SINAI Participation in SimpleText Task 2 at CLEF 2024: Zero-shot Prompting on GPT-4-Turbo for Lexical Complexity Prediction

Notebook for the SimpleText Lab at CLEF 2024

Jenny Ortiz-Zambrano[1,*,†], César Espin-Riofrio[1,†] and Arturo Montejo-Ráez[2,†]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador
[2]University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

In this article, we present our participation in Tasks 2.1 and 2.2 of the SimpleText track of CLEF 2024. Our work focused on the implementation of zero-shot learning using the GPT-4 Turbo autoregressive model. To this end, we develop and evaluate various cues to optimize the model's ability to predict lexical complexity. The results of our experiments indicated that GPT-4 Turbo can perform this task with remarkably robust performance, demonstrating its potential to assess language complexity effectively without the need for additional training.

Keywords

Lexical Complexity Prediction, Auto-regressive models, GPT-4 Turbo, Prompting, Zero-shot learning

CEUR-WS.org/Vol-3740/paper-319.pdf

1. Introduction

Readability is defined as the quality that makes a text more accessible and easier to read [1], but for many people, the way a text is written can be a barrier to understanding its content [2] due to the presence of infrequent or unknown words, and phrases with lexical and semantic complexity which drastically complicate the reader's understanding [3]; this is especially evident in cases such as children, non-native speakers [4] and people with various cognitive abilities or reading disabilities [5]. The success or failure of understanding a text will depend on the reader's prior knowledge of the meaning of the words [6].

Information technologies have facilitated access to a wide and abundant amount of information in various fields such as education, news, social networks, health, government, and also science; In the case of scientific literature to the general public, it has increased thanks to digitalization. However, this wealth of information is not available to everyone, since many people face significant obstacles to understanding , such as the complexity of grammatical structures, the use of technical language and the length of sentences, which directly affects individuals with intellectual disabilities, people with low levels of literacy, and even university students who, despite their academic training and specialized knowledge, can also be found among those who experience difficulties in reading and understanding complex texts [7]. In the case of scientific information, a significant barrier persists that makes direct access to scientific knowledge from the original sources difficult, where one of the main obstacles lies in the complexity of scientific texts, which presents difficulties for those without experience in the field due to to the lack of prior specialized knowledge [8].

The SimpleText laboratory [9] is part of the CLEF 2024 [10] initiative, which promotes the systematic evaluation of information access systems through experimentation with shared tasks. SimpleText

---

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
[*] Corresponding author.
[†] These authors contributed equally.
jenny.ortizz@ug.edu.ec (J. Ortiz-Zambrano); cesar.espinr@ug.edu.ec (C. Espin-Riofrio); amontejo@ujaen.es (A. Montejo-Ráez)
0000-0001-6708-4470 (J. Ortiz-Zambrano); 0000-0001-8864-756X (C. Espin-Riofrio); 0000-0002-8643-2714 (A. Montejo-Ráez)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

addresses the challenges of text simplification, especially in the context of making scientific information accessible. In this sense, it provides relevant data and reference points, as the general public tends to avoid reliable sources such as scientific literature due to its complex language and lack of prior knowledge. Instead, they rely on superficial and derivative sources on the web and social media, which are often published for commercial or political rather than informational purposes.

Can simplifying the text help overcome these barriers to access? That is the central question that this task addresses. Its objective is to generate a simplified summary of several scientific documents, based on a popular science query. This summary will provide the user with an accessible overview of the specific topic, with the hope of making scientific information more accessible and understandable to a broader audience.

The main objective of this research is to demonstrate the capacity of the language model based on Transformers like GPT-4 Turbo for lexical simplification. To achieve this, several variants of prompts without samples have been created and evaluated. This approach allows us to determine which concepts in scientific summaries require additional explanation and contextualization to help the reader understand the scientific text more effectively.

The article follows the following structure: first, a brief description is given of the current state-of-the-art in the prediction of lexical complexity and the identification of complex words (which are synonymous with the same natural language processing task). Section 3 presents Task 2 of the Simple Text track. Section 5 details our solution and the results obtained using different indication variants. Finally, Section 6 presents the conclusions and raises some ideas for future research.

2. Related Work

In recent years, Complex Word Identification (CWI) has attracted great interest in the scientific community and among computational linguistics researchers. These initiatives have significantly boosted development and research in the field of text simplification and improving information accessibility. The proposals include conferences, workshops and specific tasks that address challenges and promote advances in the simplification of texts, thus facilitating greater access to information for various audiences.

This advancement is reflected in the development of computational semantic analysis systems, as illustrated by several shared Complex Word Identification (CWI) tasks at notable events such as SemEval 2016 [11], NAACL-HTL 2018 [12], ALexS at IberLEF 2020 [13], the 15th edition of SemEval and the first lexical complexity prediction task [14], TSAR-2022 - the workshop on simplification, accessibility and readability of texts [15], the SimpleText tasks in CLEF in 2021 [16], 2022 [17], and 2023 [8] among other relevant initiatives in this area.

2.1. GPT-4 Turbo for solving NLP tasks

State-of-the-art deep learning models, such as BERT [18], RoBERTa [19], GPT-3 [20], and GPT-4 Turbo [21], significantly outperform existing traditional approaches. GPT-4 Turbo, the transformer-based large language model generation developed by OpenAI[1], is a prominent example of these advances. These models have demonstrated superior capabilities in various natural language processing tasks, raising the bar for performance and accuracy in the field.

GPT-3 Contains 175 billion parameters. Due to its extensive language knowledge, enormous processing power, and potential to learn from large amounts of online text data, GPT-3 can perform a wide variety of general natural language-based tasks with unprecedented ease, such as the generation and classification of text. [22]. The enormous scale of the model allows generating results with quality, precision and diversity of the generated content. This development has sparked considerable interest and concern in the field of NLP, the machine learning industry in general, the media, the AI ethics

---

[1] `https://openai.com/`

communities and civil society [23], The latest GPT-4 Turbo model of the OpenAI has been recognized for its ability to process text and images [24].

Although GTP-3 is a generative model, several approaches may be necessary to classify text, mainly ranging from zero-shot classification to single-shot and few-shot classification [22]. In zero-shot learning no samples of data are needed [25] .

3. SimpleText@CLEF-2024 Tasks

3.1. Task 2: "What is unclear?" Difficult concept identification and explanation

The goal of this task is to identify key concepts that need to be contextualized with a definition, example, and/or use-case and provide useful and understandable explanations for them. Thus, there are three subtasks:

1. To predict what are the terms in a passage of a document and their difficulty (easy/medium/difficult).
2. To generate a definition and an explanation only for the difficult terms.
3. To retrieve the provided definitions of the difficult terms in "correct" order: manual (2), generated positive 1 (1), generated positive 2 (1), generated negative 1 (0), generated negative 2 (0).

In Task 2.1, for each passage of a document, participants should provide a list of terms with corresponding difficulty scores (easy/medium/difficult). Passages (sentences) are considered to be independent, so term repetition is allowed (the same term can be detected in different sentences, even in the same document).

In Task 2.2, for each difficult term (terms that have been evaluated with the highest level of difficulty), participants should provide the definition and explanation which will be evaluated both from a qualitative point of view (manual review by terminologists) and from a quantitative point of view (applying overlapping text measures like BLUE, ROUGE, etc.).

In Task 2.3, participants should rank the set of definitions provided for the difficult terms in a way that the "best" definitions are ranked higher in the list of definitions. In particular, for each term there will be one manual definition (considered the best one) and two automatically generated good definitions that should be placed at the top of the list of retrieved definitions. Quantitative metrics (for example, P@1, P@3, rank correlation measures) will be used to evaluate participants'results.

Task 2.1 and Task 2.3 can be performed independently. Participants that want to run experiments on Task 2.2 need to accomplish Task 2.1 first.

4. Proposed system

As mentioned above, our approach is to apply learning cues without prior examples to the GPT-4 Turbo model, extracting predictions from the generated sequence. Details of the model configuration are presented in Table 1. We use the OpenAI API in Python to interact with the model, allowing us to easily integrate GPT-4 Turbo into our workflow. In addition, the OpenAI Playground offers an intuitive web interface that facilitates experimentation and rapid generation of code necessary for our tests and developments.

4.1. Experiments with zero-shot learning

We execute the model applying zero-shot learning where no examples are specified for learning and a sample of a generated output

Table 1
GPT-4 Turbo model details

Parameter	Value
Model	gpt-4-turbo
Temperature	0
Max_tokens	500
Top_p	1
Frecuency Penalty	0
Presence Penalty	0
Logprobs	5
Stop	"###"

Prompt PRM_ZS_TASK2_V1:

```
Performing the following actions according to a given text:
Select a number from 1 to 5 terms within the sentence named text
Assign the difficulty of the words: easy, medium and difficult, giving
the initial letter of each difficulty as an answer ([e]fácil/[m]edium/[d]ifficul)
Give the definition of each difficult term, add an explanation and use
case according to context of the sentences.
Provide the response in following JSON format:
{
  "word":  <word>,
  "difficulty":  <word_difficult>,
  "definition": "<definition>",
  "explanation": <explanation>
}

Context: {{context}}
Text : {{text}}

Response: .

##
Context: Digital assistant
Text : Abstract   Novel technological advances in mobile devices and
applications can be exploited in wildfire confrontation, enabling end-
users to easily conduct several everyday tasks, such as access to data
and information, sharing of intelligence and coordination of personnel
and vehicles.

Response:


json
[
    {
        "word": "exploited",
        "difficulty": "m",
        "definition": "make full use of and derive benefit from (a
        resource)",
        "explanation": "In the context, 'exploited' refers to utilizing
        the new technological capabilities of mobile devices and
        applications to their full advantage, specifically in managing
        and combating wildfires.",
        "use_case": "The novel technologies can be exploited to improve
        the efficiency and effectiveness of wildfire management."
    },
    {
        "word": "confrontation",
        "difficulty": "m",
        "definition": "a hostile or argumentative meeting or situation
        between opposing parties",
        "explanation": "Here, 'confrontation' is used metaphorically to
        describe the act of dealing with wildfires, suggesting a direct
        and active engagement with the fires.",
        "use_case": "Mobile technologies are being used in the
        confrontation of wildfires, aiding in direct engagement and
        management."
    },
    {
        "word": "coordination",
        "difficulty": "m",
        "definition": "the organization of the different elements of a
        complex body or activity so as to enable them to work together
        effectively",
        "explanation": "In this context, ’coordination’ refers to the
        organized management of personnel and vehicles to tackle
        wildfires efficiently, facilitated by mobile technology.",
        "use_case": "Effective coordination of personnel and vehicles
        is crucial for the successful management of wildfire incidents."
    }
]


###
```

Prompt PRM_ZS_TASK2_V2:

```
Depending on the text, select up to 5 terms considered complex
within the sentence, then assign the complexity of the terms to
easy, medium and difficult using the first letter of each word, for
example if the complexity is easy add the letter "e", If the
complexity of the term is medium, add the letter "m", and if the
complexity of the term is difficult, you should add the letter "d".
Then, from each complex term generate its definition, an explanation
and a use case according to the context of the sentence. Finally,
provide the response in the following JSON format:
{
  "word": <word>,
  "difficulty" <difficult_word>,
  "definition": "<definition>",
  "explanation" : <explanations>
}

##
Context: Drug discovery
Sentence: A pharmacophore analysis approach was used to investigate and
compare different classes of compounds relevant to the drug discovery
process (specifically, drug molecules, compounds in high throughput
screening libraries, combinatorial chemistry building blocks and
nondrug molecules).

Response:

json
[
    {
        "word": "pharmacophore",
        "difficulty": "d",
        "definition": "A set of structural features in a molecule that
        is recognized by a biological macromolecule and is responsible
        for its biological activity.",
        "explanation": "In drug discovery, a pharmacophore analysis is
        used to identify these structural features which can interact
        with a biological target, such as a protein, to induce a
        therapeutic effect. This helps in designing or identifying new
        drugs that can mimic or block the biological activity.",
        "use_case": "In the sentence, a pharmacophore analysis approach
        helps in comparing different classes of compounds to find
        potential drug candidates."
    },
    {
        "word": "combinatorial chemistry",
        "difficulty": "d",
        "definition": "A method in chemistry where different
        combinations of building blocks are systematically mixed to
        generate a large number of different compounds.",
        "explanation": "Combinatorial chemistry is used in drug
        discovery to rapidly synthesize and screen large libraries of
        compounds for potential drug activity. It allows researchers to
        explore a wide variety of chemical structures.",
        "use_case": "In the sentence, combinatorial chemistry building
        blocks are mentioned as part of the classes of compounds
        relevant to drug discovery."
    },
    {
        "word": "high throughput screening",
        "difficulty": "d",
        "definition": "A method used in drug discovery to quickly
        conduct millions of chemical, genetic, or pharmacological
        tests.",
        "explanation": "High throughput screening (HTS) is crucial in
        the early stages of drug discovery. It allows researchers to
        quickly identify active compounds, antibodies, or genes that
        modulate a particular biomolecular pathway.",
        "use_case": "The sentence refers to compounds in high
        throughput screening libraries, indicating these are tested in
        large-scale assays to find promising drug candidates."
    },
    {
        "word": "nondrug",
        "difficulty": "m",
        "definition": "Substances or compounds that are not considered
        drugs and do not have therapeutic effects.",
        "explanation": "In the context of drug discovery, nondrug
        molecules are those that are used as controls or are part of
        the compound libraries but are not expected to lead to a
        therapeutic drug.",
        "use_case": "The sentence compares nondrug molecules with drug
        molecules and other compounds to understand their relevance or
        irrelevance in the drug discovery process."
    },
    {
        "word": "compounds",
        "difficulty": "e",
        "definition": "Chemical substances composed of two or more
        elements chemically bonded together.",
        "explanation": "In the context of the sentence, 'compounds'
        refers to various chemical entities that are being investigated
        in the drug discovery process to find those with potential as
        drug molecules.",
        "use_case": "The sentence discusses investigating different
        classes of compounds in relation to their relevance to drug
        discovery."
    }
]

###
```

Prompt PRM_ZS_TASK2_V3

```
Identify up to a total of 5 complex words in the sentence and determine
the difficulty of each complex word by classifying it into the category
of "easy", "medium" or "difficult". For the "easy" difficulty place the
letter "e", for the "medium" difficulty place the letter "m", and for
the "hard" difficulty place the letter "d". Furthermore, in the case of
words identified as complex taking into account the context of the
sentence, three things must be generated: a definition, an explanation,
and a use case.
Finally, provide the response in the following JSON format:
{
  "word": <word>,
  "difficulty" :<hard_word>,
  "definition": "<definition>",
  "explanation" : <explanations>
}

##

Context: Humanoid robots
Sentence: In many situations, users walk on typical paths between
specific destinations at which the service of a mobile robot is needed.
```

```
Response:

json
[
    {
        "word": "humanoid",
        "difficulty": "m",
        "definition": "having an appearance or character resembling
        that of a human.",
        "explanation": "In the context of robots, ’humanoid’ refers to
        robots designed to look and sometimes behave like humans, often
        used to interact with environments built for human use.",
        "use_case": "The humanoid robot was programmed to assist
        customers in the store."
    },
    {
        "word": "destinations",
        "difficulty": "m",
        "definition": "the place to which someone or something is going
        or being sent.",
        "explanation": "In the given sentence, ’destinations’ refers to
        the specific places where the mobile robot is required to
        perform services.",
        "use_case": "The tourists set off early in the morning to reach
        their destinations before sunset."
    },
    {
        "word": "mobile",
        "difficulty": "e",
        "definition": "able to move or be moved freely or easily.",
        "explanation": "Here, ’mobile’ describes a robot that is
        capable of moving around rather than being fixed in one
        location.",
        "use_case": "The company developed a mobile app to increase
        accessibility for users on the go."
    }
]
###
```

4.2. Results

The Table 2 presents various results in response to the approach of sub-task 2.1. To automatically support this task, we have proposed an innovative approach that explores the potential of GPT-4 Turbo, an advanced large language model (LLM) developed by OpenAI as in the work done by [21]. On this occasion, we seek to evaluate the improvements and additional capabilities of GPT-4 Turbo in comparison with the proposal made for the same task in the SimpleText 2023 workshop where the results obtained were outstanding, demonstrating the effectiveness of the GPT-3 model, text-davinci-003 version [26].

The Table 3 presents several results of sub-task 2.2 applying GPT-4 Turbo with zero-shot learning according to the PRM_ZS_TASK2_V2 prompt. As we can see, the table illustrates an example that corresponds to Snt_id G01.1_1000902583_1. In the solution proposed for sub-task 2.2, the model generates a definition, an explanation, and also generates a use case illustrating in a broad way the complexity of the terms for different groups of users, making suggestions for definitions, explanations and use cases based on the context of the sentence (Abstract).

As we can see, the table illustrates an example that corresponds to Snt_id G01.1_1000902583_1. In the solution proposed for sub-task 2.2, the model generates a definition, an explanation, and also generates a use case illustrating in a broad way the complexity of the terms for different groups of users, making suggestions for definitions, explanations and use cases based on the context of the sentence (Abstract).

Table 4, we present the official results published by the organizers [27], including the scores obtained

Table 2
Predictions generated by applying GPT-4 Turbo with zero-shot learning in Sub-task 2.1-Prompt PRM_ZS_TASK2_V1.

Manual	Snt_id	Term	Difficulty
0	G01.1_1000902583_1	exploited	d
0	G01.1_1000902583_1	confrontation	m
0	G01.1_1000902583_1	coordination	m
0	G05.1_2914002216_4	CRISPR/Cas9	d
0	G05.1_2914002216_4	reagents	m
0	G05.1_2914002216_4	high-content screen	d
0	G07.2_2773680786_5	conspiracy	m
0	G07.2_2773680786_5	emergent	m
0	G07.2_2773680786_5	pervasive	d
0	G11.1_2946157960_5	consideration	m
0	G11.1_2946157960_5	applications	e
0	G11.1_2946157960_5	deployment	m
0	G11.1_2946157960_5	networked	d
0	G11.1_2946157960_5	emission	m

in our participation in subtasks 2.1 and 2.2. The meaning of each column is the following:

- **recall overall:** the proportion of terms (independently from the difficulty) that were found.
- **recall average:** the average of the recall of terms computed per sentence.
- **recall difficult terms:** the proportion of difficult terms that were found.
- **precision difficult:** the precision of terms that were labeled as difficult.
- **bleu_nx:** the BLEU score computed with ngrams n =1, 2, 3, 4.

According to the results presented in table 4, the UboNLP_Task2.1_phi3-oneshot team showed the best overall performance in terms of recall_overall (0.5424) and recall_average (0.5620), which indicates a high ability to retrieve relevant terms. However, its precision in difficult terms is relatively low (0.3743). Sharingans_Task2.2_GPT also performs well with a recall_overall of 0.4722 and a recall_average of 0.5302, in addition, it has a high recall_terms (0.5448), demonstrating a good ability to handle difficult terms.

According to the performance analysis in the difficult terms, the AIIRLab_Task2.2_LLaMAFT and SINAI_task_2_PRM_ZS_TASK2_V3 teams have a precision_difficult_terms of 1.0000 and 0.8333 respectively, which indicates that these models are very precise with difficult terms, although their recall_overall and recall_average are not as high in comparison with other equipment.

In the performance in BLEU Scores, the unipd_t21t22_chatgpt teams and their variants (mod1 and mod2) have the best bleu_n1_average, bleu_n2_average, bleu_n3_average and bleu_n4_average demonstrating good performance in terms of precision of n-grams in text generation. In the case of AIIRLab_Task2.2_Mistral, this team has a good balance with a recall_overall of 0.4118 and precision_difficult_terms of 0.4907, suggesting that their model is quite robust in general terms and in difficult terms. The SINAI team variants show variations in their results, with SINAI_task_2_PRM_ZS_TASK2_V2 having the best performance among its variants, with a recall_overall

Table 3
Predictions generated by applying GPT-4 Turbo with zero-shot learning in Sub-task 2.2 - Prompt PRM_ZS_TASK2_V1

Snt_id: G01.1_1000902583_1
Abstract: Novel technological advances in mobile devices and applications can be exploited in wildfire confrontation, enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles.

term	difficulty	Definition/Explication
exploited	d	***Definition:*** make full use of and derive benefit from (a resource). ***Explication:*** In the context of the sentence, exploited refers to utilizing the new technological capabilities of mobile devices and applications to their fullest potential in managing and combating wildfires. **Use case:** The novel technologies can be exploited to improve the efficiency and effectiveness of wildfire management.
confrontation	m	***Definition:*** a hostile or argumentative meeting or situation between opposing parties. ***Explication:*** Here, confrontation is used metaphorically to describe the act of dealing with or managing wildfires, which are seen as the opposing party to human safety and property. **Use case:** Technological advances aid in the confrontation of wildfires by providing better tools for managing the crisis.
coordination	m	***Definition:*** the organization of the different elements of a complex body or activity so as to enable them to work together effectively. ***Explication:*** In this context, coordination refers to the organized management of personnel and vehicles, ensuring they operate in a synchronized and efficient manner during wildfire emergencies. **Use case:** Effective coordination of personnel and vehicles is crucial for rapid and efficient wildfire suppression.

of 0.1556 and precision_difficult_terms of 0.7746. There are teams with values of 0 in almost all metrics, which indicates very low performance or problems in the implementation of their models.

5. Conclusions and Future Work

The model has proven to be able to generate robust responses based on the text or instruction (message) provided. We observe that GPT-4 Turbo analyzes the text, uses its extensive knowledge to identify words that can be considered complex, and categorizes them according to their complexity. Our preliminary analysis evaluates the model's ability to understand and generate arguments in specific contexts. The results show that GPT-4 Turbo is highly competent in natural language processing tasks, such as predicting lexical complexity, demonstrating its effectiveness in identifying and categorizing complex terms accurately and consistently.

Table 4
SimpleText results 2024 official results of task 2

runid	**recall_ overall**	**recall_ average**	**recall_ difficult _terms**	**precision difficult _terms**	**bleu_ n1_ average**	**bleu_ n2_ average**	**bleu_ n3_ average**	**bleu_ n4_ average**
AIIRLab_Task2.2_LLaMA	0.2792	0.3011	0.2642	0.6667	0.2883	0.1519	0.0497	0.0191
AIIRLab_Task2.2_LLaMAFT	0.0069	0.0056	0.0047	1.0000	0.2405	0.1171	0.0000	0.0000
AIIRLab_Task2.2_Mistral	0.4118	0.4415	0.1863	0.4907	0.2610	0.1338	0.0395	0.0128
Dajana&Kathy_SimpleText _Task2.2_LLAMA2_13B_CHAT	0.0118	0.0114	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FRANE_AND_ANDREA _SimpleText_Task2.2_LLAMA2 _13B_CHAT	0.0076	0.0066	0.0094	0.3636	0.0000	0.0000	0.0000	0.0000
ruby	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sharingans_Task2.2_GPT	0.4722	0.5302	0.5448	0.5954	0.2257	0.1039	0.0300	0.0160
SINAI_task_2_PRM_ZS_ TASK2_V1	0.0868	0.0872	0.1014	0.5244	0.2545	0.1579	0.0821	0.0578
SINAI_task_2_PRM_ZS_ TASK2_V2	0.1556	0.1636	0.1297	0.7746	0.2774	0.1574	0.0630	0.0443
SINAI_task_2_PRM_ZS_ TASK2_V3	0.0951	0.1045	0.0472	0.8333	0.2144	0.1113	0.0377	0.0229
team1_Petra_and_Regina_ Task2_ST	0.0042	0.0042	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Tomislav&Rowan_Task2.2_ LLAMA2_13B_CHAT	0.0069	0.0040	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Tomislav&Rowan_Task2.2_ LLAMA2_13B_CHAT_1	0.0083	0.0084	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
UAms_Task2-1_RareIDF	0.0854	0.0942	0.0259	0.0894	0.0001	0.0000	0.0000	0.0000
UboNLP_Task2.1_phi3-oneshot	0.5424	0.5620	0.3160	0.3743	0.0011	0.0000	0.0000	0.0000
unipd_t21t22_chatgpt	0.1340	0.1400	0.0825	0.6250	0.3045	0.1851	0.0905	0.0507
unipd_t21t22_chatgpt_mod1	0.2194	0.2371	0.1981	0.5957	0.3060	0.1783	0.0802	0.0430
unipd_t21t22_chatgpt_mod2	0.3146	0.3155	0.3420	0.6905	0.0302	0.0069	0.0031	0.0000

We have applied the GPT-4 Turbo model in the construction of several solutions for sub-tasks 2.1 and 2.2, where the model has demonstrated robust performance in the execution of natural language processing tasks, specifically in lexical simplification and identification of complex words. Not only has it successfully tackled these tasks, but it has also been able to generate detailed illustrations by creating definitions, explanations and use cases based on the context of the sentence (text). By identifying complex words and phrases and offering simpler versions, it facilitates access to scientific sources in an understandable way, helping readers to understand the complexity of terms in different user groups, which makes it very useful for a general audience. and especially valuable for readers with reading difficulties or cognitive disabilities.

The results showed a variety of performances among the participating teams. Some teams, such as UboNLP_Task2.1_phi3-oneshot and Sharingans_Task2.2_GPT, stand out for their overall performance and ability to handle difficult terms, while other teams show specific strengths such as accuracy on difficult terms or good BLEU scores. However, there are also teams with significantly low performances, which could indicate the need for improvements in their approaches or implementations.

6. Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by the Spanish Government.

References

[1] W. H. DuBay, The principles of readability., Online Submission (2004).

[2] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it simplext: Implementation and evaluation of a text simplification system for spanish, ACM Transactions on Accessible Computing (TACCESS) 6 (2015) 1–36.

[3] K. North, M. Zampieri, M. Shardlow, Lexical complexity prediction: An overview, ACM Computing Surveys 55 (2023) 1–42.

[4] I. Segura-Bedmar, P. Martinez, Simplifying drug package leaflets written in spanish by using word embedding, Journal of Biomedical Semantics 8 (2017). doi:10.1186/s13326-017-0156-7.

[5] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it simplext: Implementation and evaluation of a text simplification system for spanish, ACM Trans. Access. Comput. 6 (2015). URL: https://doi.org/10.1145/2738046. doi:10.1145/2738046.

[6] A. Anula, Lecturas adaptadas a la enseñanza del español como l2: variables lingüísticas para la determinación del nivel de legibilidad, La evaluación en el aprendizaje y la enseñanza del español como LE L 2 (2008) 162–170.

[7] R. Alarcón, L. Moreno, P. Martínez, Hulat-ALexS CWI Task-CWI for Language and Learning Disabilities Applied to University Educational Texts, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 2020.

[8] L. Ermakova, H. Azarbonyad, S. Bertin, O. Augereau, Overview of the clef 2023 simpletext task 2: Difficult concept identification and explanation (2023).

[9] L. Ermakova, et al., Overview of CLEF 2024 SimpleText track on improving access to scientific texts, in: L. Goeuriot, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024.

[10] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[11] G. Paetzold, L. Specia, SemEval 2016 Task 11: Complex Word Identification, 2016, pp. 560–569. doi:10.18653/v1/S16-1085.

[12] S. M. Yimam, C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, M. Zampieri, 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, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 66–78. URL: https://aclanthology.org/W18-0507. doi:10.18653/v1/W18-0507.

[13] J. Ortiz-Zambrano, A. Montejo-Ráez, Overview of ALexS 2020: First workshop on lexical analysis at SEPLN, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), volume 2664, 2020, pp. 1–6.

[14] M. Shardlow, R. Evans, G. H. Paetzold, M. Zampieri, SemEval-2021 task 1: Lexical complexity prediction, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 1–16. URL: https://aclanthology.org/2021.semeval-1.1. doi:10.18653/v1/2021.semeval-1.1.

[15] S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, W. Xu, Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022),

in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), 2022.

[16] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. Sanjuan, Overview of SimpleText CLEF 2021 workshop and pilot tasks, in: 12th Conference and Labs of the Evaluation Forum (CLEF 2021), volume 2936, CEUR-WS. org, 2021, pp. 2212–2227.

[17] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, E. Mathurin, P. Bellot, Overview of the clef 2022 simpletext lab: Automatic simplification of scientific texts, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 470–494.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[21] K. K. Shahandashti, M. Sivakumar, M. M. Mohajer, A. B. Belle, S. Wang, T. C. Lethbridge, Evaluating the effectiveness of gpt-4 turbo in creating defeaters for assurance cases, 2024. `arXiv:2401.17991`.

[22] S. Kublik, S. Saboo, GPT-3: Building innovative NLP products using large language models, O'Reilly Media (2022).

[23] A. Chan, GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry, AI and Ethics 3 (2023) 53–64.

[24] Y. Hirano, S. Hanaoka, T. Nakao, S. Miki, T. Kikuchi, Y. Nakamura, Y. Nomura, T. Yoshikawa, O. Abe, Gpt-4 turbo with vision fails to outperform text-only gpt-4 turbo in the japan diagnostic radiology board examination, Japanese Journal of Radiology (2024) 1–9.

[25] A. Bhattacharjee, R. Moraffah, J. Garland, H. Liu, Zero-shot llm-guided counterfactual generation for text, 2024. `arXiv:2405.04793`.

[26] J. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, Sinai participation in simpletext task 2 at clef 2023: Gpt-3 in lexical complexity prediction for general audience (2023).

[27] G. M. D. Nunzio, et al., Overview of the CLEF 2024 SimpleText task 2: Identify and explain difficult concepts, in: G. Faggioli, et al. (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS.org, 2024.