How Do Large Language Models Evaluate Lexical Complexity?

Anonymous ACL submission

Abstract

In this work, we explore the prediction of lexical complexity by combining supervised approaches and the use of large language models (LLMs). We first evaluate the impact of different prompting strategies (zero-shot, one-shot, and chain-of-thought) on the quality of the predictions, comparing the results with human annotations from the CompLex 2.0 corpus. Our results indicate that LLMs, and in particular gpt-4o, benefit from explicit instructions to better approximate human judgments, although some discrepancies remain. Moreover, a calibration approach to better align LLMs predictions and human judgements based on few manually annotated data appears as a promising solution to improve the reliability of the annotations in a supervised scenario.

1 Introduction

The prediction of lexical complexity is an essential task for adapting linguistic content to the specific needs of learners and educational systems. Such a task consists in predicting a numerical complexity score for a target word in a given sentence (thereafter an instance). Data annotation plays a key role in this task, directly influencing the performance of supervised models. With the emergence of large-scale language models (LLMs) the possibility of using automatically generated annotations raises new questions regarding the generalization and robustness of these models.

In this work, we focus on measuring the similarities between human annotators and generative models (LLMs) by varying the prompts. The objective is to determine whether it is possible to use LLMs as reliable annotators by measuring their level of agreement with human annotations and by analyzing the distribution of the produced annotations. We also seek to identify new perspectives for improving the alignment between these two sources of annotations using a calibration model based on few manually annotated data. We specifically apply this approach in the context of a supervised model trained on LLM-based annotated data to avoid the use of LLMs at prediction time, prioritizing time efficiency and energy conservation. All our experiments were performed on the CompLex 2.0 dataset (Shardlow et al., 2021) for English. This dataset has the advantage of including the source individual human annotations that can be used for directly comparing human and LLM annotations.

The paper is organized as follows. Section 2 presents related work with respect to lexical complexity prediction and data annotation using LLMs. Next, section 3 describes the Complex 2.0 dataset, the LLM strategies to be tested as well as the supervised model used in the final experiments. Then, sections 4 and 5 evaluate LLMs performance against human annotations. Finally, section 6 explores supervised scenarios integrating a calibration model for LLMs.

2 Related work

2.1 Lexical complexity prediction

Lexical complexity is a key issue in text simplification and accessibility. North et al. (2023) provide a comprehensive review of computational methods for predicting lexical complexity primarily in English texts. Their work aims to enhance comprehension by identifying complex words and substituting them with simpler alternatives. The review covers both traditional machine learning techniques, such as support vector machines and logistic regression, and advanced deep neural network models. Moreover, the authors emphasize the use of diverse features including psycholinguistic cues, word frequency, and word length and discuss dedicated competitions, datasets, and practical applications in readability assessment and text simplification across multiple languages.

Research emerging from shared tasks on this sub-

ject highlights the evolution of the field and significant advances in lexical complexity prediction.In 2018, the Complex Word Identification (CWI) shared task marked a turning point by proposing systems capable of identifying words that may be difficult for readers, depending on various contexts. Yimam et al. (2018) revealed that simple models based on n-grams could rival more complex approaches, emphasizing the importance of data and linguistic features in this task. In addition, Gooding and Kochmar (2018) proposed a method based on an ensemble system using majority voting among several models, demonstrating that combining diverse predictors improves overall performance and yields robust results. Furthermore, Kajiwara and Komachi (2018) explored an approach based on lexical frequency in a learner corpus, showing that this methodology is particularly well suited for educational contexts.

Research on predicting lexical complexity has progressed significantly thanks to contributions from the shared task LCP 2021 (Shardlow et al., 2021), which explored the prediction of the complexity of simple words and multiword expressions. Pan et al. (2021) proposed an approach based on a deep ensemble combining pre-trained models such as BERT with advanced techniques such as pseudo-labeling and data augmentation, achieving remarkable results, including first place for multiword expressions. Similarly, Yaseen et al. (2021) used pre-trained models BERT and RoBERTa to compute complexity scores on a continuous scale, ranking first for simple words with a Pearson correlation coefficient of 0.788. Moreover, Mosquera (2021) demonstrated that manual engineering of contextual, lexical and semantic features can still rival modern models, obtaining high correlations for both simple words and multiword expressions. In a more recent study on the LCP 2021 dataset, Kelious et al. (2024b) compared the performance of ChatGPT with that of dedicated models, showing that prompt engineering allows ChatGPT to be competitive, albeit less consistent than specialized models, which reached an R^2 score of 0.65. In parallel, the same authors explored multilingual strategies, comparing supervised and generative approaches to predict lexical complexity. The generative models, although achieving high correlations with prompting strategies (zero-shot, one-shot, etc.), are still surpassed by models optimized for specific tasks. These contributions illustrate a combination of modern and traditional approaches to address the challenges of lexical complexity in both monolingual and multilingual contexts (Kelious et al., 2024a).

Recent research on predicting lexical complexity and text simplification, particularly in multilingual contexts, demonstrates significant advances through the integration of modern techniques. The BEA 2024 shared task explored these aspects in ten languages, using open and proprietary language models, while showing the potential for improvement in complex tasks (Shardlow et al., 2024). Enomoto et al. (2024) [TMU-HIT] demonstrated the effectiveness of GPT-4 in assessing and simplifying lexical complexity in various multilingual contexts, particularly for under-resourced languages, without resorting to supervised data. Similarly, Seneviratne and Suominen (2024) used generative prompts to simplify texts in English and Sinhala, confirming the utility of generative models in less common languages. Another innovative approach used machine translation to predict lexical complexity and simplify texts, combining regressors based on linguistic features with quantized generative models to generate suitable lexical substitutions (Cristea and Nisioi, 2024).

2.2 LLMs for data annotation

Large language models (LLMs) offer significant potential to transform data annotation by reducing costs and increasing efficiency. The work of Liu et al. (2023) presents a systematic review of prompting methods based on LLMs, which allow zero-shot or few-shot learning through structured prompts and pre-trained models, thereby opening up new opportunities for automating annotation. Moreover, Tan et al. (2024) explore how LLMs, such as GPT-4, can generate annotations, classify eligible data types, and address challenges related to bias and annotation quality. Gilardi et al. (2023) show that ChatGPT outperforms human workers in text annotation tasks, with increased accuracy (25 percent higher) and costs 30 times lower. In the field of computational social science, Ziems et al. (2024) demonstrate that while LLMs do not surpass specialized models for classification, they produce qualitative explanations that can enhance research in annotation and creative generation. Other works, such as those by Farr et al. (2024), combine chains of LLMs for more robust and scalable annotation by aggregating predictions from multiple models, while Qiu et al. (2025) use ensembles of LLMs for the evaluation of unstructured textual data, thereby

improving annotation consistency. Research by Watts et al. (2024) focuses on the divergences between humans and LLMs for multilingual and multicultural data, highlighting the importance of cultural contexts in annotation. Finally, in software engineering, LLMs show their potential to replace manual annotations, though they remain limited in complex technical contexts (Ahmed et al., 2024). RED-CT, proposed by Farr et al. (2025), illustrates a hybrid approach combining LLM annotations and human interventions for linguistic classification tasks in constrained environments.

3 Data and models

This section will present the models and the data that will be used to (i) evaluate the performances of LLMs with respect to a gold standard and also with respect to individual human annotations; (ii) evaluate the impact of LLMs in a supervised scenario where the LLMs are only used to annotate the training dataset, in order to reduce the energy costs and improve response-time efficiency.

3.1 Dataset

Recently released lexical complexity datasets (Shardlow et al., 2021, 2024) usually provide for each instance a gold numerical complexity score that is the average of several numerical human annotations. In this paper, our goal is to compare LLMs and human annotations. It therefore requires the use of a dataset where all individual human annotations are available, and not only the average of their annotations. This is why for our evaluations, we use the "CompLex 2.0" dataset, an improvement over "CompLex 1.0" (Shardlow et al., 2021). This corpus contains individual human evaluations of the lexical complexity of a set of English texts, carried out using a 5-point Likert scale. The texts included in the corpus come from sources such as Wikipedia, educational books, and newspaper articles, covering a wide variety of topics. The texts were annotated by human evaluators who assessed the lexical complexity of a target word in its context (sentence) using the Likert scale. Each instance was annotated several times, and the average of these annotations was used as the complexity score for each data instance. This score, once normalized, represents a continuous value between 0 and 1. In CompLex 2.0, part of the data from CompLex 1.0 was reused, but the annotations were enriched by adding 10 additional annotations per instance, carried out via the Amazon Mechanical Turk (MTurk) platform, while keeping the same annotation instructions as before. In total, for this second phase, 523 available workers annotated the data, implying that all instances were not annotated by the same workers, which is clearly a limit for the sake of comparison. Furthermore, in the release of Complex 2.0, we only have the data provided by MTurk of the second annotation phase. Therefore, when it comes to comparing with individual human annotations, we will use this data only.

The training and test data contain 7,662 instances and 917 instances respectively.

3.2 LLMs strategies

We used three prompt approaches to evaluate the ability of large language models (LLMs) to predict in-context lexical complexity as proposed by (Kelious et al., 2024a). First, the **Zero-shot prompt (_b)** relies solely on the model's prior knowledge, without providing any specific examples. Next, the **One-shot prompt (_i)** provides a clearer framework by incorporating annotation instructions and a concrete example, allowing the model to better grasp the task at hand. Finally, the **Chain-of-thought prompt (_a)** goes further by exposing detailed instructions, a step-by-step methodology, and an illustrative example to structure the model's reasoning before producing an answer. These three strategies allow the evaluation of complexity from different angles, yielding variable results.

We will experiment with 7 different generative models with the 3 prompts: llama3:8b (Dubey et al., 2024), mistral:7b (Jiang et al., 2023), gemma:9b (Team et al., 2024), phi3:3.8b (Abdin et al., 2024), gpt-4o (January-2025) [1], falcon3:7b (Almazrouei et al., 2023), qwen2:7b (Yang et al., 2024)

For all these models we will use their 4-bit quantized version. We use Ollama[2], an open-source tool, to test these different LLMs.

3.3 Supervised model

The supervised scenario consists in using a recent system that has proven effective for predicting lexical complexity in English (Kelious et al., 2024b). The model combines a pre-trained language model with frequency-based features derived from Zipf's law. In summary, the prediction formula is:

$$\hat{y} = f\Big(W_h \cdot \sigma\big(W_e \cdot E + W_f \cdot F + b_e\big) + b_h\Big)$$

where:

[1] gpt-4o: https://openai.com
[2] https://ollama.com

- $\hat{y}$ is the predicted complexity value, between 0 and 1;
- E correspond to the lexical embeddings extracted from a transformer model (e.g., DeBerta) from the sequence: [CLS] sentence [SEP] target_word;
- F is the input frequency-based feature vector, $[F_1, F_2, F_3, F_4, F_5]$ [3];
- W_e and W_f are the weights applied respectively to the lexical embeddings (E) and the features (F);
- b_e and b_h are the bias terms for the input layer and the hidden layer;
- σ is a non-linear activation function (ReLU) applied to the combination of E and F;
- W_h corresponds to the weights of the hidden layer;
- f is the linear activation function at the output.

4 Evaluation of LLMs performances against human-based gold complexity scores

In this section, we analyze the performances of the 21 LLM systems derived from our three prompting strategies (section 3.2). We compare the predicted lexical complexity scores with the gold scores, that are, for each instance, the average of several individual human numerical annotations.

4.1 Pearson Correlation Analysis

According to Figure 1, the performance of the models follows a clear trend where the addition of structure and examples improves their ability to predict lexical complexity: on average, the **Zero-shot (_b)** strategy achieves 0.214, the **One-shot (_i)** 0.365, and the **Chain-of-Thought (_a)** [COT] 0.439, confirming the positive impact of explicit reasoning. Comparatively, gpt-4o outperforms all other models, showing high correlations even in Zero-shot (0.746) and reaching 0.780 in COT, while Llama-3 and Mistral show good performance but remain far behind, requiring more advanced prompts to improve their results. In contrast, Phi-3 and Falcon-3 are noticeably less performant, particularly in Zero-shot (respectively 0.023 and 0.088), and need the COT to reach better levels, while Gemma completely fails to capture lexical complexity, with a negative close-to-zero correlation in One-shot (-0.003). In conclusion, the advantage of advanced models like gpt-4o is undeniable, but prompt optimization remains essential to improve the performance of weaker models.

[3] F1 (the Zipf score of the word frequency), F2 (the average Zipf score in a sentence), F3 (the difference between the target word's Zipf score and the average score), F4 (the number of words with a Zipf score higher than the target word) and F5 (a binary value indicating whether the target word is considered rare with a score less than or equal to 3).

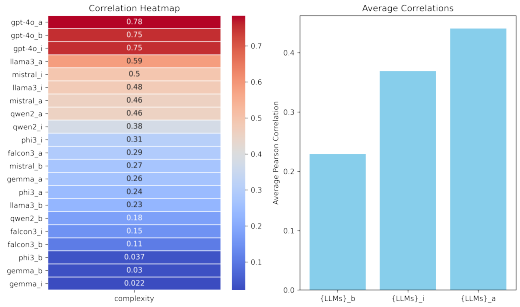

Figure 1: Pearson correlation between complexity predicted by LLMs and the gold complexity (left part); and average correlation (right part) with respect to the prompt strategy types (_b: zero-shot, _i: one-shot, _a: chain-of-thoughts).

4.2 Predicted complexity and error distributions

The violin plot in Figure 2 provides a more detailed view of the distribution of the model predictions compared to the distribution of human-based gold complexity scores on the test set. Figure 4 in Appendix A provides a complementary view showing the distributions of the residuals, i.e. the LLM errors (y_{gold} - y_{llm}).

Distribution of gold complexity scores (complexity): The distribution of values is quite spread out, meaning that the perception of lexical complexity by human annotators varies according to the instances. There is a notable concentration around specific values, which may indicate that most words have a moderately perceived complexity (neither too easy nor too difficult). Some extreme values exist, which could correspond to words that are widely considered either very simple or very complex.

Models close to gold annotations: The models gpt-4o (gpt-4o_b, gpt-4o_i, gpt-4o_a) and Llama3 (llama3_i, llama3_a) display distributions similar to human complexity. Their medians are relatively aligned with the gold annotations and their predictions cover a comparable range of values, indicating a certain consistency. Mistral (mistral_i, mistral_a) follows a similar trend with moderate dispersion, suggesting that it evaluates lexical complexity in a balanced manner, without excessively overestimating or underestimating. These trends are confirmed with the error distributions.

Models with notable discrepancies: Some models show more marked divergences compared to human annotations. Phi3 (phi3_b, phi3_i, phi3_a) and Qwen2 (qwen2_b, qwen2_i, qwen2_a) have a higher median, indicating a tendency to

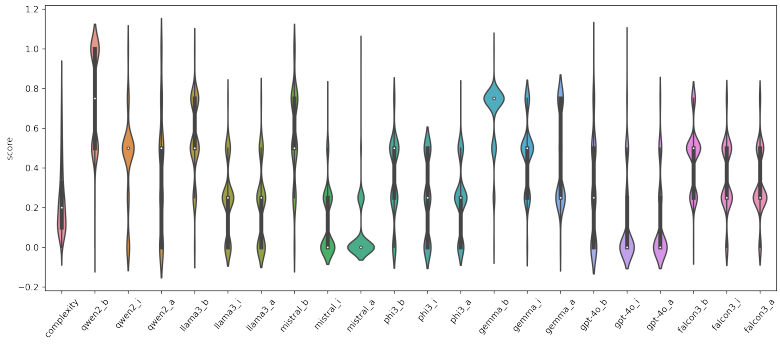

Figure 2: Distribution of lexical complexity predictions for each LLM and distribution of gold scores ("complexity" violin plot)

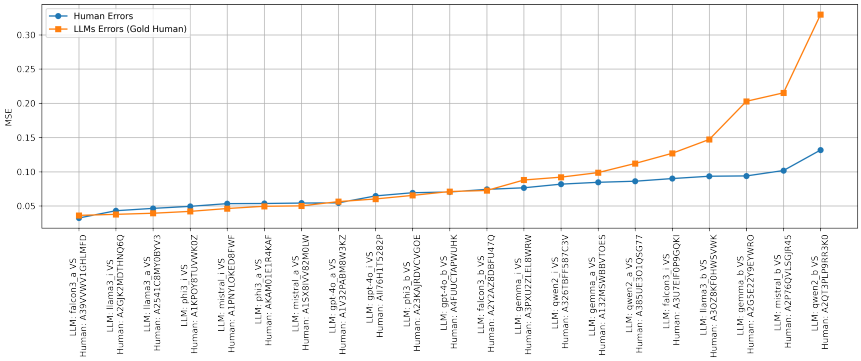

Figure 3: Comparison between individual human annotations and LLMs, by selecting the top-21 annotators with lowest MSE.

overestimate word complexity. Falcon3_i and Falcon3_a, on the other hand, display a larger dispersion, notably Falcon3_i which sometimes assigns exceptionally high values. This variability suggests a lack of stability in the predictions, which can be problematic for reliable evaluation of lexical complexity.

General insights and implications: The models gpt-4o, Llama3, and Mistral seem to be the closest to human annotations, suggesting that they could be the most reliable for predicting lexical complexity. Conversely, Phi3 and Falcon3 tend to overestimate complexity, while Qwen2 shows more rigid predictions biased toward higher values. A high dispersion in predictions, as observed in some models, may indicate inconsistency or a lack of calibration, while an overly concentrated distribution can reflect a lack of diversity in the evaluation of words. Thus, if the goal is to mimic human perception, the models most aligned with the annotations should be favored, whereas those with high variability or marked bias might require adjustment for better calibration.

5 Comparison of LLM predictions with individual human annotations

In this section, we compare LLM predictions with the individual human annotations. Unfortunately, the way the dataset is annotated using the MTurk platform with a limitation of 10 human annotations per instance makes a fair comparison difficult, whereas we have the predictions of all the LLMs per instance. Although there are clear limitations in the various provided evaluations below due to this issue, the results will reveal some trends that will pave the way for other experiments.

5.1 Comparing the LLMs with the best human annotators

Since we have 7 LLMs and 3 different prompts, making a total of 21 models, we will compare this set with the top 21 human annotators selected based on their Mean Square Error (MSE) score. Note that we did not use the Pearson correlation scores as the difference of two Pearson correlation scores is difficult to interpret with two different sets of annotated instances (all instances for LLMs vs. various numbers of instances for human annotators).

General Error Comparison (MSE): Figure 3

shows that human annotators generally have lower errors than LLMs. Indeed, most annotators display a more stable and homogeneous MSE, whereas LLMs show much more variability in their performance. Some models come close to human performance, while others have much larger discrepancies.

Error Dispersion: Human errors range between approximately 0.03 and 0.13 for the top-21, indicating a certain consistency in their annotations. In contrast, LLM errors are much more dispersed, ranging from 0.03 up to over 0.32, suggesting significant heterogeneity depending on the model used. Some LLMs are very performant, while others clearly struggle to reproduce precise annotations.

Best and Worst Performers: The best models and annotators are those that display the lowest MSE. Among the LLMs, Falcon3_a (MSE $\approx$ 0.036) and Llama3_i (MSE $\approx$ 0.038) stand out for their precision, rivaling the best human annotators, notably "A39VVWV1GHLMFD" (MSE $\approx$ 0.032) and "A2GJK2MDTHNQ6Q" (MSE $\approx$ 0.043). Conversely, some models display particularly high errors. Qwen2_b (MSE $\approx$ 0.33) is the least precise among the LLMs, followed by Mistral_b (MSE $\approx$ 0.21). On the human side, "A2QT3PLP9RR3K0" is the annotator whose annotations deviate the most from the reference values (MSE $\approx$ 0.13).

Direct Comparison between LLMs and Human Annotators: Some LLMs manage to achieve, or even surpass, the performance of the least precise human annotators among the top-21 ones. The graph shows that up to the 10th-best annotator there is more or less an equivalence between human and LLM performances.

5.2 Comparing LLMs with individual human annotations on a common set of instances

In the ideal case, comparing LLMs with individual human annotations should be performed on a common set of instances. To make the analysis manageable due to impractical combinatorics in CompLex 2.0 to find the set of annotators with the largest set of shared annotated instances, we chose to take the five annotators who annotated the largest number of instances and extract the 375 instances annotated in common. This approach reduces the scope of the problem while retaining a representative set of annotations for our analyses.

On this subset of instances, we performed an evaluation using standard evaluation metrics (R^2, Pearson Coefficient, and MSE) comparing annotators and LLMs. For each of the annotators (annot1..5), we selected the five LLM/humans whose evaluations were in the closest agreement with theirs (according to Cohen's Quadratic Kappa metric). Table 1 provides the results of the evaluation metrics.

Human	Model	R^2	Pearson	MSE	Kappa
annot1, (MSE :0.021)	gpt-4o_a	**0.4801**	**0.6929**	**0.0345**	**0.68**
	gpt-4o_i	0.4194	0.6476	0.0410	0.62
	gpt-4o_b	0.4116	0.6415	0.0518	0.60
	annot3	0.3398	0.5830	0.0610	0.57
	annot5	0.3288	0.5734	0.0542	0.57
annot2, (MSE :0.039)	annot1	**0.1612**	**0.4015**	0.0810	**0.36**
	annot3	0.1448	0.3806	0.0957	0.36
	gpt-4o_b	0.1338	0.3659	**0.0798**	0.36
	annot5	0.1481	0.3848	0.0988	0.33
	gpt-4o_a	0.0901	0.3001	0.1002	0.24
annot3, (MSE :0.034)	annot1	**0.3398**	**0.5830**	**0.0610**	**0.57**
	annot5	0.3076	0.5546	0.0722	0.54
	gpt-4o_b	0.2927	0.5410	0.0728	0.53
	gpt-4o_a	0.2982	0.5461	0.0688	0.50
	gpt-4o_i	0.2741	0.5236	0.0747	0.47
annot4, (MSE :0.039)	gpt-4o_b	**0.0529**	**0.2300**	0.0883	**0.21**
	llama3_a	0.0446	0.2113	**0.0655**	0.21
	annot2	0.0461	0.2147	0.0878	0.19
	annot1	0.0319	0.1787	0.0775	0.18
	mistral_i	0.0224	0.1497	0.0728	0.16
annot5, (MSE :0.030)	annot1	**0.3288**	**0.5734**	0.0542	**0.57**
	annot3	0.3076	0.5546	0.0722	0.54
	gpt-4o_a	0.2978	0.5457	**0.0537**	0.54
	gpt-4o_i	0.2777	0.5270	0.0562	0.51
	gpt-4o_b	0.2404	0.4903	0.0830	0.44

Table 1: Results of evaluation metrics (R^2, Pearson, MSE, Kappa) comparing annotators annot1...5 and models. The 5-closest annotators annot1...5 or LLM models are provided for each human annotator annot1...5 with respect to Cohen's Quadratic Kappa (Kappa).

Overall, we can see that for each selected human annotator there are three LLMs in its 5-closest humans/LLMs (exception: only two LLMs for annot2). It shows that we can always find an LLM closer to her/him than two other human annotators (only one for annot2). The gpt4o LLMs tend to be the closest to the selected human annotators: 3 occurrences in the top-5 for three human annotators (annot1, annot3 and annot5), 2 occurrences for annot2 and only one occurrence (zero-shot) for annot4, the latter emerging as an "outlier" (low correlation with everyone).

This view is of course partial because of the specificity of the selected human annotators (the ones who annotated the largest number of instances) that are not representative of all annotators. This should be investigated further by enlarging the set of annotators (but reducing the evaluation set), and/or by varying selection criteria in order to have more global view. Nevertheless, the preliminary investigation presented in this section show some potential for aligning individual human annotators and LLMs.

6 Train supervised model

In a real scenario, annotating an instance using 10 LLMs simultaneously would be very expensive in terms of time, money and energy cost compared to using a small supervised model. In this section, we train various supervised models on the CompLex 2.0 dataset, trying to take advantage of LLMs to annotate the training data, and therefore limiting their use to an offline setting.

6.1 Preliminary cross-evaluation

We first perform evaluations using the supervised model described in section 3.3, notably crossing the various types of annotations available. In particular, the Complex 2.0 dataset contains, for each instance, individual annotations from Amazon Mechanical Turk as well as an overall score that incorporates other inaccessible annotations. We distinguish three types of annotations:

- **llms**: the average of the annotations provided by several language models (LLMs). To simulate the Amazon Mechanical Turk approach, we randomly select 10 LLMs out of 21, recalling that MTurk selects 10 annotators from among 523.
- **mturk**: the average of the scores assigned by the human annotators from Amazon Mechanical Turk.
- **all**: the average of all annotations, that is, those from MTurk plus the additional inaccessible annotations (global score).

Train → Test	Pearson	R^2	MSE
all → all	0.79	0.62	0.0065
mturk → mturk	0.87	0.76	0.0072
llms → llms	0.78	0.62	0.0080
all → mturk	0.86	0.74	0.0100
all → llms	0.50	0.25	0.1780
mturk → all	0.79	0.63	0.1210
mturk → llms	0.53	0.28	0.3320
llms → mturk	0.57	0.33	0.0450
llms → all	0.52	0.27	0.0250

Table 2: Results of Pearson, R^2, and MSE for each *train* → *test* setting.

Intra-ensemble performance (homogeneous): When both training and testing are performed on annotations of the same type, the performance is high (Table 2). For instance, the scenario *all* → *all* (r = 0.79, R^2 = 0.62, MSE = 0.0065) illustrates good consistency when human annotators are used for both training and testing. Similarly, the *mturk* → *mturk* approach (r = 0.87, R^2 = 0.76, MSE = 0.0072) gives the highest results, reflecting the high homogeneity of MTurk annotators. Finally, in *llms* → *llms* (r = 0.78, R^2 = 0.62, MSE = 0.0080), the language models generate annotations that are globally consistent with each other, even though they remain slightly below the quality obtained with MTurk.

Cross-performance (heterogeneous): In a context where training and testing come from different sources, the generalization varies greatly. The *all* → *mturk* approach (Pearson = 0.86, R^2 = 0.74, MSE = 0.010) shows a fairly good capacity of the model to predict the MTurk-specific annotations when trained on data annotated by a larger set of human annotator. Conversely, *all* → *llms* (Pearson = 0.50, R^2 = 0.25, MSE = 0.178) results in a significant drop in performance, revealing a marked divergence between the annotations generated by LLMs and those by humans. The *mturk* → *all* option (Pearson = 0.79, R^2 = 0.63, MSE = 0.121) remains relatively satisfactory, but the increase in MSE indicates a difficulty in fully capturing the diversity of the annotations. Finally, *mturk* → *llms* (Pearson = 0.53, R^2 = 0.28, MSE = 0.332) confirms a notable incompatibility between the judgments of MTurk and those of the generative models.

Impact of LLMs with respect to human annotations: When training on annotations from LLMs to test on MTurk (*llms* → *mturk*), the performance remains modest (r = 0.57, R^2 = 0.33, MSE = 0.045), demonstrating that the models do not fully capture the complexity as perceived by human annotators. Similarly, the scenario *llms* → *all* (r = 0.52, R^2 = 0.27, MSE = 0.025) yields similar results: LLMs do not faithfully reproduce the judgments from a mixed set of human annotations.

6.2 Calibrating LLMs

The results in the previous section indicate that, despite their internal consistency, LLMs require significant adjustments to align their annotations with human judgments, especially in subjective tasks such as lexical complexity prediction. To do so, we propose the following three-step method: (1) we train a calibration model on N samples from the training set to learn how to combine the predictions from the various LLMs, (2) we directly apply this model to generate annotations on the training set; and (3) we train a supervised model (section 3.3) on these *pseudo-labels* and evaluate it on the test set produced by human annotators.

The proposed calibration model combines LLMs using a weighting scheme that can be mathematically formulated as:

$$\hat{y} = \sum_{i=1}^{n} \alpha_i \cdot x_i + b$$

where n is the number of LLMs (21 in our case) and α_i is the weight associated to the complexity score x_i predicted by the LLM LLM_i, the term b being the bias. The weights and the bias are trained by minimizing the MSE on the subset of training data annotated by humans.

Sample size (N)	Pearson	R^2	MSE
100	0.73	0.54	0.0145
500	0.75	0.55	0.0108
1 000	0.74	0.55	0.0103
2 000	0.74	0.55	0.0114
5 000	0.75	0.57	0.009
All (7 662)	0.74	0.56	0.0116
No weights (avg)	0.44	-2.90	0.0635
Model llms → all	0.52	0.27	0.0250
Model all → all	0.79	0.62	0.0065

Table 3: Evaluation of the LLMs calibration on test set.

Table 3 provides the performances of the supervised model based on the calibration model predictions to annotate the training dataset, varying the sample size N. It appears that with only few annotated data (N=100) we can observe a significant improvement of the performances with respect to using a simple average of the LLMs predictions to annotate the training set: Pearson increases from 0.44 to 0.73, MSE decreases from 0.064 to 0.015. Varying the sample size N from 100 to all instances, the performances remain mostly stable despite some little variations indicating that a N value between 100 and 500 seems sufficient to approach the results of the supervised model (all → all) which remains superior (Pearson=0.79, $R^2 = 0.62$, MSE=0.0065). Note that applying the calibration model directly on the test set yields similar trends as shown in Appendix B, confirming the validity of the approach. Moreover, the condensed error distribution around 0 for the calibrated model applied directly on the test set shows the improved alignment with human annotations (cf. "stacked_calibrated" violin plot in Appendix A). It is also interesting to note that using to simple average method with no weights tend to be better by randomly sampling ten LLMs per instance than by using the all set of LLMs (Table 3).

7 Conclusion

In this study, we explored the prediction of lexical complexity by using large language models (LLMs) with different prompting strategies (zero-shot, one-shot, chain-of-thought). Our experiments show that adding structure and explicit examples significantly improves the models' ability to approach human judgments, with gpt-4o notably standing out with high correlations and better alignment with the reference annotations.

The comparative analysis of predictions distributions and errors (MSE) highlights significant variability between LLM predictions and human evaluations. While some models (such as Llama3 and Mistral) manage to approach human performance in certain scenarios, others (such as Qwen2) exhibit marked biases or excessive dispersion in their predictions. These findings underscore the importance of precise calibration and prompt optimization to fully leverage the capabilities of generative models.

Moreover, although training a supervised model on human annotations remains the performance benchmark (Pearson=0,79, $R^2 = 0,62$, MSE=0,0065), our results show that the use of a calibration model which integrates an optimized weighting of the LLMs' predictions yields significantly higher scores than simply averaging the LLMs predictions, with Pearson coefficients reaching up to 0,75 and R^2 values of 0,57 with as few as 500 examples. This improvement, consistent across various subsets, confirms that calibration by stacking enables a better use of the combined richness of human annotations and automatic predictions, while drastically reducing the number of human annotations required.

In brief, our work shows the potential of LLMs and in particular that of the calibration models for lexical complexity prediction. However, the variability observed in certain metrics, such as the MSE, and the persistent gaps with human annotations call for continued optimization efforts, notably by refining prompting techniques and calibration strategies. Future research could focus on improving the self calibration of generative models and adapting these approaches to other languages and educational contexts, in order to fully exploit the synergy between human annotations and automatic predictions.

8 Limitations

Despite the promising results presented in this study, several limitations must be acknowledged:

- **Choice and size of language models:** The analysis was based on a limited set of models (e.g., llama3, Mistral, Gemma, Phi3, gpt-4o, Falcon3, and Qwen2), whose sizes and architectures were chosen based on practical criteria (notably the use of 4-bit quantized versions). Although this selection represents a certain segment of current LLMs, it limits the generalizability of the results. Future investigations could incorporate a greater variety of models and examine the impact of model size and parameter settings on predicting lexical complexity.

- **Focus on English and the Complex 2.0 dataset:** This study is limited to the analysis of English texts, relying exclusively on the Complex 2.0 dataset, which was chosen for the richness of its annotations. However, lexical complexity is a phenomenon that can vary significantly across languages due to structural and lexical differences. Extending the analysis to other languages, accompanied by language-specific prompts and guidelines, would help capture intercultural dynamics more accurately and broaden the scope of the conclusions.

- **Simulation of MTurk annotations:** The dataset used is based on annotations from 523 participants via Amazon Mechanical Turk. Accurately reproducing this level of heterogeneity is challenging, as simulating the equivalent of 523 annotators using LLMs is difficult. In this study, we limited our analysis to a subset of 5 annotators who annotated the highest number of common instances. Increasing this number in future research would allow for a better estimation of the variability and robustness of human judgments.

- **Calibration method:** Although the calibration method has shown its effectiveness in aligning LLM predictions with human annotations, it is only a starting point. A more comprehensive benchmark incorporating various calibration methods would be beneficial in identifying the optimal strategy and further improving the alignment between automatic predictions and human judgments.

These limitations pave the way for interesting future work, including extending the analysis to other languages, exploring a greater diversity of models and calibration methods, and incorporating a larger number of annotators to enhance the robustness and generalizability of the results.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Toufique Ahmed, Premkumar Devanbu, Christoph Treude, and Michael Pradel. 2024. Can llms replace manual annotation of software engineering artifacts? *arXiv preprint arXiv:2408.05534*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Petru Cristea and Sergiu Nisioi. 2024. Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 610–617.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. Tmu-hit at mlsp 2024: How well can gpt-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.

David Farr, Nico Manzonelli, Iain Cruickshank, Kate Starbird, and Jevin West. 2024. Llm chain ensembles for scalable and accurate data annotation. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2110–2118. IEEE.

David Farr, Nico Manzonelli, Iain Cruickshank, and Jevin West. 2025. Red-ct: A systems design methodology for using llm-labeled data to train and deploy edge linguistic classifiers. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 58–67.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana. Association for Computational Linguistics.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024a. Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches. In *Swedish Language Technology Conference and NLP4CALL*, pages 96–114.

Abdelhak Kelious, Matthieu Constant, and Christophe Coeur. 2024b. Complex word identification: A comparative study between chatgpt and a dedicated model for this task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Alejandro Mosquera. 2021. Alejandro mosquera at semeval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559.

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

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. Deepblueai at semeval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584.

Jiaxing Qiu, Dongliang Guo, Papini Natalie, Peace Noelle, Levinson Cheri, and Teague R Henry. 2025. Ensemble of large language models for curated labeling and rating of free-text data. *arXiv preprint arXiv:2501.08413*.

Sandaru Seneviratne and Hanna Suominen. 2024. Anu at mlsp-2024: Prompt-based lexical simplification for english and sinhala. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 599–604.

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

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *arXiv preprint arXiv:2406.15053*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pre-trained language models. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 661–666.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative*

Use of NLP for Building Educational Applications, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Distributions of prediction errors of the LLMs

Figure 4 provides a complementary view showing the distributions of the residuals, i.e. the LLM errors (y_{gold} - y_{llm}).

B Performances of calibrated LLMs

Table 4 shows the performances of the system combining LLMs using the calibration model on the test set.

Sample size (N)	**Pearson**	r^2	**MSE**
100	0.77	0.44	0.0169
500	0.81	0.60	0.0119
1,000	0.81	0.61	0.0118
2,000	0.82	0.64	0.0108
5,000	0.82	0.67	0.0098
All (7,662)	**0.83**	**0.68**	0.0095
No weights (avg)	0.44	-2.9	0.0635
model (all $\rightarrow$ all)	0.79	0.62	**0.0065**

Table 4: Performance metrics by sample size, applying the calibration model directly to the test set.

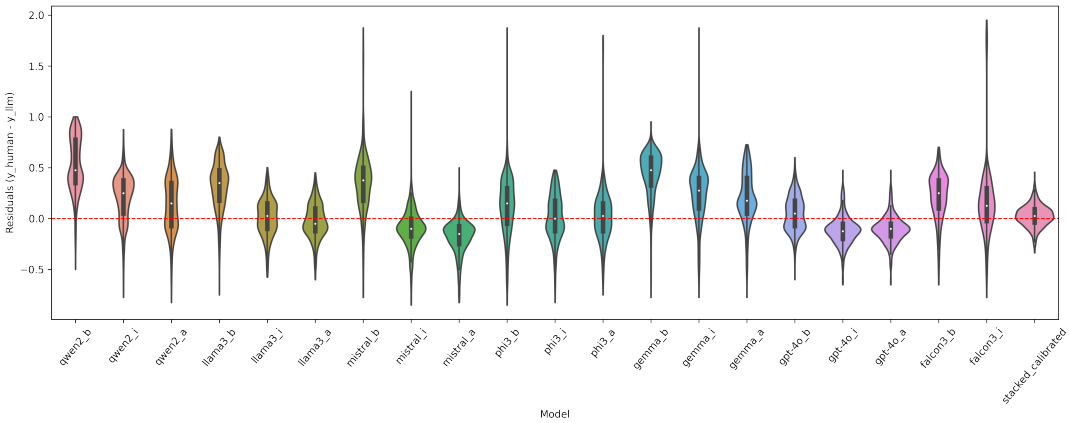

Figure 4: Distribution of errors for each LLM according to gold scores