


Exploring Few-Shot Approaches to Automatic Text Complexity Assessment in European Portuguese

Eugénio Ribeiro   [INESC-ID Lisboa & Instituto Universitário de Lisboa (ISCTE-IUL) | eugenio.ribeiro@inesc-id.pt]

David Antunes  [INESC-ID Lisboa | david.f.l.antunes@inesc-id.pt]

Nuno Mamede  [INESC-ID Lisboa & Instituto Superior Técnico, Universidade de Lisboa | nuno.mamede@inesc-id.pt]

Jorge Baptista  [INESC-ID Lisboa & Faculdade de Ciências Humanas e Sociais, Universidade do Algarve | jorge.baptista@inesc-id.pt]

 INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal.

Received: 01 April 2025 • Accepted: 15 July 2025 • Published: 21 August 2025

Abstract The automatic assessment of text complexity has an important role to play in the context of language education. In this study, we shift the focus from L2 learners to adult native speakers with low literacy by exploring the new iRead4Skills dataset in European Portuguese. Furthermore, instead of relying on classical machine learning approaches or fine-tuning a pre-trained language model, we leverage the capabilities of prompt-based Large Language Models (LLMs), with a special focus on few-shot prompting approaches. We explore prompts with varying degrees of information, as well as different example selection approaches. Overall, the results of our experiments reveal that even a single example significantly increases the performance of the model and that few-shot approaches generalize better than fine-tuned models. However, automatic complexity assessment is a difficult and highly subjective task that is still far from solved.

Keywords: Text Complexity, Readability, Few-Shot Prompting, Large Language Models

1 Introduction

Assessing text readability, complexity, or difficulty is valuable not only in education but also across various industries and everyday contexts. In education, evaluating text complexity helps educators and curriculum designers match materials to learners' proficiency levels, fostering effective language development and personalized learning experiences. Beyond education, readability assessment can play an important role in multiple sectors. For instance, in banking, ensuring that financial policies and terms are presented at an appropriate readability level enables clients to make well-informed decisions. In healthcare, readability assessment can support the creation of clear and accessible medical instructions, consent forms, and patient information materials, helping individuals with varying levels of language proficiency better understand critical information. Similarly, legal documents, government communications, and user manuals, among others, can benefit from accurate readability assessment, facilitating transparency and effective communication.

In the domain of education, the iRead4Skills¹ project [Baptista *et al.*, 2024] aims to enhance adult literacy and promote the development of reading skills through an intelligent system that evaluates text complexity and recommends reading materials suited to the user's proficiency level. This system can also assist trainers in creating or adapting texts to match individual students' needs. An important contribution of this project is the collection of corpora in three different

languages — French, Portuguese, and Spanish — annotated for complexity level by both linguistics experts [Pintard *et al.*, 2024] and trainers [Amaro *et al.*, 2024], the latter representing one of the system's primary end-user groups.

Focusing on European Portuguese, the iRead4Skills dataset provides a valuable resource for research on automatic complexity level assessment, addressing the scarcity of annotated data for this task. Previous studies have primarily relied on textual data intended for L2 proficiency evaluation (e.g., Branco *et al.*, 2014b; Curto *et al.*, 2015; Santos *et al.*, 2021; Akef *et al.*, 2024; Ribeiro *et al.*, 2024b). However, apart from a small set of texts extracted from model exams [Ribeiro *et al.*, 2024b], this data is not publicly available and the complexity levels were inferred from the Common European Framework of Reference for Languages (CEFR) [Council of Europe, 2001] levels of the exams rather than explicitly annotated. In contrast, the iRead4Skills dataset offers a larger collection of texts with direct complexity annotations by experts and trainers, ensuring higher annotation quality and reliability. Moreover, it shifts the focus from L2 learners to a native-speaking audience, introducing a new perspective to the task.

Prior work on automatic complexity level assessment in European Portuguese has extensively explored classical Machine Learning (ML) approaches based on handcrafted features (e.g. Branco *et al.*, 2014b; Curto *et al.*, 2015; Akef *et al.*, 2024) as well as the fine-tuning of foundation models (e.g. Santos *et al.*, 2021; Ribeiro *et al.*, 2024b). However, recently, we have observed the proliferation of instruction or prompt-based generative LLMs [Ouyang *et al.*, 2022], with ChatGPT

¹<https://iRead4Skills.com/>

[OpenAI, 2023] as the main driving force. These models have demonstrated strong performance across various Natural Language Processing (NLP) tasks, particularly when enhanced with Retrieval-Augmented Generation (RAG) techniques to improve example selection in few-shot learning scenarios [Gao et al., 2023; Fan et al., 2024]. In this study, we leverage these models for complexity level assessment. First, we evaluate their intrinsic ability to perform this task in a zero-shot setting using different prompts with varying degrees of specificity and detail. Then, we explore how their performance can be improved through few-shot learning and analyze the impact of different example selection strategies. Furthermore, we examine how the dual annotations of the iRead4Skills dataset — provided by linguistic experts and trainers — differ and affect the evaluation process.

In the remainder of this document, we start by providing an overview of related work on automatic text readability level assessment, with a focus on European Portuguese in Section 2. Then, in Section 3, we describe our experimental setup, including the dataset and the employed prompting, example selection, and evaluation methodologies. Next, in Section 4, we present and discuss the results of our experiments, including the observed errors and biases of the different approaches. Finally, in Section 5, we summarize the contributions of this study, discuss its limitations, and provide pointers for future research in the area.

2 Related Work

Readability assessment is a problem that has been widely explored over the years. Traditionally, the problem is addressed by creating readability formulas or indexes based on statistical information and/or domain knowledge [DuBay, 2004; Crossley et al., 2017]. Among these, the most widely used are the Flesch Reading Ease Index and the Flesch-Kincaid Grade Level [Kincaid et al., 1975].

However, considering the developments in ML, and especially in NLP, the research on automatic readability assessment shifted towards following the trends in the NLP area [Graesser et al., 2004; McNamara et al., 2014]. This trend was also followed in related tasks, such as lexical complexity assessment [North et al., 2023]. Early approaches (and many recent ones for low-resource languages) relied on handcrafted features, such as word frequency, sentence length, and syntactic complexity, combined with traditional machine learning algorithms, such as decision trees and Support Vector Machines (SVMs) (e.g., Aluisio et al., 2010; François and Fairon, 2012; Karpov et al., 2014; Curto et al., 2015; Pilán and Volodina, 2018; Forti et al., 2020; Leal et al., 2023). Then, Deep Learning (DL) approaches relying on pre-trained word embeddings, such as those generated by Word2Vec [Mikolov et al., 2013], emerged (e.g., Cha et al., 2017; Nadeem and Ostendorf, 2018; Filighera et al., 2019). Finally, more recently, research in the area shifted towards the fine-tuning of pre-trained Transformer-based foundation models, such as BERT [Devlin et al., 2019], GPT [Radford et al., 2019], and RoBERTa [Liu et al., 2019] (e.g., Santos et al., 2021; Yancey et al., 2021; Martinc et al., 2021; Mohtaj et al., 2022; Ribeiro et al., 2024b).

Similarly to most NLP tasks, a significant part of the research on automatic text readability level assessment focuses on the English language (e.g., Xia et al., 2016; Cha et al., 2017; Nadeem and Ostendorf, 2018; Filighera et al., 2019; Martinc et al., 2021). However, in this case, there are also several studies addressing the problem in other languages, many of which are low-resourced. For instance, there are studies in French (e.g., François and Fairon, 2012; François et al., 2020; Yancey et al., 2021; Wilkens et al., 2022; Hernandez et al., 2022), Chinese (e.g., Sung et al., 2015), German (e.g., Mohtaj et al., 2022), Italian (e.g., Forti et al., 2020; Santucci et al., 2020), Russian (e.g., Karpov et al., 2014; Reynolds, 2016), Swedish (e.g., Jönsson et al., 2018; Pilán and Volodina, 2018), Slovenian (e.g., Martinc et al., 2021), and Spanish (e.g., Rodríguez Rey et al., 2025).

Focusing on Portuguese, there are a few studies covering the Brazilian variety of the language (e.g., Scarton and Aluisio, 2010; Aluisio et al., 2010; Leal et al., 2023). However, in this study, we are mainly interested in the European variety. Thus, below, we describe previous studies covering this variety in further detail.

The Portuguese version of the REAP tutoring system [Marujo et al., 2009] included a readability level classifier trained on 5th to 12th-grade textbooks. The model was based on SVMs applied to lexical features, such as statistics of word unigrams, and used ordinal logistic regression [McCullagh, 1980] to capture the ordinal nature of the levels. Although this model was accurate when applied to school textbooks, its performance significantly decreased when applied to exams of the 6th, 9th, and 12th grades.

LX-CEFR [Branco et al., 2014b] is a tool designed to help language learners and teachers of Portuguese in assessing the complexity level of a text according to the CEFR. It focuses on four different features independently: the Flesch Reading Ease index, the lexical category density in terms of the proportion of nouns, the average word length in number of syllables, and the average sentence length in number of words. A corpus of 114 excerpts, extracted from the Portuguese as a foreign language proficiency exams performed by Camões, I.P.², was used to analyze the correlation between these features and compute the readability level. A subsequent study [Branco et al., 2014a] focused on the re-evaluation of the tool by human experts, as well as the re-annotation of the texts by multiple language instructors. Regarding the latter, the inter-annotator agreement was of just 0.17, which reveals the difficulty and subjectivity of the task.

Curto et al. [2015] explored the use of several traditional ML algorithms for the task. The algorithms were applied to 52 features split into 5 different groups: Part-of-Speech (POS), chunks, sentences and words, verbs, averages and frequencies, and extras. The experiments were performed on an extended version of the dataset used in the context of LX-CEFR containing 237 excerpts. The highest performance was achieved using LogitBoost [Friedman et al., 2000]. An analysis of the features revealed that text length (in words and sentences), lexical diversity (number of different words), and syntactic complexity (number of syntactic dependencies and parse tree size) were the most influential factors in clas-

²<https://www.instituto-camoes.pt/>

sification. Additionally, similarly to what was observed by Branco *et al.* [2014a], a re-annotation of the extended version of the dataset by two groups of multiple experts revealed low inter-annotator agreements of 0.19 and 0.16 [Curto, 2014].

Akef *et al.* [2024] also explored the use of traditional ML algorithms, applied to a set of 489 features split into 6 groups: count-based, lexical, syntactic, morphological, discourse, and psycholinguistic. The models were trained on multiple versions of the proficiency exam corpus, including one with 500 excerpts. In spite of considering more features, the observed performance was similar to that achieved by Curto *et al.* [2015]. Furthermore, additional experiments showed poor generalization when the models were applied to a dataset of texts extracted from books designed for Portuguese L2 learners.

Correia and Mendes [2021] explored the use of a hybrid neural network with two branches. One of the branches focused on the bidirectional processing of pre-trained word embeddings (NILC-Embeddings [Hartmann *et al.*, 2017]) and the other on a set of 14 length-based features and readability indexes. The experiments showed that the combination of the branches leads to improved performance in comparison to each individual branch, revealing the complementarity of the different representations.

Looking into approaches based on the fine-tuning of pre-trained Transformer-based foundation models, Santos *et al.* [2021] explored the fine-tuning of Portuguese versions of the GPT-2 [Radford *et al.*, 2019] and RoBERTa [Liu *et al.*, 2019] models on the multiple variants of the proficiency exam corpus. These approaches outperformed the classical ones on the larger variants of the dataset, with the GPT-2 model achieving the top performance. Akef *et al.* [2024] also explored the use of a fine-tuned version of the GPT-3.5-Turbo model, reporting better performance than that of GPT-2, albeit with a different evaluation methodology. In our studies, we explored the use of several additional foundation models [Ribeiro *et al.*, 2024a]. The highest performance was achieved using a model of the Albertina PT-* family [Rodrigues *et al.*, 2023] and an approach to leverage the ordinal relation between the CEFR levels [Ribeiro *et al.*, 2024c]. However, our experiments also showed that the fine-tuned models lack generalization ability to types of text that were not observed during training [Ribeiro *et al.*, 2024b]. Furthermore, by relying on annotations by linguistics experts, we have observed that the approximation of the complexity level given by the level of the exam is not particularly accurate, with a Cohen’s κ agreement of just 0.29.

Finally, we conducted preliminary experiments using prompt-based generative LLMs in a zero-shot scenario [Ribeiro *et al.*, 2024b]. Although these models are able to deal with diverse types of text and appear to be less biased than the fine-tuned models, they exhibited erratic behavior across different models and prompts. Overall, except for a few specific cases, the models revealed a lack of intrinsic knowledge to accurately assess text complexity and were outperformed by the fine-tuned models. However, incorporating relevant examples into the prompts is expected to enhance performance and that is the main approach we explore in this study.

3 Experimental Setup

In this section, we describe our experimental setup. We start by describing the dataset used in our experiments in Section 3.1. Then, in Section 3.2, we present the training-based baselines used for result comparison. In Sections 3.3 and Sections 3.4, we describe the prompting and example selection approaches explored in this study. Section 3.5 summarizes our evaluation approach. Finally, in Section 3.6, we provide implementation details that enable the future reproduction of our experiments.

3.1 Dataset

The iRead4Skills project targets adult native speakers with low literacy and their educators. In this context, three complexity levels were defined — very easy, easy, and plain — which roughly correspond to the A1, A2, and B1 levels of the CEFR [Monteiro *et al.*, 2023]:

Very Easy: Texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish the primary school) and almost no reading experience. It roughly corresponds to CEFR A1 level.

Easy: Texts that are fully or almost fully understood by people with low schooling (i.e., that completed the primary school but do not have more than the 9th year) and have poor reading experience. It roughly corresponds to CEFR A2 level.

Plain: Texts that are understood the first time they are read by people that completed the 9th year and have a functional-to-average reading experience. It roughly corresponds to CEFR B1 level.

Corpora featuring textual examples of these levels, as well as a more complex one, were collected and annotated by linguistics experts [Pintard *et al.*, 2024]. The corpora were collected for three languages — French, Portuguese, and Spanish — and were designed to cover eleven communication domains — personal communication, institutional/professional communication, social media communication, commercial communication/dissemination, non-fiction books, fiction books, didactic books, academic communication, political communication, legal communication, and religious communication. Additionally, different document types were considered for each communication domain, leading to 66 different domain-type pairs and ensuring diversity. Subsets of each language corpus were also annotated by the trainers of adults with low literacy, providing additional insight by one of the target user groups [Amaro *et al.*, 2024].

Focusing on Portuguese data, the iRead4Skills dataset contains 2933 examples [Reis *et al.*, 2024]. For the development of automatic approaches to text complexity assessment based on this data, we split it into three subsets. As a test set, we opted for using the subset of the corpus that is also annotated by trainers. This way, we can analyze the alignment of the developed approaches with both the more general perspective on complexity provided by the linguistics experts and the more target-specific one provided by the trainers. To obtain

Table 1. Distribution of the texts in the dataset by partition and complexity level. The (E) and (T) annotations on the test set refer to the expert and trainer assessments, respectively.

	V. Easy	Easy	Plain	+Complex	Total
Train	523	418	490	555	1986
Val.	142	111	123	143	519
Test (E)	112	137	139	40	428
Test (T)	94	129	154	51	428

the subsets for training and validation/development, we applied an 80/20 approach stratified by complexity level, communication domain, and document type. Singletons were attributed to the validation set in an attempt to reward approaches that generalize better. However, considering that this study focuses on prompt-based approaches and not on model training, we mainly rely on the training set as an example pool and on the test set for evaluation purposes. Still, the validation set is used during the development of baselines, as described in Section 3.2.

Table 1 shows the distribution of the examples across partitions and complexity levels. We can see that the distribution across complexity levels is fairly balanced in the train and validation sets. On the test set, there are fewer examples of the more complex level because the focus of the trainer annotation process was on the low literacy target levels. Furthermore, we can see significant differences between the expert and trainer annotations. Looking into the differences in more detail, the classification differs for 186 of the 428 examples, with the trainers tending towards higher complexity. However, only 25% of the changes are by more than one level. Overall, this confirms the difficulty and the subjectivity of text complexity assessment, as well as the need for insight regarding the target audience. Examples of each level are provided in Appendix A.

3.2 Baselines

As baselines for comparison, we rely on two training-based approaches: one based on classical ML applied to hand-crafted features and the other on the fine-tuning of a foundation model. As these baselines take advantage of all the annotated data available to train models specialized on the task, we do not expect the prompt-based approaches explored in this study to outperform them on data with similar characteristics, as they rely at most on just a small set of examples for guidance. However, increased specialization typically comes at the cost of decreased generalization. Thus, it is possible that the prompt-based approaches can outperform the baselines on data with different characteristics, such as the test set with trainer annotations.

Both baselines come from previous research on text complexity classification in European Portuguese. The first consists of a gradient boosting classifier [Friedman, 2001] applied to a set of 631 descriptive, lexical, syntactic, and discursive features identified as potential complexity indicators in the context of the iRead4Skills³ project [Ribeiro et al., 2024d]. Boosting approaches have proved to be top performers among classical ML approaches on this task [Curto et al.,

2015] and we have also used this approach as a baseline in our previous study on text readability assessment on proficiency exam data [Ribeiro et al., 2024b]. The second baseline consists of a fine-tuned version of a model of the ALbertina PT-* family [Rodrigues et al., 2023], which proved to be the top performer for automatic text readability assessment in the same study.

3.3 Prompting Strategies

The performance of prompt-based LLMs is significantly impacted by the quality of the prompts or instructions [Mishra et al., 2022; Giray, 2023; Chen et al., 2023]. In our preliminary experiments on proficiency exam data [Ribeiro et al., 2024b], we explored the use of two prompts: one targeting the complexity level of the texts directly and the other approaching the problem from the perspective of the proficiency level required to understand the texts. In general, the former led to higher performance. Thus, in this study, we focus on prompts that target the complexity level directly. We explore the use of three different prompts, with increasing specificity and detail. They are described below.

3.3.1 Generic

The first prompt is the same one used in our preliminary experiments on proficiency exam data. It targets the complexity level directly in an attempt to leverage any intrinsic knowledge that the LLM may have regarding the task. The model is not expected to have any knowledge regarding the complexity levels defined in the context of the iRead4Skills project. On the other hand, there is a higher probability that it came across documentation about CEFR levels during training, as they are widely used across multiple languages. As there is a rough correspondence between iRead4Skills and CEFR levels, we rely on the latter in the prompt.

PT: *És um classificador do nível de complexidade de textos de acordo com o Quadro Europeu Comum de Referência para as Línguas (CEFR). Responde apenas com o nível: A1, A2, B1, B2+.*

EN: *You are a classifier of the complexity level of texts according to the CEFR. Answer with just the level: A1, A2, B1, B2+.*

3.3.2 Target Audience

The CEFR is typically used in the context of L2 learning. On the other hand, the data used in this study is targeted at adult native speakers with low literacy. Thus, the second prompt has this target audience explicit.

PT: *És um classificador do nível de complexidade de textos para serem lidos por adultos falantes nativos de português europeu, mas com baixa literacia. Usa como referência os níveis do Quadro Europeu Comum de Referência para as Línguas (CEFR): A1, A2, B1, B2+. Responde apenas com o nível.*

³<https://gitlab.hlt.inesc-id.pt/iread4skills/docs/>

EN: *You are a classifier of the complexity level of texts intended for adult native speakers of European Portuguese with low literacy. Use the levels of the CEFR as a reference: A1, A2, B1, B2+. Answer with just the level.*

3.3.3 Descriptive

The final prompt is the most descriptive, as it complements the previous one with a short description of the levels as defined by Monteiro *et al.* [2023]. This aims to provide additional guidance to the model without being overly extensive.

PT: *És um classificador do nível de complexidade de textos para serem lidos por adultos falantes nativos de português europeu, mas com baixa literacia. Usa como referência os níveis do Quadro Europeu Comum de Referência para as Línguas (CEFR) e a seguinte descrição:*

A1: Textos que são compreendidos total ou quase totalmente por todas as pessoas, incluindo aquelas com escolaridade muito baixa (ou seja, que não concluíram o ensino primário) e quase nenhuma experiência de leitura.

A2: Textos que são compreendidos total ou quase totalmente por pessoas com baixa escolaridade (ou seja, que concluíram o ensino primário, mas não têm mais do que o 9.º ano) e pouca experiência de leitura.

B1: Textos que são compreendidos à primeira leitura por pessoas que concluíram o 9.º ano e têm uma experiência de leitura funcional a média.

Responde apenas com o nível, usando B2+ para textos com dificuldade superior.

EN: *You are a classifier of the complexity level of texts intended for adult native speakers of European Portuguese with low literacy. Use the levels of the CEFR and the following description as a reference:*

A1: Texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish the primary school) and almost no reading experience.

A2: Texts that are fully or almost fully understood by people with low schooling (i.e., that completed the primary school, but do not have more than the 9th year) and have poor reading experience.

B1: Texts that are understood the first time they are read by people that completed the 9th year and have a functional-to-average reading experience.

Answer with just the level, using B2+ for texts with a higher complexity.

3.4 Example Selection

Although LLMs possess extensive intrinsic knowledge due to the vast amount of data used in their training, they have been shown to perform better on specific tasks when provided with a small set of examples in a few-shot setting [Brown *et al.*, 2020; Touvron *et al.*, 2023]. Furthermore, example selection typically plays an important role in this context, as providing more relevant examples allows the model

to make more informed decisions. To address this aspect, we explore the use of different selection strategies, as well as a varying number of provided examples.

3.4.1 Random Sample

The simplest way to select examples is by performing a random sample of the available pool. That is, in a k -shot setting, when prompting the model for the classification of a text, it is paired with k labeled examples selected at random. This provides generic guidance about the task, but the examples may be unrelated to the input text.

3.4.2 Semantic Search

A more informed approach to example selection involves retrieving semantically similar examples from the available pool. This approach, which has increasingly become a core component in the context of RAG [Sawarkar *et al.*, 2024], ensures that the retrieved examples are contextually relevant, potentially improving classification performance by providing the model with more useful references. Unlike random sampling, which may introduce irrelevant examples, semantic search helps maintain task consistency by selecting examples that share linguistic or structural similarities with the input text.

In most implementations of this strategy, both the input text and the pool of available examples are embedded into a high-dimensional vector space using a pre-trained language model focused on sentence similarity [Reimers and Gurevych, 2019]. A nearest-neighbor search is then performed to select the k examples that are most similar to the input text based on cosine similarity or another appropriate distance metric.

Looking specifically into text complexity, in many cases, documents may have similar semantics, but different complexity. As revealed by the features identified as potential complexity indicators in the context of the iRead4Skills project, complexity is, at many times, introduced by syntactic structures or uncommon vocabulary. Thus, in addition to finding the nearest neighbors in the embedding space of a pre-trained language model, we also explore the space generated by these task-specific features.

3.4.3 Stratification

In addition to applying the previous strategies to the example pool without restrictions, we explore their pairing with two stratification approaches based on communication domains and complexity levels.

As described in Section 3.1, the iRead4Skills dataset covers multiple communication domains, each featuring specific document types. As text complexity may be perceived differently across different communication domains, we explore whether restricting the example pool to examples of the same domain as the input text impacts performance. In a real-world scenario, this would also require automatic communication domain classification. However, we leave that for future work.

The second stratification approach has a different rationale and implementation. If no restrictions are applied to the

example selection process, the provided examples may not cover every complexity level or even be all of the same level. When using semantic search to select the examples, this can provide information regarding the most probable complexity level for the input text. However, that is not the case when using random sampling. Furthermore, as previously discussed, documents may have similar semantics, but different complexity. Thus, in addition to selecting k examples without restrictions in terms of level, we also explore the selection of k examples per level. This ensures diversity when using a random sampling strategy. When using semantic search, this informs the model on how complexity changes for similar semantics.

3.5 Evaluation Metrics

We adopt some of the most common evaluation metrics across previous studies on automatic readability level classification, namely accuracy, adjacent accuracy, and the macro F_1 score. Accuracy evaluates the precise identification of a text's complexity level, while adjacent accuracy also considers neighboring levels, offering further insight into the identification of texts slightly easier or harder than the assigned level. As the distribution of examples across levels is not balanced in test set, especially when considering trainer annotations, the macro F_1 score is also a relevant metric to identify potential biases.

3.6 Implementation Details

Starting with the baselines, we rely on the gradient boosting implementation provided by scikit-learn⁴ library [Pedregosa et al., 2011]. For fine-tuning, we use the HuggingFace Transformers⁵ library [Wolf et al., 2020]. As foundation model, we use the Albertina PT-PT⁶ model with 100M parameters [Rodrigues et al., 2023].

For semantic search, we rely on the Sentence Transformers⁷ library [Reimers and Gurevych, 2019]. To generate document embeddings, we use the Serafim⁸ model with 900M parameters tuned for Information Retrieval (IR) [Gomes et al., 2025]. When using the iRead4Skills features, to avoid biasing the similarity calculation toward specific features, we use only the mean aggregator to combine word- and sentence-level features into document-level representations, and we normalize the features before computing similarity.

In the few-shot setting, the descriptive prompt is extended with a sentence stating that examples will be provided together with each text:

PT: *Em conjunto com cada texto, vais receber alguns exemplos para ajudar a guiar a classificação.*

EN: *Along with each text, you will receive some examples to help guide the classification.*

The examples and input text are provided sequentially using the following format:

PT:

Exemplo: [TEXTO]

Nível: [NÍVEL]

...

Texto: [TEXTO]

Nível:

EN:

Example: [TEXT]

Level: [LEVEL]

...

Text: [TEXT]

Level:

Regarding the base LLM, in our preliminary experiments on proficiency exam data, we were not able to identify a clear top performer. Due to the computational requirements required to run large open-weight models locally, in this study, we mainly rely on the GPT-4o mini model [OpenAI et al., 2024], accessible through the OpenAI API⁹. Still, we also perform the experiments using two smaller open-weight models for comparison: Llama 3.1 8B¹⁰ [Llama Team, 2024] and EuroLLM 9B¹¹ [Martins et al., 2024]. We opted for these models due to their moderate size and multilingual performance. Considering that both the answers of the models and some of the example selection approaches involve some level of non-determinism, we perform three runs of each experiment. The cost of using the OpenAI API to perform the experiments was approximately \$50 USD and the experiments using the open-weight models took approximately 12 hours on an NVIDIA A100 80GB Tensor Core GPU. Unless stated otherwise, the results reported in Section 4 and Appendix B correspond to the average and standard deviation across the three runs. All of the metrics are reported in percentage form.

4 Results

Considering that our experiments were performed in multiple steps, we organize the presentation of the results accordingly. In Section 4.1, we present the results of training-based baselines. In the following sections, we focus on the results achieved using GPT-4o mini as the base LLM: in Section 4.2, we discuss the results of the experiments in the zero-shot setting and in Sections 4.3 and 4.4, we cover our few-shot experiments, using random sampling and semantic search for example selection, respectively. Then, in Section 4.5, we compare those results with the ones achieved using open-weight models. Finally, in Section 4.6, we perform a more in-depth error analysis of the top performers, in order to identify the main points of confusion and potential approaches to mitigate them.

⁴<https://scikit-learn.org/>

⁵<https://huggingface.co/docs/transformers>

⁶<https://huggingface.co/PORTULAN/albertina-100m-portuguese-ptpt-encoder>

⁷<https://sbnet.net/>

⁸<https://huggingface.co/PORTULAN/serafim-900m-portuguese-pt-sentence-encoder-ir>

⁹<https://openai.com/api/>

¹⁰<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

¹¹<https://huggingface.co/utter-project/EuroLLM-9B-Instruct>

Table 2. Baseline results.

Baseline	Accuracy	Experts	Macro F ₁	Accuracy	Trainers	Macro F ₁
		Adj. Acc.			Adj. Acc.	
Gradient Boosting	51.71±0.48	88.94±0.22	49.79±0.39	32.87±0.11	84.58±0.19	32.13±0.16
Fine-Tuning	51.87±0.00	92.99±0.00	50.79±0.00	37.62±0.00	86.68±0.00	36.34±0.00

4.1 Baselines

Table 2 shows the results achieved using the baseline approaches. Considering that both baselines are training-based approaches, it is important to remember that the training and validation sets were annotated by the linguistics experts. Consequently, as expected, the performance on the test set is significantly higher when considering expert annotations in comparison to trainer annotations. Still, even when considering expert annotations, the performance is significantly lower than that observed in previous studies using proficiency exam data (e.g. Curto *et al.*, 2015; Santos *et al.*, 2021; Akef *et al.*, 2024). This confirms that the diversity in terms of communication domains and document types of the iRead4Skills dataset introduces a more complex problem that highlights the generalization issues of the models.

Although both baselines achieve similar results when considering the expert annotations, the fine-tuned language model outperforms the gradient boosting classifier when trainer annotations are considered. This suggests that the former is more accurate at predicting the complexity level of the examples with agreement between both groups of annotators. On the other hand, the latter, which does not benefit from the general knowledge provided by pre-training, is more overfit to the expert annotations.

Overall, an adjacent accuracy of at least 84% shows that, similarly to what was observed on exam data, most misclassifications fall on neighboring complexity levels. Furthermore, the closeness between accuracy and F₁ scores suggests that the performance across all classes is relatively balanced. This is consistent with the subjectivity of the task.

4.2 Zero-Shot

Our experiments in a zero-shot setting aimed at assessing how increasing specificity in the prompt impacts performance. The results in Table 3 show that specifying the target audience of the text significantly improves the performance in comparison to the generic prompt that relies only on the model’s intrinsic knowledge of the CEFR and its levels. The benefits of providing the descriptions of the levels are less pronounced. While it led to improved performance in terms of all metrics when considering the expert annotations, it only led to a slight increase in terms of F₁ score at the expense of accuracy when considering trainer annotations.

Overall, the performance is not satisfactory in any case, suggesting that the LLM does not have an intrinsic ability to assess text complexity. Still, the model seems to be more aligned with the trainers than the experts. This exacerbates the differences in comparison with the baselines, with the performance in terms of accuracy and F₁ score being half of that of the baselines when considering expert annotations.

4.3 Few-Shot: Random Sample

Moving into few-shot experiments, Table 4 shows the results achieved when using random sampling to select the examples. The first main conclusion that can be drawn is that even a single example significantly increases performance in comparison to the zero-shot setting. In fact, when considering the trainer annotations, the performance equals that of the fine-tuned model.

In general, the performance seems to increase with the number of examples, but with diminishing returns, especially when considering trainer annotations. Albeit to a lesser extent than in the baselines, this is also probably due to the examples not being annotated by the trainers. Still, providing five random examples allowed the prompt-based approach to outperform the fine-tuned model by an average of 2.57 and 4.32 percentage points in terms of accuracy and F₁ score, respectively. Furthermore, when considering expert annotations, the 5-shot approach leads to an improvement of at least 12 percentage points in terms of every metric in comparison to the zero-shot setting.

Restricting the example pool to the same communication domain as the input text does not significantly affect the results, especially considering the high variability across runs introduced by the random selection of examples. However, this variability confirms that some examples are more relevant than others, supporting the use of the semantic search strategy for example selection.

In this set of experiments, the best results were achieved when three examples of each complexity level were provided to the system. This suggests that highlighting the potential differences between the levels provides important information to the model. However, further analysis is required to confirm this, as the improvement may be caused by the sheer amount of examples or even by more favorable samples.

4.4 Few-Shot: Semantic Search

Table 5 shows the results achieved when using semantic search for example selection. Even though this strategy leads to the highest performance among the prompt-based approaches, there are no major differences in comparison to when the examples are randomly chosen. This suggests that either the used embeddings are not able to capture the complexity of the texts, we are reaching the performance limits of the base LLM, or the task is so difficult and subjective that we are reaching the ceiling even in comparison to human performance. Still, we can observe some interesting patterns. First, as expected, the variability across runs is lower in this setting, as the provided examples for each text do not change. Second, using the Serafim model for document embedding seems to lead to the selection of examples that align better with the linguistics experts, while the selection based

Table 3. Zero-shot results using different prompts.

Prompt	Accuracy	Experts		Accuracy	Trainers	
		Adj. Acc.	Macro F ₁		Adj. Acc.	Macro F ₁
Generic	17.52±0.19	65.19±0.00	16.12±0.17	22.12±0.22	70.64±0.11	20.19±0.17
Target Audience	24.53±0.33	76.79±0.11	22.95±0.30	30.14±0.50	79.75±0.11	27.25±0.49
Descriptive	25.23±0.83	77.73±0.40	25.60±1.02	27.88±0.48	79.21±0.50	27.88±0.66

Table 4. Few-shot results using random samples.

Source	k	Accuracy	Experts		Accuracy	Trainers	
			Adj. Acc.	Macro F ₁		Adj. Acc.	Macro F ₁
Global	1	37.85±2.10	89.02±1.19	37.76±2.23	37.54±1.35	87.07±0.44	37.91±1.59
	3	42.13±0.48	89.49±0.19	42.12±0.18	39.95±2.85	86.84±0.67	40.60±3.02
	5	42.37±0.67	89.95±0.83	42.34±0.64	40.19±0.50	87.31±0.22	40.66±0.60
Comm. Domain	1	38.47±0.48	88.47±0.29	38.44±0.73	37.54±1.46	85.98±1.01	37.94±1.63
	3	42.99±2.20	90.97±0.44	42.68±2.31	39.88±1.12	87.85±0.50	40.18±1.21
	5	42.06±0.99	91.90±0.96	41.86±1.31	40.97±1.60	87.85±0.50	41.67±1.51
Level	1	40.03±0.55	89.02±1.01	40.30±0.48	40.03±0.96	86.45±0.19	40.75±0.89
	3	44.39±1.53	90.58±0.11	44.20±1.53	41.04±1.17	88.24±0.67	41.60±1.21
Level + Comm. Domain	1	39.33±2.24	88.63±0.22	39.74±2.14	38.32±1.75	85.51±0.83	39.25±1.81
	3	45.02±1.08	91.67±0.98	44.70±1.36	42.06±1.38	86.92±0.87	42.14±1.49

Table 5. Few-shot results using semantic search for example selection.

Source	Embedding	k	Accuracy	Experts		Accuracy	Trainers	
				Adj. Acc.	Macro F ₁		Adj. Acc.	Macro F ₁
Global	Serafim	1	41.36±0.33	89.72±0.33	40.04±0.36	37.62±0.38	88.08±0.19	36.93±0.41
		3	41.36±0.33	90.65±0.19	40.89±0.34	39.64±0.44	87.62±0.19	39.52±0.45
		5	42.29±0.57	92.68±0.22	42.03±0.56	39.88±0.11	89.41±0.40	40.00±0.14
	iR4S	1	38.71±0.40	89.17±0.22	37.11±0.37	42.52±0.83	88.86±0.29	41.49±0.84
		3	43.54±0.22	89.95±0.50	42.17±0.22	41.28±0.40	88.47±0.22	40.17±0.36
		5	40.81±0.67	90.11±0.11	38.95±0.99	40.81±0.40	88.47±0.11	39.80±0.59
Comm. Domain	Serafim	1	41.36±0.33	90.26±0.22	40.52±0.34	37.85±0.19	88.08±0.19	37.24±0.18
		3	42.99±0.69	90.73±0.29	41.95±0.60	38.71±0.44	88.55±0.19	37.89±0.26
		5	44.78±0.40	90.73±0.22	43.95±0.38	38.79±0.19	88.79±0.00	38.74±0.22
	iR4S	1	41.28±0.96	90.65±0.50	40.75±1.06	41.82±0.76	88.16±0.29	41.01±0.83
		3	42.99±0.69	91.04±0.11	42.02±0.54	40.89±0.76	89.49±0.19	39.97±0.47
		5	44.86±0.19	90.97±0.22	44.06±0.15	40.26±0.58	88.32±0.57	39.58±0.60
Level	Serafim	1	41.67±0.72	92.45±0.29	41.19±0.70	39.56±0.48	89.72±0.19	39.40±0.43
		3	44.86±0.50	92.60±0.11	44.07±0.43	42.29±0.50	88.63±0.40	42.33±0.33
	iR4S	1	39.64±0.29	88.08±0.33	38.73±0.30	41.67±0.29	86.84±0.44	40.94±0.31
		3	42.99±0.19	91.51±0.22	41.67±0.31	40.19±0.95	89.64±0.44	38.97±1.13
Level + Comm. Domain	Serafim	1	45.25±0.11	92.13±0.40	44.40±0.08	40.73±0.29	88.40±0.29	40.46±0.27
		3	45.64±0.48	92.99±0.33	45.44±0.44	41.12±0.83	89.17±0.40	41.55±0.82
	iR4S	1	40.58±0.40	89.88±0.11	40.18±0.55	40.89±0.50	86.68±0.19	40.62±0.52
		3	45.02±0.22	91.36±0.19	44.26±0.26	39.41±0.11	88.40±0.22	38.24±0.27

on the iRead4Skills features seems to align better with the trainers. Finally, the importance of providing examples of all levels is confirmed when considering the expert annotations, as higher performance is achieved when using the most similar example of each level than when using the three most similar examples regardless of the level.

Overall, using the prompt-based approach paired with semantic search for example selection, we achieved adjacent accuracy scores in line with or surpassing those of the baselines. Furthermore, as the model relies only on a small number of examples for each prediction, it is not as impacted by the differences between expert and trainer annotations as the training-based baselines. Consequently, when considering the trainer annotations, this approach surpasses the best baseline by around five percentage points in terms of both accuracy and F_1 score. However, as expected, that improvement is reverted when considering the expert annotations. Still, this highlights the higher generalization capabilities of prompt-based approaches.

4.5 Model Comparison

Table 6 summarizes the best results across all metrics for each base LLM. For simplicity, we mostly rely on this summary in the discussion below. The full set of results achieved using open-weight LLMs is presented in Appendix B.

Overall, GPT-4o mini achieved the best performance across all metrics. While the Llama 3.1 8B model can be considered competitive, especially in terms of accuracy, and showed patterns similar to GPT-4o mini, the same cannot be said for the EuroLLM 9B model, despite its specialization in European languages, including European Portuguese.

The lower performance of EuroLLM is primarily due to its difficulty in handling longer contexts. In many cases, the input examples did not fit within the model’s context window. When this occurred, we applied a fallback strategy that used the prediction of the model in the same setting, but with fewer examples. Nonetheless, even when examples fit within the context window, EuroLLM was more sensitive to the quality of example selection. In particular, performance dropped significantly when examples were selected at random.

Compared to GPT-4o mini, both open-weight models benefited more from example selection based on the similarity in terms of iRead4Skills features. Additionally, they had greater difficulty adhering to the instruction to answer with just the predicted level of complexity.

4.6 Error Analysis

Accuracy and F_1 scores below 50% reveal that automatic text complexity assessment is far from a solved task. Thus, it is important to perform a more fine-grained analysis of the results to assess whether there are relevant error patterns. Tables 7 and 8 show the confusion matrices of the best runs considering the annotations of linguistics experts and trainers, respectively. First of all, it is important to refer that both these runs relied on random sampling for example selection. This shows that the semantic search strategy is not identifying the most relevant examples and, thus, that there is room for improvement in that area. Potential approaches include

training a document embedding model focused on complexity aspects or selecting a curated subset of the iRead4Skills features that captures the most relevant aspects for similarity in terms of textual complexity.

Looking into the matrices, we can identify similar patterns in terms of precision and recall. The more complex level is that with both the lowest precision and the highest recall. Very easy texts are those with highest precision and plain texts are those with the lowest recall. Overall, these confusion matrices suggest that the model tends to overshoot the complexity level. This is confirmed by the confusion matrix shown in Table 9, which only includes the texts for which there is an agreement between the linguistics experts and the trainers. The performance on this subset is higher than on the full test set and only 1.5 percentage points below the performance of the fine-tuned model on the same subset in terms of every metric. Once again, this contributes to the argument that a significant part of the observed performance impairment is due to the high subjectivity of the task.

Finally, it is interesting to observe that nearly half of the examples with total disagreement between experts, trainers, and the model predictions correspond to texts for which there is a discrepancy of more than one level between experts and trainers. In these cases, the model tends to predict a level that is between the two annotations, which seems appropriate. Furthermore, most of these texts for which there is a disagreement are of the social media communication domain. Prominent examples include an article about vegetarianism and another telling the story of an immigrant family, both of which were classified as very easy by the experts and as more complex by the trainers. These examples are shown in Appendix A.5.

5 Conclusion

In this article, we have explored a new perspective on automatic text complexity assessment in European Portuguese. On the one hand, the iRead4Skills dataset allowed us to shift the focus from L2 learners, which were extensively targeted in previous research, to adult native speakers with low literacy. On the other hand, we relied on prompt-based LLMs for the task, with a special focus on few-shot prompting approaches, which rely on a small set of examples to guide the models towards more informed predictions.

Our experiments in a zero-shot setting showed that it is particularly important to explicit the target audience in the prompt, while providing descriptions of the complexity levels can improve performance but not to the same extent.

The few-shot experiments revealed that even a single example significantly increases the performance of the model and it keeps increasing with the number of examples, albeit with diminishing returns. Furthermore, it is important to provide examples of the multiple complexity levels, to ensure diversity and allow the model to infer which aspects contribute to the complexity. On the other hand, even though, on average, the highest performance was achieved when semantic search was used to select the examples, it did not represent a major increase in comparison to random sampling. This suggests that the embedding approaches used to generate the

Table 6. Comparison of the best results achieved using different base LLMs.

Model	Experts			Trainers		
	Accuracy	Adj. Acc.	Macro F ₁	Accuracy	Adj. Acc.	Macro F ₁
GPT-4o mini	45.64±0.48	92.99±0.33	45.44±0.44	42.52±0.83	89.72±0.19	42.33±0.33
Llama 3.1 8B	43.22±0.00	87.85±0.00	40.29±0.00	41.59±0.00	86.45±0.00	38.37±0.00
EuroLLM 9B	38.32±0.00	91.67±0.27	29.76±0.00	32.71±0.00	86.45±0.00	26.69±0.22

Table 7. Confusion matrix of the best run considering expert annotations. (Accuracy: 46.50%, Adjacent Accuracy: 91.36%, Macro F₁: 46.52%)

		Predicted			
		V. Easy	Easy	Plain	+Complex
Experts	V. Easy	55	45	12	0
	Easy	14	61	47	15
	Plain	2	29	58	50
	+Complex	1	7	7	25

Table 8. Confusion matrix of the best run considering trainer annotations. (Accuracy: 43.92%, Adjacent Accuracy: 86.92%, Macro F₁: 44.70%)

		Predicted			
		V. Easy	Easy	Plain	+Complex
Trainers	V. Easy	46	30	13	5
	Easy	14	58	37	20
	Plain	8	50	54	42
	+Complex	3	7	11	30

Table 9. Confusion matrix of the best run considering only the examples for which there is agreement between the experts and trainers. (Accuracy: 50.41%, Adjacent Accuracy: 94.21%, Macro F₁: 48.47%)

		Predicted			
		V. Easy	Easy	Plain	+Complex
Annotation	V. Easy	39	22	3	0
	Easy	8	33	26	9
	Plain	1	16	41	31
	+Complex	0	1	3	9

representation of the documents are not the most appropriate to capture complexity information.

Overall, the results of our experiments indicate that the problem posed by the iRead4Skills dataset is more challenging than that posed by the proficiency exam data commonly used in previous studies. This increased difficulty stems not only from the dataset’s diversity in communication domains and document types, but also from the dual annotations of the test set, provided by both linguistics experts and adult literacy trainers. These two annotation sources reflect distinct perspectives on textual complexity and diverge in multiple instances. As a result, models fine-tuned on expert annotations tend to generalize poorly to trainer annotations. In contrast, the few-shot prompting approach demonstrates better generalization, as it only relies on a small set of examples for each prediction instead of learning a fixed mapping.

Automatic text complexity assessment remains a challenging and largely unsolved task, offering several promising directions for future research. Within the few-shot prompting paradigm, future work should investigate alternative example selection strategies, either by identifying more effective similarity features or by training models specifically tailored to the task. Moreover, the performance of LLMs is progressing at a rapid pace, with frequent advances leading to improvements across a wide range of language understanding tasks. Thus, it is important to evaluate the performance of additional base LLMs, particularly larger models, to determine whether their enhanced capabilities on other language tasks translate into gains for text complexity assessment.

Beyond performance optimization, the inherently subjective nature of text complexity, often influenced by reader background, context, and purpose, highlights the need for models that are not only accurate but also interpretable. Developing models capable of providing transparent justifications for their assessments is especially valuable in educational contexts, where understanding why a text is deemed suitable or too difficult for a given audience is as important as the prediction itself. In this regard, the integration of chain-of-thought prompting and reasoning-based models [Qiao *et al.*, 2022; Wei *et al.*, 2022] represents a promising direction, as these models can provide a rationale for their assessment and insights for adapting texts to diverse proficiency levels and target audiences.

Declarations

Authors’ Contributions

All authors contributed to the conception of this study. JB acquired funding and supervised the project. ER performed the experiments

and is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

Availability of data and materials

The datasets analyzed during the current study are available in Zenodo¹². Additional materials, including the code used to perform the experiments, will be made available upon request.

References

- Akef, S., Mendes, A., Meurers, D., and Rebuschat, P. (2024). Investigating the Generalizability of Portuguese Readability Assessment Models Trained Using Linguistic Complexity Features. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 332–341. Available at: <https://aclanthology.org/2024.propor-1.34.pdf>.
- A Luisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability Assessment for Text Simplification. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Available at: <https://aclanthology.org/W10-1001.pdf>.
- Amaro, R., Monteiro, R., François, T., and de Deuxchaisnes, J. N. (2024). iRead4Skills Dataset 2: Annotated Corpora by Level of Complexity for FR, PT and SP. Number D3.7. DOI: 10.5281/zenodo.14653180.
- Baptista, J., Ribeiro, E., and Mamede, N. (2024). iRead4Skills @ IberSPEECH 2024: Project Presentation and Developments for the Portuguese Language. In *Proceedings of the IberSPEECH Conference*, pages 297–299. DOI: 10.21437/IberSPEECH.2024-62.
- Branco, A., Rodrigues, J., Costa, F., Silva, J., and Vaz, R. (2014a). Assessing Automatic Text Classification for Interactive Language Learning. In *Proceedings of the International Conference on Information Society (i-Society)*, pages 70–78. DOI: 10.1109/i-society.2014.7009014.
- Branco, A., Rodrigues, J., Costa, F., Silva, J., and Vaz, R. (2014b). Rolling out Text Categorization for Language Learning Assessment Supported by Language Technology. In *Proceedings of the International Conference on the Computational Processing of the Portuguese Language (PROPOR)*, pages 256–261. DOI: 10.1007/978-3-319-09761-9_29.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sasstry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. pages 1877–1901. DOI: 10.48550/arxiv.2005.14165.
- Cha, M., Gwon, Y., and Kung, H. (2017). Language Modeling by Clustering with Word Embeddings for Text Readability Assessment. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 2003–2006. DOI: 10.1145/3132847.3133104.
- Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2023). Unleashing the Potential of Prompt Engineering in Large Language Models: a Comprehensive Review. *Computing Research Repository*, arXiv:2310.14735. DOI: 10.48550/arXiv.2310.14735.
- Correia, J. and Mendes, R. (2021). Neural Complexity Assessment: A Deep Learning Approach to Readability Classification for European Portuguese Corpora. In *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 300–311. DOI: 10.1007/978-3-030-91608-4_30.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press. Available at: <https://rm.coe.int/1680459f97>.
- Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., and Kyle, K. (2017). Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, 54(5-6):340–359. DOI: 10.1080/0163853x.2017.1296264.
- Curto, P. (2014). Classificador de Textos para o Ensino de Português como Segunda Língua. Master’s thesis. Available at: <https://fenix.tecnico.ulisboa.pt/cursos/meic-a/dissertacao/565303595499701>.
- Curto, P., Mamede, N., and Baptista, J. (2015). Automatic Text Difficulty Classifier. In *Proceedings of the International Conference on Computer Supported Education (CSEDU)*, volume 1, pages 36–44. DOI: 10.5220/0005428300360044.
- Devlin, J., Chang, M.-W., Kenton, L., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, volume 1, pages 4171–4186. DOI: 10.18653/v1/N19-1423.
- DuBay, W. H. (2004). *The Principles of Readability*. Impact Information. Available at: <https://files.eric.ed.gov/fulltext/ED490073.pdf>.
- Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501. DOI: 10.1145/3637528.3671470.
- Filighera, A., Steuer, T., and Rensing, C. (2019). Automatic Text Difficulty Estimation Using Embeddings and Neural Networks. In *Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL)*, pages 335–348. DOI: 10.1007/978-3-030-29736-7_25.

¹²<https://zenodo.org/communities/iread4skills/>

- Forti, L., Grego Bolli, G., Santarelli, F., Santucci, V., and Spina, S. (2020). MALT-IT2: A New Resource to Measure Text Difficulty in Light of CEFR Levels for Italian L2 Learning. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 7204–7211. Available at: <https://aclanthology.org/2020.lrec-1.890>.
- François, T. and Fairon, C. (2012). An “AI Readability” Formula for French as a Foreign Language. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 466–477. Available at: <https://aclanthology.org/D12-1043>.
- François, T., Müller, A., Rolin, E., and Norré, M. (2020). AMesure: A Web Platform to Assist the Clear Writing of Administrative Texts. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL-IJCNLP): System Demonstrations*, pages 1–7. DOI: 10.18653/v1/2020.aacl-demo.1.
- Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–407. DOI: 10.1214/aos/1016218223.
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5). DOI: 10.1214/aos/1013203451.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *Computing Research Repository*, arXiv:2312.10997. DOI: 10.48550/arXiv.2312.10997.
- Giray, L. (2023). Prompt Engineering with ChatGPT: a Guide for Academic Writers. *Annals of Biomedical Engineering*, 51(12):2629–2633. DOI: 10.1007/s10439-023-03272-4.
- Gomes, L., Branco, A., Silva, J., Rodrigues, J., and Santos, R. (2025). Open Sentence Embeddings for Portuguese with the Serafim PT* Encoders Family. In *Proceedings of the EPIA Conference on Artificial Intelligence*, pages 267–279. DOI: 10.1007/978-3-031-73503-5_22.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202. DOI: 10.3758/BF03195564.
- Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., and Aluísio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 122–131. DOI: 10.48550/arxiv.1708.06025.
- Hernandez, N., Oulbaz, N., and Faine, T. (2022). Open Corpora and Toolkit for Assessing Text Readability in French. In *Proceedings of the Workshop on Tools and Resources to Empower People with REading Difficulties (READI)*, pages 54–61. Available at: <https://aclanthology.org/2022.readi-1.8>.
- Jönsson, S., Rennes, E., Falkenjack, J., and Jönsson, A. (2018). A Component Based Approach to Measuring Text Complexity. In *Proceedings of the Swedish Language Technology Conference (SLTC)*, pages 58–61.
- Karpov, N., Baranova, J., and Vitugin, F. (2014). Single-sentence Readability Prediction in Russian. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)*, pages 91–100. DOI: 10.1007/978-3-319-12580-0_9.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. DOI: 10.21236/ada006655.
- Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese. *Language Resources and Evaluation*. DOI: 10.1007/s10579-023-09693-w.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository*, arXiv:1907.11692. DOI: 10.48550/arXiv.1907.11692.
- Llama Team (2024). The Llama 3 Herd of Models. *Computing Research Repository*, arXiv:2407.21783. DOI: 10.48550/arXiv.2407.21783.
- Martinc, M., Pollak, S., and Robnik-Šikonja, M. (2021). Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179. DOI: 10.1162/coli_a_00398.
- Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J. G. C., Birch, A., and Martins, A. F. T. (2024). EuroLLM: Multilingual Language Models for Europe. *Computing Research Repository*, arXiv:2409.16235. DOI: 10.48550/arXiv.2409.16235.
- Marujo, L., Lopes, J., Mamede, N., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J., and Viana, C. (2009). Porting REAP to European Portuguese. In *Proceedings of the International Workshop on Speech and Language Technology in Education (SLaTE)*, pages 69–72. DOI: 10.21437/slate.2009-28.
- McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127. DOI: 10.1111/j.2517-6161.1980.tb01109.x.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press. DOI: 10.1017/CBO9780511894664.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS*, pages 3111–3119. Available at: <https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>.

- Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. (2022). Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3470–3487. DOI: 10.18653/v1/2022.acl-long.244.
- Mohtaj, S., Naderi, B., and Möller, S. (2022). Overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text. In *Proceedings of the GermEval Workshop on Text Complexity Assessment of German Text*, pages 1–9. Available at: <https://aclanthology.org/2022.germeval-1.1/>.
- Monteiro, R., Amaro, R., Correia, S., Pintard, A., Gauthola, R., Moutinho, M., and Blanco Escoda, X. (2023). iRead4Skills Complexity Levels. Number D3.1. DOI: 10.5281/zenodo.10459090.
- Nadeem, F. and Ostendorf, M. (2018). Estimating Linguistic Complexity for Science Texts. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55. DOI: 10.18653/v1/W18-0505.
- North, K., Zampieri, M., and Shardlow, M. (2023). Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42. DOI: 10.1145/3557885.
- OpenAI (2023). ChatGPT. Available at: <https://chat.openai.com/>.
- OpenAI et al. (2024). GPT-4o System Card. *Computing Research Repository*, arXiv:2410.21276. DOI: 10.48550/arXiv.2410.21276.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744. DOI: 10.48550/arxiv.2203.02155.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. Available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page.
- Pilán, I. and Volodina, E. (2018). Investigating the Importance of Linguistic Complexity Features Across Different Datasets Related to Language Learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58. Available at: <https://aclanthology.org/W18-4606>.
- Pintard, A., François, T., Nagant de Deuxchaisnes, J., Barbosa, S., Reis, M. L., Moutinho, M., Monteiro, R., Amaro, R., Correia, S., Rodríguez Rey, S., García González, M., Mu, K., and Blanco Escoda, X. (2024). iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP. Number D3.2. DOI: 10.5281/zenodo.13768477.
- Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. (2022). Reasoning with Language Model Prompting: A Survey. *Computing Research Repository*, arXiv:2212.09597. DOI: 10.48550/arXiv.2212.09597.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. DOI: 10.18653/v1/D19-1410.
- Reis, M. L., Barbosa, S., Moutinho, M., Monteiro, R., Correia, S., and Amaro, R. (2024). Intelligent Support for Low Literacy Adults: The European Portuguese iRead4Skills Corpus. *International Journal of Emerging Technologies in Learning (iJET)*, 19(8):61–81. DOI: 10.3991/ijet.v19i08.52023.
- Reynolds, R. (2016). Insights from Russian Second Language Readability Classification: Complexity-Dependent Training Requirements, and Feature Evaluation of Multiple Categories. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300. DOI: 10.18653/v1/W16-0534.
- Ribeiro, E., Mamede, N., and Baptista, J. (2024a). Automatic Text Readability Assessment in European Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 97–107. Available at: <https://aclanthology.org/2024.propor-1.10.pdf>.
- Ribeiro, E., Mamede, N., and Baptista, J. (2024b). Avaliação Automática do Nível de Complexidade de Textos em Português Europeu. *Linguamática*, 16(2):121–145. DOI: 10.21814/lm.16.2.449.
- Ribeiro, E., Mamede, N., and Baptista, J. (2024c). Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 551–557. Available at: <https://aclanthology.org/2024.propor-1.59.pdf>.
- Ribeiro, E., Wilkens, R., Braña, A. B., Bolos, A. C., Mamede, N., Baptista, J., González, M. G., Amaro, R., and François, T. (2024d). ICA API Documentation. Project Deliverable D4.2, iRead4Skills. Available at: <https://gitlab.hlt.inesc-id.pt/iread4skills/docs/>.
- Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. In *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*, page 441–453. DOI: 10.1007/978-3-031-49008-8_35.
- Rodríguez Rey, S., Bernárdez Braña, A., and García, M. (2025). Exploring Linguistic Features in

- a New Readability Corpus for Spanish. *Procesamiento del Lenguaje Natural*, 74:221–239. Available at: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6677>.
- Santos, R., Rodrigues, J., Branco, A., and Vaz, R. (2021). Neural Text Categorization with Transformers for Learning Portuguese as a Second Language. In *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*, pages 715–726. DOI: 10.1007/978-3-030-86230-5_56.
- Santucci, V., Santarelli, F., Forti, L., and Spina, S. (2020). Automatic Classification of Text Complexity. *Applied Sciences*, 10(20):7285. DOI: 10.3390/app10207285.
- Sawarkar, K., Mangal, A., and Solanki, S. R. (2024). Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. In *Proceedings of the International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 155–161. DOI: 10.1109/MIPR62202.2024.00031.
- Scarton, C. E. and Aluísio, S. M. (2010). Análise da Inteligibilidade de Textos via Ferramentas de Processamento de Língua Natural: Adaptando as Métricas do Coh-Metrix para o Português. *Linguamática*, 2(1):45–61. Available at: <https://linguamatica.com/index.php/linguamatica/article/view/44>.
- Sung, Y. T., Lin, W. C., Dyson, S. B., Chang, K. E., and Chen, Y. C. (2015). Leveling L2 Texts through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, 99(2):371–391. DOI: 10.1111/modl.12213.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *Computing Research Repository*, arXiv:2302.13971. DOI: 10.48550/arXiv.2302.13971.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 24824–24837. Available at: <https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html>.
- Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K., and François, T. (2022). FABRA: French Aggregator-Based Readability Assessment Toolkit. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1217–1233. Available at: <https://aclanthology.org/2022.lrec-1.130>.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). HuggingFace’s Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 38–45. DOI: 10.48550/arXiv.1910.03771.
- Xia, M., Kochmar, E., and Briscoe, T. (2016). Text Readability Assessment for Second Language Learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22. DOI: 10.18653/v1/W16-0502.
- Yancey, K., Pintard, A., and Francois, T. (2021). Investigating Readability of French as a Foreign Language with Deep Learning and Cognitive and Pedagogical Features. *Lingue e Linguaggio*, 20(2):229–258. DOI: 10.1418/102814.

A Examples

This appendix shows an example of a text of each complexity level, randomly selected from those in which there is agreement between the linguistics experts and the trainers. Additionally, in Appendix A.5, we provide some prominent examples of classification disagreements between the experts and the trainers, which were also a source of confusion for automatic classification. The communication domain and document type of each example are provided in addition to the complexity level. Considering the length of some of the examples and the fact that translation may impact the complexity level, we only provide the original version in Portuguese.

A.1 Very Easy

Complexity Level: Very Easy

Communication Domain: Non-fiction Book

Document Type: (Auto)biography

A minha amiga Rita trabalha num Banco em Santarém. Como ela mora em Lisboa, tem de se levantar todos os dias muito cedo, pois começa a trabalhar antes das 8.30. Ela normalmente vai de carro para o trabalho, mas às vezes prefere ir de comboio. À tarde ela nunca tem uma hora fixa para sair. Muitas vezes só consegue sair do Banco entre as 7 e as 8 da noite. A Rita gostava de trabalhar em Lisboa, mas sabe que não vai ser fácil. Aos fins-de-semana ela gosta de ir jogar ténis comigo e ao sábado à noite sai sempre com os amigos: vão ao cinema, jantam fora, vão à discoteca, etc.

A.2 Easy

Complexity Level: Easy

Communication Domain: Social Media (Newspapers/Magazines)

Document Type: Reportage

Um dos problemas mais comuns de quem vive num prédio é o barulho dos vizinhos. Quer esteja em teletrabalho ou não, é importante saber como lidar com estas situações.

Estão sempre a martelar e não a deixam dormir? O vizinho do lado gosta de ouvir música até de madrugada? A vizinha do terceiro direito só aspira depois das 21 horas? Se é vítima de alguma destas situações de ruído de vizinhança, ou de outras, deve saber o que a lei diz sobre elas.

A.3 Plain

Complexity Level: Plain

Communication Domain: Non-fiction Book

Document Type: Travel Report/Diary

Diz-se que, quando em Nova Iorque são três da tarde, na Europa são nove horas há dez anos atrás. Talvez assim seja. A voracidade do tempo foi avançando para ocidente e fechou o círculo no Oriente: o futuro de hoje ruge em Xangai. Não sei se Nova Iorque vai uma década à frente. O cinema da nossa memória torna-a tão conhecida, que já faz parte do passado. Mas não interessa, cheguei com atraso e tenho ideias confusas sobre o progresso. Podendo escolher, teria visto pela primeira vez os molhes do Hudson por volta de 1960, do convés de um transatlântico, e teria desembarcado numa cidade onde não havia almoço sem antes se tomarem três martinis, nem taxistas sem gravata, onde se fumava sem filtro, num tempo em que Times Square era a Babilónia e não uma encruzilhada barulhenta envolta em anúncios luminosos. A cidade de 1960 era jovem e cínica, arrogante, intacta.

Como segunda opção, escolheria o longo Verão dos anos vinte, corrupto e turbulento, com uma viagem num navio da marinha mercante e uma chegada nocturna aos molhes industriais do East River. Da ponte de Brooklyn, com o sol nascente nas costas, teria visto o amanhecer reflectido numa silhueta urbana que não era a mais célebre do mundo, nem tinha tantas torres de vidro como hoje. O lado oriental de Manhattan, com as planícies de Greenwich, os cumes de cimento e mármore de Midtown, e as colinas de Battery, ainda em construção.

Outra hipótese seria chegar hoje mesmo. Os Yankees ganharão e os Mets perderão em circunstâncias escandalosas; as pessoas passarão junto de um terreno em construção onde antes havia duas torres muito altas e olharão, como de costume, para as montras da Century 21; no Holland Tunnel continuará assinalada a fronteira com Nova Jérсия e os Estados Unidos, esse país imenso, absorto nos seus centros comerciais, nas suas bíblias, nos seus revólveres e nos seus fantasmagóricos inimigos exteriores; e em Washington Square alguém se sentará diante do tabuleiro ocupado por Bobby Fischer e deslocará, como ele, o peão do bispo da rainha preta para construir uma defesa siciliana. Em Nova Iorque, que nada sabe da nossa memória sentimental nem do nosso calendário, é sempre hoje e todos os momentos contam.

«Em Nova Iorque, o presente é tão poderoso, que o passado se perdeu.» Quem o disse foi John Jay Chapman, ensaísta nova-iorquino que em 1900 proferiu o discurso de graduação no Hobart College com a seguinte recomendação: «Esqueçam a reputação. Deixem-se odiar, deixem-se ridicularizar, deixem-se ter medo e deixem-se duvidar, mas não se deixem amordaçar. Façam o que bem vos apetece, mas digam sempre o que pensam.» Ignoro o que terão feito aqueles jovens na vida. Se deram ouvidos a Chapman e se negaram a estar calados, foram típicos cidadãos nova-iorquinos.

Discordo, no entanto, da afirmação de que o «passado se perdeu». Não. O passado é esquecido sem contudo se perder.

O passado de Nova Iorque é dominado pela Holanda, a potência fundadora, e diverge dos restantes passados norteamericanos. Nova Iorque não foi puritana como as outras colónias; Nova Iorque nasceu do comércio, não da agricultura, e acreditou mais em piratas do que em pregadores; Nova Iorque teve uma breve relação com a escravatura (já com o dinheiro dos escravagistas, a coisa foi diferente), teve pouca fé na independência e nunca mostrou um respeito exemplar pela autoridade no seio da União. Nova Iorque, nascida Nova Amesterdão, foi e ainda é refúgio de livres-pensadores, charlatães, inadaptados e gente excêntrica. Os primeiros quatrocentos habitantes de origem europeia falavam dezoito idiomas distintos, apesar de serem quase todos oriundos de Amesterdão.

Se disso restassem dúvidas, a bandeira da cidade de Nova Iorque exhibe as cores azul, branca e laranja, as mesmas da bandeira holandesa no século XVII. No brasão, vêem-se pás de moinho, um marinheiro, um índio, dois castores e uns barris de farinha.

A.4 More Complex

Complexity Level: More Complex

Communication Domain: Didactic Book

Document Type: Encyclopedia/Dictionary

TETRÁPODES

Alguns dos peixes atuais, como os dipnoicos, possuem quatro barbatanas fortes e carnudas, que são quase como pernas. Há cerca de 380 milhões de anos, alguns destes peixes com barbatanas lobadas viviam em pântanos de água doce e começaram a rastejar para fora de água em busca de alimento, regressando à água para desovar, como o fazem os atuais anfíbios. Foram os primeiros tetrápodes e os antepassados dos vertebrados terrestres.

ESQUELETOS FORTES

O corpo de um vertebrado aquático, como os répteis marinhos, é suportado pela água. Logo, a principal função do esqueleto é ancorar os músculos. Todavia, o mesmo tipo de esqueleto pode também suportar o peso de um animal terrestre, só que os ossos são muito mais fortes e estão interligados por articulações que suportam o peso. Esta adaptação permitiu a evolução dos vertebrados terrestres, incluindo os dinossauros gigantes.

ANIMAIS GIGANTES

Os maiores animais terrestres sempre foram vertebrados. Assim e porque os pesados animais terrestres carecem de esqueletos internos fortes para suportar o seu peso. Há porém um limite, e é provável que o dinossauro gigante *Argentinosaurus* tenha sido o mais pesado no que diz respeito a animais terrestres. O único vertebrado que pesa mais do que ele vive nos oceanos: é a baleia azul.

A.5 Disagreements

Expert Classification: Very Easy

Trainer Classification: More Complex

Communication Domain: Social Media (Newspapers/Magazines)

Document Type: News

Porque (não) somos vegetarianos?

Comer é um acto de prazer. É uma condição para nos mantermos vivos e saudáveis. É também sinal dos tempos, uma forma de estar e socializar. Deixar de comer carne é, cada vez para mais pessoas, uma opção inteligente e sensata. Não concorda? Então reveja os seus argumentos.

Já reparou que há cada vez mais restaurantes vegetarianos, lojas de produtos naturais e secções destinadas a estes alimentos nos supermercados, hipermercados e feiras de alimentação? Têm rótulos sugestivos de “natural” “sem açúcar”, “sem lactose”, “sem colesterol”, “menos gordura”, “menos calorias”, e são muitos os que os procuram para perder aqueles quilinhos a mais ou para controlar o colesterol...

Mas, a par dos consumidores “sazonais”, encontram-se vegetarianos convictos, movidos pela crescente sensibilização para as questões éticas e ecológicas e pela noção de que a dieta vegetariana é mais saudável.

Entre os vegetarianos distinguem-se dois grandes grupos: os que consomem ovos e lacticínios e os mais radicais (os vegetarianos puros ou vegans, que excluem todos os produtos de origem animal da alimentação, e também do vestuário, recusando-se a usar lãs, sedas e todo o tipo de peles, por considerarem que o ser humano não tem o direito de usar os animais).

Muitos vegetarianos “queixam-se” de que são olhados com estranheza e suspeição, de que as pessoas não estão informadas, incluindo muitos profissionais de saúde e da área alimentar. A pressão para o consumo de carne e produtos de origem animal é muito grande em Portugal. No entanto, até há cinquenta anos, poucas pessoas consumiam regularmente carne. Comia-se mais feijão, grão, vegetais, pão escuro... e de vez em quando peixe.

A carne está associada a uma imagem de abundância, poder e força.

Os vegetarianos não comem só “ervas”, nem têm obrigatoriamente pratos estranhos, com nomes esquisitos e difíceis de confeccionar. A cozinha vegetariana é mais fácil do que se pensa e muito mais diversificada e saborosa do que imagina. Vale a pena experimentar!

Expert Classification: Very Easy

Trainer Classification: More Complex

Communication Domain: Social Media (Newspapers/Magazines)

Document Type: Profile

Família Reis

São Tomé e Príncipe

Ana Paula tem 46 anos, 19 dos quais passados em Portugal. Foi a primeira da família a emigrar e tem trabalhado no bairro da Mouraria desde que chegou ao nosso país. A sobrinha Jusnéria, de 26 anos, partiu de São Tomé há cinco anos e meio e tem trabalhado com a tia nos dois últimos anos. Da ilha natal têm sentido saudades do peixe fresco. Comem, com frequência, calulu, um prato típico de São Tomé, mas sentem a falta das várias verduras frescas que este prato leva: mussua, maquequê, quiabo, ossami, pau-pimenta e óleo de palma feito em casa com acompanhamento de farinha de mandioca e banana (angu).

B Results using Open-Weight LLMs

This appendix shows the full set of results achieved using the two open-weight LLMs: Llama 3.1 8B and EuroLLM 9B.

B.1 Llama 3.1 8B

Starting with the zero-shot setting, in Table 10, we can see that, similarly to GPT-4o mini, this model seems to be more aligned with the trainers than the experts and the performance increases with the specificity and detail of the prompt. However, in this case, the performance differences are more pronounced, with the descriptive prompt significantly outperforming the others. This leads to this model of the Llama family outperforming GPT-4o mini in the zero-shot setting and even the classical ML baseline when considering trainer annotations.

Table 10. Llama 3.1 8B: Zero-shot results using different prompts.

Prompt	Accuracy	Experts		Accuracy	Trainers	
		Adj. Acc.	Macro F ₁		Adj. Acc.	Macro F ₁
Generic	14.64±0.13	56.85±0.13	12.20±0.13	20.09±0.40	62.31±0.27	16.36±0.37
Target Audience	23.91±0.27	74.84±0.13	19.05±0.22	27.34±0.00	77.10±0.40	20.29±0.09
Descriptive	31.46±0.13	84.27±0.27	23.37±0.06	34.42±0.27	86.99±0.13	24.92±0.13

Moving into the few-shot setting, in Table 11, we can see that providing examples selected at random can improve performance, but not to the extent observed when using GPT-4o mini. As the examples were annotated by the experts, a more pronounced improvement is observed when considering expert annotations than when considering trainer annotations. For the same reason, while the performance increases with the number of examples when considering expert annotations, the same is not as clear when considering trainer annotations. Furthermore, while the stratification by communication domain led to improved performance, once again especially when considering expert annotations, the stratification by level significantly impacted performance, bringing it to levels below the zero-shot setting. This suggests that the model is more influenced by the quality of the context than GPT-4o mini.

Table 11. Llama 3.1 8B: Few-shot results using random samples.

Source	k	Accuracy	Experts		Accuracy	Trainers	
			Adj. Acc.	Macro F ₁		Adj. Acc.	Macro F ₁
Global	1	34.74±1.15	81.07±0.62	25.49±2.36	36.84±1.59	83.57±0.88	26.63±3.40
	3	35.28±0.62	83.10±1.59	29.88±0.87	37.07±1.15	85.20±0.36	31.46±1.96
	5	34.58±0.84	84.11±1.24	31.08±1.52	33.80±1.15	83.80±1.05	29.39±0.26
Comm. Domain	1	35.67±2.98	82.63±1.33	28.86±2.88	35.98±1.53	83.80±0.27	27.68±2.04
	3	36.06±0.88	83.10±0.94	29.34±2.46	36.76±1.50	83.64±0.84	29.30±0.17
	5	37.15±1.91	84.50±0.27	32.09±2.42	36.68±3.67	84.81±1.17	32.55±3.70
Level	1	26.56±2.12	72.35±1.56	22.36±3.21	31.78±1.62	75.23±1.42	26.17±4.25
	3	29.91±0.84	70.56±3.71	26.73±0.92	27.88±1.64	67.45±1.50	26.24±1.50
Level + Comm. Domain	1	26.95±2.50	70.95±0.27	24.61±2.47	29.21±1.46	73.68±0.94	25.58±1.53
	3	27.57±0.70	69.78±1.10	24.24±0.81	27.26±1.56	67.52±1.53	25.45±1.31

Table 12 shows the results achieved when using semantic search for example selection. We can observe significant improvements in comparison to when using random example selection, which confirms that the model’s sensitivity to the quality of the provided examples. In this case, when considering expert annotations, the best accuracy was achieved using the per-level stratification approach. This suggests that by receiving examples of each level that are similar to the text it is assessing, the model is able to identify the subtleties that distinguish the different complexity levels. Furthermore, in comparison to GPT-4o mini, the Llama 3.1 8B model seems to benefit more from example selection based on similarity in terms of the iRead4Skills features. As this model seems to be more sensitive to the provided context, this suggests that the handcrafted features can actually be more informative for the task than generic embeddings used for semantic similarity.

Overall, the Llama 3.1 8B model can be considered competitive with GPT-4o mini for this task, especially when considering trainer annotations, as it also outperforms both baselines in that setting.

Table 12. Llama 3.1 8B: Few-shot results using semantic search for example selection.

Source	Embedding	k	Accuracy	Experts	Macro F ₁	Accuracy	Trainers	Macro F ₁
				Adj. Acc.			Adj. Acc.	
Global	Serafim	1	38.01±0.13	82.24±0.00	31.48±0.12	36.06±0.13	83.18±0.00	27.23±0.07
		3	38.71±0.13	86.92±0.00	33.65±0.38	39.64±0.13	83.88±0.00	34.89±0.31
		5	41.12±0.00	87.15±0.00	40.29±0.00	40.19±0.00	86.45±0.00	37.32±0.00
	iR4S	1	34.97±0.13	83.33±0.13	29.89±0.11	35.59±0.13	83.02±0.13	28.93±0.04
		3	41.20±0.13	84.81±0.00	36.76±0.16	39.10±0.13	84.58±0.00	33.20±0.19
		5	41.36±0.00	86.21±0.00	38.51±0.00	39.49±0.00	83.88±0.00	35.09±0.00
Comm. Domain	Serafim	1	38.40±0.13	83.49±0.13	32.38±0.17	36.21±0.40	84.03±0.13	27.74±0.34
		3	38.24±0.13	86.37±0.13	33.54±0.15	35.90±0.13	86.14±0.13	29.92±0.18
		5	42.29±0.00	86.21±0.00	39.95±0.00	38.08±0.00	85.28±0.00	34.34±0.00
	iR4S	1	38.94±0.13	83.33±0.13	26.73±0.17	38.24±0.54	82.17±0.13	24.42±0.36
		3	41.82±0.00	87.85±0.00	38.95±0.00	39.72±0.00	85.75±0.00	32.63±0.00
		5	42.29±0.00	87.62±0.00	38.85±0.00	41.59±0.00	86.45±0.00	38.37±0.00
Level	Serafim	1	37.62±0.00	83.64±0.00	32.45±0.00	38.55±0.00	83.18±0.00	32.77±0.00
		3	41.82±0.00	82.71±0.00	36.80±0.00	35.75±0.00	75.23±0.00	31.16±0.00
	iR4S	1	31.31±0.00	76.09±0.13	24.88±0.00	35.20±0.13	79.91±0.00	28.42±0.13
		3	43.22±0.00	85.05±0.00	35.48±0.00	38.55±0.00	78.27±0.00	31.06±0.00
Level + Comm. Domain	Serafim	1	41.59±0.00	86.21±0.00	33.73±0.00	41.36±0.00	84.11±0.00	34.84±0.00
		3	36.45±0.00	77.80±0.00	31.50±0.00	36.21±0.00	72.43±0.00	33.56±0.00
	iR4S	1	33.41±0.00	77.80±0.00	26.81±0.00	35.51±0.00	80.14±0.00	28.03±0.00
		3	41.59±0.00	83.88±0.00	34.45±0.00	35.98±0.00	79.91±0.00	31.55±0.00

B.2 EuroLLM 9B

Starting with the zero-shot setting, in Table 13, we can see that EuroLLM is the highest performer among the three models when considering the expert annotations and, similarly to the model of the Llama family, it outperforms GPT-4o mini when considering trainer annotations. This can be due to the model’s specialization in European languages, including European Portuguese. However, this model exhibits the most erratic result patterns among the three. First, it does not seem to be more aligned with the trainer annotations. Second, there does not seem to be a clear improvement with the more detailed prompts. When considering trainer annotations, each prompt leads to the highest performance in one of the metrics. When considering expert annotations, the description of the complexity levels leads to a decrease in performance. This may be due to the model’s difficulty in handling longer contexts, as discussed in Section 4.5.

Table 13. EuroLLM 9B: Zero-shot results using different prompts.

Prompt	Accuracy	Experts	Macro F ₁	Accuracy	Trainers	Macro F ₁
		Adj. Acc.			Adj. Acc.	
Generic	28.35±0.27	73.05±0.13	25.17±0.25	28.74±0.40	73.75±0.13	26.69±0.22
Target Audience	33.57±0.67	91.67±0.27	19.33±0.55	32.24±0.40	86.45±0.00	19.00±0.23
Descriptive	31.93±0.27	84.81±0.00	15.07±0.27	32.63±0.54	81.70±0.13	15.92±0.31

The model’s inability to handle longer contexts is highlighted by the results in Table 14, which reveal a significant performance impairment when randomly selected examples are provided as context. Furthermore, the performance decreases with the number of examples. Still, similarly to when using the other base LLMs, selecting examples of the same communication domain improves the performance in comparison to a completely random selection.

Table 15 shows the results achieved when the examples are selected using semantic search. We can see that, in this case, the model is actually able to take some advantage of the provided examples and exceed the performance of the zero-shot setting when considering expert annotations. This highlights the importance of selecting appropriate examples to guide the model. However, in most cases, the performance of the model is still hindered when more than one example is provided. Similarly to the model of the Llama family, EuroLLM seems to benefit more from example selection based on the iRead4Skills features.

Overall, although EuroLLM is specialized in European languages, including European Portuguese, which may benefit the model’s performance in a zero-shot setting, the model is significantly impacted by its difficulty in handling longer contexts. This limitation prevents the model from fully leveraging the examples provided in the few-shot setting. As a result, it underperforms in comparison to the other two base LLMs evaluated in this study.

Table 14. EuroLLM 9B: Few-shot results using random samples.

Source	k	Experts			Trainers		
		Accuracy	Adj. Acc.	Macro F ₁	Accuracy	Adj. Acc.	Macro F ₁
Global	1	28.19±2.09	80.22±2.25	18.84±2.37	27.26±1.29	75.39±1.69	18.16±1.11
	3	27.26±1.75	69.24±0.88	19.44±0.95	26.95±1.50	67.13±1.56	19.69±0.73
	5	23.91±2.17	60.36±4.12	18.63±1.45	22.43±2.83	55.92±4.41	17.82±2.64
Comm. Domain	1	31.93±2.17	79.67±1.21	20.98±1.68	29.44±1.46	75.00±0.47	19.38±1.56
	3	26.95±0.59	70.79±2.14	18.59±0.36	23.29±1.66	67.37±1.69	16.08±1.23
	5	26.71±3.75	64.49±4.46	20.55±2.96	21.50±3.46	60.20±2.63	16.63±1.98
Level	1	23.99±0.49	57.79±1.95	15.73±1.01	21.03±1.30	54.75±0.88	14.43±1.18
	3	23.91±0.36	57.63±1.89	15.70±0.95	20.95±1.41	54.52±0.67	14.40±1.28
Level + Comm. Domain	1	24.07±0.84	56.15±0.94	15.31±0.40	20.72±2.12	51.95±1.72	13.98±0.83
	3	23.44±0.94	54.98±0.94	15.03±0.34	20.48±2.09	51.01±2.11	14.00±0.89

Table 15. EuroLLM 9B: Few-shot results using semantic search for example selection.

Source	Embedding	k	Experts			Trainers		
			Accuracy	Adj. Acc.	Macro F ₁	Accuracy	Adj. Acc.	Macro F ₁
Global	Serafim	1	32.71±0.00	83.02±0.13	21.52±0.00	31.07±0.00	78.04±0.00	21.51±0.00
		3	31.62±0.13	77.80±0.00	24.06±0.08	28.82±0.13	70.79±0.00	21.55±0.08
		5	26.40±0.00	64.49±0.00	20.00±0.00	24.07±0.00	59.81±0.00	19.40±0.00
	iR4S	1	38.32±0.00	85.98±0.00	27.11±0.00	30.61±0.00	80.37±0.00	22.11±0.00
		3	38.32±0.00	81.07±0.00	29.76±0.00	29.44±0.00	73.83±0.00	21.92±0.00
		5	36.45±0.00	76.64±0.00	28.01±0.00	28.74±0.00	70.79±0.00	22.31±0.00
Comm. Domain	Serafim	1	35.05±0.00	83.64±0.00	24.03±0.00	32.71±0.00	80.61±0.00	22.91±0.00
		3	33.80±0.13	79.60±0.13	23.83±0.31	29.52±0.13	75.78±0.13	21.43±0.08
		5	30.61±0.00	66.36±0.00	22.96±0.00	24.07±0.00	59.11±0.00	19.38±0.00
	iR4S	1	36.37±0.27	85.83±0.13	25.97±0.38	27.26±0.13	79.21±0.00	19.05±0.08
		3	37.69±0.13	80.92±0.13	28.49±0.25	29.44±0.00	73.83±0.00	21.92±0.00
		5	33.96±0.13	76.71±0.13	25.75±0.25	27.80±0.00	71.26±0.00	20.81±0.00
Level	Serafim	1	26.71±0.13	68.46±0.00	17.64±0.41	23.13±0.00	63.01±0.13	15.61±0.00
		3	26.56±0.27	68.07±0.13	18.13±0.48	23.44±0.13	62.23±0.13	16.40±0.09
	iR4S	1	29.13±0.13	79.05±0.27	16.44±0.09	30.69±0.13	76.48±0.27	17.62±0.07
		3	30.61±0.00	72.90±0.00	18.84±0.02	26.64±0.00	67.60±0.27	17.58±0.01
Level + Comm. Domain	Serafim	1	26.79±0.13	65.19±0.00	17.26±0.06	24.22±0.13	60.98±0.00	17.70±0.07
		3	26.79±0.13	64.25±0.00	17.35±0.06	24.22±0.13	60.05±0.00	17.70±0.07
	iR4S	1	27.80±0.00	75.16±0.13	16.00±0.00	31.78±0.00	73.75±0.13	18.53±0.01
		3	27.57±0.00	66.74±0.13	17.88±0.01	27.57±0.00	64.64±0.13	19.23±0.00