Enhancing Lexical Complexity Prediction Through Few-Shot Learning with GPT-3

Jenny Ortiz-Zambrano[1],César Espín-Riofrío[2], Arturo Montejo-Ráez [3]
[1,2]Guayaquil University, [3]Jaen University
[1,2]Av. Delta S/N y Av. Kennedy - Guayaquil - Ecuador,
[3]Campus Las Lagunillas s/n. 23071 - Jaén - Spain
{jenny.ortizz, cesar.espinr}@edu.ec, amontejo@ujaen.es

Abstract

This paper describes an experiment to evaluate the ability of the GPT-3 language model to classify terms regarding their lexical complexity. This was achieved through the creation and evaluation of different versions of the model: text-Davinci-002 y text-Davinci-003 and prompts for few-shot learning to determine the complexity of the words. The results obtained on the CompLex dataset achieve a minimum average error of 0.0856. Although this is not better than the state of the art (which is 0.0609), it is a performing and promising approach to lexical complexity prediction without the need for model fine-tuning.

Keywords: GPT-3, Few-shot Learning, Lexical Complexity Prediction

1. Introduction

Reading involves a complex process that goes beyond coming across words or sections that are difficult for the reader to understand. Therefore, it is essential to properly understand the content of the texts to build coherent mental representations and fully understand their meaning (van den Broek, 2010).

Advancements in information technologies enable individuals to access a wealth of information across diverse domains, including education, information, social, health, government, and even scientific literature. Nonetheless, a considerable portion of the population faces obstacles in accessing this information due to significant reading challenges. These hurdles include lengthy sentences, technical jargon, and hard linguistic constructions that impede their comprehension of the text. Among those particularly impacted are individuals with intellectual disabilities and those with limited education. Surprisingly, even university students, with their advanced education and specialized knowledge in various subjects, can be part of groups struggling with reading disabilities (Alarcón García, 2022).

Lexical simplification (LS) is an automated process that substitutes words considered challenging for a particular target audience with easier alternatives while maintaining the original sentence's meaning intact. LS has an important role in Text Simplification (TS) and aims to enhance text accessibility for diverse groups of individuals (North et al., 2023a). Deep learning and, more recently, large language models (LLM) and prompt learning, have transformed our approach to various natural language processing (NLP) tasks including lexical simplification (LS) (North et al., 2023b).

The main objective of this article is to demonstrate how the Transformers GPT-3 based language model can classify text in terms of lexical complexity. This was achieved through the creation and evaluation of different versions of the model and prompts for few-shot learning to determine the complexity of the words.

This paper is organized as follows: in Section 2, a brief overview of the state-of-art is provided in complex word identification is provided. Section 3 explains GPT-3 for solving NLP tasks. Section 4 presents the experimental settings. Section 5 our solution and the results obtained with different variations on prompting are detailed. Section 6 presents a discussion about the results obtained compared to those proposed in SemEVal 2021, allowing us to present an analysis of our findings and highlight its importance and the contributions of the model in the field of predicting lexical complexity. Finally, in Section 7, conclusions and some insights on planned work are provided.

2. Previous work

Previous innovative forms of lexical simplification involved complicated systems with multiple components, each requiring extensive technical mastery and fine-tuned interaction to achieve maximum performance (Aumiller and Gertz, 2023). Recent advances in deep learning, particularly with the advent of large language models (LLMs) can be fine-tuned quickly. The high performance of these models sparked renewed interest in LS (North et al., 2023b). More advanced deep learning models,

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 68–76
21 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020) and others, are capable of automatically generating, selecting, and classifying candidate substitutions with performance superior to traditional approaches (North et al., 2023b).

With a capacity of 175 billion parameters, GPT-3 stands out for its deep knowledge of the language, its processing power, and its ability to learn from large volumes of online text data. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks (Brown et al., 2020). Due to these qualities, GPT-3 can perform a wide variety of natural language-related tasks with never-before-seen ease, including text generation and classification (Kublik and Saboo, 2022). The immense magnitude of the model allows it to produce results of high quality, precision, and diversity in the generated content. This development has raised a great deal of interest and concern in various fields, including Natural Language Processing (NLP), the machine learning industry, the media, the AI ethics communities, and society at large (Chan, 2023).

Despite being a generative model, GPT-3 can take different approaches to classify text, including zero-shot classification (where no examples are provided to the model), as well as one or few-shot classification (where some examples are presented to the model). In zero-shot learning, no prior training or adjustment to the labeled data is required. Currently, GPT-3 produces results for invisible data, but to perform zero-shot classification with GPT-3, we must provide you with a compatible prompt (Kublik and Saboo, 2022). In the few-shot learning, some examples of the task to be solved are provided. GPT-3's exceptional ability to learn in just a few tries, which is unprecedented in Natural Language Processing (NLP) models, is a prominent and notable feature (Chan, 2023).

SimpleText@CLEF-2022 Task[1] investigates the barriers that ordinary citizens face when accessing scientific literature head-on, by making available corpora and tasks to address different aspects of the problem (Ermakova et al., 2022). Mostert et al. (2022) ran a GPT-2-based text simplification model in a zero-shot way, resulting in conservative rewriting of abstracts, able to significantly reduce the text complexity. The findings indicate that taking text complexity into account is crucial for enhancing the accessibility of scientific information for non-experts.

Aumiller and Gertz (2023) in TSAR-2022 Shared Task on Multilingual Lexical Simplification presented two systems (Saggion et al., 2023). The initial system involved a zero-shot prompted GPT-3, where a prompt was used to request simplified synonyms based on a specific context, and the resulting simplifications were ranked. The second system was an ensemble comprising six distinct GPT-3 prompts/configurations, using average rank aggregation. Remarkably, the second system achieved the highest score for English across all metrics.

Traditional approaches are outperformed by the most advanced state-of-the-art deep learning models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020) and others. GPT-3 known as Generative Pre-trained Transformer 3, is a next-generation language model based on large-scale transformers, created by OpenAI[2].

(Ortiz-Zambrano et al., 2023) participated in the CLEF 2023 Simple@Text[3] track's Task 2.1 and 2.2. In their approach, they explore zero-shot and few-shot learning strategies over the auto-regressive GPT-3 model. Several prompts to achieve those strategies were tested. The results were ranked among the top submitted runs and demonstrated a solid performance for the task of lexical complexity prediction.

(Wei et al., 2022) investigated how generating a thought sequence, composed of a series of intermediate reasoning steps, significantly improves the ability of large language models to perform complex reasoning. Specifically, we demonstrated how these reasoning abilities develop naturally in sufficiently broad language models through a simple approach called "chain-of-thought prompting" where some chain-of-thought demonstrations are provided as examples of provocation. Experiments with three large language models reveal that chain-of-thought elicitations improve performance on a variety of reasoning tasks, including arithmetic problems, common sense questions, and symbol manipulation.

According to (Zhang et al., 2022), the superior performance of the Manual-CoT approach is based on manually building proofs. For this reason, they proposed Auto-CoT as an alternative to eliminate this manual task and automatically generate demos. The Auto-CoT method presents a wide variety of questions and creates chains of reasoning to construct corresponding proofs. Results from experiments on ten publicly available reasoning datasets show that using GPT-3, Auto-CoT consistently matches or exceeds the performance of the conventional CoT approach, which requires manual proof construction.

[1] https://simpletext-project.com/2022/clef/en/task2

[2] https://openai.com/

[3] SimpleText@CLEF-2023. Available in https://simpletext-project.com/2023/clef/

3. Experimental Settings

3.1. Dataset

We used the *CompLex* corpus proposed by (Shardlow et al., 2020). *CompLex*[4] is the first English multiple domain dataset, where words are scored in a context concerning their complexity using a five-point Likert scale to label complex words in texts of three sources/domains: the Bible, Europarl and Biomedical texts. The corpus is split into single-word and multiple-word annotations. The corpus contains a total of 9,476 sentences, each annotated by approximately 7 annotators, see Table 1.

	All	Single words	Multiple words
Europarl	3,496	2,896	600
Biomed	2,960	2,480	480
Bible	3,020	2,600	420
All	9,476	7,974	1,500

Table 1: Volumetric information for the single and multiple words in each subcollection of the CompLex dataset.

Each entry contains the name of the source corpus, the sentence representing the context of the word, the targeted word to classify, and a score (in training data) that is an average of the scores given by different annotators, taking into account that each class has a predefined weight between 0 (very easy) and 1 (very difficult) in a linear distribution over possible classes, see Table 2. Sample entries are shown in Table 3. This resource was used as a benchmark collection for SemEval 2021 Task 1: Prediction of Lexical Complexity of the 15th International Workshop on Semantic Evaluation[5] (Shardlow et al., 2021).

4. Methodology

We based our work on the one presented by Mostert et al. (2022), applying GPT-3 in this text classification task, specifically, a task of identification of complex words in texts and the respective categorization of them. The experiments were carried out applying a *few-shot classification* where the purpose was to carry out an analysis of the content of the texts coming from diverse sources such as the Bible, Biomedical, and Europarl to determine that GPT-3 was able to predict the complexity degree of the word. Various experiments were carried out applying different contents in the prompt, this is because GPT-3 has no concrete way of verifying the truth, logic, or meaning of any of the millions of lines of text it generates daily. The setting for the model parameters for GPT-3 is given in Table 4.

The steps that were carried out during the development of the experimentation are the following:

- The respective model configuration was specified.
- Different prompts were built and executed with the data set in its training and evaluation phases, to ask the model to return a "soft" response in the Likert scale.
- The probabilities were obtained to know what is the priority with which the model determines its result.
- The respective evaluation metrics were calculated: MAE, MSE, and RMSE, to determine the accuracy of the results.
- The respective comparison of the results generated by the model versus the data set was performed.

4.1. Few shot prompting for Lexical Simplification

We applied the *few shot prompting* strategy by providing the model with a few samples of what we wanted it to do. In Table 5 the prompt provided to the model is shown. The prompt contains several examples that are complemented with the word to be evaluated from the text, and it also indicates which resource the text corresponds to. After prompt specification, the values are replaced by the CompLex corpus dataset with the sentence and word to be evaluated.

4.2. Construction of the different prompt

To facilitate comparison, analysis, and interpretation of the results of the model runs against the test data set assign a name for each execution as is in the Prompt Variants column, as can be seen in See Table 8. The differentiation in the construction of the various prompts intended for the GPT-3 model was based on their size, determined by the number of examples integrated for construction and training during their learning phase. This distinction was indicated by the first letter of the respective name: S to indicate small applications (Small) with an average of 2 to 4 examples, M for medium-sized ones (Medium) with an average of 5 to 6 examples, and L for those of large magnitude

[4]CompLex: Is available at https://github.com/MMU-TDMLab/CompLex

[5]Semeval 2021 - LCP SHARED TASK 2021 - https://sites.google.com/view/lcpsharedtask2021

Scale	Description	Complexity
Very Easy	Words which were very familiar to an annotator.	0
Easy	Words with which an annotator was aware of the meaning.	0.01 - 0.25
Neutral	A word which was neither difficult nor easy.	0.26 - 0.50
Difficult	Words in which an annotator was unclear of the meaning, but may have been able to infer the meaning from the sentence.	0.51 - 0.75
Very Difficult	Words that an annotator had never seen before, or were very unclear.	0.76 - 1.00

Table 2: Categories on the Likert scale proposed by (Shardlow et al., 2020)

.

Corpus	Sentence	Token	LCP score
Bible	*He sees the place of **stones**.*	stones	0.3421
Bible	*But I will stay at Ephesus until **Pentecost**,*	Pentecost	0.6250
Bible	*These are the **families** of the Levites.*	families	0.2205
Bible	*The seeds rot under their **clods**.*	clods	0.6250
Biomed	*p150CAF-1 knockdown in **ES** cells was quantified.*	ES	0.6944
Biomed	*The 2P unique region (Region I) contains an **hg***	hg	0.7500
Biomed	*on behalf of the PPE **Group**.*	Group	0.1527
Europarl	*We have taken note of your **comment**, Mr Helmer.*	comment	0.0499
Europarl	*Country **Strategy** Papers - Malaysia, Brazil*	Strategy	0.2894
Europarl	***Documents** received: see Minutes*	Documents	0.2000
Europarl	***Situation** in Darfur (vote)*	Situation	0.2115

Table 3: Examples form the CompLex dataset where the complex word is highlighted in bold.

Parameter	Values
model	text-davinci-003
prompt	orden
temperature	0
maximum tokens	5
top_p	1
presence_penalty	0
logprobs	5

Table 4: GPT-3 Model Configuration

(Large) with an average of examples between 9 and 12 examples included. This process aimed to generate multiple prompts that enable the model to offer more precise results during its evaluation. This process aimed to generate multiple prompts that enable the model to offer more precise results during its evaluation.

Next, SO comes from Source, that is, whether or not the source from which the text to be evaluated came was included in the prompt specification. We also include the *nor* operator (the result of the negation of the OR operator) and *neither* to indicate the denial of the alternatives presented, translating to “NOR” and “neither” which would mean “none” or “none”, and we have used them to express that the application does not consider any of the two previous options mentioned. See Table 9 in the *Prompt Variants* column.

4.3. Methods used to calculate the complexity level of words

To calculate the level of complexity of complex words generated by GPT-3 as a value within the range [0, 1], we explored three ways based on the categories of the complex words, as detailed below. In this way, the linguistic responses of the model are transformed into numerical values. In Table 2 the range of values that correspond to the complexity of each category was presented.

1. **Method #1 - *Middle of the range***
   Half between the lower limit and the upper limit of each range of complexity values. Scores are fixed on a per category basis. For example:

   Neutral = (0.26 + 0.50) / 2
   Neutral = 0.375

   The calculated values for each category are:

   Very Easy = 0
   Easy = 0.125
   Neutral = 0.375
   Difficult = 0.625
   Very Difficult = 0,875

   Table 6 presents in the column *Method #1* the results of an execution carried out with a total of 30 records where it can be seen that there is a large number of coincidences with the categories that correspond to the complex

```
I'm reading fragments from some sources such as the Bible, Biomed, and Europarl,
and some words are not easy to understand.  I'm classifying these words into "very
easy", "easy", "neutral", "difficult" and "very difficult".  The sentence is "neu-
tral" when it is neither "very easy", nor "easy", nor "difficult", nor "very dif-
ficult".  Several examples are:  " However, no defects in axon pathfinding along
the monosynaptic reflex arc or in muscle spindle differentiation have been noted in
PV KO mice, which develop normally and show no apparent changes in their behavior
or physical activity (Schwaller et al.  1999).  ".  I find that word "spindle" is
neutral
###
The following fragment comes from the "bible" and after reading the fragment " I
will sprinkle clean water on you, and you shall be clean:  from all your filthiness,
and from all your idols, will I cleanse you.  ".  I find that the word "filthiness"
is easy
###
The following fragment comes from the "biomed" and after reading the fragment "
Moreover, acute dosing does not recapitulate the marked learning deficits produced
in rodents [15,16] by chronic exposure to dopamine D2R antagonists [6,7] ".  I find
that the word "antagonists" is difficult
###
The following fragment comes from the "biomed" and after reading the fragment "
Thrombus formation on fissured atherosclerotic plaques is the precipitating event
in the transition from a stable or subclinical atherosclerotic disease and leads to
acute myocardial infarction, ischemic stroke or peripheral arterial occlusion.  ".
I find that word "Thrombus" is very difficult
###
The following fragment comes from the "bible" and after reading the fragment " Mount
Sinai, all it, smoked, because Yahweh descended on it in fire; and its smoke as-
cended like the smoke of a furnace, and the whole mountain quaked greatly.  ".  I
find that the word "fire" is very easy
###
The following fragment comes from the @recurso and after reading the fragment @ora-
cion I find that word @aEvaluar is
```

Table 5: Prompt example

word of the CompLex corpus. The highlighted values correspond to matches. After the execution with the test data, it was obtained a MAE=0.1293, MSE=0.0258, RMSE=0.1608.

2. **Method #2 - *Average by category***
   The average of complexity values of a category is used as the complexity degree for that category. Therefore, the scores are fixed on a per category basis. Again, scores are fixed on a per category basis. The table 6 presents in the *Method #2* column the results of the execution carried out with the test records, a total of 30, where you can see the value calculated for the level of complexity of the categories generated by the model for the complex words of the texts. After the execution with the test data, it was obtained a MAE=0.086, MSE=0.016, RMSE=0.125.

   The corresponding calculated values for each category would be the following:

   Very Easy = 0
   Easy = 0.189
   Neutral = 0.351
   Difficult = 0.588
   Very Difficult = 0.811

3. **Method #3 - *The Confidence of GPT-3***
   The Confidence Level of the model corresponds to the high percentage of precision and coherence with which the model has made use of its attention mechanism and the context to select the category to which the complex word in the text corresponds.

   We consider for the assignment of the category generated by the model the one whose confidence level is the highest. For example: If the confidence level is 90% for the Easy category, the complexity closest to the left limit of the category range is taken. If the confidence level is 80% for the Difficult category, the complexity level that is furthest to the right of the category range is assigned. In the case of Very Difficult, the same procedure as the previous ones is considered. The table 6 presents in the Method #3 column the results of the execution carried out with the test records, a total of 30, where it can be seen, the value calculated

for the level of complexity of the categories generated by the model for the words complexities of the texts according to the level of confidence of the model. After the execution with the test data, a MAE=0.191, MSE=0.047, RMSE=0.216 was obtained.

It is important to note that in the table 6, in the "Category" column of the GPT-3 section, the complexity category generated by the GPT-3 model was selected based on the highest level. high confidence in percentage terms of the probabilities associated with the predictions generated by the model that corresponds to the category of the complex word. This is observed in the execution of the model on the corpus, using a test data set composed of 30 records, as detailed in the table 7.

5. Results

Our goal is to advance research on the use of the GPT-3 model to predict word complexity in the English language by adopting a few-shot examples learning approach. We have carried out multiple iterations with the objective that GPT-3 generates more precise and coherent answers with quality and relevance, we have formulated 19 several prompts pretending to optimize the performance of the model. Through this approach, we aspire to achieve greater precision in our predictions, approaching the results obtained by the winners of the lexical complexity prediction task proposed in the framework of SemEval 2021[6].

We experimented with different prompts issued to OpenAI's largest available model: text-davinici-002 and davinici-003 as evidenced by Table 9. Our first approach uses a singular prompt template in a few-shot setting to obtain the category of word complexity: *easy - very easy - neutral - difficult - very difficult*; we further improve upon these results by combining predictions from different prompt templates as can be seen in the Table 8, the application of different runs performed with the evaluation data set. The results derived from our approach toward single word prediction yielded the following values: MAE = 0.0875, MSE = 0.0131, and R2 = 0.1930.

A test was carried out by taking a sample of 30 records to train the model applying the few-shot learning technique. The data in the column *GPT-3 Confidence Level* represents the level of lexical complexity generated by the GPT-3 model for each token in the corpus. In the table 9, we can see that the "Score Type" column in the last three rows shows runs where this strategy was applied to assign complexity levels to complex words in the corpus. This is complemented by the results presented in the table 10 in the "GPT-3 Complexity" column, which refers to the complexity generated by the model. It is from these values that the strategy for calculating lexical complexity was derived, called *GPT-3 Confidence Level*. Additionally, the table **??** shows matches where the model's complexity prediction for a token matches the complexity assigned to the token in the CompLex corpus.

[6] https://sites.google.com/view/lcpsharedtask2021

6. Discussion

In this article, we present a system proposal to resolve the task of lexical complexity prediction. Table 11 shows the results achieved by the first five classified in the evaluation carried out by the organizing entity (Shardlow et al., 2021). It is important to note that the competition involved a large number of participants, specifically 54 teams. In contrast to the performance of the first-place winner, who achieved an MAE of 0.0609, we see relatively little difference in our results in terms of the linguistic categories considered. This fact gives a dose of confidence to our approach, which, despite its simplicity, proved to be competitive compared to the proposals of several teams that opted for more complex approaches. Among these more complex approaches, the use of deep neural networks such as the BERT and ROBERTa models stands out, evidenced in teams such as JUST BLUE, RG PA, Andi, CS-UM6P, OCHADAI-KYOTO, to mention just a few examples. It is worth mentioning that only one team used a GPT model (GPT-2) in their approach.

The results generated with GPT-3 would have reached an MAE = 0.0882 as presented in Table 9. It should be noted that when running the GPT-3 model, the approach *few-shot learning* used 4 to 6 examples in various experiments so that the model can learn and then generate its response.

The best result is achieved by using the combination M-SO-05, which corresponds to 5 examples sent to the model. This practice is highly beneficial to the model, as it allows it to generate more accurate predictions. To evaluate performance, the *Means* type score was used for the *davinci-003* model which yielded the following results: MAE=0.0882, MSE=0.0136, RMSE=0.1165, and Pearson=0.5776. These indicators highlight the effectiveness of the strategy used and the model's ability to provide high-quality results. The results achieved are very encouraging since they show that the model can understand the requests made by humans in a considerable way and without much effort as when applying other models.

7. Conclusions and future Works

Using GPT-3 to classify complex words involves finding a balance between your capacity and ability

Comparing the results applying GPT-3 with Few-Shot learning in corpus CompLex							
#	Corpus CompLex			GPT-3			
	Complex word	Category	Range of Values	Category	Complexity Level		
					Method #1	*Method #2*	*Method #3*
10	voice	neutral	0.01 - 0.25	easy	0.125	0.189	87.07% - 0.032
11	darkness	**easy**	0.01 - 0.25	**easy**	0.125	0.189	76.91% - 0.058
12	behold	easy	0.26 - 0.50	neutral	0.375	0.351	29.40% - 0.381
13	camp	**easy**	0.01 - 0.25	**easy**	0.125	0.189	81.11% - 0.045
14	bonds	**easy**	0.01 - 0.25	**easy**	0.125	0.189	54.29% - 0.115
15	statutes	neutral	0.01 - 0.25	easy	0.125	0.189	51.43% - 0.127
16	snares	**easy**	0.01 - 0.25	**easy**	0.125	0.189	54.95% - 0.112
17	exhortation	**difficult**	0.51 - 0.75	**difficult**	0.189	0.588	61.30% - 0.665
18	River	**easy**	0.01 - 0.25	**easy**	0.125	0.189	86.88% - 0.033
19	generation	**easy**	0.01 - 0.25	**easy**	0.125	0.189	85.27% - 0.037
20	dainties	**difficult**	0.51 - 0.75	**difficult**	0.189	0.588	58.36% - 0.657

Table 6: Methods applied to calculate the level of lexical complexity.

Probabilities associated with the complexity category predictions generated by the GPT-3 model The results applying the *davinci-002* model and *few shots learning* approach						
#	Token	GPT-3 complexity	CompLex complexity	GPT-3 confidence level		
				Option #1	Option #2	Option #3
20	dainties	difficult	difficult	difficult 59.39%	neutral: 25.25%	easy: 10.23%
21	subjection	difficult	neutral	difficult: 92.44%	neutral: 5.0%	easy: 1.16%
22	perverseness	difficult	neutral	difficult: 92.35%	neutral: 5.26%	very: 1.3%
23	grasshoppers	easy	easy	easy: 72.01%	neutral: 13.88%	very: 9.34%
24	signet	difficult	neutral	difficult: 74.86%	neutral: 20.39%	very: 2.74%
25	snare	easy	neutral	easy: 65.44%	neutral: 18.25%	difficult: 13.4%
26	Asher	easy	neutral	easy: 76.68%	very: 9.64%	difficult: 7.44%
27	demons	difficult	easy	difficult: 59.93%	easy: 30.62%	neutral: 7.4%
28	prophet	easy	easy	easy: 88.42%	neutral: 5.55%	difficult: 3.21%
29	lion	easy	neutral	easy: 90.23%	very: 4.46%	neutral: 4.32%
30	Lion	easy	easy	easy: 84.1%	very: 12.29%	difficult: 2.07%

Table 7: Probabilities associated with the predictions generated by the GPT-3 model that correspond to the category of the complex word.

Standard applied for the construction of the *Prompt Variants*				
Size	**Source**	**Connector Logical**	**# exp**	**Emphasis**
L-M-S	**SO**	**NOR**	**05**	**Em**
M	SO		05	
M	SO	NOR	05	Em
M	SO		06	
S		NOR	04	
S		NOR	04	Em
S	SO	NOR	05	
L	SO	NOR	09	

Table 8: Standard applied for the construction of the prompt variants.

to obtain more accurate results. The result opens new perspectives in the investigation of lexical complexity. Several experiments were carried out running various prompts, a few-shot with various models of the GPT-3 Family. We have applied three strategies to calculate the level of complexity of the complex words applied in the SemEval 2021 data set. Furthermore, we found some responses where learning from a few GPT-3 examples still presents difficulties, the responses generated by the model did not match the data sets in the work proposed by (Brown et al., 2020).

The best result was generated by the text-Davinci-003 model with an MAE of 0.0882. The model has been able to interpret and generate its responses based on a few examples and complex instructions, demonstrating that the text-Davinci-003 version provides better results than text-Davinci-002.

Nowadays, GPT-3 has been intensively used and tested on many different tasks using zero-shot and few-shot learning (Huang et al., 2023). Some of them found that this model is not that good. As new models are appearing, we plan to explore how these new models Claude 2 (Wu et al., 2023), GPT-4, or LLaMA 2 (Fan et al., 2023) perform on lexical complexity prediction.

Besides, an interesting research topic is to study

Final results generated with GPT-3 The results applying the *davinci-002* and *davinci-003* models and *few shots learning* approach							
#	**Prompt Variants**	**Score Type**	**Model Version**	**Metrics**			
				MAE	**MSE**	**RMSE**	**Pearson**
1	**M-SO-05**	**Means**	**davinci-003**	**0.0882**	**0.0136**	**0.1165**	**0.5776**
2	M-SONOR-05-Em	Means	davinci-003	0.0956	0.0153	0.1238	0.5103
3	M-SONOR-05-Em	Means	davinci-002	0.1011	0.0170	0.1305	0.4661
4	S-SONOR-05	Means	davinci-003	0.1057	0.0190	0.1378	0.5016
5	M-SO-06	Means	davinci-003	0.1074	0.0199	0.1412	0.4924
6	S-SONOR-05	Means	davinci-002	0.1098	0.0208	0.1442	0.5086
7	S-NOR-04	Means	davinci-002	0,1143	0.0229	0.1512	0.4919
8	S-NOR-04	Means	davinci-003	0.1725	0.0440	0.2099	0.3826
9	S-NOR-04-Em	Means	davinci-002	0.1793	0.0512	0.2262	0.3524
10	S-NOR-04-Em	Means	davinci-003	0.1875	0.0503	0.2242	0.4477
11	M-SO-05	Half of the range	davinci-003	0.1212	0.0219	0.1480	0.5730
12	M-SONOR-05-Em	Half of the range	davinci-003	0.1292	0.0239	0.1546	0.5099
13	S-SONOR-05	Half of the range	davinci-003	0.1475	0.0310	0.1761	0.5136
14	M-SO-06	Half of the range	davinci-003	0.1555	0.0345	0.1859	0.4944
15	S-NOR-04-Em	Half of the range	davinci-003	0.2164	0.0603	0.2456	0.4650
16	S-NOR-04	Half of the range	davinci-003	0.2106	0.0580	0.2409	0.3806
17	S-SONOR-05	GPT-3 Confidence level	davinci-003	0.2333	0.0655	0.2559	0.5600
18	M-SO-06	GPT-3 Confidence level	davinci-003	0.2658	0.0816	0.2857	0.5247
19	L-SONOR-09	GPT-3 Confidence level	davinci-003	0.2431	0.0708	0.2662	0.5241

Table 9: The results applying the davinci-002 and davinci-003 models and few shots learning approach.

Probabilities associated with the level of complexity predictions generated by the GPT-3 model The results applying the davinci-002 model and few shots learning approach							
#	**Token**	**GPT3 category**	**GPT3 range**	**GPT3 complexity**	**CompLex complexity**	**CompLex range**	**Match**
20	dainties	**difficult**	0.51 - 0.75	0.5880	0.5625	**difficult**	**Yes**
21	subjection	difficult	0.51 - 0.75	0.5880	0.4375	neutral	No
22	perverseness	difficult	0.51 - 0.75	0.5880	0.4166	neutral	No
23	grasshoppers	easy	0.01 - 0.25	0.1896	0.25	easy	Yes
24	signet	difficult	0.51 - 0.75	0.5880	0.4687	neutral	No
25	snare	easy	0.01 - 0.25	0.1896	0.3194	neutral	No
26	Asher	easy	0.01 - 0.25	0.1896	0.4285	neutral	No
27	demons	difficult	0.51 - 0.75	0.5880	0.125	easy	No
28	prophet	easy	0.01 - 0.25	0.1896	0.2222	easy	Yes
29	lion	easy	0.01 - 0.25	0.1896	0.2812	neutral	No
30	Lion	easy	0.01 - 0.25	0.1896	0.1710	easy	yes

Table 10: The results of the probabilities associated with the level of complexity predictions generated by the GPT-3 model applying the davinci-002 and few shots learning approach.

#	**Team Name**	**MAE**	**MSE**	R^2
1	JUST_Blue	0.0609	0.0062	0.6172
2	DeepBlueAI	0.0610	0.0061	0.6210
3	OCHADAI-KYOTO	0.0617	0.0065	0.6015
4	ia pucp	0.0618	0.0066	0.5929
5	Alejandro M.	0.0619	0.0064	0.6062
	FSL with GPT-3	0.0882	0.0136	0.1613

Table 11: Results achieved by the first five classified in the SemEval 2021 International workshop.

how large language models learn about "metalinguistic" knowledge, such as lexical complexity. Is it inferred from the enormous collection of texts due to explicit references to complexity? Is it, instead, a knowledge that "emerges" from the comprehension of language itself? These are captivating questions that, in the era of large language models, could be considered central for current research in lexical complexity prediction.

8. Bibliographical References

Rodrigo Alarcón García. 2022. Lexical simplification for the systematic support of cog-

nitive accessibility guidelines. *https://doi.org/10.1145/3471391.3471400*.

Dennis Aumiller and Michael Gertz. 2023. Unihd at tsar-2022 shared task: Is compute all we need for lexical simplification. *arXiv preprint arXiv:2301.01764*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Anastasia Chan. 2023. Gpt-3 and instructgpt: technological dystopianism, utopianism, and "contextual" perspectives in ai ethics and industry. *AI and Ethics*, 3(1):53–64.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liana Ermakova, Irina Ovchinnikova, Jaap Kamps, Diana Nurbakova, Sílvia Araújo, and Radia Hannachi. 2022. Overview of the clef 2022 simpletext task 2: complexity spotting in scientific abstracts. In *Proceedings of the Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th-to-8th, 2022*.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.

Cynthia Huang, Yuqing Xie, Zhiying Jiang, Jimmy Lin, and Ming Li. 2023. Approximating human-like few-shot learning with gpt-based compression.

Sandra Kublik and Shubham Saboo. 2022. Gpt-3: Building innovative nlp products using large language models. *O'Reilly Media*.

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

Femke Mostert, Ashmita Sampatsing, Mink Spronk, and J Kamps. 2022. University of amsterdam at the clef 2022 simpletext track. *Proceedings of the Working Notes of CLEF*.

Kai North, Alphaeus Dmonte, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Alexsis+: Improving substitute generation and selection for lexical simplification with information retrieval. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 404–413.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023b. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Jenny Ortiz-Zambrano, César Espin-Riofrio, and Arturo Montejo-Ráez. 2023. Sinai participation in simpletext task 2 at clef 2023: Gpt-3 in lexical complexity prediction for general audience.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2023. Findings of the tsar-2022 shared task on multilingual lexical simplification.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. Complex: A new corpus for lexical complexity predicition from likert scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*.

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

Paul van den Broek. 2010. Using texts in science education: Cognitive processes and knowledge representation. *Science (New York, N.Y.)*, 328:453–6.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. 2023. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.