CEUR-WS.org/Vol-3497/paper-255.pdf

SINAI Participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for General Audience

Notebook for the SimpleText Lab at CLEF 2023

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 paper our participation in the CLEF 2023 SimpleText track's Task 2.1 and 2.2 is described. In our approach, we explore zero-shot and few-shot learning strategies over the auto-regressive model GPT-3. Several prompts to achieve those strategies were tested. Our results were ranked among top submitted runs and demonstrated a solid performance for the task of lexical complexity prediction.

Keywords

Lexical Complexity Prediction, Auto-regressive models, GPT-3, Prompting, Zero-shot learning, Few-shot learning

1. Introduction

Reading involves a complex process that goes beyond dealing with words or sections that are difficult for the reader to understand. Therefore, it is necessary to have an adequate understanding of the content of the texts in order to build coherent mental representations and thus fully grasp their meaning [1].

Information technologies make it possible for people to access abundant information in different fields such as education, information, social, health or government and even science. However, this information is not accessible to many, since some people face great reading barriers such as long sentences, technical language or complex linguistic structures that do not allow them to understand the content of the texts, being directly affected people with intellectual disabilities and people with low learning level; including university students, who are people with a high level of education and specialized knowledge in different subjects of study but, even so, could be part of groups of people with reading disabilities [2].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] 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)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

The task of detecting in the content of the documents the words that are difficult or complex to understand for a certain group is called Complex Word Identification. This task is essential in many applications related to Natural Language Processing (NLP), such as Text Simplification [3].

The SimpleText[1] lab at CLEF2023 [4] aims to generate a concise summary of various scientific documents, based on a popular scientific query, in order to provide the user with an easily understandable overview of a specific topic, as people generally avoid reliable sources, such as scientific literature, due to their complex language and lack of prior knowledge. Instead, they prefer to rely on superficial, web- and social media-generated sources, which are often motivated by commercial or political interests, rather than providing quality information. Can text simplification help to remove some of these access barriers? The purpose of the task is to generate a shortened version of several scientific papers, based on a popular scientific query, in order to provide the user with an easy-to-understand overview of a specific subject.

The main purpose of this article is to demonstrate the ability of the GPT-3 Transformers-based language model to perform text classification in the realm of lexical simplification achieved by creating and evaluating several prompt variants for few-shot and zero-shot learning examples of short sentences to determine word complexity.

The paper is organized as follows: first, a brief overview of the state-of-art is provided in lexical complexity prediction and complex word identification (which are synonyms for the same NLP task). Then, large language models are introduced, with special emphasis on auto-regressive ones (generative models). This is complemented with a brief description of zero-shot and few-shot learning approaches usually applied with these models. In Section 4, Task 2 in SimpleText track is introduced. Section 5 presents our solution and the results obtained with different variations on prompting. Finally, conclusions and some insights on planned work is provided in Section 6.

2. Related Work

In recent years, Complex Word Identification (CWI) has aroused great interest in the scientific community and in computational linguistics researchers. These initiatives contribute to development and research in the field of simplifying texts and improving the accessibility of information. These proposals include conferences, workshops and specific tasks that seek to address the challenges and promote advances in the field of text simplification. This is evidenced in the development of computational semantic analysis systems, as demonstrated in CWI shared tasks at events such as SemEval 2016 [5], NAACL-HTL 2018 [6], ALexS at IberLEF 2020 [7], the 15th edition of SemEval and the first Lexical Complexity Prediction task [8], TSAR-2022 - the workshop on Simplification, Accessibility and Readability of Texts [9], the SimpleText task at CLEF in 2021 [10] and 2022 [11], among other initiatives in this regard.

---

[1] texttthttp://simpletext-project.com/2023/clef/

2.1. GPT-3 for solving NLP tasks

State-of-the-art deep learning models such as BERT [12], RoBERTa [13], GPT-3 [14] and others outperform traditional approaches. Last generation GPT-3, or Generative Pre-trained Transformer 3, is a transformer-based large language model developed by OpenAI[2]. It contains 175 billion parameters. Due to its extensive language knowledge, enormous processing power, and potential to learn from vast 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 the text [15]. The enormous scale of the model allows to generate results with quality, precision and diversity of the generated content. This breakthrough has sparked considerable interest and concern in the NLP field, the machine learning industry in general, the media, the AI ethics communities, and civil society [16].

Although GTP-3 is a generative model, it can take several approaches to get to classify text, which are mainly zero shot classification (in which no examples are provided to the model) to single shot and few shot classification (in which it shows some examples to the model) [15].

In zero-shot learning, no prior training or adjustment to the labeled data is required. Currently, GPT-3 produces results for invisible data, but in order to perform zero-shot classification with GPT-3, we must provide you with a compatible prompt [15].

In few-shot learning, the prompt is composed including some examples of the task to be solved. The ability to learn with few attempts, never before seen in Natural Language Processing (NLP) models, is a prominent feature of GPT-3 [16].

Another method of classifying text is by fine-tuning an AI model with one or a few training examples, known as one or few-shot text classification. By providing examples of how to classify text, the model can learn information about object categories based on those examples [15].

3. SimpleText task at CLEF 2023

The SimpleText track at CLEF faces the challenges of text simplification methods with the purpose of promoting access to scientific information. Provided a corpus of scientific literature abstracts and popular science requests. Three tasks were proposed. First, the content selection task (what is in, or out?) poses a challenge to systems, as they must choose which passages to include in a simplified summary in response to a query, deciding what information should be present, and what information should be omitted. Second, the goal of complexity detection (what is unclear?) is to identify what is unclear in a passage and a query. It seeks to classify the terms or concepts that require additional explanation to understand the passage, including definitions, context, and applications. Third, the text simplification task (rewrite this!) is to simplify passages of scientific abstracts in response to a query, while keeping the essential content intact [11].

We have only participated in task 2, complexity spotting.

---

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

3.1. Task 2: What is unclear?

Complexity Spotting: Identifying and explaining difficult concepts for general audience. The goal of this task is:

1. to decide which terms (up to 5) require explanation and contextualization to help a reader to understand a complex scientific text – for example, with regard to a query, terms that need to be contextualized (with a definition, example and/or use-case)
2. to provide short (one/two sentence) explanations/definitions for the detected difficult terms. For the abbreviations, the definition would be the extended abbreviation.

For each piece of text, participants must submit an ordered list of terms that are difficult, along with scores on a scale of 1 to 3 (where 3 indicates the most difficult terms, and 1 are easiest to deduce or guess), and also provide definitions for these terms. The text fragments (sentences) are considered independently, which means that the repetition of difficult terms in different fragments is allowed.

4. Proposed system

As stated before, our approach consists in applying zero-shot or few-shot prompting to GPT-3 model, extracting the predictions from the generated sequence. The details of the model are given in Table 1. The model is queried through its Python API and OpenAI playground is a convenient web interface that can be used to generate the needed code in a fast and easy way.

Table 1
GPT-3 model details

Parameter	Value
Model	text-davinci-003
Temperature	0.7
Max_tokens	250
Top_p	1
Frecuency Penalty	0
Presence Penalty	0
Logprobs	5
Stop	"###"

4.1. Experiments with few-shot learning

We apply three executions with few-shot learning with different prompts. The values of the settings also varied, it was observed that in the determination of the level of complexity (difficult, very difficult or neutral), the model tended to identify terminology beyond its verbatim expression in the text. After the experiments, we show that this approach is important because the model can be adapted to a specific task. Next is the detail of the different prompts tested:

Prompt PRM_FS_TASK2_1_V1:

```
To retrieve up to 5 difficult terms in a given passage from a
scientific abstract. Rank the list of  difficult terms with
corresponding scores on the scale 1-3 where 3  is the most
difficult term.  You have a maximum of 250 tokens you can t go over
that. (here is an example)
context: quantum computing
Text: This paper describes a tool that converts Unreal levels to web-ready
environments in VRML and X3D.
##
Term 1: X3D
Difficulty:2

Term 2: vrml
Difficulty:2

Now it is your turn

##
Context: drones
Text: In an attempt to achieve the above mentioned tasks, we
propose an imitation learning based, data-driven solution to UAV
autonomy for navigating through city streets by learning to fly
by imitating an expert pilot.
###
```

Prompt PRM_FS_TASK2_1_V2:

```
To decide which terms (up to 5) require explanation and
contextualization to help a reader to understand a complex
scientific text. Rank the list of difficult terms with
corresponding scores on the scale 1-3 where 3 is the most
difficult term. You have a maximum of 250 tokens you can t go
over that. (here is an example)
context: quantum computing
Text: This paper describes a tool that converts Unreal levels to web-ready
environments in VRML and X3D.
##
Term 1:X3D
Difficulty:2

Term 2: vrml
Difficulty:2

Now it is your turn

##
Context: misinformation
Text: We find that Italy is the most polarized country, followed
by France, Germany and lastly Spain.
###
```

Prompt PRM_FS_TASK2_2_V1:

```
Select a number from 1 to 5 terms within the sentence named
text, rank them from 1 to 3 is the most difficult and finally
give the meaning of each difficult term and add an example and
an use-case. Use next form (term, difficulty, 'definition,
example, use-case'). You have a maximum of 250 tokens you can t go
over that. Here is an example:
context: quantum computing
```

```
Text: This paper describes a tool that converts Unreal levels to
web-ready environments in VRML and X3D.
##

Term 1:3D
Difficulty:2
Definition: X3D is an ISO-ratified, file format and run-time
architecture to represent and communicate 3D scenes and objects.
X3D fully represents 3-dimensional data.

Term 2:vrml
Difficulty:2
Definition: Virtual Reality Modeling Language (VRML) is a
standard file format for representing 3-dimensional interactive
vector graphics, designed particularly with the World Wide Web
in mind. It has been superseded by X3D.

Now it is your turn

##
Context: drones
Text: This paper discusses preliminary work regarding robot
companionship applications by using a miniature humanoid capable
of fetching different toys based on voice command.
###
```

The Table 2 presents the results generated by applying few-shot learning, where the model has been able to learn from a very small number of training examples.

4.2. Experiments with zero-shot learning

In the case of zero-shot learning, no examples are provided in the prompt.

Prompt PRM_ZS_TASK2_1_V1:

```

To retrieve up to 5 difficult terms in a given passage from a
scientific abstract. Rank the list of difficult terms with
corresponding scores on the scale 1-3 where 3 is the most
difficult term.  you have a maximum of 250 tokens you can t go over that.

##
Context: drones
Text: With the ever increasing number of unmanned aerial
vehicles getting involved in activities in the civilian and
commercial domain, there is an increased need for autonomy in
these systems too.
###
```

Prompt PRM_ZS_TASK2_1_V2:

```
To decide which difficult terms (up to 5) are in the complex
scientific text. Rank the list of difficult terms with
corresponding scores on the scale 1-3 where 3 is the most
difficult term.  you have a maximum of 250 tokens you can t go over that.

##
Context: drones
```

Table 2
Predictions generated by applying GPT-3 with *few-shot learning*

# Execution	Prompt	Term_rank_snt	Term	Difficulty
1 - Task 2.1	PRM_FS_TASK2_1_V1	1	UAV	2
		2	Autonomy	3
		3	Imitation Learning	3
		4	Data-Driven	2
		5	Expert Pilot	2
2 - Task 2.1	PRM_FS_TASK2_1_V2	1	Polarized	3
		2	Italy	1
		3	France	1
		4	Germany	1
3 - Task 2.2	PRM_FS_TASK2_2_V1	1	Voice command ***Definition:*** *Voice command* is a type of technology that enables a user to control devices or services using voice commands. ***Example:*** *Voice command* technology is used in Amazon's Alexa to control smart home devices. ***Use Case:*** *Voice command* technology can be used to control home appliances, such as lights, thermostats, and security systems.	2
		2	Robot companionship ***Definition:*** *Robot companionship* is a type of human-robot interaction in which robots are used to provide companionship to humans. ***Example:*** A robotic pet can be used to provide companionship to elderly people who live alone. ***Use Case:*** *Robot companionship* can be used to provide comfort and companionship to people who are isolated or have difficulty forming relationships with other people.	3

```
Text: In an attempt to achieve the above mentioned tasks, we
propose an imitation learning based, data-driven solution to UAV
autonomy for navigating through city streets by learning to fly
by imitating an expert pilot.
###
```

Prompt PRM_ZS_TASK2_2_V1

```
Select a number from 1 to 5 terms within the sentence named
text, rank them from1 to 3 is the most difficult and finally
give the meaning of each difficult term and add an example and
an use-case. Use next form (term, difficulty, 'definition,
example, use-case'). you have a maximum of 250 tokens you can t go over that.

##
Context: drones
Text: Derived from the classic image classification algorithms,
our classifier has been constructed in the form of a fast 39-
layered Inception model, that evaluates the presence of roads
using the tomographic reconstructions of the input frames.
###
```

The Table 3 presents the results generated by applying zero-shot learning, where the model has been able to generate its own results without the need for it to learn from a small number of examples.

In the example of one of the texts presented by the Table 3 for the task 2.1 on first run - PRM_ZS_TASK2_1_V1, the model has identified a total of 5 words (corresponds to the column *Term_rank_snt*) achieving to generate its response according to the request made. The model has also managed to identify the complex terms of the text (corresponds to the column *Term*), The model has also been able to provide a complexity value for each term (corresponds to the column *Difficulty*).

In the second execution for the generation of results of task 2.1 - PRM_ZS_TASK2_1_V2, the model has not only been able to identify a total of 5 difficult terms from the content of the text, but it has also been able to generate a brief description of the difficult terms; we must emphasize that it had not been specified in the request.

Finally, taking as a sample one of the examples resulting from the third execution - PRM_ZS_TASK2_2_V1, and as a requirement specified in task 2.2, the model has generated a definition, an example, and a use case to better illustrate the reader about the term identified as difficult. In the run example, the model has identified 4 complex words in the text.

According to the examples of the analyzed sample, GPT-3 has shown an outstanding ability to recognize the definitions of acronyms, generating coherent definitions, examples and use cases, the issue of identification and recognition of complex words have been addressed with success.

5. Results

In Tables 4 and 5 we show the results obtained in our participation at task 2.1 and 2.2 respectively, according to the scores officially released by the organizers. The runs were named against our team ID (SINAI) together with the prompt strategy used from the ones described in previous sections. We have omitted our team ID for simplification.

Our results in task 2.1 show a significant performance of the proposed approach close to top performing systems. It is interesting to note that, in general, version 1 for the prompt is better and that zero-shot learning prompts are better at detecting a difficult word, while few-shot learning prompts are better at determining a difficulty score. A mixture of both approach could be worth exploring.

In task 2.2 we obtain high values at semantic match. For providing definitions or explanations over difficult terms, few-shot learning prompting is consistently better than zero-shot learning prompts.

6. Conclusion and future work

We present the results obtained by a auto-regressive language model with 175 billion parameters that demonstrates strong performance in solving tasks 2.1 and 2.2 proposed in SimpleText@CLEF-2023.

The model demonstrated a solid performance in obtaining the results in the experiments applying zero-shot and few-shot approaches, generating high-quality samples. In addition,

Table 3
The Experiments applying GPT-3 with Zero-shot learning

# Execution	Prompt	Term_rank_snt	Term	Difficulty
1 - Task 2.1	PRM_ZS_TASK2_1_V1	1	Autonomy	3
		2	Unmanned Aerial Vehicles (UAVs)	2
		3	Civilian	1
		4	Commercial	1
		5	Activities	1
2 - Task 2.1	PRM_ZS_TASK2_1_V2	1	UAV autonomy:	3
		2	Imitation learning: A type of machine learning algorithm that learns from observing and mimicking an expert.	2
		3	Data-driven Using data to inform a decision-making process.	2
		4	City streets Urban roadways.	1
		5	Expert pilot A highly skilled and experienced pilot.	1
3 - Task 2.2	PRM_ZS_TASK2_2_V1	1	Inception model ***Definition:*** A type of convolutional neural network, often used for image classification and object ecognition. ***Example:*** GoogleNet, a popular *Inception model*, was used to classify images in the ImageNet challenge ***Use Case:*** Drones can use *Inception models* to identify roads in tomographic reconstructions of input frames.	3
		2	Image classification ***Definition:*** The process of recognizing the content of an image. ***Example:*** A photo of a dog may be classified as an image of an animal. ***Use Case:*** Drones can use *image classification* algorithms to identify features in tomographic reconstructions of input frames.	2
		3	Tomographic reconstructions ***Definition:*** The process of reconstructing an image from a series of slices taken from different angles. ***Example:*** A computed tomography (CT) scan is a type of *tomographic reconstruction*. ***Use Case:*** Drones can use *tomographic reconstructions* to analyze the presence of roads in input frames.	3
		4	ImageNet challenge ***Definition:*** A competition to classify images into different categories. ***Example:*** The *ImageNet challenge* involves classifying images into categories like animals, plants, and vehicles. ***Use Case:*** Drones can use the results of the *ImageNet challenge* to enhance their capabilities in various ways.	1

Table 4
SINAI results for task 2.1. The ranking reached over the 34 runs submitted by different teams is shown. The results obtained by the best submitted run is at the last row.

Run	total	evaluated	TERM LIMITS OK	Correct difficulty scores	Correct LIMITS & Correct difficulty scores
PRM_ZS_TASK2_1_V1	11081	1044	970 (5th)	450 (3rd)	428 (3rd)
PRM_FS_TASK2_1_V1	10768	1025	942 (6th)	498 (2^{d})	471 (2^{d})
PRM_ZS_TASK2_1_V2	10952	861	794 (7th)	426 (4th)	402 (4th)
PRM_FS_TASK2_1_V2	8836	800	739 (8th)	397 (5th)	378 (5th)
UAms_Task_2_RareIDF	675090	1200	1068 (1st)	503(1st)	476 (1st)

Table 5
SINAI results for task 2.2.

Run	total	BLEU BLEU	ROUGE precision	ROUGE recall	ROUGE fmeasure	semantic match
Abbreviation extension results:						
SINAI_task_2.1_PRM_FS_TASK2_2_V1	228	1.35	0.05	0.81	0.1	0.53
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	228	1.61	0.07	0.76	0.13	0.55
Definition generation results:						
SINAI_task_2.1_PRM_FS_TASK2_2_V1	75	4.17	0.19	0.37	0.24	0.78
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	89	3.2	0.17	0.29	0.21	0.68
Results on the extended test set:						
SINAI_task_2.1_PRM_FS_TASK2_2_V1	257	4.64	0.17	0.39	0.22	0.77
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	649	03.08	0.16	0.34	0.19	0.69

the model shows that learning carried out with few samples has the ability to generalize from limited information. Few-shot learning has been applied to a never-before-performed task where it has been shown to overcome the limitation of large training data sets allowing the model to learn quickly and effectively with few examples.

Despite the fact that the model presents several limitations and weaknesses, these results indicate that extremely large language models as GPT-3 can play a crucial role in the development of accessibility related solutions.

As future work, we plan to explore more strategies on prompting, like Chain-of-Thought or adding more examples.

Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

We appreciate Arteaga Burgos Anthony Alexander, graduate of the Computer Systems Engineering degree from the University of Guayaquil, for his valuable contribution to the development of our work.

References

[1] P. van den Broek, Using texts in science education: Cognitive processes and knowledge representation, Science (New York, N.Y.) 328 (2010) 453–6. doi:`10.1126/science.1182594`.

[2] 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.

[3] A. Rico-Sulayes, General lexicon-based complex word identification extended with stem n-grams and morphological engines, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 2020.

[4] M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos, Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS, 2023.

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

[6] 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`.

[7] 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.

[8] 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`.

[9] 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.

[10] 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.

[11] 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.

[12] 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).

[13] 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).
[14] 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.
[15] S. Kublik, S. Saboo, GPT-3: Building innovative NLP products using large language models, O'Reilly Media (2022).
[16] A. Chan, GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry, AI and Ethics 3 (2023) 53–64.