Alan Bailin and Ann Grafstein. 2001. The linguistic assumptions underlying readability formulae: a critique. *Language & Communication*, 21(3):285–301.

Richard Bamberger and Erich Vanacek. 1984. *Lesen-Verstehen-Lernen-Schreiben*. Diesterweg.

Rebekah George Benjamin. 2012. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24(1):63–88.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Lärarbiblioteket. Liber, Stockholm.

Patrick Gustav Blaneck, Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2022. Automatic readability assessment of German sentences with transformer ensembles. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 57–62, Potsdam, Germany. Association for Computational Linguistics.

Thorben Bonn and Aurin Gaida-Albers. 2024. Does the Interaction of Informativeness, Readability, and Sentiment within Company's Sustainability Disclosure Shape an Entity's ESG Score? – Evidence from Germany.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

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 A. Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. 2007. Toward a New Readability: A Mixed Model Approach. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29(29).

Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32(1):19.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Johannes Florstedt, Jonas Fahlbusch, and Moritz Sontheimer. 2025. Detecting greenwashing in ESG reports: A comparative analysis of machine learning methods in traffic-related emissions disclosure. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 25–30, Winterthur, Switzerland. Association for Computational Linguistics.

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

Lars Hillebrand, Maren Pielka, David Leonhard, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Milad Morad, Christian Temath, Thiago Bell, Robin Stenzel, and Rafet Sifa. 2023. sustain.AI: a recommender system to analyze sustainability reports. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 412–416, New York, NY, USA. Association for Computing Machinery.

Jie Huang, Derek D. Wang, and Yiying Wang. 2024. Textual Attributes of Corporate Sustainability Reports and ESG Ratings. *Sustainability*, 16(21):9270.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-lite - a new corpus of easy German: Compilation, exploration, analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, St. Julian's, Malta. Association for Computational Linguistics.

Eric Jones, Travis Oliphant, Pearu Peterson, and others. 2001. SciPy: Open source scientific tools for Python.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce

Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Juyeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle K. Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry (Dima) Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

David Kauchak, Gondy Leroy, and Alan Hogue. 2017. Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.

Maurice G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251.

Jan Kercher. 2013. *Verstehen und Verständlichkeit von Politikersprache: Verbale Bedeutungsvermittlung zwischen Politikern und Bürgern*. Springer Fachmedien, Wiesbaden.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Steffen Kleinle, Jakob Prange, and Annemarie Friedrich. 2024. OMoS-QA: A dataset for cross-lingual extractive question answering in a German migration context. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 231–248, Vienna, Austria. Association for Computational Linguistics.

Fengkai Liu, Tan Jin, and John SY Lee. 2025. Automatic readability assessment for sentences: neural, hybrid and large language models. *Language Resources and Evaluation*, 59:2265–2296.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yunfang Luo, Xiling Cui, Qiang Liu, Qiang Zhou, and Yingxuan Zhang. 2025. Identifying exaggeration in ESG reports using machine learning techniques. *Data and Information Management*, 9(2):100084.

G. Harry McLaughlin. 1969. SMOG Grading—A New Readability Formula. *Journal of Reading*, 12(8):639–646.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 shared task on text complexity assessment of German text. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 1–9, Potsdam, Germany. Association for Computational Linguistics.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann

machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA. Omnipress.

Miguel Pombinho, Ana Fialho, and Jorge Novas. 2024. Readability of sustainability reports: A bibliometric analysis and systematic literature review. *Sustainability*, 16(1).

Jakob Prange, Charlott Jakob, Patrick Göttfert, Raphael Huber, Pia Wenzel Neves, and Annemarie Friedrich. 2025. Overview of the SustainEval 2025 shared task: Identifying the topic and verifiability of sustainability report excerpts. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 229–238, Hannover, Germany. HsH Applied Academics.

Abigail Razon and John Barnden. 2015. A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 521–528, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Nils Smeuninx, Bernard De Clerck, and Walter Aerts. 2020. Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP. *International Journal of Business Communication*, 57(1):52–85.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Hou-Chiang Tseng, Hsueh-Chih Chen, Kuo-En Chang, Yao-Ting Sung, and Berlin Chen. 2019. An Innovative BERT-Based Readability Model. In *Innovative Technologies and Learning*, pages 301–308, Cham. Springer International Publishing.

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.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Sondre Wold, Petter Mæhlum, and Oddbjørn Hove. 2024. Estimating lexical complexity from document-level distributions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6309–6318, Torino, Italia. ELRA and ICCL.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Victor H. Yngve. 1960. A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

12. Language Resource References

Prange, Jakob and Jakob, Charlott and Göttfert, Patrick and Huber, Raphael and Wenzel Neves, Pia and Friedrich, Annemarie. 2025. *SustainEval 2025 Data*. Available on GitHub.

A. Systematic Literature Search

To find existing research on the readability of German ESG reports, we queried several research databases: We searched in the ACL Anthology[15], which focuses on NLP research. We queried the DBLP[16], a German computer science bibliography, to include computer science research in general. We searched the EBSCOhost database[17] and Scopus[18] to include research from the area of business informatics. Finally, we searched the Web of Science[19] as it includes research from various disciplines. All databases were queried using with the following search string:

```
(readability OR understandability OR ((text OR sentence) AND (complexity or difficulty))) AND ("Environment* Social* Governance" OR ESG OR "Corporate Social Responsibility" OR CSR OR "sustainability report" OR "company climate report") AND German
```

This search string thereby combines several terms that describe readability with terms for German ESG reports. We searched the title, keywords, and abstract fields on the 24th of September in 2025.

B. Implementation Details for the Syntax Model

We load the *de_dep_news_trf* POS-tagging model for the German language by the python library `spacy`. It allows extracting the POS-tags of each word, identifying the root of a sentence and its POS-tag, and extracting the depth of the dependency tree. The mean dependency distance is calculated by an extension to the `spacy` library called `textdescriptives`. If `spacy` detects a passivized subject or the sentence includes a participle with a form of "werden" as its head, it is considered passive. Finally, a sentence has a subordinate clause if `spacy` finds a subordinate conjunction.

The features are aggregated in a neural network. The first layer is split into two parts. In the first part, the n-grams vector is compressed to 500 neurons to reduce its impact on the model and the remaining features are expanded to 25 neurons for the second part of the first layer. Then, we concatenate the two parts and pass them to the second layer consisting of 256 neurons. The third layer compresses the model down to 128 neurons before the model outputs the regression value in the single output neuron. After each layer, except for the output layer, we add the ReLU activation function (Nair and Hinton, 2010) and 10% dropout (Srivastava et al., 2014). The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2019; Kingma and Ba, 2015). The following hyperparameters were identified using grid search: batch size: 20; training epochs: 40; learning rate: 0.01; early stopping patience: 15.

C. Implementation Details for the Readability Formulae Baseline

Based on their historical relevance and novelty, we decided to use the following models that are applicable to German sentences:

Flesch-Reading-Ease Test. Since its introduction by Flesch (1948), the Flesch-Reading-Ease test has often been used to rate the readability of English sentences (Kauchak et al., 2017). It calculates a readability score based on the number of words per sentence and number of syllables per word. Amstad (1978) recalculated its factors to fit the formula to German sentences:

$$\text{Flesch Reading Ease} = 180 - \left(\frac{\text{\# Words}}{\text{\# Sentences}}\right) - 58.5 \cdot \left(\frac{\text{\# Syllables}}{\text{\# Words}}\right) \quad (2)$$

Hohenheim Complexity Index. The HKPS (Kercher, 2013) is based on articles on politics from the German newspaper *BILD* and on dissertations on politics from PhD students. Shallow sentence and word features are weight against each other based on their importance in the two text groups. If a text is more similar to a BILD article it is easier-to-read for laypersons, whereas dissertations are harder-to-read for laypersons.

Polysyllabic Proportion. The idea that sentences containing many long words tend to be more complex has been often applied in research. This is, for example, one of the core ideas shared by the SMOG index (McLaughlin, 1969) and the Gunning Fox Index (Gunning, 1952). However, both are neither designed nor adapted to German texts. Therefore, we use the simple polysyllabic proportion as a feature for our readability formulae-based model and follow McLaughlin's definition of polysyllabic words as words with at least three syllables.

[15] Available at https://aclanthology.org/, last accessed Sep 24, 2025.

[16] Available at https://dblp.uni-trier.de/, last accessed Sep 24, 2025.

[17] Available at https://research.ebsco.com/, last accessed Sep 24, 2025.

[18] Available at https://scopus.com/, last accessed on 24.9.2025.

[19] Available at https://webofscience.com/, last accessed Sep 24, 2025.

$$\text{Polysyllabic Proportion} = \frac{\text{\# Polysyllabic Words}}{\text{\# Words}} \quad (3)$$

Vienna Educational Text Formula. The Vienna formula was specifically designed for German scientific texts by Bamberger and Vanacek (1984). It considers the proportion of polysyllabic words, the length of sentences, long words, and the proportion of monosyllabic words. The authors supply three versions of the WSTF (Wiener Sachtext Formel in German). We use the first one as it is the most accurate one, according to the authors.

$$\begin{aligned} \text{WSTF} &= 0.1935 \cdot \text{MS} \\ &+0.1672 \cdot \text{Average Words per Sentence} \\ &+0.1297 \cdot \text{IW} - 0.0327 \cdot \text{ES} - 0.875, \end{aligned} \quad (4)$$

where MS is the percentage of polysyllabic words; IW is the percentage of words with more than six characters; and ES is the percentage of monosyllabic words. The scale represents the expected year of full-time education that is required to understand the text and ranges from 4 (easy) to 15 (very difficult).

Swedish Readability Index. Finally, the Läsbarhetsindex (LIX) was designed for the Swedish language by Björnsson (1968) and has already been successfully applied to other Germanic languages (e.g., Wold et al., 2024).

$$\text{LIX} = \frac{\text{Number of words}}{\text{Number of sentences}} \cdot \frac{\text{Number of long words} \cdot 100}{\text{Number of words}} \quad (5)$$

Long words are defined as words with more than six characters. Björnsson (1968) also provides a table that assigns readability classes for the score, but we use the raw LIX score to retain all information.

Model-specific implementation details. The Flesch Reading Ease and the first Vienna Eucational Text Formula are calculated using the python library `textstat`. The formulae for the polysyllabic proportion, the LIX score, and the HKPS are implemented by the authors of this work. The XGBoost aggregation method is trained with the MSE objective and the following parameters: number of boosted trees: 100; learning rate: 0.1; maximum tree depth: 5. These parameters were identified using a simple grid search,

D. Implementation Details for the Experiments with XLM-RoBERTa

The RoBERTa checkpoints are loaded using the library `transformers`. The following hyperparameters were identified using grid search: batch size: 20; training epochs: 30; learning rate: 0.0001; weight decay: 0.001; gradient disabled for the first five layers.

E. LLM Prompting

The LLMs are loaded using the `transformers` library with disabled sampling for reproducibility. They are prompted to rate the readability of sentences as follows: First, a system prompt outlines the general task and indicates to the model that it is supposed to rate the readability of German sentences (see figure 5). Second, a user prompt outlines the structure in which the model is supposed to output. This includes describing the rating scale from 1 to 4 and telling to model to only output a single digit. Furthermore, the user prompt includes a single shot based on the training split of our dataset including the majority vote of the human annotators delineated by placeholder tokens. The user prompt ends with the sentence that is supposed to be rated. Third, the LLM's output begins with a placeholder token for the score, followed by the score generated by the model.

System Prompt

You are a helpful assistant that rates the readability of German sentences.

User Prompt

I will give you a sentence and I want you to rate it's readability. If the sentence has a low readability, choose 1. If the sentence has a low to medium readability, choose 2. If the sentence has a medium to high readability, choose 3. If the sentence has a high readability, choose 4. Please only answer with a single digit corresponding to the readability level.
[Sentence] Dazu wurden die Aufgaben des Strategiekreises auf die Vertriebslinien übertragen, die Nachhaltigkeitsabteilung und Nachhaltigkeitssäulen gestärkt und Nachhaltigkeit in die regulären Entscheidungsprozesse integriert.
[Readability Score] 3
[Sentence] In diesen Trainings werden direkte Dialoge simuliert und konkrete Fälle mit den Mitarbeitenden bearbeitet.

Assistant

[Readability Score] 4

Figure 5: Prompt for the LLM-based ARA model using single-shot prompting.

F. Experimental Setup

Unless otherwise specified, we always use the default hyperparameters.

Hardware. The results to the experiments listed in this work were all created in a local workstation with an NVIDIA RTX 5080 with 16GB of VRAM, paired with an AMD 9800x3D CPU.

Software. All experiments were carried out using Python 3.12 with separate Conda environments for each model. CUDA version 12.8 was used. The experiments were run on Ubuntu 24.04 LTS within Windows Subsystem for Linux 2 on Microsoft Windows 11. We always set a seed for reproducibility.