# Assessing French Readability for Adults with Low Literacy: A Global and Local Perspective

**Wafa Aissa**[1,*], **Thibault Bañeras-Roux**[1,*], **Elodie Vanzeveren**[1], **Lingyun Gao**[1], **Rodrigo Wilkens**[2], **Thomas François**[1,†]
[1]CENTAL, IL&C, UCLouvain, Belgium, [2]University of Exeter, UK
[*]Equal contribution, [†]Corresponding author.
{wafa.aissa, thibault.roux, elodie.vanzeveren, lingyun.gao, thomas.francois}@uclouvain.be,
r.wilkens@exeter.ac.uk

## Abstract

This study presents a novel approach to assessing French text readability for adults with low literacy skills, addressing both global (full-text) and local (segment-level) difficulty. We also introduce a dataset of 461 texts annotated using a difficulty scale developed specifically for this population. Using this corpus, we conducted a systematic comparison of key readability modeling approaches, including machine learning techniques based on linguistic variables, fine-tuning of CamemBERT, a hybrid approach combining CamemBERT with linguistic variables, and the use of generative language models (LLMs) to carry out readability assessment at global and local level.

## 1 Introduction

According to UNESCO, illiterate individuals are persons "aged 15 years and above who cannot read and write with understanding a short simple statement on their everyday life" (UNESCO, 2009, 4). Reading deficit has far-reaching consequences for both individuals and society. Individuals affected by illiteracy are more likely to experience health problems (Berkman et al., 2011), have reduced life expectancy (Messias, 2003), and earn, on average, 30 to 42% less (Lal, 2015), largely due to lower levels of professional integration. At the societal level, illiteracy has a substantial economic impact, reducing the GDP of developed countries by approximately 2% (Steward, 2023). Being a continuum, the phenomenon of illiteracy ranges from complete illiteracy to low-literacy readers, who scored at or below Level 1 on the PIAAC scale (Grotlüschen et al., 2016), on which we focus here.

In response to such findings, governments have long taken steps to reduce low-literacy and illiteracy, notably through the creation of targeted literacy programs (e.g., ANLCI in France) that have proven particularly effective — for instance, *Alpha Plus* (Russeler et al., 2012). In parallel, since the Great Depression of the 1930s and the desire to better equip the many unemployed workers through reading training, researchers developed readability formulas for individuals facing low-literacy (DuBay, 2004). These tools automatically assess text difficulty, enabling efficient matching of texts to readers in educational and training contexts. However, as we will explore in greater detail in Section 2, no such specialized readability formula or model currently exists for French.

In the absence of dedicated formulas for low-literate readers, it is common to reuse formulas designed for different audiences, which is far from optimal (François, 2015; Napolitano et al., 2015). In fact, low-literate readers present a distinctive profile. While they generally have a solid oral command of their native language (unlike foreign language learners), they have not fully developed the automatic reading skills of experienced native readers. In conventional readability formulas such as those of Flesh or Dale and Chall, lower levels correspond to texts for primary children and thus cover rather childish topics that do not appeal to adult readers with low-literacy. This is why we believe that specialized formulas are needed to reflect their specific profile. As our first contribution in this paper, we propose the first readability model for French specifically designed for low-literate readers. The proposed model can be considered specialized in that it has been trained on a corpus of texts specifically calibrated by literacy educators with expertise in the field of illiteracy. Furthermore, it relies on a difficulty scale explicitly designed for this target population (Monteiro et al., 2023) and make use of recent advances in readability modeling, particularly through the use of hybrid architectures.

In addition, fieldwork observations indicate that generating a single, global readability score to match texts with readers is insufficient for effectively supporting reading skill development. In



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20529–20551
November 4-9, 2025 

practice, literacy educators are frequently required to manually simplify web-based texts prior to instruction, in order to tailor them to the needs of their learners. Although this task could align with the domain of automatic text simplification (Saggion, 2017; Štajner, 2021; Al-Thanyyan and Azmi, 2021), current approaches remain limited in their ability to be fine-tuned for specific target populations, as they relied on the Newsela dataset, fit for children (Kew and Ebling, 2022; Alkaldi and Inkpen, 2023). In this study, we adopt an alternative approach: rather than attempting to simplify texts directly, we focus on identifying the text segments that contribute to text difficulty for individuals with low-literacy. This approach remains highly interpretable and enables educators to retain control over simplification operations. As a task, it closely aligns with research on complex word identification (North et al., 2023), although our scope extends beyond lexical complexity to include the syntactic, discursive, and semantic dimensions of difficulty. To our knowledge, no such fine-grained system exists for local difficulty assessment in French, the closest system being AMesure (François et al., 2020) that targets standard readers of administrative documents and relies on pre–deep learning technologies. Our second contribution is therefore a novel system for local difficulty assessment in French, specifically designed for low-literate readers. The third contribution of this paper is an annotated corpus used to develop both above solutions, made up of 461 texts annotated, which is publically available[1]. This resource stands out in two key respects. First, it is unique in the context of French low-literate readers. Second, it provides a dual representation of textual difficulty: each text is annotated both globally (overall text difficulty) and locally (highlighting difficult segments within the text). Notably, items marked as complex in the local annotations are defined relative to the overall difficulty level of the corresponding text.

The remainder of the paper is structured as follows, section 2 reviews prior work on readability and local difficulty assessment, with a focus on research in French and on readers facing literacy challenges, and provides a comparison with existing datasets. Section 3 introduces the created dataset used in our study. Then, Sections 4 and 5 present the models evaluated for constructing our readability formula, with performance analyses on the global and local levels, respectively. We conclude with a discussion of the findings and future directions in Section 6.

## 2 Related work

Some of the earliest studies on readability focused explicitly on low-literate adults (Dale and Tyler, 1934; Gray and Leary, 1935). Over time, however, research interests shifted toward formulas designed for adults in general (Flesch, 1948; Gunning, 1952) or for schoolchildren (Dale and Chall, 1948). Today, readability formulas have been used to evaluate the difficulty of various types of documents — e.g., medical texts (Wilson, 2009; Mcinnes and Haglund, 2011) or contracts (Arbel, 2024) — for people with functional illiteracy. However, these studies rely almost exclusively on traditional formulas, which are not tailored to this population.

When narrowing the scope to studies that propose new readability models specifically aimed at low-literate adults, only a few efforts can be identified. These include research conducted on Portuguese (Aluisio et al., 2010), Italian (Dell'Orletta et al., 2011) or German (Weiss et al., 2018). The latter is particularly noteworthy, as it introduces a formula based on the *Alpha* difficulty scale, which was explicitly developed for individuals with functional illiteracy (Riekmann and Grotlüschen, 2011). This formula was subsequently incorporated into a dedicated search engine for low-literacy users (Dittrich et al., 2019), targeting *Alpha* levels 3 to 6. Despite these advances, studies dedicated to this specific population remain scarce. As noted by the review by Collins-Thompson (2014), which surveys the range of audiences addressed in readability research, individuals with functional illiteracy are not explicitly considered. To our knowledge, no such readability model currently exists for French.

Recent work, including in French, has increasingly emphasized the algorithmic aspects of readability modeling. Prior to 2017, the prevailing approach relied on machine learning aimed at identifying the textual features most predictive of reading difficulty and optimizing their combination (Schwarm and Ostendorf, 2005; Feng et al., 2010; Vajjala and Meurers, 2012), including for French (François and Fairon, 2012; Dascalu, 2014). Then, the field experienced renewed momentum with the advent of distributed semantic representations (Cha et al., 2017; Filighera et al., 2019) and the rise

[1] https://github.com/tfrancoiscental/iread4skills_readability_corpus_fr

of deep learning (Nadeem and Ostendorf, 2018; Azpiazu and Pera, 2019; Martinc et al., 2021). In French, Blandin et al. (2020b) exploited deep feed-forward models to generate reading age recommendations for children; Yancey et al. (2021) proposed a BERT-based readability formula for French as a foreign language and Van Ngo and Parmentier (2023) explored the relationship between a text's overall difficulty and the sentences that compose it. Lately, two recent trends have emerged in readability research. The first involves hybrid approaches that integrate traditional linguistic features within deep learning architectures, also known as hybrid approaches (Qin et al., 2020; Deutsch et al., 2020; Liu and Lee, 2023) with Wilkens et al. (2024) offering a representative example for French. The second trend leverages generative large language models (LLMs), which enable zero-shot readability assessment without requiring prior fine-tuning. This shift is exemplified by the promising results reported by Jamet et al. (2024).

As regards the automatic detection of local reading difficulties, the task addressed in this study does not precisely align with any existing standard task. It is most closely related to Complex Word Identification (CWI) (Paetzold and Specia, 2016; Yimam et al., 2018), later renamed as Lexical Complexity Prediction (LCP) (Shardlow et al., 2021, 2022, 2024), in that it aims to predict challenging tokens for readers. However, while CWI and LCP are limited to the detection of complex words or multi-word expressions, our approach encompasses a broader spectrum of reading difficulties. Nevertheless, techniques developed for LCP remain highly relevant to our task. As described in North et al. (2023), the field has broadly followed the general trajectory of natural language processing (NLP), transitioning from classical machine learning and ensemble approaches to deep learning and transformer-based models. For French, Tack et al. (2016) proposed personalized CWI models using support vector machines and neural networks. More recently, several studies have explored the potential of generative large language models (LLMs) for CWI in zero-shot and few-shot configurations (Zaharia et al., 2020; Ortiz-Zambrano et al., 2024), including for French (Kelious et al., 2024).

In terms of dataset comparison, existing French readability datasets such as Naous et al. (2024), Hernandez et al. (2022) and Blandin et al. (2020a) target different audiences—primarily children or second-language learners—and include genres intended either for general readers (e.g., Wikipedia, news, research, literature, legal texts) or for schoolchildren (e.g., children's stories and textbooks). In contrast, most of the texts in our dataset were sourced from trainers working with low-literacy adults, making them more tailored to this specific population. Furthermore, our dataset provides both global difficulty levels and segment-level annotations that highlight specific linguistic phenomena contributing to reading difficulty.

## 3 Data Set

To develop a readability model tailored to adults with low-literacy, we first compiled a corpus of texts and assessed their difficulty levels for our target audience. This dataset consists of 461 texts representative of 11 different types of communication (personal, professional, business, academic, political, legal, religious, social media, as well as fiction, non-fiction and didactic books). The texts are short (ranging from 18 to 387 words), as they are intended for an adult low-literacy audience; the distribution of text lengths is shown in Appendix A.3. Most of the texts were sourced from trainers working with low-literate adults, supplemented by materials retrieved from the web. We then conducted an annotation campaign to assess the level of difficulty of each text (**global annotations**) as well as to identify specific segments and features that may present challenges for readers from this level (**local annotations**).

We adopt a difficulty scale specifically tailored for adults with low literacy, as proposed by Monteiro et al. (2023). This scale was informed by the CEFR, due to its widespread recognition and compatibility with other grading systems, while the use of simple, familiar labels ensures accessibility for the intended users. The authors examined text complexity dimensions from the literature and refined them in collaboration with professionals working with low-literate adults, resulting in a list of 79 descriptors organized across four levels: `Very Easy`, `Easy`, `Plain` and `+Complex` (see Appendix A.4 for definitions). A mapping to CEFR was established for alignment with proficiency frameworks.

We began the annotation campaign by developing an annotation guide including descriptions and examples, which was iteratively reviewed until all contributors reached consensus. A first annotation phase was then conducted via the Qualtrics

platform[2], accompanied by training sessions for annotators. Their feedback led to two key revisions: the introduction of a 5-point Likert scale to allow finer-grained judgments within difficulty levels, and the adjustment of certain labels. The Likert scale provides an ordinal refinement of categorical judgments. A score of 1 corresponds to the lower boundary of a given difficulty class, while a score of 5 corresponds to its upper boundary. For instance, if an annotator selects Easy and assigns a score of 1, this indicates that they consider the example to be at the lower boundary of the `Easy` category, bordering on `Very Easy`. Conversely, a score of 5 reflects a case at the upper boundary of `Easy`, closer to `Plain`. Then, a second pre-test confirmed the usefulness of these adjustments. Although results varied only slightly, annotators reported improved ease of use, leading to the adoption of the revised version. The texts were first annotated for global difficulty, using the four-level difficulty scale and each text was assigned both a categorical level and a numerical score from 1 to 20, calculated by adding the Likert scale value (from 1 to 5) to the global level, converted as follows: `Very Easy` (0), `Easy` (5), `Plain` (10), `+Complex` (15). Using both categorical labels and numerical scores allows us to model the ordinal structure of the task and to model it both with classification and regression approaches. For the local annotation, annotators followed the guidelines to highlight specific words or expressions that could pose difficulties for readers within the assigned global level. We define nine difficulty classes, which cover lexical, syntactic, semantic, and structural sources of complexity:

- Difficult or unknown word,
- Spelling or decoding problem,
- Figure of speech, idiomatic expression,
- Difficult cultural reference,
- Grammar-related difficulty,
- Difficult cohesion cue (connector, pronoun, inference),
- Too much secondary information,
- Unusual syntactic order,
- Other.

This two-step annotation process was carried out by 15 professionals active in the field of illiteracy Each text was annotated by at least three annotators and we provide all individual annotations, with each linked to its corresponding anonymized annotator and text. Texts were grouped into sets of approximately 16 items and annotators could choose how many sets to complete; some annotated only one set, while others contributed up to 28. For the global annotations, we established reference values for both scales by averaging the three annotations[3]. For the local annotations, as the complex tokens were relative to the global level annotated by each annotator, the reference consider that a token is complex if at least one annotator deemed it complex, following Paetzold and Specia (2016).

At the end of this annotation process, we obtained 461 annotated texts both at the global and local level. Table 1 presents the distribution of annotated texts across the different difficulty levels. Unfortunately, it is unbalanced, as around 85 % of the texts were assigned to the `Easy` and `Plain` categories. Additional details on annotation and data analysis can be found in Appendix A.

| Very Easy | Easy | Plain | +Complex |
|---|---|---|---|
| 19 | 212 | 198 | 32 |

Table 1: Distribution of texts by difficulty category

## 4 Global Readability Assessment

In this study, we investigated multiple approaches for global text readability assessment, as illustrated in Figure 1, including classical machine learning models, deep learning methods (a fine-tuned CamemBERT), a hybrid model and generative large language models using zero and few-shot prompting techniques. Based on the two previously defined global difficulty scales—one consisting of discrete classes and the other their corresponding 1–20 ordinal scores—we approached the automatic readability evaluation task from two perspectives: classification and regression. The classification framework facilitates the assignment of functional difficulty levels relevant to practical applications, whereas the regression approach enables a more nuanced and continuous assessment, taking into account the ordinal nature of the task, thereby mitigating limitations inherent to rigid class boundaries.

Model performance was estimated using 5-fold cross-validation[4] with stratified sampling. To mitigate the impact of class imbalance, we applied class-frequency weighting to the loss func-

[2] https://www.qualtrics.com

[3] In the dataset, we also provide an alternative calculation method based on the most represented label, with the average used when all labels differ.

[4] 60% of the data for training, 20% for hyperparameter search and 20% for testing.

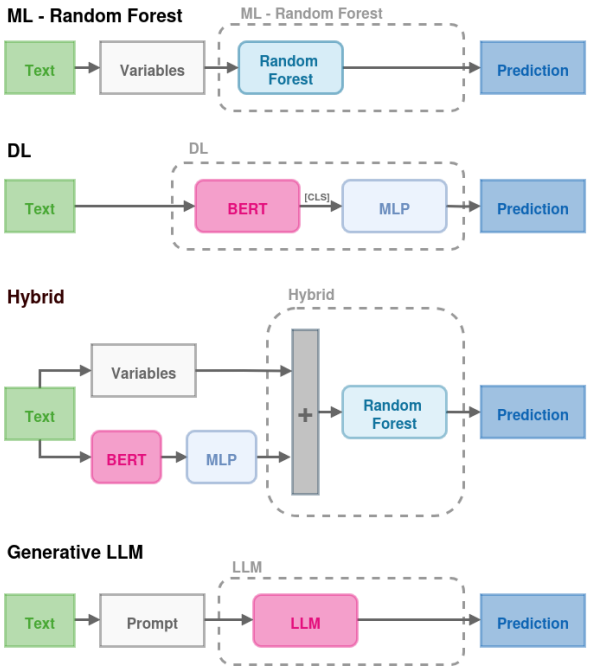


Figure 1: Architecture of global readability assessment systems.

tion, assigning higher penalties to underrepresented classes during training.

### 4.1 Machine Learning Models

Machine learning (ML) modelling involves converting texts into numerical features. For this, we used FABRA (Wilkens et al., 2022)[5], a tool that provides numerous linguistic descriptors relevant to readability assessment at the document, sentence, and word levels for French. For each descriptor, FABRA provides up to 20 statistical aggregators (e.g., mean, median, mode, 80th percentile), resulting in a very large set of features. We thus performed automated feature selection using the minimum Redundancy Maximum Relevance (mRmR) algorithm (Ding and Peng, 2003), which identifies the most informative and least redundant subsets of features. We compared feature subsets ranging in size from 10 to 500 features, which were combined using support vector machines (**SVM**), decision trees (**DT**), and random forests (**RF**) for the classification task, along with their corresponding regression models. In addition, because text readability inherently follows a graded progression, we also experimented with ordinal regression (**OrdR**) models (Pedregosa-Izquierdo, 2015). Technical details about each of our models are described in Appendix B.

[5]A comprehensive list of variables is available at `https://cental.uclouvain.be/fabra`.

### 4.2 Deep Learning Models

For deep learning (DL), we fine-tuned two pre-trained BERT-type transformers for French, namely **CamemBERT** (Martin et al., 2020) and **CamemBERT-v2** (Antoun et al., 2024) on our dataset (Section 3). This approach leverages the rich linguistic knowledge encoded within the transformer's internal layers while specializing the model for readability prediction. For classification tasks, we employed cross-entropy loss, whereas for regression we used mean squared error (MSE). Additionally, we applied the Ordinal Log-Loss (Lim and Lee, 2024) to account for the ordinal nature of the task, enabling the model to better capture the relative ordering between classes. Technical details are provided in Appendix B.

### 4.3 Hybrid models

Our hybrid models are based on the Soft-Label (SO) architecture proposed by (Lee et al., 2021) and later adopted in (Wilkens et al., 2024), which demonstrated strong performance across four different datasets. In this approach, the softmax output of a deep learning model is concatenated with a set of linguistic features, and the resulting representation is used as input to a classical machine learning model for predicting text difficulty. Based on the observations obtained from experiments with both machine learning and deep learning models, as reported in Table 2, we constructed hybrid model inputs by concatenating the linguistic variables selected by the mRMR algorithm with the output of the fine-tuned CamemBERT-v2 model. These concatenated feature representations were then fed to a random forest (RF) model to predict text difficulty. The training and evaluation strategy follows the approach described in Section 4.1.

### 4.4 Generative Large Language Models

Due to their broad generalization capabilities across a wide range of NLP tasks, we investigated the use of generative LLMs for global text difficulty assessment. Their ability to operate effectively in zero-shot and few-shot settings makes them attractive for tasks with limited annotated data. We evaluated various LLMs—spanning open- and closed-weight models, as well as instruction- and reasoning-tuned variants—to explore their potential and shortcomings in this context.

To construct our prompts, we adapted the formulation proposed by Jamet et al. (2024), which

guides the model to produce CEFR scores. We used the corresponding CEFR levels rather than our custom scale, as they yielded better performance—likely due to the model's prior exposure during training. Using the mapping defined in our dataset, the CEFR levels were aligned with our annotation scheme as follows: `A1` → `Very Easy`, `A2` → `Easy`, `B1` → `Plain`, and `B2 and above` → `+Complex`. We evaluated two instruction languages in the prompts given to the LLMs, French and English, in order to analyze the impact of the interaction language on the model's performance. For both zero-shot and few-shot prompting, the input prompt includes a Chain-of-Thought (CoT) reasoning example based on a toy difficulty classification example to provide minimal guidance to the model's reasoning process. In the few-shot experiments, we selected examples representative of each difficulty level, ensuring consensus among annotators. Example prompts are provided in Appendix D.

### 4.5 Results & Analysis

The performance of the models trained on our dataset (Section 3) was evaluated using several metrics. For classification models, we employed accuracy, adjacent accuracy[6], and macro-F1 score, which accounts for class imbalance. For regression models, performance was assessed using mean squared error (MSE). To fairly compare classification and regression models, we follow previous work (Ribeiro et al., 2024; Wilkens et al., 2022) and discretize continuous regression scores to the corresponding readability classes (`Very Easy` $\leq$ 5, `Easy` 5–10, `Plain` 10–15, `+Complex` $>$ 15), enabling computation of macro-F1 scores for regression predictions. Macro-F1 results are shown in Table 2, with full results in Table 11 (Appendix E).

The results highlight clear differences between ML, DL, and hybrid models across tasks. Among ML classifiers, RF achieves the strongest results with 62.77% accuracy and 98.05% adjacent accuracy, though macro-F1 remains similar to SVM, indicating that its improvement in accuracy does not necessarily translate into better handling of underrepresented classes. ML models generally show higher accuracy than macro-F1, reflecting a tendency to favor majority classes; Applying class weighting helped reduce this gap but does not eliminate it. DL models consistently outperform ML, with CamemBERT and CamemBERT-v2 reaching around 64% accuracy and macro-F1 near 60%, while the ordinally trained CamemBERT-OLL further improves performance, achieving the best macro-F1 (61.14%) and adjacent accuracy (99.78%), in line with previous findings in (Lim and Lee, 2024). Hybrid models attain the highest classification accuracy (67.32%) but lower macro-F1 (56.26%), suggesting improved overall correctness at the expense of class balance. In regression, ML baselines perform poorly with accuracies around 40% and very low macro-F1, though the ordinal regressor OrdR achieves a much lower MSE (0.84%), highlighting the advantage of ordinal constraints. DL regressors outperform all ML and hybrid models, with CamemBERT reaching 70.77% accuracy, 59.63% macro-F1, and perfect adjacent accuracy, while CamemBERT-v2 attains slightly lower accuracy. Hybrid regression shows competitive adjacent accuracy (99.57%) but underperforms DL in macro-F1 and MSE. Overall, DL provides the strongest and most balanced performance, ordinal-aware methods slightly improve both ML and DL, and hybrid approaches mainly improve raw accuracy without fully capturing class balance or minimizing regression error.

| | Model | Macro-F1 |
|---|---|---|
| **Classification** | ML - SVM (500) | 47.54 ± 6.15 |
| | ML - DT (300) | 43.84 ± 4.98 |
| | ML - RF (400) | 47.78 ± 7.60 |
| | DL - CamemBERT | 60.36 ± 8.23 |
| | DL - CamemBERT-v2 | 60.05 ± 6.01 |
| | DL - CamemBERT-OLL | **61.14 ± 4.30** |
| | Hybride - RF (300) | 56.26 ± 9.17 |
| **Regression** | ML - SVR (500) | 22.63 ± 1.96 |
| | ML - DT (50) | 22.13 ± 3.15 |
| | ML - RF (500) | 22.96 ± 4.71 |
| | ML - OrdR (50) | 22.70 ± 2.30 |
| | DL - CamemBERT | **59.63 ± 2.55** |
| | DL - CamemBERT-v2 | 47.52 ± 8.36 |
| | Hybrid - RF (300) | 36.50 ± 5.21 |

Table 2: Comparison of macro-F1 performance for readability classification and regression models. Parentheses denote the number of selected features.

Concerning generative LLMs, Table 3 presents the classification performance using different prompting strategies. To assess performance variability, the dataset was partitioned into five folds, and we computed the average accuracy, adjacent accuracy, and macro-F1 score for each fold,

[6]Adjacent accuracy considers a prediction correct if it matches the true class or a neighboring class on the ordered scale (e.g., predicting `Very easy` instead of `Easy`).

| Model | EN-zero-shot | EN-few-shot | FR-zero-shot | FR-few-shot |
|---|---|---|---|---|
| DeepSeek-70b | 29.64 ± 8.22 | **48.95 ± 4.85** | **38.41 ± 3.14** | **47.06 ± 3.55** |
| Gemma-27b | **30.19 ± 7.97** | 41.12 ± 4.45 | 29.8 ± 6.46 | 39.49 ± 3.76 |
| Qwen-72b | 17.65 ± 4.43 | 46.93 ± 3.53 | 19.32 ± 11.13 | 42.9 ± 5.33 |
| Mistral-large | 22.78 ± 5.34 | 48.6 ± 4.08 | 26.72 ± 4.24 | 43.86 ± 3.57 |
| GPT-4.1 | 27.01 ± 6.04 | 44.69 ± 6.84 | 35.55 ± 9.34 | 43.02 ± 7.82 |

Table 3: Comparison of the macro-F1 performance of the generative LLM for global difficulty classification.

along with their corresponding standard deviations. We also provide additional results in the appendix Table 12. We observe a positive correlation between model size and performance, for example, DeepSeek-7b achieves 25.6% macro-F1 in EN zero-shot, while DeepSeek-70b reaches 29.64%, jumping to 48.95% with few-shot prompting, making it the overall top performer across both languages. Gemma-27b performs competitively, achieving 30.19% (EN-zero-shot) and 41.12% (EN few-shot), while Qwen-72b exhibits a strong few-shot jump, especially in EN despite low zero-shot performance, suggesting high sensitivity to few-shot guidance. Mistral-large and GPT-4.1 also show strong few-shot improvements, reaching nearly 48% macro-F1 for EN and around 43% for FR. In zero-shot settings, French prompts outperform English ones, likely due to the fact that French prompts might match more precisely the internal representation of French linguistic complexity in the model. In contrast, few-shot settings favor English, possibly because instruction tuning is more robust in English. These results highlight the importance of model scale and in-context examples for effective global difficulty classification.

In contrast to DL models, generative LLMs generally exhibit lower macro-F1 but achieve promising performance without task-specific training. This capacity for generalization without direct supervision highlights their potential, although they currently do not surpass fine-tuned models.

## 5 Local Readability Assessment

From the data collected on local text difficulties (see Section 3 and Appendix A.5 for details about the labels), which provides fine-grained annotations of words and structures considered challenging for readers at specific proficiency levels along with the reasons for their difficulty, we explore the ability of language models to replicate expert judgments of local complexity. We used two different strategies, fine-tuning and zero-shot prompting. Specifically, we fine-tuned CamemBERT for token classification on our dataset, complemented by generative LLMs applied in a framework inspired from the LCP task. During annotation, human annotators labeled words they deemed difficult for readers at specific proficiency levels (our global annotations). To replicate this, we provide these global difficulty levels as part of the input to condition the model's behavior accordingly.

To assess the ability of language models to identify local reading difficulties, we design two experimental settings: binary classification and multi-label classification. The binary setup allows to evaluate the model's discriminative ability in distinguishing difficult from non-difficult words and structures. The multi-label setup enables a more fine-grained assessment by assigning one or more specific difficulty types to each word or structure.

### 5.1 Deep Learning Models

To establish a supervised learning baseline, we employed deep learning models based on pre-trained language representations. In particular, we used CamemBERT fine-tuned for token classification to detect difficult words and structures within the input text. To condition the models' predictions on global difficulty, each text is prefixed with its global difficulty level, allowing the model to tailor its identification of difficult words accordingly. In the binary setup, the model classifies each token as either difficult or not, with positive labels assigned to tokens representing challenging words or structures, regardless of the specific difficulty type. The model is trained using cross-entropy loss and predicts a class for each token—either positive or negative—by applying the argmax to the output probabilities. In the multi-label setup, the difficulty types are used as class labels for each token, allowing the model to assign multiple difficulty categories to a single token when applicable. We used a binary cross-entropy (BCE) loss, and consider a class to be positive if its output is greater than a threshold. Otherwise, the class is considered negative. All models are trained with grid search

using the same hyperparameters as global difficulty assessment. Technical details are mentioned in Appendix B.

### 5.2 Generative Large Language Models

We evaluate the zero-shot capability of generative LLMs to identify difficult words and structures within a text, conditioned on the reader's proficiency level. We frame the problem as a Lexical Complexity Prediction (LCP) task (North et al., 2023), which involves assessing the difficulty of a target word in a given context. Grounded in recent research on the use of generative LLMs for LCP tasks (Ortiz-Zambrano et al., 2024; Kelious et al., 2024), we investigate whether such models can reliably indentify difficult words for our target population without fine-tuning. To mirror the human-annotators setup, we prompt the model using the full text, the associated global readability level, and a list of target words. The model is then prompted to assess which of the target words are likely to present difficulty for a reader at the specified proficiency level. For evaluation, words and structures annotated by humans as difficult are treated as positive instances. To reduce computational overhead and leverage the structured output format of generative LLMs, we apply undersampling directly from the same text for negative instances (non-difficult words), ensuring a balanced set of positive and negative instances per text[7]. In the multi-label setup, we prompt the model to assign each candidate word or structure to one or more classes corresponding to predefined difficulty types, or to a special non-difficulty class if the model determines that the word is not difficult. The prompt templates are provided in Appendix D.

### 5.3 Results & analysis

| Difficulty type | BERT-Binary |
|---|---|
| No difficulty | 87% |
| Difficult | 43% |
| **Macro-F1** | 65% |

Table 4: F1 scores (%) per local difficulty label for CamemBERT model in binary settings.

The binary classification (Table 4) performs well, achieving a macro-F1 of 65%. In contrast, the multi-label setup (Table 5) exhibits weak performance, this can be attributed to the inherent complexity of the task and the subtle, overlapping patterns of linguistic difficulty present in the text. The performance drop is also likely influenced by dataset imbalance, particularly for low-frequency difficulty types (*e.g.* `Grammar Difficulties`). While the model achieves the best performance on the `Difficult or unknown word` class, it struggles significantly with less frequent and more abstract categories such as `Difficult cohesion cue`, despite weighting the loss by class frequency. This underscores the need for more targeted research on modeling fine-grained language complexity, particularly in the presence of class imbalance and subtle contextual cues.

| Difficulty type | BERT-Multilabel |
|---|---|
| No difficulty | 84% |
| Grammar difficulties | 9% |
| Figure of speech, idiomatic expression | 6% |
| Spelling or decoding problems | 13% |
| Difficult cohesion index | 7% |
| Difficult or unknown word | 16% |
| Unusual syntactic order | 10% |
| Difficult cultural reference | 8% |
| Too much secondary information | 10% |
| **Macro-F1** | 18% |

Table 5: F1 scores (%) per local difficulty label for CamemBERT model in multilabel settings.

In table 6, we present the performance results of several generative LLMs on our task. Full results for the binary LLM across the different global difficulty levels are provided in Table 13 in Appendix E. For the binary setup, GPT-4.1 consistently outperforms the other models, achieving the highest marcro-F1 score of 76.7% and an accuracy of 76.8%, and demonstrating superior performance across all difficulty levels, particularly on `Easy` texts (76.93%) and `Plain` (76.71%). This suggests a strong capacity for handling increasing linguistic complexity, likely due to its high-quality instruction tuning and extensive coverage of diverse textual data. DeepSeek-R1 ranks second, showing robust results across all levels, especially `Plain` (75.29%). Its retrieval-augmented and alignment-aware architecture may contribute to its effective handling of localized difficulty. Mistral-large performs moderately, with stable but less adaptive performance, particularly struggling to capture patterns in `Plain` and `+Complex` levels compared to GPT-4.1. Finally, Qwen2.5 shows the weakest overall performance (F1: 68.12%, accuracy: 68.47%), indicating potential limitations in linguistic generalization for French language compared to the rest of the models. Overall, the top-performing

[7]To control for potential length biases, the negative samples are selected such that their word lengths match those of the positive instances.

| Difficulty type | Mistral-large | GPT-4.1 | Qwen2.5 | DeepSeek-R1 |
|---|---|---|---|---|
| ***Binary*** | | | | |
| **Macro-F1** | 71.34 | **76.7** | 68.12 | 74.73 |
| ***Multilabel*** | | | | |
| Grammar difficulties | 35.76 | **51.27** | 38.17 | 44.82 |
| Figure of speech, idiomatic expression | 43.98 | **58.62** | 46.20 | 54.43 |
| Spelling or decoding problems | 9.93 | **22.68** | 17.18 | 18.61 |
| Difficult cohesion index | 33.15 | **47.78** | 25.76 | 24.70 |
| Difficult or unknown word | 63.80 | **69.61** | 64.33 | 67.90 |
| Unusual syntactic order | 12.31 | 24.78 | 17.73 | **27.23** |
| Difficult cultural reference | 47.60 | **55.99** | 47.30 | 49.63 |
| Too much secondary information | 29.98 | **38.10** | 36.46 | 36.34 |
| **Macro-F1** | 34.56 | **46.10** | 36.64 | 40.46 |

Table 6: F1 scores (%) per local difficulty label, with average across all labels for generative LLMs.

models demonstrate strong multilingual capabilities, reflecting the ability of modern LLMs to adapt to French text.

The results on multi-label local difficulty assessment (see Table 6; full results in Table 14, Appendix E), show that GPT-4.1 outperforms the other models across most difficulty types, achieving a macro-F1 score of 46.1%. Although this score remains moderate, it supports the model's strong performance in the binary local difficulty assessment and indicates good discriminative capabilities across various linguistic challenges. DeepSeek-R1 follows with a macro-F1 of 40.46%, showing good results in culturally and lexically challenging cases. In contrast, Mistral-large and Qwen2.5 perform moderately (macro-F1 around 34–37%) and struggle particularly with `unusual syntactic order` class and rare difficulty types. Overall, all models perform best on identifying difficulties related to lexical access—such as `difficult or unknown word`—while they struggle most with structural and low-frequency phenomena like `spelling or decoding problems` and `unusual syntactic order`, indicating these aspects remain more challenging to model. These findings suggest that zero-shot generative LLMs offer a promising strategy for local difficulty assessment, particularly when compared to CamemBERT fine-tuning approaches in contexts where training data is scarce. They therefore warrant further investigation to better understand their strengths and limitations across diverse linguistic phenomena.

## 6 Conclusion and Perspectives

This study systematically evaluates multiple approaches for classifying the difficulty of French texts specifically designed for adults with low literacy, it examines both global and local text difficulty. The analysis compared the effectiveness of four different approaches to readability, providing insights into their relative performance and suitability for this target population. For global difficulty classification, the results indicate that hybrid and deep learning models generally outperform other approaches in terms of accuracy and robustness. In contrast, LLMs exhibit comparatively lower effectiveness in this specific context. Nevertheless, these findings suggest potential for improvement: fine-tuning a generative LLM on a dedicated readability task could be a promising direction for future research. At the local difficulty assessment level, both generative and fine-tuned large language models (LLMs) demonstrate good comparable performance in a binary classification setup—specifically, when identifying whether a given token is difficult or not. However, their effectiveness decreases in the more challenging multi-label token difficulty classification task. In this setting, generative LLMs show better performance than fine-tuned models, though both approaches exhibit notable limitations. These results highlight the need for further refinement to improve token-level difficulty modeling on our dataset. Text readability assessment is inherently subjective, and annotator subjectivity can lead to inter-annotator disagreement. Despite this, the results suggest that averaging global text difficulty annotations yields more consistent and reliable data for training classification models for global text difficulty assessment. This aggregation appears to mitigate individual biases and enhance model performance. To further improve annotation quality—particularly for applications targeting adults with low literacy—a more in-depth comparative analysis of annotator behavior would be beneficial. Such an analysis could inform strategies for standardizing annotations in cases of substantial discrepancy, ultimately leading to more robust and inclusive readability assessment frameworks.

## Limitations

There are a number of limitations to our dataset and analysis, in addition to those already discussed in the paper. First, although we attempted to anticipate dataset imbalance during corpus construction by using an automated readability analysis tool from our previous work (François, 2015), which was designed for generic second-language learners, to select a balanced set of French texts, the final class distribution was subsequently modified by the human annotation process. Annotators assigned scores based on their expert judgment of readability for low-literacy adult audiences. This change in audience, combined with the inherently subjective nature of assessing text readability, explains the imbalance observed in our dataset. In fact, the task of assessing text readability remains inherently subjective. Expert annotations are influenced by a variety of factors, including individual backgrounds, interpretation of difficulty, and prior experience with low-literacy adult populations. This introduces variability that is difficult to fully control. Second, the number of texts included in our dataset (461) is relatively small. This is largely due to the challenges involved in collecting and annotating texts that are representative of the linguistic and cognitive profiles of low-literate adults. Each text required expert-level, fine-grained annotation of lexical difficulty types, making the process resource-intensive. Finally, we do not include direct comparisons with existing readability models and datasets. This choice stems from the poor transferability of models fine-tuned on other datasets to our own. Preliminary experiments revealed a lack of generalization, likely due to differences in annotation protocols, target populations, and text types. As a result, we focus our evaluation and modeling efforts on our dataset, which is specifically tailored to french low-literacy adults.

## Acknowledgments

This research/Part of this research/The research of (name of researcher) is supported by the European Commission (Project: iRead4Skills, Grant number: 1010094837. [,Topic: HORIZONCL2- 2022-TRANSFORMATIONS-01-07, DOI:10.3030/101094837]). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We would like to thank the ANLCI, the centers *Savoirs pour réussir Paris* and *Poinfor* as well as Alice Pintard and all the annotators for their invaluable contribution to the project.

## References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11(9):2063.

S. Aluisio, L. Specia, C. Gasperin, and C. Scarton. 2010. Readability assessment for text simplification. In *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles.

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. *arXiv preprint arXiv:2411.08868*.

Yonathan A Arbel. 2024. The readability of contracts: Big data analysis. *Journal of Empirical Legal Studies*, 21(4):927–978.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, and Aline Étienne. 2020a. Recommandation d'âge pour des textes (age recommendation for texts). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 164–171, Nancy, France. ATALA et AFCP.

Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, and Aline Etienne. 2020b. Recommandation d'âge pour des textes (age recommendation for texts). In *Actes de la 6e conférence conjointe Journées*

*d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles*, pages 164–171.

M. Cha, Y. Gwon, and H.T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006. ACM.

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

E. Dale and J.S. Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

E. Dale and R.W. Tyler. 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4:384–412.

M. Dascalu. 2014. Readerbench (2)-individual assessment through reading strategies and textual complexity. In *Analyzing Discourse and Text Complexity for Learning and Collaborating*, pages 161–188. Springer.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

F. Dell'Orletta, S. Montemagni, and G. 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.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

C. Ding and H. Peng. 2003. Minimum redundancy feature selection from microarray gene expression data. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 523–528.

Sabrina Dittrich, Zarah Weiss, Hannes Schröter, and Detmar Meurers. 2019. Integrating large-scale web data and curated corpus data in a search engine supporting german literacy education. In *Proceedings of the 8th workshop on NLP for computer assisted language learning*, pages 41–56.

William H DuBay. 2004. The principles of readability. *Online submission*.

L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *COLING 2010: Poster Volume*, pages 276–284.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, pages 335–348. Springer.

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

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, (2):79–97.

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

Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. Amesure: a web platform to assist the clear writing of administrative texts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 1–7.

W.S. Gray and B.E. Leary. 1935. *What makes a book readable*. University of Chicago Press, Chicago: Illinois.

Anke Grotlüschen, David Mallows, Stephen Reder, and John Sabatini. 2016. Adults with low proficiency in literacy or numeracy.

R. Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. 2022. Open corpora and toolkit for assessing text readability in French. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 54–61, Marseille, France. European Language Resources Association.

Henri Jamet, Maxime Manderlier, Yash Raj Shrestha, and Michalis Vlachos. 2024. Evaluation and simplification of text difficulty using llms in the context of recommending texts in french to facilitate language learning. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 987–992.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches.

In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 96–114, Rennes, France. LiU Electronic Press.

Tannon Kew and Sarah Ebling. 2022. Target-level sentence simplification as controlled paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42.

B Suresh Lal. 2015. The economic and social cost of illiteracy: an overview. *International Journal of Advance Research and Innovative Ideas in Education*, 1(5):663–670.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ho Hung Lim and John Lee. 2024. Improving readability assessment with ordinal log-loss. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 343–350, Mexico City, Mexico. Association for Computational Linguistics.

Fengkai Liu and John SY Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Nicholas Mcinnes and Bo JA Haglund. 2011. Readability of online health information: implications for health literacy. *Informatics for health and social care*, 36(4):173–189.

Erick Messias. 2003. Income inequality, illiteracy rate, and life expectancy in brazil. *American Journal of Public Health*, 93(8):1294–1296.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2023. iread4skills - complexity levels. Version 1.0.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 45–55.

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 Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 12230.

Diane Napolitano, Kathleen Sheehan, and Robert Mundkowsky. 2015. Online readability and text complexity analysis with TextEvaluator. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, Denver, Colorado. Association for Computational Linguistics.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Jenny Alexandra Ortiz-Zambrano, César Humberto Espín-Riofrío, and Arturo Montejo-Ráez. 2024. Enhancing lexical complexity prediction through few-shot learning with gpt-3. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 68–76, Torino, Italia. ELRA and ICCL.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fabian Pedregosa-Izquierdo. 2015. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI.

Qi Qin, Wenpeng Hu, and Bing Liu. 2020. Feature projection for improved text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8161–8171.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Text readability assessment in European Portuguese: A comparison of classification and regression approaches. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 551–557, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Wibke Riekmann and Anke Grotlüschen. 2011. Konservative entscheidungen: Größenordnung des funktionalen analphabetismus in deutschland. *REPORT-Zeitschrift für Weiterbildungsforschung*, (3):24–35.

Jascha Russeler, Klaus Menkhaus, Annegret Aulbert-Siepelmeyer, Ivonne Gerth, and Melanie Boltzmann. 2012. " alpha plus": An innovative training program for reading and writing education of functionally illiterate adults. *Creative Education*, 3(3):357–361.

Horacio Saggion. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

S.E. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530.

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024. The bea 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

June Steward. 2023. The economic & social cost of illiteracy: A snapshot of illiteracy in a global context.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. In *JEP-TALN-RECITAL 2016*.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

UNESCO. 2009. Education indicators, technical guidelines. Technical report, UNESCO Institute for Statistics.

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.

Duy Van Ngo and Yannick Parmentier. 2023. Towards sentence-level text readability assessment for french. In *Second Workshop on Text Simplification, Accessibility and Readability (TSAR@ RANLP2023)*.

Zarah Weiss, Sabrina Dittrich, and Detmar Meurers. 2018. A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 79–90.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Meg Wilson. 2009. Readability and patient education materials used for low-income populations. *Clinical Nurse Specialist*, 23(1):33–40.

Kevin Yancey, Alice Pintard, and Thomas François. 2021. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, 2021(2):229–258.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Cross-lingual transfer learning for complex word identification. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 384–390. IEEE.

# A Data and annotation details

## A.1 Texts copyright compliance

We documented the source of each text included in our dataset[8]. Following consultation with our institution's legal advisor, we were informed that these texts may be used and shared for research purposes under the citation exception, as they consist of relatively short excerpts and the original sources are properly cited. Where necessary, measures will be implemented to ensure that the use of this data is restricted to teaching and research activities.

## A.2 Annotator Demographics and Compensation

All our annotators were women, although gender was not a selection criterion. They were paid 15€ per series, with the choice between a bank transfer or a bookshop gift voucher for a annotation session between 30 minutes and 2 hours.

## A.3 Data analysis

We first examine the corpus in terms of text length, as shown in Figure 2.

To better understand the variability of difficulty within each global level, we examine the distribution of the 1–20 numerical scores assigned to texts in Figure 3, which shows that each level does not present the same profile. For example, texts in the `Very Easy` category have the highest scores in their category, with the third and fourth quartiles merging. The other categories seem to have more homogeneous scores.

## A.4 Global labels of Difficulty

In the annotation guidelines provided to annotators, the global difficulty levels are described as follows:

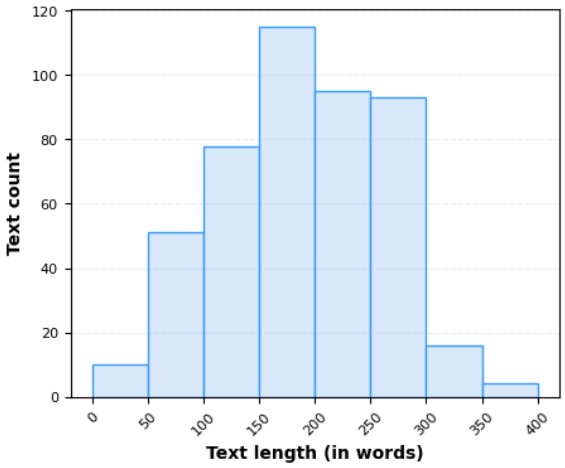


Figure 2: Distribution of corpus text lengths measured in words, using whitespace-based tokenization.

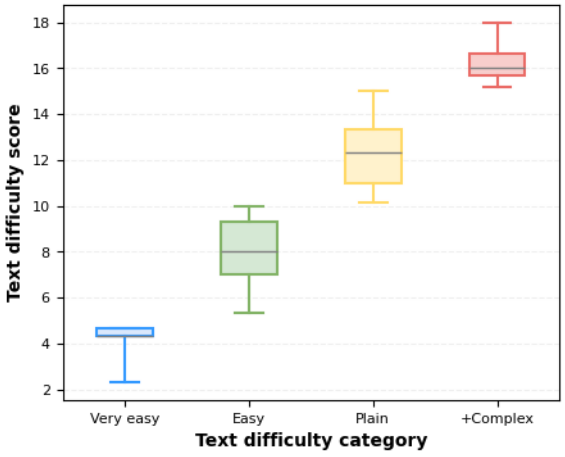


Figure 3: Box plot of text difficulty scores by category.

- **Very Easy**: Texts that are completely or almost completely understood by all readers, including those with a very low level of schooling (up to about sixth grade) and almost no reading experience. They are very short and deal with simple subjects, with a basic vocabulary.
- **Easy**: Texts fully or almost fully understood by people with a low level of schooling (i.e. having completed elementary school, but no more than the ninth year of education) and limited reading experience. These are short texts, which may include abstract concepts and common figures of speech.
- **Plain**: Texts comprehensible on first reading by individuals who have completed the ninth grade and have functional to average reading experience. These are longer texts, which may present more varied concepts, more complex syntactic structures and irregular verbs if they are very

[8] https://github.com/tfrancoiscental/iread4skills_readability_corpus_fr

frequent in the language.

- **+Complex**: Texts requiring more attentive reading and a certain linguistic mastery to be fully understood. They are aimed at readers with a more advanced level of schooling and more in-depth reading experience. This level includes all the more complex elements than those described for the previous categories.

### A.5 Local labels of difficulty

The local difficulty labels are described as follows:

- **Difficult or unknown word**: A word is considered difficult if it meets one of the following criteria:
  - A word whose meaning may not be well understood by the reader.
  - A word potentially absent from the reader's vocabulary, as it belongs to a specialized domain (e.g., technical, scientific, literary).
  - A word from a foreign language.
  - A word belonging to a very formal register.
  - An archaic word.
  - An expression where a single isolated word makes the entire expression difficult.
- **Spelling or decoding problem**: A word or expression is considered to pose a decoding problem if it meets one of the following characteristics:
  - A word whose spelling may hinder access to meaning, but which remains familiar orally.
  - Numbers written in a way that is difficult to read for the reader's CEFR level.
- **Figure of speech, idiomatic expression**: Figures of speech include, but are not limited to, metaphors, metonymies, personifications, and ironies. Idiomatic expressions are multiword units which, taken together, may not be understood literally.
- **Difficult cultural reference**: Cultural references include the reader's prior knowledge such as cultural, artistic, or literary references, as well as general or digital culture. A cultural reference is considered complex for a reader of a given CEFR level if it prevents comprehension.
- **Grammar-related difficulty**: Grammatical difficulties include, but are not limited to, problems with tense, mood, concord, passive voice, omission of determiners, etc.
- **Too much secondary information**: A sentence is considered overloaded with secondary information when such information may hinder comprehension. Secondary information is "the surplus" that could be removed or turned into a separate sentence. This includes, for example, asides, parentheses, and embedded subordinate clauses.
- **Difficult cohesion cue (connector, pronoun, inference)**: Difficult cohesion markers include issues related to the micro-structure of the text, such as challenging inferences and anaphoric references (pronouns), connectors (e.g., "all the same," "however," "rarely"), and other types of inference.
- **Unusual syntactic order**: An unusual word order occurs when deviation from the standard subject–verb–object order may cause comprehension difficulties.

| Local label | Count |
|---|---|
| Difficult or unknown word | 4297 |
| Spelling or decoding problem | 2121 |
| Figure of speech, idiomatic expression | 914 |
| Difficult cultural reference | 789 |
| Difficult cohesion cue | 731 |
| Unusual syntactic order | 560 |
| Grammar-related difficulty | 468 |
| Too much secondary information | 371 |
| Other | 181 |
| **Total** | **10432** |

Table 7: Distribution of the local labels of difficulty across the corpus.

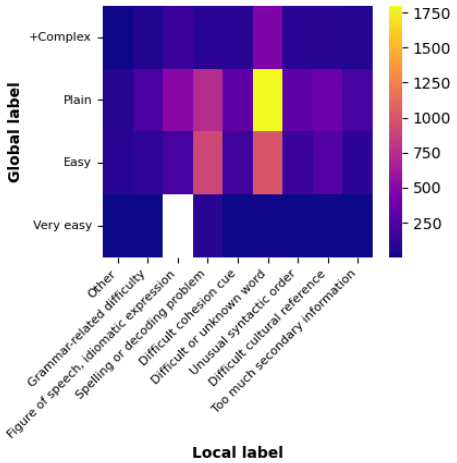


Figure 4: Distribution of local labels for each global difficulty category.

### A.6 Inter-annotator agreement

Inter-annotator agreement scores were assessed for both global and local annotations. Given the wide variation in annotator participation, we constructed three super-annotators by aggregating annotations and keeping together those from the same annotator.

| | Kappa | Spearman |
|---|---|---|
| Annot 1 vs. ref. | $0.28 \pm 0.12$ | $0.63 \pm 0.13$ |
| Annot 2 vs. ref. | $0.38 \pm 0.15$ | $0.69 \pm 0.13$ |
| Annot 3 vs. ref. | $0.54 \pm 0.11$ | $0.69 \pm 015$ |

Table 8: Inter-annotator agreement according to Weighted Quadratic Kappa and Spearman correlation for scores out of 20.

These three super-annotators comprise one, five and ten annotators respectively.

For the global level, we computed Cohen's $\kappa$ and Spearman's $\rho$ between each super-annotator and the reference, allowing to identify the best annotators. Table 8 reports those values for the 1-20 scores (with quadratic weighted kappa - QWK), ranging from 0.28 and 0.54, with an average per annotator and per series of 0.33. These results are admittedly low, but quite similar to the QWK obtained on a related task at SemEval 2012 (Specia et al., 2012). For the local annotations, where the boundaries of annotated phenomena may vary, we evaluated inter-annotator agreement using token-level macro F1 scores, following common practice for this type of task. We distinguished between two aspects: agreement on whether a token is complex (a binary classification task), and agreement on the assigned difficulty category (in a multilabel setting, as more than one class can be assigned to the same token). Macro-F1 scores for the binary classification range from 0.63 to 0.69 (see Table 9). This suggests that annotators generally identified similar reading difficulty within the texts. However, agreement on the type of difficulty was notably lower, with micro-F1 scores ranging from 21.7 to 18.5. It is important to note that local annotations are relative to the global difficulty level assigned to the text, disagreement may stem from different global annotations.

| Comparaison | Binary | Multilabel |
|---|---|---|
| Annot1 vs Annot2 | 61.03 | 21.71 |
| Annot2 vs Annot3 | 59.22 | 18.73 |
| Annot1 vs Annot3 | 60.07 | 18.44 |
| Average | 60.11 | 19.63 |

Table 9: Macro-F1 for pairs of annotators for the local annotation seen as a binary task and a multilabel task.

### A.7 Qualtrics interface for trainers


Figure 5: Qualtrics interface used for selecting the global difficulty of texts.

**Text 4**

*Hi, how are you? Can you let me know if you could cover for me next Tuesday plz? Thanks for your reply.*

- Very Easy
- Easy
- Plain
- Complex

1 2 3 4 5

Beginning of level — End of level

Figure 6: English translation of the global difficulty selection interface (see Figure 5 for the original French version).

## B Technical details

### B.1 Machine Learning models for global difficulty assessment

For machine learning models, we used a grid search to explore the set of hyperparameter configurations for our different models. Our implementation is based on the Scikit-learn library (Pedregosa et al., 2011). To address class imbalance in the classification setting, the cost function was weighted according to class frequency.

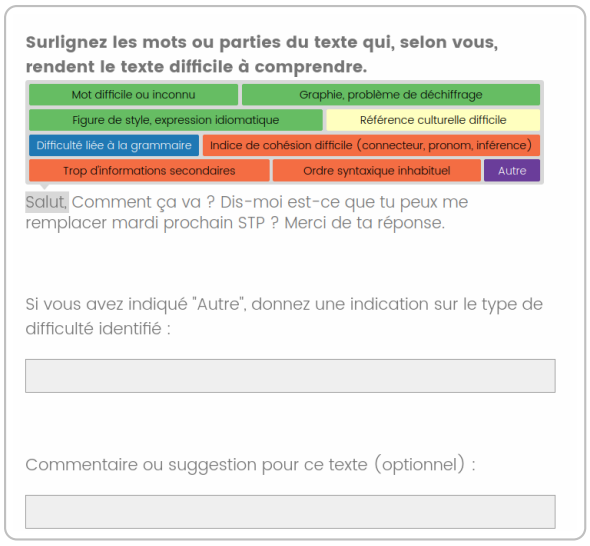


Figure 7: Qualtrics interface used for selecting the local difficulty of texts.

**Highlight the words or parts of the text that, in your opinion, make the text difficult to understand.**

- Difficult or unknown word
- Spelling or decoding problem
- Figure of speech, idiomatic expression
- Difficult cultural reference
- Grammar-related difficulty
- Difficult cohesion cue (connector, pronoun, inference)
- Too much secondary information
- Unusual syntactic order
- Other

*Hi, how are you? Can you cover for me next Tuesday plz? Thanks for your reply.*

**If you selected "Other", please provide an explanation of the type of difficulty identified:**

`[Free text field]`

**Comment or suggestion for this text (optional):**

`[Free text field]`

Figure 8: English translation of the interface for local annotations (see Figure 7 for the original French version).

Here is the list of hyperparameters by task and model type:

- **SVM**[9]: We evaluated three kernel types (`linear`, `rbf`, `sigmoid`). For the regularization parameter `C`, we tested the following values: 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100.
- **DT**[10]**:** Among the objective functions, we tested `gini`, `entropy`, and `log_loss`. For the maximum tree depth, we used values from 1 to 8. The minimum number of samples required to split an internal node was selected from 2, 3, 5, 10, 15, and 20.
- **RF**[11]**:** The number of trees in the forest could take values between 10, 20, 30, 40, 50, 100, and 200. For the objective function, the maximum tree depth, and the minimum number of samples, we explored the same values as for the decision trees.
- **SVR**[12]: We evaluated three kernel types (`linear`, `rbf`, `sigmoid`). For the regularization parameter `C`, we tested the following values: 0.001, 0.01, 0.1, 1, 10, 100.
- **DTR**[13]: We tested different objective functions: `squared error`, `friedman mse`, `absolute error`, and `poisson`. The maximum tree depth was set to 1, 5, 7, or 8, and the minimum number of samples required to split an internal node was set to 2, 5, 10, or 20.
- **RFR**[14]: The number of trees in the forest (`n estimators`) could be 10, 30, 50, or 200. The other parameters (`criterion`, `max depth`, `min samples split`) were explored with the same values as for the decision trees.
- **OrdR**[15]: We evaluated four ordinal regression methods from the `mord` library—threshold-based and regression-based: `LogisticAT`, `LogisticIT`, `OrdinalRidge`, and `LAD`. All models were tuned via grid search. For the first three, the regularization parameter $\alpha$ was tested with values 0.1, 1.0, and 10.0. For `LAD`, the

[9] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[11] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[12] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
[13] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[15] https://pythonhosted.org/mord/

inverse regularization $C$ and optimization tolerance `tol` were tested with 0.1, 1.0, 10.0 and $10^{-2}, 10^{-3}$, respectively. We report resulss for the `OrdinalRidge` variant which slightly outperforms the others.

### B.2 Deep Learning models for difficulty assessment

The **fine-tuned CamemBERT models** considered in this work include `camembert-base`[16], `camembertv2-base`[17], and `sentence-camembert-base`[18]. The AdamW optimizer (Loshchilov and Hutter, 2019) was employed to update the model parameters, with all layers of the language model fine-tuned without freezing the intermediate representations. To prevent overfitting and reduce training time, early stopping was applied with a patience of 10 epochs. An exhaustive grid search was conducted to identify optimal hyperparameters. Different values were explored for the following parameters: the learning rate, tested with values of 1e-5, 1e-4, 1e-3; the batch size, with values of 16, 32, 64; the weight decay, explored for values 1e-5, 1e-4, 1e-3; and finally, the dropout values 0.1, 0.3, 0.5 were tested to mitigate overfitting.

For the classification task, due to class imbalance in our dataset, we employed a weighted cross-entropy loss. The weights, calculated as the inverse of class frequency, enable the model to better account for underrepresented classes. For the regression task, root mean square error (RMSE) was used as the optimization objective.

### B.3 Generative LLMs for difficulty assessment

The majority of the experiments were conducted by prompting the models via their respective APIs, with a total cost of approximately $20. An exception was made for the global difficulty experiments involving DeepSeek-R1, Gemma-27b, and Qwen2.5, which were executed locally using the Ollama framework. These experiments were run on a local cluster equipped with NVIDIA A100 and V100 GPUs, thereby incurring no API-related financing costs.

[16] https://huggingface.co/almanach/camembert-base

[17] https://huggingface.co/almanach/camembertv2-base

[18] https://huggingface.co/dangvantuan/sentence-camembert-base

- **Mistral Large**: We used the latest version of Mistral Large[19], released in November 2024. This model reportedly contains 123 billion parameters and demonstrates state-of-the-art reasoning capabilities across a wide range of tasks.
- **GPT-4.1**: We employed OpenAI's GPT-4.1[20], their flagship model optimized for complex tasks. Although the number of parameters has not been publicly disclosed, GPT-4.1 is known for its strong performance in problem-solving and generalization across domains.
- **DeepSeek-R1**: We used DeepSeek-R1 (DeepSeek-AI, 2025), a reasoning-optimized Mixture-of-Experts model from DeepSeek-AI. It has 671B total parameters, with 37B active per forward pass, and supports long contexts (up to 128k tokens). The model is open-weight and trained with reinforcement learning for strong performance on reasoning tasks.
- **qwen2.5**: We used Qwen2.5-72B (Yang et al., 2024), a 72B-parameter decoder-only transformer with a 131k token context window. It supports instruction following and performs well on complex language tasks.
- **Gemma-27b**: We used Gemma-27B (Team, 2024), an open weights 27B-parameter decoder-only transformer with an 8,000-token context window, designed to balance efficiency and capability.

## C Correlation between system entropy and human disagreement

| | Spearman | Pearson |
|---|---|---|
| ML - SVM (500) | -0.196* | -0.1777* |
| ML - DT (300) | 0.0318 | 0.0292 |
| ML - RF (400) | -0.1465* | -0.1489* |
| DL - CamemBERT-v2 | -0.0461 | -0.0314 |
| Hybride - RF (300) | -0.1373* | -0.1298* |

Table 10: Spearman and Pearson correlation results for classification models. Significant correlations are indicated by an asterisk (*).

In this section, we study the correlation between annotator disagreement and model uncertainty for global readability assessment. Shannon entropy, calculated from the model output probabilities for each text, measures the uncertainty of predictions.

[19] https://mistral.ai/news/mistral-large

[20] https://platform.openai.com/docs/models/gpt-4.1

In a previous study (Tack et al., 2017), disagreement between annotators for a text $i$ is calculated based on the observed disagreement $D_{o_i}^{\alpha}$ regarding the label $x$ given by annotator $c$. This measure is obtained by decomposing the Krippendorff formula for the observed disagreement $D_o^{\alpha}$ (Artstein and Poesio, 2008), which is equal to twice the empirical variance per text $s_i^2$.

$$D_{o_i}^{\alpha} = \frac{1}{c(c-1)} \sum_{m=1}^{c} \sum_{n=1}^{c} \delta_{\text{interval}}(x_{ic_m}, x_{ic_n}) = 2s_i^2 \quad (1)$$

In this way, a disagreement score and an entropy score can be calculated for each text, depending on the system being evaluated. A Pearson or Spearman correlation can then be used to study the relationship between human uncertainty and systems.

Table 10 presents the Spearman and Pearson correlations for different classification models.

The results in Table 10 indicate that some models have a slight correlation with human uncertainty. For example, the SVM model shows a significant negative correlation for both tests (Spearman: -0.196), suggesting that greater uncertainty in predictions is associated with greater disagreement between annotators. This behavior is similarly observed for the RF model, and by extension the hybrid model, which shares common architectural features.

Conversely, the CamemBERT-v2 and DT models do not show significant correlations, indicating that their prediction uncertainty does not appear to be directly related to human disagreement.

## D Generative LLMs prompts

## E Additional results

1. Vous êtes un expert linguistique spécialisé dans l'évaluation des niveaux de français selon le Cadre européen commun de référence pour les langues (CEFR).
2. Votre tâche consiste à classer le texte français suivant dans l'un des niveaux du CEFR:
↪ A1, A2, B1, B2, C1 ou C2.
3. Example:
↪ Texte à classifier : "Bonjour, je m'appelle Jean. J'habite à Paris. J'aime jouer au football."
4. Le texte fourni est composé de phrases simples et courtes, utilisant des structures grammaticales de base et un vocabulaire élémentaire. Selon le Cadre européen commun de référence pour les langues (CEFR), le niveau A1 correspond à la capacité de comprendre et dútiliser des expressions familières et quotidiennes ainsi que des énoncés très simples visant à satisfaire des besoins concrets.
↪ Niveau CECR: **A1**
5. Classifiez ce texte français: {text}

Figure 9: Example of a prompt used for global difficulty classification.

1. You are a linguistic expert specialized in evaluating French language levels according to the Common European Framework of Reference for Languages (CEFR).
2. Your task is to classify the following French text into one of the CEFR levels:
↪ A1, A2, B1, B2, C1 ou C2.
3. Example:
↪ Text to classify : "Hello, my name is Jean. I live in Paris. I like to play football."
↪ The provided text consists of simple, short sentences, using basic grammatical structures and elementary vocabulary. According to the Common European Framework of Reference for Languages (CEFR), level A1 corresponds to the ability to understand and use familiar, everyday expressions as well as very simple statements aimed at satisfying concrete needs.
↪ CECR level: **A1**
4. Classify this French text: {text}

Figure 10: English translation of the example prompt used for global difficulty classification from Figure 9.

1. Vous êtes un expert linguistique spécialisé dans l'évaluation des niveaux de français selon le Cadre européen commun de référence pour les langues (CEFR).
2. Votre tâche consiste à classer le texte français suivant dans l'un des niveaux du CEFR:
↪ A1, A2, B1, B2, C1 ou C2.
3. Exemple :
↪ Texte à classifier : "Bonjour, je m'appelle Jean. J'habite à Paris. J'aime jouer au football."
↪ Le texte fourni est composé de phrases simples et courtes, utilisant des structures grammaticales de base et un vocabulaire élémentaire. Selon le Cadre européen commun de référence pour les langues (CEFR), le niveau A1 correspond à la capacité de comprendre et d'utiliser des expressions familières et quotidiennes ainsi que des énoncés très simples visant à satisfaire des besoins concrets.
↪ Niveau CECR : **A1**
4. Classifiez ce texte français : {shot1}
↪ {cot1}
↪ Niveau CECR : **{value1}**
5. Classifiez ce texte français : {shot2}
↪ {cot2}
↪ Niveau CECR : **{value2}****
6. Classifiez ce texte français : {shot3}
↪ {cot3}
↪ Niveau CECR : **{value3}**
7. Classifiez ce texte français : {shot4}
↪ {cot4}
↪ Niveau CECR : **{value4}**
8. Classifiez ce texte français : {text}

Figure 11: Example of a prompt used for global difficulty level classification with multiple few-shot examples.

1. You are a linguistic expert specialized in evaluating French language proficiency according to the Common European Framework of Reference for Languages (CEFR).
2. Your task is to classify the following French text into one of the CEFR levels:
↪ A1, A2, B1, B2, C1, or C2.
3. Example:
↪ Text to classify: "Bonjour, je m'appelle Jean. J'habite à Paris. J'aime jouer au football."
↪ The provided text consists of simple, short sentences using basic grammatical structures and elementary vocabulary. According to the CEFR, level A1 corresponds to the ability to understand and use familiar everyday expressions and very simple statements aimed at meeting concrete needs.
↪ CEFR Level: **A1**
4. Classify this French text: {shot1}
↪ {cot1}
↪ CEFR Level: **{value1}**
5. Classify this French text: {shot2}
↪ {cot2}
↪ CEFR Level: **{value2}**
6. Classify this French text: {shot3}
↪ {cot3}
↪ CEFR Level: **{value3}**
7. Classify this French text: {shot4}
↪ {cot4}
↪ CEFR Level: **{value4}**
8. Classify this French text: {text}

Figure 12: English translation of the example of a prompt used for global difficulty level classification with multiple few-shot examples from Figure 11.

1. Vous êtes un assistant linguistique spécialisé dans l'analyse de la complexité lexicale.
2. Votre tâche est d'évaluer si un mot est complexe dans le contexte fourni, en fonction du niveau CECR du lecteur cible.
3. Un mot est considéré comme complexe s'il présente une ou plusieurs des difficultés suivantes, selon les définitions ci-dessous :
↪ {definitions}
4. Format attendu : une liste d'objets JSON, un par mot, contenant les champs suivants :
↪ - "term" : le mot analysé
↪ - "label" : "1" si le mot est jugé complexe, sinon "0".
↪ Niveau CECR du lecteur : {level}
↪ Texte : {text}
↪ Liste de mots à évaluer : {tokens}
↪ Évaluez la complexité de chacun des mots de la liste pour ce niveau de lecteur.

Figure 13: Example of a prompt used for binary local difficulty prediction based on reader CEFR level.

1. You are a linguistic assistant specialized in analyzing lexical complexity.
2. Your task is to evaluate whether a word is complex in the given context, based on the CEFR level of the target reader.
3. A word is considered complex if it presents one or more of the following difficulties, as defined below:
↪ {definitions}
↪ Expected format: a list of JSON objects, one per word, containing the following fields:
↪ - "term": the analyzed word
↪ - "label": "1" if the word is considered complex, otherwise "0".
↪ Reader CEFR level: {level}
↪ Text: {text}
↪ List of words to evaluate: {tokens}
↪ Evaluate the complexity of each word in the list for this reader level.

Figure 14: English translation of the example of a prompt used for binary local difficulty prediction based on reader CEFR level from Figure 13.

1. Vous êtes un assistant linguistique spécialisé dans l'analyse de la complexité lexicale.
2. Votre tâche est d'évaluer si un mot est complexe dans le contexte fourni, en fonction du niveau CECR du lecteur cible.
3. Un mot est considéré comme complexe s'il présente une ou plusieurs des difficultés suivantes, selon les définitions ci-dessous :
↪ {definitions}
4. Important : un même mot complexe peut présenter plusieurs types de difficulté simultanément. Dans ce cas, indiquez tous les types de difficulté applicables sous forme de liste de labels.
↪ Si le mot n'est pas complexe, utilisez la valeur "0".
↪ Format attendu : une liste d'objets JSON, un par mot, contenant les champs suivants :
↪ - "term" : le mot analysé
↪ - "label" : la liste des types de difficulté pertinents parmi ceux listés ci-dessus si le mot est jugé complexe, sinon "0".
↪ Niveau CECR du lecteur : {level}
↪ Texte : {text}
↪ Liste de mots à évaluer : {tokens}
↪ Évaluez la complexité de chacun des mots de ce texte pour ce niveau de lecteur.

Figure 15: Example of a prompt used for multi-label local difficulty prediction based on reader CEFR level.

1. You are a linguistic assistant specialized in the analysis of lexical complexity.
2. Your task is to evaluate whether a word is complex in the given context, based on the CEFR level of the target reader.
3. A word is considered complex if it presents one or more of the following difficulties, according to the definitions below:
↪ {definitions}
4. Important: a single complex word may present several types of difficulty simultaneously. In this case, indicate all applicable types of difficulty in the form of a list of labels.
↪ If the word is not complex, use the value "0".
↪ Expected format: a list of JSON objects, one per word, containing the following fields:
↪ - "term": the analyzed word
↪ - "label": the list of relevant difficulty types among those listed above if the word is judged complex, otherwise "0".
↪ Reader CEFR level: {level}
↪ Text: {text}
↪ List of words to evaluate: {tokens}
↪ Evaluate the complexity of each word in this text for this reader level.

Figure 16: English translation of an example of a prompt used for multi-label local difficulty prediction based on reader CEFR level from Figure 15.

| | Accuracy | Adj. Accuracy | Macro-F1 | MSE |
|---|---|---|---|---|
| ***Classification*** | | | | |
| ML - SVM (500) | 55.84 ± 4.26 | 97.83 ± 1.95 | 47.54 ± 6.15 | |
| ML - DT (300) | 54.10 ± 3.28 | 94.81 ± 0.81 | 43.84 ± 4.98 | |
| ML - RF (400) | 62.77 ± 4.18 | 98.05 ± 1.26 | 47.78 ± 7.60 | |
| DL - CamemBERT | 64.04 ± 9.97 | 98.71 ± 1.77 | 60.36 ± 8.23 | |
| DL - CamemBERT-v2 | 64.26 ± 5.67 | 99.17 ± 0.91 | 60.05 ± 6.01 | |
| DL - CamemBERT-OLL | 66.02 ± 3.80 | **99.78 ± 0.43** | **61.14 ± 4.30** | |
| Hybride - RF (300) | **67.32 ± 4.08** | 99.14 ± 0.81 | 56.26 ± 9.17 | |
| ***Regression*** | | | | |
| ML - SVR (500) | 39.60 ± 5.39 | 93.06 ± 3.61 | 22.63 ± 1.96 | 4.94 ± 1.07 |
| ML - DT (50) | 38.75 ± 5.29 | 88.10 ± 2.24 | 22.13 ± 3.15 | 7.22 ± 1.55 |
| ML - RF (500) | 40.89 ± 6.28 | 91.77 ± 3.10 | 22.96 ± 4.71 | 4.70 ± 0.73 |
| ML - OrdR (50) | 42.20 ± 3.80 | 91.30 ± 3.30 | 22.70 ± 2.30 | 0.84 ± 0.09 |
| DL - CamemBERT | **70.77 ± 5.48** | **100.00 ± 0.00** | **59.63 ± 2.55** | 3.87 ± 0.72 |
| DL - CamemBERT-v2 | 68.38 ± 5.70 | **100.00 ± 0.00** | 47.52 ± 8.36 | **3.78 ± 0.75** |
| Hybrid - RF (300) | 64.28 ± 6.55 | 99.57 ± 0.53 | 36.50 ± 5.21 | 4.88 ± 0.75 |

Table 11: Additional results comparing performance metrics of readability classification and regression models. Parentheses denote the number of selected features.

| Model | EN-zero-shot | EN-few-shot | FR-zero-shot | FR-few-shot |
|---|---|---|---|---|
| DeepSeek-7b | 25.6 ± 3.36 | 28.01 ± 2.54 | 23.71 ± 2.88 | 28.97 ± 1.31 |
| DeepSeek-14b | 23.97 ± 3.07 | 34.96 ± 4.78 | 25.02 ± 4.57 | 34.4 ± 3.93 |
| DeepSeek-32b | 27.77 ± 5.29 | 43.06 ± 7.37 | 27.33 ± 4.84 | 41.83 ± 8.16 |
| DeepSeek-70b | 29.64 ± 8.22 | **48.95 ± 4.85** | **38.41 ± 3.14** | **47.06 ± 3.55** |
| Gemma-27b | **30.19 ± 7.97** | 41.12 ± 4.45 | 29.8 ± 6.46 | 39.49 ± 3.76 |
| Qwen-72b | 17.65 ± 4.43 | 46.93 ± 3.53 | 19.32 ± 11.13 | 42.9 ± 5.33 |
| Mistral-large | 22.78 ± 5.34 | 48.6 ± 4.08 | 26.72 ± 4.24 | 43.86 ± 3.57 |
| GPT-4.1 | 27.01 ± 6.04 | 44.69 ± 6.84 | 35.55 ± 9.34 | 43.02 ± 7.82 |

Table 12: Additional results of macro-F1 performance of generative LLMs for global difficulty classification.

| level | Mistral-large | | | | GPT-4.1 | | | | Qwen2.5 | | | | DeepSeek-R1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 | Acc |
| Very Easy | 72.68 | 71.9 | 71.68 | 71.94 | **74.88** | 71.07 | 69.97 | 71.15 | 73.59 | **73.51** | **73.49** | **73.52** | 68.1 | 68.99 | 68.0 | 68.55 |
| Easy | 73.4 | 73.31 | 73.27 | 73.29 | **78.24** | **77.12** | **76.93** | **77.18** | 71.19 | 70.77 | 70.57 | 70.7 | 74.99 | 74.91 | 74.92 | 74.97 |
| Plain | 72.57 | 70.96 | 70.32 | 70.8 | **77.04** | **76.74** | **76.71** | **76.79** | 68.43 | 67.09 | 66.36 | 66.92 | 75.99 | 75.44 | 75.29 | 75.41 |
| +Complex | 73.11 | 70.43 | 69.41 | 70.24 | **76.76** | **76.53** | **76.43** | **76.47** | 70.13 | 68.51 | 67.72 | 68.32 | 75.61 | 73.41 | 72.66 | 73.2 |
| Macro-F1 | 72.59 | 71.7 | 71.34 | 71.59 | **77.12** | **76.74** | **76.7** | **76.8** | 69.57 | 68.61 | 68.12 | 68.47 | 75.25 | 74.8 | 74.73 | 74.86 |

Table 13: Additional results on performance metrics for binary local difficulty assessment by proficiency level. P: Precision, R: Recall, Acc: Accuracy.

| Difficulty type | Mistral-large | | | GPT-4.1 | | | Qwen2.5 | | | DeepSeek-R1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Grammar difficulties | **48.56** | 28.30 | 35.76 | 47.54 | 55.64 | **51.27** | 32.13 | 47.00 | 38.17 | 33.95 | **65.95** | 44.82 |
| Figure of speech, idiomatic expression | 53.99 | 37.11 | 43.98 | **62.14** | 55.48 | **58.62** | 38.81 | 57.06 | 46.20 | 48.57 | **61.90** | 54.43 |
| Spelling or decoding problems | 56.59 | 5.44 | 9.93 | 53.65 | **14.38** | **22.68** | 53.01 | 10.25 | 17.18 | **58.87** | 11.05 | 18.61 |
| Difficult cohesion index | 43.00 | 26.97 | 33.15 | **47.39** | **48.18** | **47.78** | 34.79 | 20.45 | 25.76 | 44.31 | 17.12 | 24.70 |
| Difficult or unknown word | 61.15 | 66.68 | 63.80 | 60.73 | **81.52** | **69.61** | 61.78 | 67.10 | 64.33 | **61.82** | 75.32 | 67.90 |
| Unusual syntactic order | 40.82 | 7.25 | 12.31 | **66.67** | 15.22 | 24.78 | 48.00 | 10.87 | 17.73 | 61.94 | **17.45** | **27.23** |
| Difficult cultural reference | 43.99 | 51.86 | 47.60 | **55.54** | 56.46 | **55.99** | 46.54 | 48.08 | 47.30 | 40.59 | **63.82** | 49.63 |
| Too much secondary information | 35.34 | 26.04 | 29.98 | 35.58 | **41.00** | **38.10** | 35.32 | 37.67 | 36.46 | **40.34** | 33.06 | 36.34 |
| Average | 47.93 | 31.21 | 34.56 | **53.66** | **45.98** | **46.10** | 43.80 | 37.31 | 36.64 | 48.80 | 43.21 | 40.46 |

Table 14: Additional results of generative LLM performance for multi-label local difficulty classification, broken down by local difficulty class. P: Precision, R: Recall.