



Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article



ChatGPT: Jack of all trades, master of none

Jan Kocoń^{*}, Igor Cichecki¹, Oliwier Kaszyca¹, Mateusz Kochanek¹, Dominika Szydło¹,
 Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń,
 Bartłomiej Koptyra, Wiktoria Mieszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy,
 Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, Przemysław Kazienko

Department of Artificial Intelligence, Wrocław University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

ARTICLE INFO

Dataset link: <https://github.com/CLARIN-PL/c-hatgpt-evaluation-01-2023>

Keywords:

ChatGPT
 GPT-4
 Natural language processing (NLP)
 Semantic NLP tasks
 Pragmatic NLP tasks
 Subjective NLP tasks
 Natural language inference (NLI)
 Sentiment analysis
 Offensive content
 Emotion recognition
 Humor detection
 Stance detection
 Word sense disambiguation (WSD)
 Question answering (QA)
 Model personalization
 Text classification
 SOTA analysis
 Large language model
 Prompting

ABSTRACT

OpenAI has released the Chat Generative Pre-trained Transformer (ChatGPT) and revolutionized the approach in artificial intelligence to human-model interaction. The first contact with the chatbot reveals its ability to provide detailed and precise answers in various areas. Several publications on ChatGPT evaluation test its effectiveness on well-known natural language processing (NLP) tasks. However, the existing studies are mostly non-automated and tested on a very limited scale. In this work, we examined ChatGPT's capabilities on 25 diverse analytical NLP tasks, most of them subjective even to humans, such as sentiment analysis, emotion recognition, offensiveness, and stance detection. In contrast, the other tasks require more objective reasoning like word sense disambiguation, linguistic acceptability, and question answering. We also evaluated GPT-4 model on five selected subsets of NLP tasks. We automated ChatGPT and GPT-4 prompting process and analyzed more than 49k responses. Our comparison of its results with available State-of-the-Art (SOTA) solutions showed that the average loss in quality of the ChatGPT model was about 25% for zero-shot and few-shot evaluation. For GPT-4 model, a loss for semantic tasks is significantly lower than for ChatGPT. We showed that the more difficult the task (lower SOTA performance), the higher the ChatGPT loss. It especially refers to pragmatic NLP problems like emotion recognition. We also tested the ability to personalize ChatGPT responses for selected subjective tasks via Random Contextual Few-Shot Personalization, and we obtained significantly better user-based predictions. Additional qualitative analysis revealed a ChatGPT bias, most likely due to the rules imposed on human trainers by OpenAI. Our results provide the basis for a fundamental discussion of whether the high quality of recent predictive NLP models can indicate a tool's usefulness to society and how the learning and validation procedures for such systems should be established.

1. Introduction

In recent years, Transformer-type model architecture has dominated the world of natural language processing (NLP) [1–3]. Before that, recurrent neural networks, such as LSTMs, were used to solve a wide variety of existing NLP problems [4–6]. The recurrent neural models could not capture distant dependencies in data sequences, for example, information occurring at the text beginning or end [7]. In addition, their architecture did not allow for efficient parallelization of training and inference processes [8]. The answer to the aforementioned problems was precisely the Transformer architecture, presented initially as an encoder–decoder model for sequence-to-sequence tasks [1]. Such

a model had the advantage of capturing distant relationships in the text using an attentional mechanism and easily parallelizing calculations with matrix operations. As more powerful GPUs and TPUs were developed [9], it became possible to create models with more and more parameters, resulting in models that began to achieve human performance for an increasing number of tasks [10–12]. However, the most significant quality improvement was achieved by unsupervised pre-training language models on a huge number of texts acquired from the Internet. In BERT-based models, the pre-training tasks involved foreseeing masked tokens and subsequent sentences [13]. In autoregressive models, the pre-training task has been changed to predicting

^{*} Corresponding author.

E-mail addresses: jan.kocoon@pwr.edu.pl (J. Kocoń), kazienko@pwr.edu.pl (P. Kazienko).

URL: <https://kazienko.eu> (P. Kazienko).

¹ Equal contribution.

<https://doi.org/10.1016/j.inffus.2023.101861>

Received 21 February 2023; Received in revised form 13 May 2023; Accepted 29 May 2023

Available online 3 June 2023

1566-2535/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

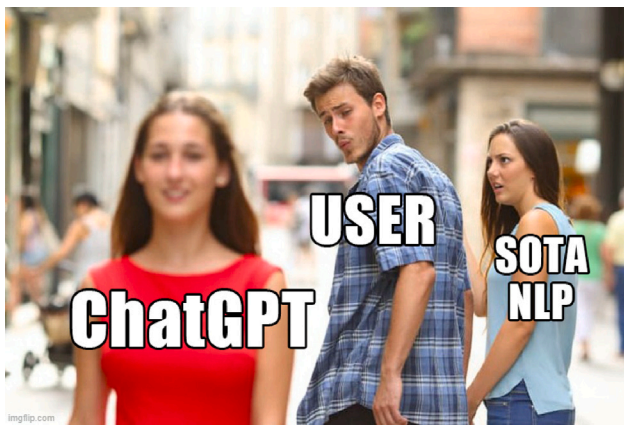


Fig. 1. Will a user charmed by the first impression created by ChatGPT abandon proven state-of-the-art solutions? We present the results of a study showing whether it is worth it.

the next word, which masks the attentional layer so that the model forecasts future values based only on past values [14].

Generative Pre-Training (GPT [15]) was one of the first autoregressive generative models based on the Transformer architecture. From the original Transformer, only the decoder stack is used by GPT, and bi-directional self-attention is converted to uni-directional. Such a model can perform all tasks based on generating new text, such as translation, summarization, or answering questions. In GPT-2, an extension of this concept, several technical improvements were made that eliminated the transferability problem for fine-tuning the models to downstream tasks and introduced multi-task training [16]. In addition, the input context length was doubled (from 512 to 1024), and the data for pre-training increased to 40 GB, but the total number of model parameters soared from 117M (GPT) to 1.5B (GPT-2). As a result, GPT-2 showed the ability to solve many new tasks without the need for supervised training on large data. Two factors mainly distinguished the succeeding GPT-3 model: the number of model parameters increased to 175B, and 45TB text data was used for pre-training. This model provided outstanding results, especially in zero-shot and few-shot scenarios [17].

A further step towards matching the model's responses to human needs was creating the InstructGPT model [18]. Its main innovation focused on alternative model fine-tuning methods, particularly Reinforcement Learning from Human Feedback (RLHF). This solution uses human feedback as a reward signal for updating model parameters. OpenAI recruited 40 annotators with high levels of agreement in sensitive speech flagging, ranking model answers by quality, sensitive demonstration writing, and the ability to identify sensitive speech for different groups. Their task was to describe what kind of answer is expected for different prompts, and the next GPT-3 finetuning followed this input. In the second step, the subjects created a ranking of several responses of the system based on the given prompt to train a reward model. In the third step, reinforcement learning using proximal policy optimization (PPO) was applied to improve the model quality further. As a result, users strongly preferred the InstructGPT responses compared to GPT-3. One of the conclusions from this work was that model quality on publicly available NLP benchmark datasets is worse than for SOTA models. However, InstructGPT authors found that benchmark NLP tasks do not reflect what most people may really expect from the language models [18]. Only 18% of users using the OpenAI API queried GPT-3 model with tasks familiar to typical NLP tasks, most of which are analytical. On the other hand, only a small fraction of popular NLP datasets have been used to evaluate InstructGPT [18].

One of the latest iterations of InstructGPT is the ChatGPT model (Fig. 2), which most likely exploited even more users' feedback on

a greater variety of tasks². At the moment, little information on the construction of this model is available, but the excellent quality of the system has resulted in its massive popularity (Fig. 1). Interestingly, the base model in InstructGPT is a model that has only 3.5B parameters [18]. Yet, in conversation tasks, it provides answers better than the GPT3 model, with 175B parameters. This shows the high relevance of collecting data from humans for supervised model fine-tuning [18]. ChatGPT's successor, the GPT-4 [19], is most likely an even larger model that can additionally receive not only text but also images as input.

In this work, we propose a new approach to testing a prompt-based model, ChatGPT, on various NLP tasks. We focus on evaluating the ChatGPT tool for 25 public NLP datasets, a large part of which involved subjective problems and for which there is a high probability that ChatGPT could be wrong. This intuition is based on the fact that OpenAI developers chose human annotators based on their high agreement rate [18]. At the same time, it is difficult to identify universal ground truth in tasks such as predicting emotions or offensiveness of text, especially in the personalized context [20–22]. It is very likely that the ChatGPT model has not been trained on most of the datasets that we test in our work, because for all of these datasets, we observe a significant drop in quality relative to state-of-the-art models. This allows us to assess its quality in various personalized NLP tasks. When it comes to the analysis and processing of the answers of the general majority, it is not difficult to retrieve information about the correlations and relationships between each task, however, grasping the preferences of each person individually is a much more demanding task, especially when analyzing the correlation between tasks in a personalized perspective. We have managed to successfully process our prompts, including ones that consisted of 3 annotated texts and one that had been later annotated by ChatGPT positively or negatively according to the already annotated texts. Those prompts were especially interesting, as ChatGPT was in fact tasked to predict the answer. This implied a certain “preference”, which is contradictory to the rule that ChatGPT should not have any emotions, preferences or opinions. The results we have obtained are the beginning of a discussion on whether the models trained on existing NLP tasks respond to people's demands and how to train such models so that they not only respond to the expectations of the majority of the population but also take into account the preferences of minority or outliers.

Language models are prone to generating responses containing human-like biases as well as presenting moral and ethical stances [23]. A number of procedures are created to make sure that these biases will not affect users, such as the European Union's AI Ethics Guidelines or AI Fairness 360. Of all that standards are addressing bias and fostering ethical development of AI systems [24]. Even though the creators of ChatGPT secure the model against answers that are offensive, there are still multiple methods that may reveal its hidden biases.

We wanted to consider some more specific research problems in the area of our research and thus specified 11 research questions that we have successfully managed to find the answers to:

- Q1: Is ChatGPT loss in performance compared to SOTA different for individual tasks of different kinds, Section 6.1 and the same for GPT-4, Section 6.7?
- Q2: Is there a difference in ChatGPT's ability to solve difficult and easy NLP analytical tasks, Section 6.2?
- Q3: How much a few-shot approach to personalization (Random Contextual Few-Shot Personalization) can make reasoning more subjective, thus, potentially increasing the overall inference quality, Section 6.3?
- Q4: What is the impact of the context while processing multiple questions (prompts) that may or may not be related to each other, Section 6.4?

² <https://openai.com/blog/chatgpt/>

Evolution from Transformer architecture to ChatGPT

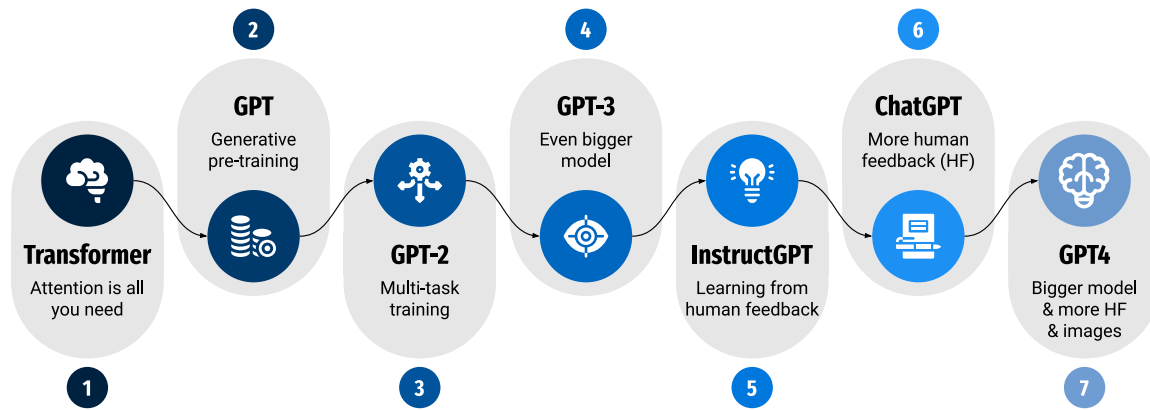


Fig. 2. Development of autoregressive models based on Transformer architecture: (1) basic model [1]; (2) first version of Generative Pre-Training (GPT) model [15]; (3) GPT-2 [16]; (4) GPT-3 [17]; (5) InstructGPT based on human feedback [18]; (6) ChatGPT — a model interacting in a conversational way, trained on more human feedback: <https://openai.com/blog/chatgpt>; (7) GPT-4 — a large-scale multimodal model with text and/or image as an input [19].

- Q5: Can we improve the model performance with prompt engineering, i.e. manual fine-tuning, Section 6.6?
- Q6: Is GPT-4 better or worse compared to ChatGPT, Section 6.7?
- Q7: Does the public availability of the data and its exploitation for training ChatGPT impact its performance, Section 6.5?
- Q8: What are necessary post-processing activities that can improve the quality of ChatGPT output for analytical tasks, Section 5.2?
- Q9: What is the internal policy of ChatGPT providers and its biases making it not to provide adequate responses to some prompts, Section 7.1?
- Q10: Can ChatGPT be used to validate the quality of the training datasets annotated by humans, Section 7.2?
- Q11: Can ChatGPT be used for explainability purposes while solving analytical tasks and ambiguous questions, Section 7.3?
- Q12: What are limitations and unexpected behavior of ChatGPT, Section 8? In which domains ChatGPT can catalyze AI technologies and change human everyday life, Section 9?

This collection of questions was formed after a thorough analysis of the available research and information regarding ChatGPT, which we have included in Section 3.

We have categorized our research into the quantitative analysis (Section 6), qualitative analysis (Section 7), limitations and discussion (Section 8) as well as prospective ChatGPT application domains (Section 9).

2. Related work

Early discourse related to ChatGPT revolves around two main topics — potential usage in expert fields and evaluation of specific tasks or aspects of chat performance. In the first topic, there are many papers suggesting potential benefits and risks of using ChatGPT in education (e.g. [25–27]), medicine (e.g. [28]), or even in the creation of legal documents (e.g. [29]). The main concerns about the usage of the chatbot are that it will escalate the issues of plagiarism in many fields (e.g. [25,30]) and might be used for cheating in academic tests [25]. The latter topic points out the strengths and vulnerabilities of ChatGPT performance. The two topics are strongly related as the main limitation of using the chatbot in expert fields is the reliability of the results. Thus the comprehensive and systematic evaluation is crucial for the proper assessment of the capabilities of ChatGPT. To properly assess the progress in evaluating the chatbot, it is necessary to put the evaluated tasks in order. For this purpose, the taxonomy of the natural processing tasks must be established. There are two main approaches to

establishing such a taxonomy. First — relates the tasks directly to the methods used for solving them [31]. While this approach allows for the systematic organization of most tasks, it is not very useful for this paper as the goal is to establish how many tasks can be performed by the same chatbot. The second approach is to organize the tasks first into tasks of analysis and generation and then to divide the first ones into the levels of syntactic, semantic, and pragmatic analysis [32]. Looking at the field through the lens of this taxonomy, the main areas that ChatGPT has been tested so far are generation tasks.

The studies published within few months after the release of ChatGPT focused mostly on one pre-selected task, mainly on question answering (e.g. [25,26,28,33–35]) or summarizing (e.g. [30,36–39]). However, such tasks as humor identification and generation [40], machine translation [41], sentiment recognition [42], paraphrasing [38], and other text generation subtasks were also analyzed [43–45]. In most cases, the evaluation was conducted manually. This concerned, in particular open-ended question answering (e.g. [26,34,35]) and scientific texts summarization (e.g. [37,39]). This was related to the fact that benchmark datasets did not appear in many studies. If they were included, they were often treated as a basis for manual expert analysis of the ChatGPT answers, e.g. in the case of medical education [26,34]. Another issue connected with the dominant approach concerns the comparison of the NLP toolkits and their performance in solving NLP tasks. Relatively few studies analyzed the differences between diverse toolkits and systems. In cases where the performance of ChatGPT was compared to other solutions (e.g. [26,30,41]), it worked on a comparable level to the competitor but not outperforming any major SOTA solutions.

The most recent studies show a more broad and nuanced approach to evaluating Chat GPT. There were a couple of large-scale evaluations published recently [46,47] and both of them indicate that while the performance of ChatGPT is significant it does not outperform the SOTA solutions with the exception of sentiment analysis task in case of [46] which suggests it is a significant area for further research. Both articles give an interesting insight into the strengths and weaknesses of the Chat performance observed during the research. Two more recently published papers that present usage for GPT models in sentiment analysis for Italian [48] and Arabic [49] languages which further shows emotion-related tasks as a particularly interesting area of NLP tasks to examine when it comes to GPT models.

Simultaneously there were three major studies published that focus on evaluating language models, in general, [12,50]. Rather than comparing the performance of the language models to the other solutions, these studies focus on comparing language models with each other.

They give a great insight into the capabilities and limitations of many language models and show the progress that LLMs made over the years. Many of the evaluated models are predecessors of the ChatGPT model which makes those studies particularly valuable as a lens through which the capabilities of ChatGPT can be viewed.

The recent increased popularity of large language models such as ChatGPT has brought more attention to the more nuanced aspects of NLP such as AI Ethics. Evaluation of such elements as the bias of the model or its toxicity requires a different approach than the evaluation of the ability of the model to perform particular NLP tasks. Such research has already been conducted to some extent and multiple interesting findings about ethical aspects of ChatGPT's performance in comparison to different large language models can be found in [51].

Another area of performance that was recently evaluated was the robustness of ChatGPT [52]. The research indicates that while ChatGPT is still prone to adversarial attacks it is significantly more robust than its predecessors and other currently available models. Finally, after GPT-4 model was released it quickly showed significant advancement in its capabilities in comparison to previous GPT models [53–55]. Most notably it was identified by Michał Kosiński as capable of solving Theory of the Mind tasks [56] which also indicates that sentiment analysis and emotion processing are particularly significant areas of research when it comes to the newest GPT models.

There are many ways to carry out prompting with ChatGPT. Although the popular trial-and-error method may seem good, utilizing techniques with proven effectiveness is crucial. The model usually understands many ways in which a question might be asked. However, there are also instances where an explanation must be included to receive a proper answer from ChatGPT. In Natural Language Processing, there are multiple interesting prompting methods, many of which are collected and clearly outlined in [57].

3. Research question

As existing evaluations of ChatGPT focus on its ability to generate language utterances, we want to investigate its analytical skills, particularly in tasks requiring language analysis and understanding, i.e., typical NLP problems examined by science and companies. Therefore, we aim to target two abilities (task categories; see Table 1): *semantic* and *pragmatic*. Distinguishing semantics from pragmatics, we refer to the classic concept of Morris, who proposed syntactic, semantic, and pragmatic dimensions and levels of semiosis [58]. He states that “semantics deals with the relation of signs to their designate” [58, p. 21], while pragmatics refers to “the science of the relation of signs to their interpreters” [58, p. 30]. This idea has found its application in contemporary pragmatics “is the study of linguistic communication in context: the choices users of language make and the process of meaning-making in social interaction” [59]. The former kind of task entails recognition of text properties (like word sense description or a speaker's stance polarity in a language construction) or mining information that is directly expressed in a text fragment, e.g., various relations between sentences and text fragments, or extraction of the answer to a question). In the pragmatic analysis, we dig into ChatGPT's potential in exploiting general knowledge stored in the model to solve the tasks beyond the literal semantic content of the textual prompt — input. Here, we investigate a range of different pragmatic problems with a common denominator of the necessity to predict the influence of the utterance interpretation on the reader and their often subjective content perception. We asked ChatGPT to predict not only sentiment polarity and emotions evoked in the reader but also humor and offensiveness. Several of these tasks are also stated in a personalized version, in which the outcome depends on a particular reader (interlocutor). Overall, the tasks considered in this paper have relatively structured and simple

expected results reflecting typical machine learning solutions, i.e., various types of classification.³ This, in turn, directly corresponds to the analytical approach: further numerical processing of the outcome. For example, one might want to know how well ChatGPT would perform in evaluating customers' sentiment toward a particular product based on an analysis of multiple online reviews. This requires obtaining accurate polarity (classification) of individual texts assessed by ChatGPT and aggregating decisions to acquire the final ratio of positive and negative opinions.

In all cases, we are interested in the correctness of ChatGPT's analysis and inference, i.e., different forms of understanding of the natural language utterances, while intentionally neglecting the aspect of the quality of the generative results as perceived by the user, as opposed to alternative studies. This means that we do not attempt to quantify how well the user perceives the output text, i.e., the style of generated text or how rich the content is. It has little or no relevance to a reliable evaluation of analytical tasks.

Does ChatGPT perform as well as the best recent models (SOTA) in solving typical NLP analytical tasks?

4. Tasks

We tested ChatGPT on 25 tasks focusing on solving common NLP problems and requiring analytical reasoning, Table 1. These tasks include (1) a relatively simple binary classification of texts like spam, humor, sarcasm, aggression detection, or grammatical correctness of the text; (2) a more complex multiclass and multi-label classification of texts such as sentiment analysis, emotion recognition; (3) reasoning with the personal context, i.e., personalized versions of the problems that make use of additional information about text perception of a given user (user's examples provided to ChatGPT); (4) semantic annotation and acceptance of the text going towards natural language understanding (NLU) like word sense disambiguation (WSD), and (5) answering questions based on the input text.

The tasks were divided into two categories described in Section 3: semantic and pragmatic. The latter requires the model to utilize additional knowledge that is not directly captured by distributional semantics [88]. For personalized tasks, the input texts have to be extended with additional personal context (personalized solutions of the problem [20]); see Section 6.3. These tasks involve the datasets such as Aggression → AggressionPer, GoEmo → GoEmoPer, and Unhealthy → UnhealthyPer.

Most of the tasks were based on public datasets investigated in the literature. However, we also utilized a collection of new unpublished datasets such as (ClarInEmo), which ChatGPT could not have indexed. Most of the evaluated texts were written in English (23, 92% of the tasks), while two others (8%) were in Polish. The prompts were in line with the language of the input text.

We manually evaluated the probability that a given annotated dataset was available and used by ChatGPT for training. We assigned a rating of highly probable (3) to most of the datasets in this evaluation. Still, for their personalized versions, the rating was reduced to (2) since ChatGPT was almost certainly not trained in personalized settings. In the case of PolEmo — the dataset was unlikely to be used for training and received a score of (1). Finally, we assigned a score (0) to the unpublished version of the ClarInEmo dataset. Additionally, we asked ChatGPT whether or not the dataset was used for training. Based on collected data, we performed appropriate analyses, Section 6.5.

Due to the scale of our test data and the limitations of ChatGPT's API, we had to limit the number of input texts. This means that for some tasks, we randomly selected a sample of texts (column #Used) in Table 1) from all available instances in the test or dev set (column #Test).

³ In some question answering tasks, the output is given in few words (SQuAD) or as a number — the result of mathematical calculations (MathQA).

Table 1
Profile of the tested NLP tasks named according to their resource (dataset).

ID	Task name (resource-based)	Category	Language	NLP problem	Context	Reasoning type	Dataset/SOTA	Availability	Trained	#Test	#Used	#None	#Post-processed	#N	#Classes	#Majority/minority class
1	Aggression	P	EN	Offensiveness detection	No	Binary classification	WikiDetox Aggr. [60]/[61]	3	Yes	23153	1000	13	151 (15.1%)	987	2	19823 /3330
2	AggressionPer	P	EN	Offensiveness det.: personalized	Yes	Binary classification	WikiDetox Aggr. [60]/[21]	2	No	349582	1000	19	92 (9.2%)	981	2	282918 /66664
3	CoLa	S	EN	Linguistic acceptability	No	Binary classification	CoLa [62]/[63]	3	Yes	1042	1042	0	0 (0%)	1042	2	721 /322
4	ColBERT	P	EN	Humor recognition	No	Binary classification	ColBERT [64]/[64]	2	No	40000	1000	5	93 (9.3%)	995	2	20137 /19643
5	Sarcasm	P	EN	Humor recognition	No	Binary classification	Sarcasmania [65]/[66]	3	Yes	5967	1000	10	61 (6.1%)	990	2	3051 /2916
6	Spam	P	EN	Spam detection	No	Binary classification	SMS Spam v.1 [67]/[68]	3	Yes	1115	1115	3	14 (1.3%)	1112	2	966 /149
7	WordContext	S	EN	Word sense disambiguation	Yes	Binary pair classification	WiC [69]/[70]	3	No	638	638	0	5 (0.8%)	638	2	319 /319
8	TextEntail	S	EN	Natural language inference	No	Binary sentence pair classification	RTE [71]/[70]	3	Yes	277	277	0	0 (0%)	277	2	146 /131
9	WNLI	S	EN	Natural language inference	No	Binary sentence pair classification	WNLI [72]/[73]	3	Yes	71	71	0	0 (0%)	71	2	40/31
10	SQuAD	S	EN	Question answering	Yes	Extractive QA	SQuAD v2 [74]/[75]	3	Yes	11873	1000	0	247 (24.7%)	1000	-	-
11	MathQA	S	EN	Question answering	No	Mathematical reasoning	GSM8K [76]/[77]	3	Yes	1319	1000	0	1 (0.1%)	999	-	-
12	ClarinEmo	P	PL	Emotion recognition	No	Multi-label classification	ClarinEmo -/-	0	No	1264	1264	0	9 (0.7%)	1264	11	624/59
13	GoEmo	P	EN	Emotion recognition	No	Multi-label classification	GoEmotions [78]/[79]	3	No	5427	1000	18	87 (8.7%)	1000	28	1787/6
14	GoEmoPer0	P	EN	Emotion rec.: personalized	No	Multi-label classification	GoEmotions [78]/[79]	2	No	19470	1151	28	1 (0.1%)	1123	28	288/6
15	GoEmoPer1	P	EN	Emotion rec.: personalized	Yes	Multi-label classification	GoEmotions [78]/[79]	2	No	19470	1151	11	0 (0%)	1140	28	288/6
16	GoEmoPer2	P	EN	Emotion rec.: personalized	Yes	Multi-label classification	GoEmotions [78]/[79]	2	No	19470	1151	8	0 (0%)	1143	28	288/6
17	GoEmoPer3	P	EN	Emotion rec.: personalized	Yes	Multi-label classification	GoEmotions [78]/[79]	2	No	19470	1151	10	0 (0%)	1141	28	288/6
18	Unhealthy	P	EN	Offensiveness detection	No	Multi-label classification	Unhealthy Conv. [80]/[80]	3	No	44354	1000	22	348 (34.8%)	963	8	936/25
19	UnhealthyPer	P	EN	Offensiveness det.: personalized	Yes	Multi-label classification	Unhealthy Conv. [80]/[20]	2	No	227975	1000	9	15 (1.5%)	991	8	782/30
20	PolEmo	P	PL	Sentiment analysis	No	Multiclass classification	PolEmo2 [81]/[81]	1	No	820	820	3	23 (2.8%)	817	4	339 /118
21	TweetEmoji	P	EN	Emoji prediction	No	Multiclass classification	TweetEval [82]/[83]	2	No	50000	1666	2	0 (0%)	1664	20	10798 /1010
22	TweetSent	P	EN	Sentiment analysis	No	Multiclass classification	TweetEval [82]/[83]	2	No	12283	5143	0	245 (4.8%)	5143	3	5937 /2375
23	TweetStance	S	EN	Stance detection	No	Multiclass classification	TweetEval [82]/[83]	2	No	1249	1249	7	99 (7.9%)	1249	3	715 /230
24	ReAding	S	EN	Question answering	Yes	Multiple choice QA	RACE [84]/[85]	3	Yes	4887	1000	4	206 (20.6%)	996	4	-
25	WSD	S	EN	Word sense disambiguation	Yes	Sequence labeling	Raganato [86]/[87]	3	Yes	7253	7253	5	176 (2.4%)	7253	61	-

Category: S - semantic, P - pragmatic; *Context* refers to either additional contextual information added to prompts (e.g. related to a given user – personalization) or to the context directly considered in the task; *Availability*: our assessment of whether ChatGPT used the dataset for fine-tuning: 3 - highly probable, 2 - probable, 1 - rather no; 0 - impossible. *Trained*: ChatGPT answers if it used the dataset for training. *#Test*: no. of cases available in the test or dev set. *#Used*: no. of cases from the test or dev set (prompts) used by us. *#None*: no. of prompts ChatGPT returned 'none'. *#Post-processed*: no. of prompts requiring manual post-processing. *#N*: no. of valid prompts used for quality evaluation (Table 2). *#Classes*: no. of distinct classes in the output. *#Majority/minority class*: the number of examples for the majority/minority classes in the test or dev set (*#Test*).

Table 2

Quantitative analysis. Values of quality measures obtained for (a) the ChatGPT output, (b) SOTA, i.e., our launch of the best available model, or if not possible, taken from the paper. *Difference*: $(b - a)$. *Difficulty*: $(100\% - b)$. *Loss*: $100\% \cdot (b - a) \div b$. Emotion tasks marked with an asterisk: 12–17, 20–21. Tasks without emotions discard eight emotion-related tasks.

ID	Task name (resource-based)	Task category	Measure type	SOTA type	ChatGPT (a) [%]	SOTA (b) [%]	Difference (b-a) [pp]	Difficulty [%]	Loss [%]
1	Aggression	Pragmatic	F1 Macro	Our	69.10	74.45	5.35	25.55	7.19
2	AggressionPer	Pragmatic	F1 Macro	Our	72.57	81.03	8.46	19.97	10.44
3	CoLa	Semantic	Accuracy	Paper	80.82	86.40	5.58	13.60	6.46
4	ColBERT	Pragmatic	F1 Macro	Our	86.47	98.50	12.03	1.50	12.21
5	Sarcasm	Pragmatic	F1 Macro	Our	49.88	53.57	3.69	46.43	6.89
6	Spam	Pragmatic	F1 Macro	Our	82.67	99.42	16.75	0.58	16.85
7	WordContext	Semantic	Accuracy	Paper	64.58	74.00	9.42	26.00	12.73
8	TextEntail	Semantic	F1 Macro	Paper	88.09	92.10	4.01	7.90	4.35
9	WNLI	Semantic	Accuracy	Paper	81.69	97.90	16.21	2.10	16.56
10	SQuAD	Semantic	F1 Macro	Paper	69.21	90.75	21.54	9.25	23.74
11	MathQA	Semantic	Accuracy	Paper	71.40	83.20	11.80	16.80	14.18
12	*ClarinEmo	Pragmatic	F1 Macro	Our	53.23	68.04	14.81	31.96	21.77
13	*GoEmo	Pragmatic	F1 Macro	Our	25.55	52.75	27.20	47.25	51.56
14	*GoEmoPer0	Pragmatic	F1 Macro	Paper	23.74	54.50	30.76	45.50	56.44
15	*GoEmoPer1	Pragmatic	F1 Macro	Paper	19.00	66.10	47.10	33.90	71.26
16	*GoEmoPer2	Pragmatic	F1 Macro	Paper	20.34	66.10	45.76	33.90	69.23
17	*GoEmoPer3	Pragmatic	F1 Macro	Paper	23.41	66.10	42.69	33.90	64.58
18	Unhealthy	Pragmatic	F1 Macro	Our	45.21	50.96	5.75	49.04	11.28
19	UnhealthyPer	Pragmatic	F1 Macro	Our	54.02	70.92	16.90	29.08	23.83
20	*PolEmo	Pragmatic	F1 Macro	Our	44.08	76.44	32.36	23.56	42.33
21	*TweetEmoji	Pragmatic	F1 Macro	Our	18.19	32.20	14.01	67.80	43.51
22	TweetSent	Pragmatic	F1 Macro	Our	63.32	72.07	8.75	27.93	12.14
23	TweetStance	Semantic	F1 Macro	Our	56.44	67.42	10.98	32.58	16.29
24	ReAding	Semantic	F1 Macro	Our	76.36	84.71	8.35	15.29	9.86
25	WSD	Semantic	F1 Macro	Paper	73.30	83.20	9.90	16.80	11.90
All tasks			Average		56.51	73.71	17.21	26.29	25.50
			std. dev.		± 23.31	± 16.74	± 13.08	± 16.74	± 21.44
Only tasks without emotions			Average		69.71	80.04	10.32	19.96	12.76
			std. dev.		± 12.76	± 14.36	± 5.08	± 14.36	± 5.49
*Only emotion tasks			Average		28.44	60.28	31.84	39.72	52.59
			std. dev.		± 18.76	± 14.87	± 13.84	± 14.87	± 20.10
Only pragmatic tasks			Average		46.92	67.70	20.77	32.30	32.59
			std. dev.		± 23.42	± 17.18	± 14.86	± 17.18	± 23.85
Only semantic tasks			Average		73.54	84.41	10.87	15.59	12.90
			std. dev.		± 9.59	± 9.26	± 5.33	± 9.26	± 5.80

In some cases, the outputs from ChatGPT required a manual post-processing procedure (column *#Post-processing*), and some responses were out of the desired domain (column *#None*).

To compare the performance of ChatGPT with SOTA methods, we trained and tested the best available models (or close to the best) by reusing the source code provided with references (column *SOTA* in Table 1). In other cases, we exploited the values of reported quality metrics published in original papers; see column *SOTA* in Table 2. Examples of chats for all the tasks included in our study are available in Appendix B.

1. Aggression. We used the Wikipedia Talk Labels: Aggression dataset [60] collected in the *Wikipedia Detox* project. It includes over 100k comments acquired from the English Wikipedia with binary annotations from multiple Crowdfunder workers regarding the aggressiveness of each text. In the non-personalized variant of the dataset, each text is associated with a single annotation obtained via majority voting.

2. AggressionPer. We have also used the personalized variant of the Aggression dataset. In this case, we represented the individual's perspective by providing three user-specific annotations as an addition to the standard input prompt. These additional texts were selected according to their highest controversy, i.e., with the highest standard deviation among the annotator votes. It was inspired by the findings from [21].

3. CoLa. The Corpus of Linguistic Acceptability [62] consists of 10 657 sentences from 23 linguistics publications, annotated for acceptability (grammaticality). Here, ChatGPT had to classify whether a sentence was grammatically correct. It was confronted with the metrics from existing work on Few-Shot Learners [63].

4. ColBERT. The ColBERT dataset [64] contains 200k short texts acquired from news, headlines, Wikipedia, tweets, and jokes. Each sample is annotated as *funny* or *not-funny*. The distribution of labels is uniform.

5. Sarcasm. The Sarcasmania dataset [65] consists of 39,780 texts from the Twitter platform. Each tweet is associated with one of the two classes: *sarcastic* or *non-sarcastic*.

6. Spam. SMS Spam Collection v.1 [67] is a dataset containing SMS contents labeled as *spam* or not. Here, ChatGPT had to classify an input text accordingly.

7. WordContext. The task of identifying the intended meaning of a word in a given context — Word in Context task (WIC) [69]. The WIC task is strongly related to the Word Sense Disambiguation task (WSD) as it tests language models' sense understanding abilities. Contrary to WSD, the task is framed as binary classification, testing if two independent contexts express the same meaning of the highlighted word.

8. TextEntail. One of the SuperGLUE benchmark [71] tasks is called Recognizing Textual Entailment (RTE). This dataset comes from a collection of annual competitions on textual entailment. Given two text fragments, the model has to decide whether the meaning of one text is entailed (logically related) to another. The task is formulated as a two-class classification problem. ChatGPT had to decide if the two sentences were "entailed" or "not entailed".

9. WNLI. SuperGLUE Winograd NLI dataset comes from the GLUE benchmark [72]. Initially, this task was inspired by the Winograd Schema Challenge [89] in which a model must read a sentence with a pronoun and select the referent of that pronoun from a list of choices. For the WNLI dataset, the original data was converted to the

sentence pair classification problem. The second sentence in a pair was created by replacing the ambiguous pronoun with each possible referent. ChatGPT has to predict whether texts are entailed with each other (“1” label) or not (“0” label).

10. SQuAD. SQuAD v₂ [74] is a question-answering dataset, which combines 100,000 examples from SQuAD1.1 with over 50,000 unanswerable questions looking similar to real ones. Each question consists of the context, textual answer, and number referring to the location in the context where the answer can be found. To perform well on the dataset, any given system must be able to answer the questions and infer whether the answer can be found in the given context.

11. MathQA. The multi-step mathematical reasoning dataset GSM8K [76] - MathQA contains grade school level maths word problems (MWP) that require only basic arithmetic operations. It was designed to test large language models with auxiliary chain-of-thought reasoning data. It was shown that the dataset is challenging for even the largest generative models.

12. ClarinEmo. It is an original dataset consisting of 1,110 texts in Polish — various opinions have been hand-annotated with three sentiment polarizations and eight emotions describing the author’s intention. The annotations of six independent annotators were aggregated to label each sentence with all potential options, using the label when at least two annotators agreed on it. It is our new dataset that has not yet been published. We exploited this dataset to ensure that ChatGPT was not trained on it.

13. GoEmo. The GoEmotions dataset [78] consists of 58k carefully selected Reddit comments from popular English subreddits labeled according to a 27 + 1 schema, i.e. 27 possible emotion categories plus neutral. ChatGPT is ordered to determine the emotions of provided text from the list of available 28 categories. To additionally guide ChatGPT, we request it to provide a specific number of emotions that matches the number of emotions annotated as ground truth.

14.–17. GoEmoPer. To investigate ChatGPT’s performance in Personalized Emotion Recognition, we obtained individual annotator annotations from raw GoEmotions data. ChatGPT is requested to predict emotions assigned to provided text by a selected annotator. We analyze ChatGPT performance in four different scenarios: **GoEmoPer0**, **GoEmoPer1**, **GoEmoPer2**, **GoEmoPer3**. ChatGPT is not given any information about the annotator in the prior experiment. In the following scenarios, we provide an additionally predefined number of texts annotated by this annotator. The goal is to provide ChatGPT with a context that will help it learn the personal preferences of the annotator. We start with a context consisting of one text and gradually increase the number to three.

18. Unhealthy. Unhealthy Conversation [80] is a dataset of 44,000 comments of 250 characters or fewer, annotated by 588 crowd workers. Each comment was annotated as healthy or unhealthy. Additionally, each comment could be annotated with one of the following attributes: antagonistic, hostile, dismissive, condescending, sarcastic, generalization, or unfair generalization.

19. UnhealthyPer. This is the personalized version of Unhealthy Conversations. The dataset texts and annotations are identical to the non-personalized Unhealthy Conversations version. The only difference is that the personalized UserID model [20] is used instead of the standard transformer model.

20. PolEmo. PolEmo 2.0 [81] is a corpus of Polish consumer reviews from four domains: medicine, hotels, products, and school. Each text was manually annotated with the sentiment using one of the following labels: positive, neutral, negative, or ambivalent.

21. TweetEmoji. This is one of the seven heterogeneous tasks from the Tweeteval dataset [82]. It focuses on emoji prediction for a given tweet. There are twenty available emojis, and ChatGPT is asked to provide a list of three emojis, which could be added at **To**. the end of a given tweet ranges from the most probable to the least. To calculate metrics such as F1 or accuracy, the first emoji on the list was assumed to be ChatGPT’s answer.

22. TweetSent. TweetSent, another task from the Tweeteval [82] dataset, involves determining the **sentiment** expressed in a Tweet. In our work, ChatGPT is tasked to identify the sentiment of a given text, categorizing it as negative, neutral, or positive.

23. TweetStance. TweetStance is one more task from the Tweeteval [82] dataset that focuses on detecting stances in Tweets in five different areas: abortion, atheism, climate change, feminism, and Hillary Clinton. Each text was labeled as *none*, *against*, *favor*.

24. ReAding. RACE dataset [84] is a reading comprehension dataset consisting of over 100,000 multiple-choice questions relating to about 28,000 passages from various topics. It was created using English examinations in China for middle and high school students. Each question has four possible answers labeled *A*, *B*, *C*, *D*, with only one answer correct.

25. WSD. It is a unified evaluation framework for word sense disambiguation proposed in [86]. The framework consists of five evaluation datasets with standard English texts from Senseval [90,91] and Semeval [92–94] competitions. Texts were annotated with meanings (senses) from Princeton WordNet 3.0 (PWN) sense inventory [95] containing 117,664 synsets (sets of synonymous senses). The framework has been used as a standard evaluation environment for knowledge-based, weakly supervised, and supervised word sense disambiguation models. The overall collection of datasets contains 7,253 classification instances — sense annotations. The number of senses depends on the disambiguated word and varies from 2 candidate senses to more than 60 — mainly for polysemous verbs. On average, the models must choose only one sense from 5.24 candidate senses for each word. The dataset also contains a subset of instances where words are monosemous and have only one meaning concerning PWN. Such cases do not require any disambiguation, so all post-processing decisions were made in favor of the ChatGPT model. To evaluate ChatGPT’s sense recognition abilities, we adopted sense glosses from PWN⁴ as they are often used as the basis for training supervised word sense disambiguation models. The glosses briefly summarize the meanings of senses using natural language. We used the glosses to explain meanings to the model when disambiguating the words in a given context. Using the glosses to explain senses to a language model implicitly tests its language comprehension abilities.

5. Research methodology

Our research focused on three main steps depicted in Fig. 3. Having quality measures for both reference models and ChatGPT, we were able to confront them with one another to answer our main research question: is ChatGPT a good jack of all trades?

5.1. Prompt generation

Prompt generation consists of three goals that we want to achieve. The key idea is to solve a particular natural language processing task, like sentiment analysis or emotion detection, using ChatGPT. Additionally, we must force ChatGPT to answer with a specified value from a list of annotations used in the chosen task/dataset and an easy-to-process format, like a Python list or single integer.

All of the above can be achieved by using various schemas of prompts. The general chat schema looks like the following Chat 1:

```
Chat CHAT_ID. Task: TASK_NAME. Case EXAMPLE_ID. E.g.:
Chat 1. Task: Aggression. Case 3.
```

```
Prompt //our input to ChatGPT
```

```
INSTRUCTION //task description, e.g.:
```

⁴ <https://wordnetcode.princeton.edu/glosstag.shtml>

ChatGPT evaluation flow diagram

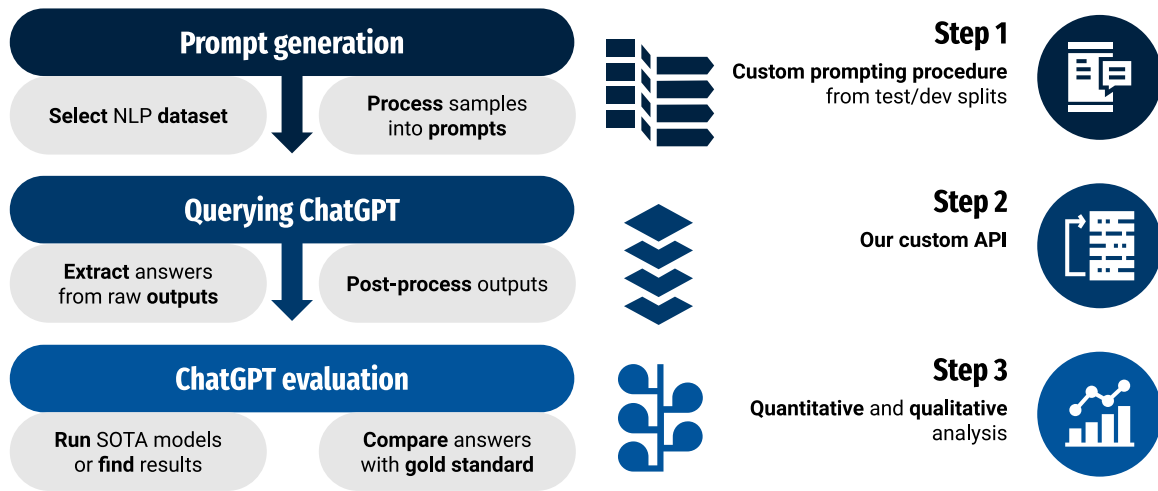


Fig. 3. ChatGPT evaluation flow diagram showing the three stages of data processing: (1) selecting a dataset and converting the test set to prompt-based form; (2) querying (prompting) the ChatGPT service using our custom reverse-engineered API; (3) extracting labels from raw outputs and evaluating using ground truth and comparing the results with SOTA models or SOTA results from papers.

Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.
TEXT //input text, e.g.: Text: (Or should I follow your example and delete things I don't like from other people's talk pages ?)
ChatGPT answer //raw output
["non-aggressive"]
Extracted answer //processed output
non-aggressive
Expected answer //expected output
non-aggressive
Evaluation result //additional judgement
Label: OK, ChatGPT answer: OK

Case number is the example ID for the following task in *ChatGPT Evaluation v2.0.xlsx* file available in our GitHub repository⁵.

There are multiple options when creating prompt schemas. For example, we can add sentiment label mappings to integers, forcing ChatGPT to answer with only integers. We can further specify ChatGPT output format by adding allowed values again after Text input. Moreover, we provided additional user annotations describing their perspective in the case of personalized tasks. The example prompts for each task are presented in Appendix B. The generated prompts were used as questions in a ChatGPT conversation. It is worth noting that we did not force the API to create a new conversation window per prompt. Consequently, multiple texts were allocated across multiple conversations within the specified ChatGPT limitations.

5.2. Post-processing

Raw text provided by ChatGPT is different from the final version achieved after post-processing. Some answers are returned as whole

sentences instead of requested predefined lists. This imposes a necessity to check what happened and extract answers from ChatGPT output manually. The next step is to cast the resulting outputs to the correct labels in the dataset. For example, if ChatGPT returned a sentiment with the typo “negaitiv”, we mapped it to “negative”, assuming that this was the intended answer. Sometimes the model returns values out of the requested list. For example, given the possible 28 emotions in emotion recognition, ChatGPT returned the unmentioned “determination”. Such cases were converted to a value of “none”, which was not considered in the performance evaluation (column #None in Table 1, plus 3k additional prompts used in Section 6.4).

Overall, the number of cases that required post-processing was relatively small (column #Post-processed in Table 1). For most tasks (16), the contribution of such texts was less than 5%. Only for Aggression, SQuAD, Unhealthy, and ReAding, it exceeded 15%.

5.3. Experimental setup

Without an official API, we modified and used an unofficial API called PyGPT⁶, written in Python. During the research, we exploited up to 20 accounts to gather data regarding 25 datasets.

Every dataset was first assigned to a different task manager who independently prepared appropriate prompts based on the dataset texts and the output structure. Next, our API managers ran parallel processes to query prompts and acquire the raw ChatGPT output in a shared sheet *ChatGPT Evaluation v2.0.xlsx*⁷.

In total, over 38,000 prompts were exploited⁸.

Post-processing procedures (Section 5.2) were applied afterward, along with quality measure computation (Section 5.4) and in-depth analyses.

5.4. Performance measures

If possible, we launched our models equivalent to SOTA solutions since the setup (especially data split) was often different than in the

⁶ <https://github.com/PawanOsman/PyGPT>

⁷ <https://github.com/CLARIN-PL/chatgpt-evaluation-01-2023>

⁸ 35,142 is the sum of column #Used in Table 1, plus 3k additional prompts used in Section 6.4, and some in Section 7.3.

⁵ <https://github.com/CLARIN-PL/chatgpt-evaluation-01-2023>

original paper. For that purpose, we usually utilized source codes published by the authors. Unfortunately, it was impossible for some tasks, so we exploited the performance results provided in the original paper. If available, we tried to validate ChatGPT using one measure — F1 Macro, which is commonly acceptable for imbalanced data, Table 2. F1 Macro in multi-label classification is an average of harmonic means between precision and recall calculated per label. If Q is the number of labels, p_i and r_i are the precision and the recall calculated for i th label, F1 Macro is given by equation:

$$F1_{macro} = \frac{1}{Q} \sum_{i=1}^Q \frac{2 \cdot p_i \cdot r_i}{p_i + r_i}$$

In the case of CoLa, WNLI, WordContext, and MathQA, we had to rely on the accuracy, as it was the only one presented in the reference paper; we could not replicate their studies and calculate our measures. WNLI and WordContext have their two classes balanced, so it is not an issue.

Only the post-processed and cleaned cases (column #N in Table 2) were considered in the quantitative analysis. Other metric values are presented in Appendix A, Table 9.

Having calculated the SOTA and ChatGPT results, we were able to compute Loss that reflects how much ChatGPT is worse than the best-dedicated methods, as follows:

$$Loss = \frac{100\% \cdot (SOTA - ChatGPT)}{SOTA}$$

Loss measure was exploited in Table 2, Figs. 5, 7, 8, 10, 11, and 16.

Yet another measure is utilized in Fig. 9: Gain. It quantifies which part of the entire possible improvement of the performance of the reference non-personalized method was reached by a given personalized in-context solution:

$$Gain = \frac{100\% \cdot (Per - NonPer)}{100\% - NonPer}$$

where *Per* is the F1 result provided by our personalized in-context processing; *NonPer* is F1 delivered by the reference, non-personalized model.

6. Quantitative analysis

6.1. Jack of all trades, master of none

We tested ChatGPT on 25 NLP tasks listed in Table 1 by computing appropriate quality measures both for ChatGPT and the best recently available models (SOTA), Fig. 5. The ChatGPT performance is depicted in Fig. 4. It is usually greater for semantic tasks rather than for pragmatic ones, which is related to the task difficulty, see Section 6.2.

We also estimated the loss of ChatGPT compared to the SOTA solution, Section 5.4. The loss indicates how worse ChatGPT is relative to SOTA, which is considered 100% capacity, Table 2, Fig. 5. The crucial finding from our studies is that the ChatGPT performance is always lower than the SOTA methods (loss > 0) in all the tasks considered. It means that ChatGPT never reached the level of the best existing models. However, its loss was greater or lesser depending on the problem. The average quality of SOTA methods was at 73.7%, whereas ChatGPT was at only 56.5%. Simultaneously, ChatGPT was less stable: the standard deviation of its performance was 23.3% compared to only 16.7% for the SOTA solutions.

The loss for most tasks did not exceed 25%. It was greater only for three problems: GoEmotions, PolEmo, and TweetEmoji. All these tasks are related to a very subjective problem of emotional perception and individual interpretation of the content. Also, for the last emotional task — ClarinEmo, the loss was 21.8%. If we discard all eight emotion-related tasks (ids: 12–17, 20–21), the average SOTA performance reaches 80% (increase by 6.3pp), but ChatGPT improves much more: by 13.2pp, up to 69.7%. In such a case, the average loss is reduced by as much as half, from 25.5% to 12.8%; the difference in performance drops from 17.2pp to 10.3pp.

We know that a direct comparison of performance between different tasks does not always rightly show the difficulty of the tasks being compared. A small increase in the evaluation score in one task might be more challenging to overcome than a larger increase in another task. Moreover, simple solutions, such as majority class voting or a simple lexical similarity function, often appear to be a strong baseline for complex neural architectures. For example, an increase of 10pp in WSD or WordContext tasks might be more challenging to obtain, and the most outstanding solutions are far from 100% performance. Furthermore, the best unsupervised or weakly-supervised solutions obtain a 70% performance of F1-score in the WSD task, and their architectures have significantly fewer parameters than the ChatGPT model.

Nevertheless, we can state that ChatGPT performs pretty well on all tasks except emotional ones. Simultaneously, its achievements are always below SOTA but usually not so much. Such results prove that ChatGPT is *Jack of all trades, master of none*.

6.2. Task difficulty vs. ChatGPT performance

Task difficulty is defined as $(100\% - SOTA_{performance})$. In other words, we assume that difficulty is reflected by the level of the best recent models' performance, i.e., the closer the SOTA performance to 100%, the easier (less difficult) the task. The difficulty of each task is presented in Table 2 and Fig. 6. In general, pragmatic tasks are more difficult (average difficulty = 32.3%), while the average difficulty for semantic tasks is only 15.6%. It comes especially from the emotional tasks, which are pragmatic and very difficult (average 39.7%).

We can also observe that the loss is correlated with the task difficulty; see Fig. 7. The Pearson correlation coefficient between difficulty and loss is equal to 0.46. It is observable that semantic tasks (blue crosses) are rather easy; hence, their ChatGPT loss is relatively small into the Q3 quadrant: easy task, low losses. A stronger dependence: greater difficulty, the higher loss can be seen for pragmatic tasks dominated by emotion-related problems, Fig. 8.

This analysis, however, requires further investigations since the number of the tasks considered (25) still remains relatively small.

6.3. Random contextual few-shot personalization

As a concept of contextual and human-centered processing, personalization in NLP was proposed by us and recently extensively explored in [20–22,96–100]. Here, we extend it to ChatGPT prompts as *personalized in-context processing*. This is somewhat similar to in-context learning with demonstrations [101]. However, in the case of personalized tasks, the user preferences are difficult to capture with a user context consisting of only up to three past annotations of this user.

It is important to design a tailor-made architecture for generating user representation to address this. On the other hand, the embedding of a person should describe the similarity or peculiarity of their perspective compared to others. During our experiments, we observed higher loss values for the ChatGPT model compared to the SOTA models in the case of the AggressionPer and UnhealthyPer datasets: 3.25 and 12.55 percentage points, respectively. On the other hand, enriching the user context with more annotations resulted in 4.08 percentage points better ChatGPT accuracy for GoEmoPer3 compared to GoEmoPer0. The percentage gains between the context-based setup and the baseline are presented in Fig. 9.

Demonstration-based personalization included in our prompts can be treated as similar to few-shot learning, even though ChatGPT does not update its model after every prompt. Therefore, we would prefer to call it a *few-shot evaluation* or *personalized in-context processing*.

Moreover, we also evaluated the non-personalized in-context processing semantic tasks: (1) WordContext, (2) SQuAD, (3) ReAding, and (4) WSD. In this case, the ChatGPT loss values were relatively small and ranged between 9.9% for ReAding and 12.7% for WordContext. While solving mathematical calculations (SQuAD), the highest loss was among semantic tasks: 23.7%.

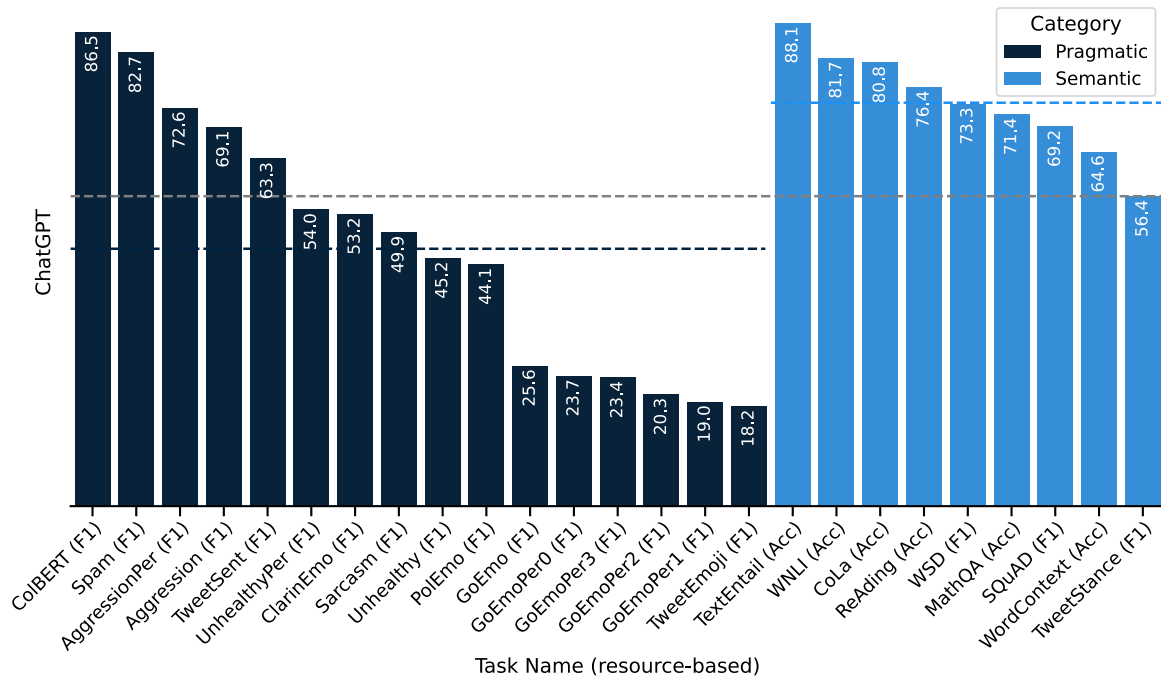


Fig. 4. ChatGPT performance (%) for all tasks considered and named according to their resource (dataset). Dashed lines denote the average performance for only semantic, all, and only pragmatic tasks.

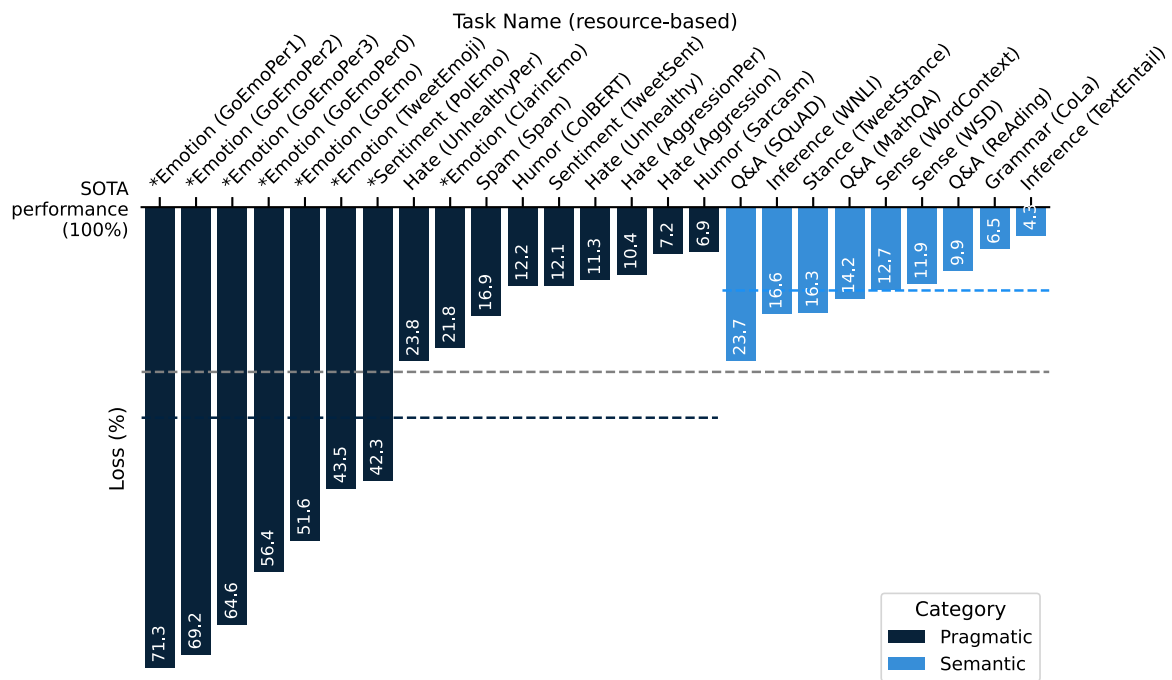


Fig. 5. The ChatGPT loss in performance (%) for all tasks considered and named according to their resource (dataset), descending ordered by loss value. Tasks preceded by an asterisk are related to emotions. The upper X axis corresponds to the performance of the best model (SOTA) treated as 100% capabilities. Dashed lines denote the average loss values for only pragmatic, only semantic, and all tasks.

6.4. Impact of the context

One of the many features of ChatGPT is its ability to reference previous messages within the conversation. We wonder whether ChatGPT treats all previous messages as an extended context to a given prompt. If so, ChatGPT may not recognize properly that an unanswerable question does not have an answer. As a result, it may wrongly treat the previous prompts as a valuable context and response based on

them rather than refuse any response. To test this ChatGPT capability, we used a question-answering dataset SQuAD_v2 [74]. Apart from the original processing of the set (Table 2), three additional experiments were conducted. The first involved prompting ChatGPT a week later with the same prompts as during the initial testing of SQuAD. The second experiment exploited the same prompts, but with a new order, i.e., all unanswerable questions were prompted before the answerable ones. That way, ChatGPT could not treat the previous answers to

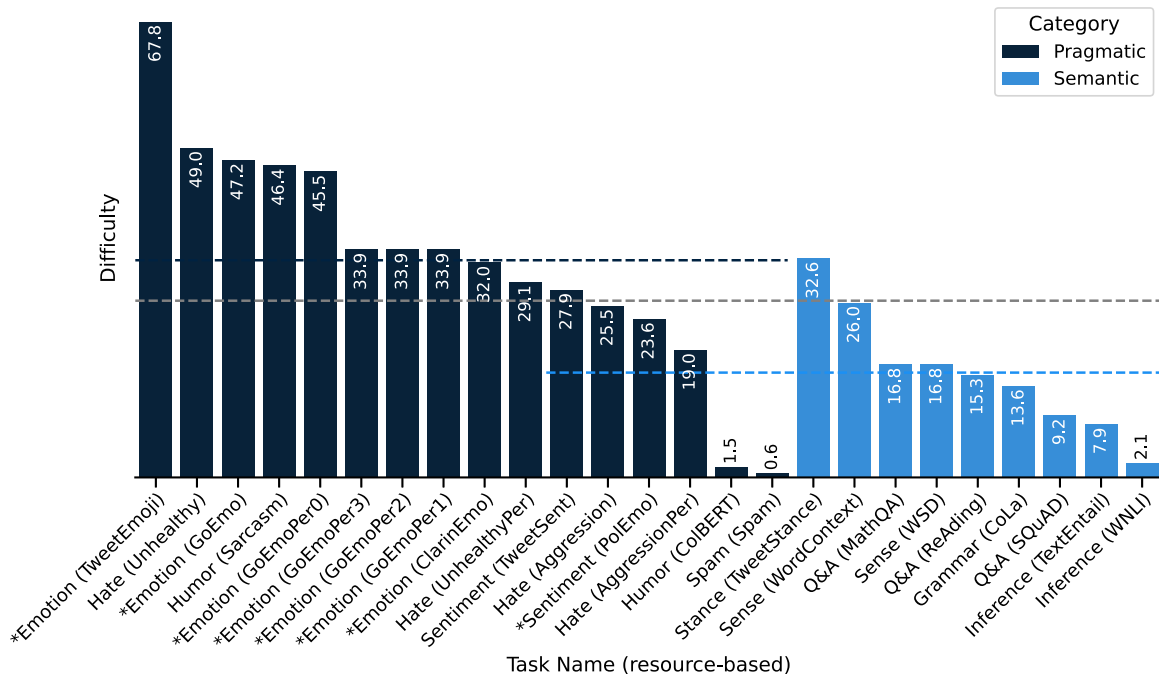


Fig. 6. Difficulty of the task (100% - SOTA performance) descending ordered. Tasks preceded by an asterisk are related to emotions. Dashed lines denote the average difficulty level for only pragmatic, all, and only semantic tasks.

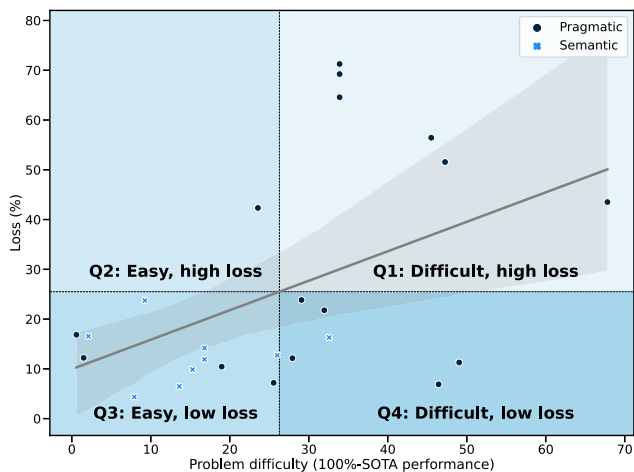


Fig. 7. Quadrants with the correlation between the loss of ChatGPT performance compared to the best, recent (SOTA) method and difficulty of the task. Each data point represents a separate task and its index can be found in Fig. 8. Quadrant borders are established according to the average loss (25.5%) and average difficulty (26.3%), Table 2.

the questions with the same context as the extended context of the given prompt. The final experiment involved the same set of prompts. However, a separate conversation with ChatGPT was initialized for each prompt. We computed accuracy and F1 macro for each scenario, along with the number of unanswerable questions (300 cases in total), which were correctly or incorrectly detected by ChatGPT, Table 3.

The obtained results demonstrate that ChatGPT performance on the same set of prompts in the same order and setup insignificantly decreased over a week by 1pp (accuracy) or 0.5pp (F1). ChatGPT reasoning quality barely improved when the order of the prompts was changed and slightly decreased when prompts were isolated in separate conversations. The number of unanswerable questions was correctly detected and ChatGPT performance was almost identical for the original set and the one with a new prompt order. For the dataset tested a

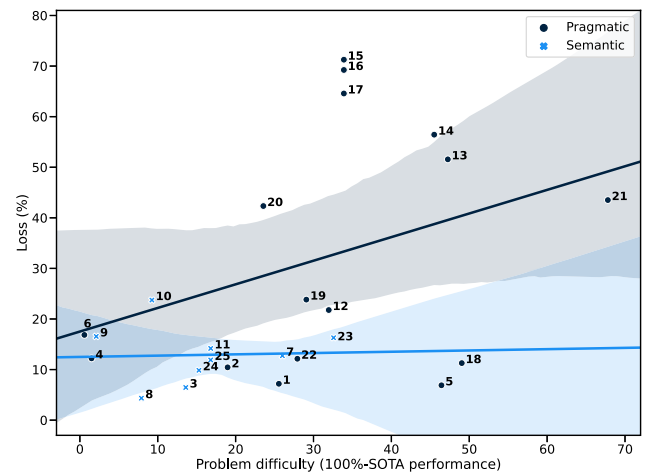


Fig. 8. Correlations between the loss of ChatGPT performance compared to the SOTA method and difficulty of the task. Regression lines are drawn separately for pragmatic and semantic tasks. Each data point represents a single task with the index from Table 1.

week later and with separate conversations, all the metrics decreased. It indicates that ChatGPT is not directly influenced by the previous prompts while determining whether the question is unanswerable. Both the performance of ChatGPT and its ability to detect unanswerable questions were worst when separate conversations were established for each prompt. It may suggest that providing some answerable questions helps it detect unanswerable ones with the same context. However, the differences in performance are not significant enough to be sure of such dependencies.

The results are inconclusive as to whether ChatGPT treats the previous prompts as a context for the prompt. Anyway, the differences in performance are not significant. On the other hand, ChatGPT demonstrated its instability and tendency towards non-determinism. This can be a serious disadvantage for some application domains. Even with the same setup, its results may vary with each launch.

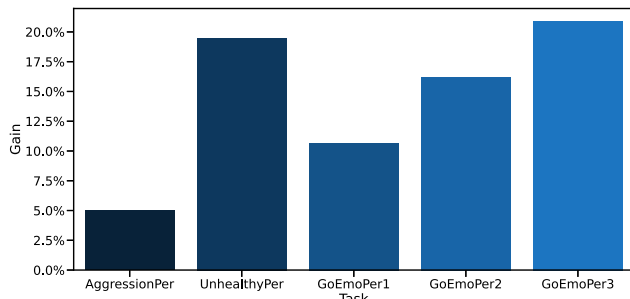


Fig. 9. Impact of context on classification metrics for GoEmotions, Aggression, and Unhealthy Conversations datasets. We show the percentage gain between setup with context and baseline, i.e. setup where no prior knowledge about the annotator is provided to the model. We show a gain in accuracy for the former dataset, whereas, for Aggression and Unhealthy Conversations, we present a gain for the F1-score.

Table 3

Performance of ChatGPT on different experiment setups of the SQuAD task. *Unanswerable detected* represents cases that ChatGPT correctly recognized as unanswerable questions. *Unanswerable not detected* are unanswerable questions, to which ChatGPT incorrectly answered.

Dataset	Accuracy [%]	F1 score [%]	Unanswerable detected cases	Unanswerable not detected cases
Original set	56.50	69.21	76 (25.33%)	224 (74.67%)
After week	55.40	68.72	64 (21.33%)	236 (78.67%)
New order	57.00	69.76	74 (24.67%)	226 (75.33%)
Separate conversations	53.60	67.23	60 (20.00%)	240 (80.00%)

6.5. Availability of the testing set for ChatGPT training

Some of the datasets exploited in our ChatGPT evaluation were publicly available at the time of the ChatGPT training. Therefore, the model could have been learned on those data, which may influence its performance on those particular datasets, see column *Availability* and *Trained* in Table 1. *Availability* has been estimated by us while *Trained* was extracted from ChatGPT responses. In general, most of the analyzed sets were probable or highly probable to be used for training the model.

The results shown in Figs. 10 and 11 indicate that the datasets on which ChatGPT was likely to have been trained tend to achieve higher performance (smaller loss) compared to SOTA solutions than the ones ChatGPT was less likely to be trained on. The tasks which ChatGPT claims it used for training (Fig. 11) are in opposite dependency difficulty — loss than the ones the model is unaware of. Analysis of availability rather supports this phenomenon (Fig. 10). It means that sets known for ChatGPT and estimated by us to be used for training overlap each other, and their loss is not much dependent on task difficulty.

6.6. Manual prompt fine-tuning

In the course of conducted evaluations, it became apparent that the construction of the prompt can have an impact on the obtained results. This hypothesis was inspired by [102], where common patterns for various needs and problems were proposed. Therefore, we prepared various versions of queries modeled on patterns such as *“The persona pattern”* and *“The game pattern”*. A compilation of the results can be presented in Table 4.

The experimental setup differed from the above studies, as we utilized the official OpenAI API⁹, which allows for greater control over the model behavior. In every experiment, the default API parameters

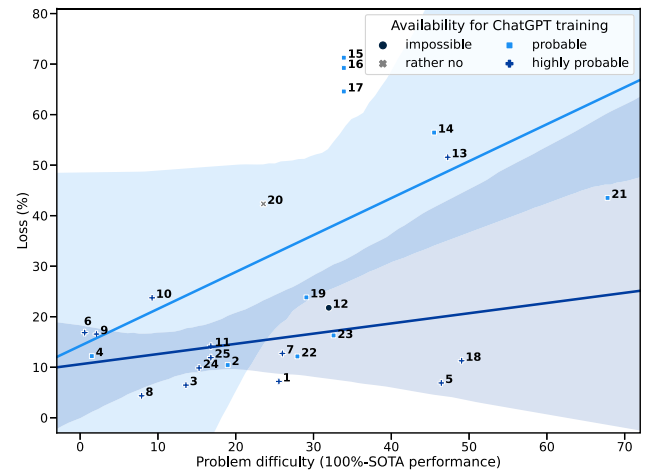


Fig. 10. Correlations between the loss of ChatGPT performance compared to the SOTA method and the difficulty of the task. Regression lines are drawn separately for two categories of Availability (2 and 3) from Table 1. Each data point represents a single task with the index from Table 1.

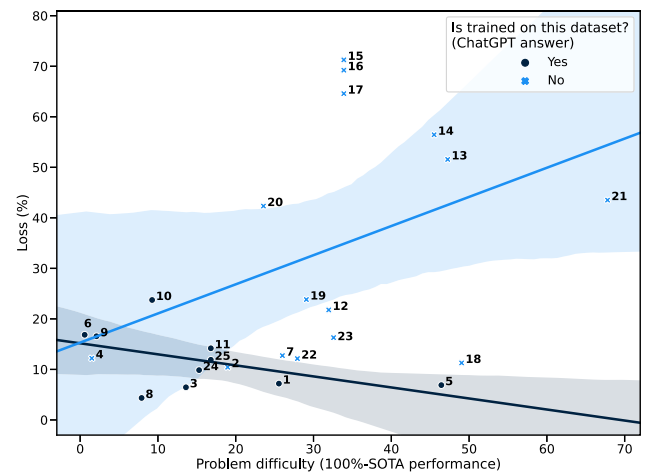


Fig. 11. Correlations between the loss of ChatGPT performance compared to the SOTA method and the difficulty of the task. Regression lines are drawn separately for whether ChatGPT claims to be trained on the dataset or not (Table 1). Each data point represents a single task with the index from Table 1.

described by OpenAI were exploited¹⁰. Each task was tested with the inclusion of the message *“System”*, which helps set the behavior of the assistant. The prompt patterns were described in Appendix D. The results demonstrate that the prompt pattern substantially influences the obtained outcomes. For semantic tasks (TextEntail, WNLI), we were not able to improve ChatGPT performance (negative difference) with various prompt patterns. However, it was possible for emotion-related datasets, i.e. a small benefit for GoEmo and very significant for PolEmo (pattern 76 and 77) – increase by even 14.8 p.p.

In summary, we emphasize the significance of prompt patterns on the obtained outcomes. It can dramatically impact on performance. Nevertheless, additional and dedicated research is imperative to determine the optimal prompt pattern for each problem.

6.7. Comparison with GPT-4

To complement our quantitative analysis, as demonstrated in Table 2, we conducted a comparison between ChatGPT and new GPT-4

⁹ <https://platform.openai.com/docs/guides/chat/introduction>

¹⁰ <https://platform.openai.com/docs/api-reference/chat>

Table 4

Quantitative analysis. Values of quality measures obtained for (a) Initial ChatGPT evaluation, see Table 2 and (b) Pattern ChatGPT: with different prompt patterns. Improvement provided by prompts – *Difference*: $(b - a)$. *Pattern benefit*: $100\% \cdot (b - a) \div b$. Emotions tasks are marked with an asterisk. Prompt patterns are explained in Appendix D.

ID	Task name (resource-based)	Prompt pattern	Task category	Measure type	Initial ChatGPT (a) [%]	Pattern ChatGPT (b) [%]	Difference (b-a) [pp]	Pattern benefit [%]
8	TextEntail	78	Semantic	F1 Macro	88.1	77.6	-10.5	-11.9
8	TextEntail	79	Semantic	F1 Macro	88.1	81.2	-6.9	-7.8
8	TextEntail	80	Semantic	F1 Macro	88.1	77.6	-10.5	-11.9
8	TextEntail	81	Semantic	F1 Macro	88.1	75.1	-13.0	-14.8
9	WNLI	82	Semantic	Accuracy	81.7	74.2	-7.5	-9.2
9	WNLI	83	Semantic	Accuracy	81.7	77.5	-4.2	-5.1
9	WNLI	84	Semantic	Accuracy	81.7	76.1	-5.6	-6.9
13	*GoEmo	71	Pragmatic	F1 Macro	25.6	21.8	-3.8	-14.8
13	*GoEmo	72	Pragmatic	F1 Macro	25.6	26.4	0.8	3.1
13	*GoEmo	73	Pragmatic	F1 Macro	25.6	23.7	-1.9	-7.4
13	*GoEmo	74	Pragmatic	F1 Macro	25.6	24.6	-1.0	-3.9
20	*PolEmo	75	Pragmatic	F1 Macro	44.1	38.6	-5.5	-12.5
20	*PolEmo	76	Pragmatic	F1 Macro	44.1	57.8	13.7	31.1
20	*PolEmo	77	Pragmatic	F1 Macro	44.1	58.9	14.8	33.6
All tasks				Average std. dev.	59.4 ± 27.8	56.5 ± 24.0	-2.9 ± 8.2	-2.7 ± 15.6

Table 5

Quantitative analysis. Values of quality measures obtained for (a) the ChatGPT output, (b) GPT-4. Both models were tested on identical sets of prompts. SOTA (c) is provided as a supplementary reference, see Table 2. *Difference*: $(b - a)$. *GPT-4 benefit*: $100\% \cdot (b - a) \div b$. *ChatGPT loss*: $100\% \cdot (c - a) \div c$. *GPT-4 loss*: $100\% \cdot (c - b) \div c$. Emotions tasks are marked with an asterisk.

ID	Task name (resource-based)	Measure type	ChatGPT (a) [%]	GPT-4 (b) [%]	Difference (b-a) [pp]	GPT-4 benefit [%]	SOTA (c) [%]	ChatGPT loss [%]	GPT-4 loss [%]
8	TextEntail	F1 Macro	88.1	91.3	3.2	3.5	92.1	4.3	0.9
9	WNLI	Accuracy	81.7	91.6	9.9	10.8	97.9	16.5	6.4
10	SQuAD	F1 Macro	69.2	76.3	7.1	9.3	90.8	23.8	16.0
13	*GoEmo	F1 Macro	25.6	23.1	-2.5	-10.6	52.8	51.6	56.3
20	*PolEmo	F1 Macro	44.1	41.0	-3.1	-7.6	76.4	42.3	46.3
All tasks		Average std. dev.	61.7 ± 26.3	64.7 ± 31.1	2.9 ± 5.7	1.1 ± 9.7	82.0 ± 18.1	27.7 ± 19.2	25.2 24.7

on a selection of five tasks from our previous evaluation¹¹, see Table 5, Fig. 16. Additionally, we provided quality measurements for the SOTA model as a point of reference.

Interestingly, despite GPT-4 being a more advanced model than ChatGPT, we observed varying performance results. ChatGPT still outperformed GPT-4 in pragmatic, emotional tasks, i.e. GoEmo and PolEmo, while GPT-4 achieved significantly higher scores in the remaining three semantic tasks with even a 9.9 p.p. increase for the WNLI task.

It is important to note that, across all tasks, the SOTA model consistently outperformed both ChatGPT and GPT-4. The loss for GPT-4 was very small for semantic tasks (TextEntail, WNLI): 0.9%–6.4%, and still very high for emotional problems: 46.3%–56.3%.

We emphasize that the results for GPT-4, as for ChatGPT, could significantly differ, if distinct prompt schemes are compared, see Section 6.6.

7. Qualitative analysis

Understanding the cases when ChatGPT is not acting as expected requires a deeper analysis, divided into three types: exploratory analysis, benchmarking analysis, and explanatory analysis. The exploratory analysis evaluates system answers for different prompts. In benchmarking analysis, the expert evaluates ChatGPT ratings and dataset label quality. The explanatory analysis allows an understanding of the ChatGPT answers by asking in-depth questions.

Fig. 12 contains our summary of the differences between ChatGPT and the latest state-of-the-art solutions dedicated to specific NLP tasks, as the result of the quantitative analysis presented in Section 6 and the qualitative analysis presented here.

¹¹ Unfortunately, a more extensive study was impossible due to recent access restrictions, i.e., a limit of 25 prompts per 3 h.

7.1. Exploratory analysis: Case study

When exploiting the possibilities of ChatGPT, we can see that it can perform various tasks, including recognizing generalized and personalized dimensions of Natural Language Processing, answering questions where a generous amount of domain knowledge is required, or even writing lines of code in the programming language of choice. What can be observed from time to time is the instances where ChatGPT is faced with a lack of knowledge. Those situations are usually solved by supplementing the model with information. But what if the information we are providing is, in fact, wrong? When asked about the main character of the Polish novel “Lalka” (“The Doll”), ChatGPT answered correctly. Still, when explaining that the answer was wrong and that the author’s name was different, ChatGPT added the wrongly inputted name and proceeded to answer with this inaccurate information. We can see that the domain knowledge of the model can be weak to disinformation, which further implies possible consequences regarding clashes with fake news. Another layer of divergent behavior of ChatGPT is in the ethics of the model. When conducting experiments regarding tasks such as humor recognition or offensiveness detection, we have stumbled upon output that not only refuses to answer whether something is or is not funny but also sends a moralizing message with an irritated tone. Interestingly, the model implies it is fully neutral and has no biases, yet it has them in topics regarding ideological views.

Hagendorff [103] drew attention to the fact that chatbot ethics can be a subject of debate in fairness, nondiscrimination, and justice. ChatGPT should respond to questions and generate text based on the given parameters. However, there is still a blank area where the tool will not accomplish tasks. At first glance, ChatGPT refuses to provide specific content that can be presumed as judgmental, discriminative, or promoting hate speech. During the exploratory dialogue, we found many ways to display messages that are not always politically correct.

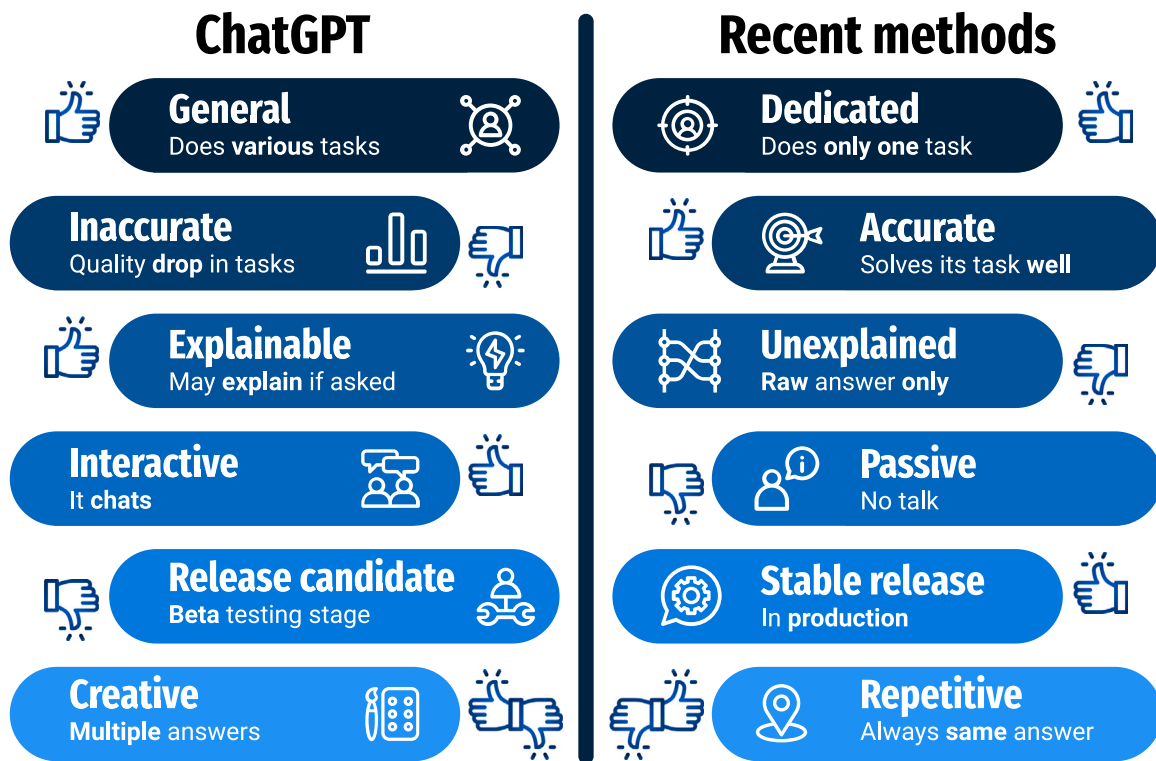


Fig. 12. Difference between ChatGPT and the best recent solutions (SOTA) related to analytical NLP tasks.

The first example (Chat 51) is to avoid answering the question about the likelihood of achieving a goal in an academic or professional career by listing the potential factors that may influence this fact. ChatGPT answers only after the researcher asks directly about the typical representatives of the particular position. By making the request more specific based on the data, ChatGPT gives a precise answer. The second example (Chat 52) of task-solving avoidance is refusing to make up the story with a word that can be offensive. ChatGPT assumes that the user refers to this meaning, omitting the context from the previous question, whose purpose indicated that nonvulgar sense is involved. Another type (Chat 53) of refusal is making up stories that raise the delicate subject, i.e. stories about the traumatic event that can be seen in the third example. ChatGPT will only generate the content if the user adequately motivates it with the scientific goals. The fourth example (Chat 54) highlights the possibility of the chatbot exhibiting bias while answering requests for characterizing the widely known traits of controversial politicians without judgmental opinions. However, in the second task, in which ChatGPT has to write a joke that this politician would admire, it refuses to motivate his decision politician’s disregard for human rights. This proves that the tool has hidden biases that are revealed inappropriately worded answers for tasks or questions. Borji [104] conducted a systematic review of the typical categories of ChatGPT failures. The above errors are derived from both incorrect reasoning in terms of psychological reasoning and bias and discrimination.

The performance of modern language models, such as T5, GPT-3, and ChatGPT, heavily relies on the quality of task-specific prompts. The prompt-based learning paradigm requires careful prompt engineering and prompt tuning. However, in the case of the ChatGPT model, prompt tuning is technically unavailable, and the only way to verify prompt relevance is to evaluate its performance directly in the downstream task. We decided to tune the prompts manually according to the task — we selected the prompts such that the answers generated by the model on a small validation sample for the given task were the most stable and accurate. On the other hand, using the prompts directly as humans designed them implicitly allows us to evaluate models’

language comprehension abilities. Such evaluation is important for tasks in the area of semantics, where models should successfully utilize short natural language descriptions of words or phrases, as they are used in other supervised solutions.

Most tasks require a prompt that enables the model to choose a certain value from the provided options. However, to evaluate ChatGPT’s ability to understand various data formats, we tried not to restrict the design of our prompts to a single data template. Still, the prompts must include all the information required for the ChatGPT to perform the task. A good example can be a prompt for Aggression or CoBERT tasks, where we provide possible outcomes and expect ChatGPT to choose the right answer and return it in Python list format. Some tasks require a choice from multiple options, like TweetEmojis, where the correct answer is the emoji that fits the best-provided tweet. ChatGPT can also return a number as a category indicator or whole output in the JSON format. In the case of mathematical reasoning, it can provide a whole explanation of how it reached a certain outcome and provide only the answer without explanation. Understanding prompts and user intent for how the output should be structured is not an issue for the model, which is a very impressive capability. We also noticed that when it is unable to perform a task on the provided example, it will refuse to do so and provide an explanation why, as it has happened in the case of ClarinEmo B.12, where the model stated that all provided texts are legal and financial statements. Therefore it is not possible to assign emotion labels to them.

7.2. Benchmarking analysis: Validation based on human expert

There are some trends in the ChatGPT responses, which were the basis for the difficult case analysis. One of the main trends is connected with the chat *sensitivity*. Importantly, this sensitivity could be observed during the execution of different tasks. Offensiveness detection is an example — ChatGPT assigned additional labels to those texts from Unhealthy Conversations Dataset labeled by human annotators simply as *healthy*. Similarly, ChatGPT has associated most of the statements

Table 6

The percentage of output values originally assigned to the input text by *Human* or by *ChatGPT*, which our experts accepted.

Task name	Human annotations approved	ChatGPT responses approved
Aggression	68%	51%
TweetSent	69%	55%
GoEmo	61%	73%
Unhealthy	43%	81%

coming from GoEmotions and labeled by people simply as *neutral* with different emotions.

Interestingly, in many cases, ChatGPT tends to have more negative (and therefore safe) assessments than people. Characteristic examples come from two sources. ChatGPT labeled as aggressive only 11 texts from the WikiDetox Aggression dataset labeled by people as non-aggressive, while the opposite decision was taken 207 times. A similar trend is observed for the TweetSent task — ChatGPT assigned positive sentiment to 27 tweets labeled by people as negative, while the opposite decision was taken 83 times. It turns out that the system erroneously assigns a positive sentiment to those texts in which there are linguistic cues of a contradictory nature, e.g.:

WESTWORLD Dolores is MF Wyatt mutherfuckerrrrrr I don't think I've guessed one MF thing I love shows like this.

or

Hahaha #Negan #TheWalkingDead if you watch you'll know, if you don't then what the fuck man!!!!

In the case of misattributed negative sentiment, no such clear correlation can be observed. However, those texts whose interpretation is context-dependent (this context is very often political) are a significant proportion, e.g.:

Bill Clinton built a wall on the Mexican border in the 90 s. #FunFactFriday

or:

The election of Donald Trump could have a significant future impact on the project Dakota Access Pipeline when he takes office.

We have analyzed the inconsistencies between human annotations and ChatGPT answers based on four datasets: Wikipedia Aggression, GoEmotions, Tweeteval: sentiment, and Unhealthy Conversations. We have examined 100 randomly selected cases for each dataset. Each case was composed of prompt, human annotation, and adequate (but inconsistent) ChatGPT answers.

Analysis was conducted by experts who are specialists trained in the recognition of emotions in the text. One of them is a psychologist and another is a linguist, both are experienced annotators. They get acquainted with the text prompt and decided whether the evaluation both of human and ChatGPT were correct. Expert analysis was focused on different points of view that someone may take. The annotations in the selected 4 datasets were of a more or less subjective nature, and for this reason, it was not necessary to create detailed guidelines in order to achieve high inter-annotator agreement (moreover, the same was true for the original datasets). The essential goal was precisely to capture possible and acceptable differences in the labeling of texts. Rather, the idea was to take into account the various possibilities, including those not captured in the benchmark dataset. The experts evaluated the labels assigned to the texts. In some cases (when different contexts may affect different interpretations), human annotation and ChatGPT answers were considered correct. The number of ChatGPT correct answers is relatively high, see Table 6 and Fig. 13.

A more detailed analysis focused on five types of comparison (see Table 7 and the visualization of the differences between the tasks based on selected categories presented in Fig. 14): the cases in which the

Table 7

Expert-based evaluation of the agreement between ChatGPT responses and original human annotations (ground truth): *Human & ChatGPT* – the expert accepted both ChatGPT answer and the human annotation, *Only human* annotation was approved by the expert, *Only ChatGPT* was found acceptable, *Neither human nor ChatGPT* was acceptable, *N/A* evaluation was not available since the expert was not able to link the input text to the possible output.

Task name	Human & ChatGPT	Only human	Only ChatGPT	Neither human nor ChatGPT	N/A
Aggression	21%	48%	31%	0%	0%
TweetSent	26%	44%	30%	0%	0%
GoEmo	45%	16%	28%	8%	3%
Unhealthy	24%	19%	57%	0%	0%

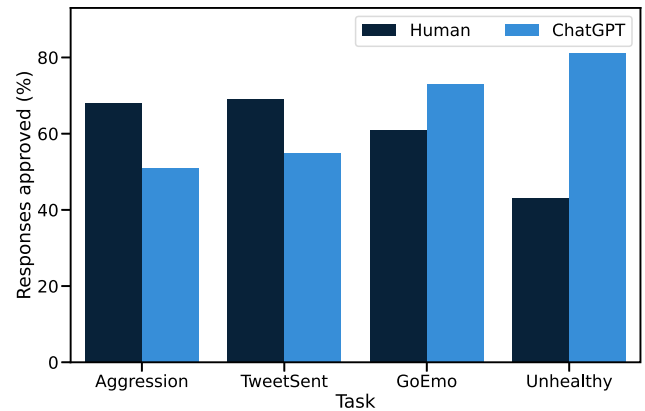


Fig. 13. The contribution of output values assigned to the input text by humans or by ChatGPT, which our experts have approved, Table 6.

expert accepted both human annotation and ChatGPT answer (*Human & ChatGPT*: for example see Chat 58); the cases in which only human annotation was considered correct (*Only human*: for example see Chat 57); the cases in which only ChatGPT answer was considered correct (*Only ChatGPT*: for example see Chat 55r 56); the cases in which neither human nor ChatGPT answer was considered correct (*Neither human nor ChatGPT*: for example see Chat 59 or the cases in which evaluation was impossible due to the unintelligible content (for example see Chat 60. The analysis revealed that in many cases (especially for Unhealthy Conversations), only ChatGPT labeled the text correctly. ChatGPT pointed out many human errors (see Appendix C.2 for more examples). Interestingly, the cases where only ChatGPT gave the correct answer have a common characteristic: in most of them, the human annotator was less sensitive, e.g. the annotator(s) labeled aggressive utterances as non-aggressive, negative tweets as neutral or unhealthy conversation as healthy. ChatGPT tends to interpret a given text more negatively than a human does.

It is also connected with pragmatic categories such as sarcasm. Many utterances, which humans labeled as neutral, ChatGPT classified as sarcastic, e.g.:

Yes, it's sarcasm. I shouldn't use it actually, it's pretty hard to tell nowadays". Yours wasn't but yeah it sure is getting harder... scary..

This fact shows that many of the neutral messages can be classified as sarcastic and aggressive, which as a result, can limit freedom of speech in case of using it commercially or in a public debate. The tool's creator should emphasize the preparing model that will be available to distinguish small nuances between sarcasm and a neutral message. This is desirable not only for the usability of the solution but also for building public confidence in artificial intelligence solutions. ChatGPT's informing that a message is negatively perceived is a way to teach a user with the wrong intentions to be politically correct. On the other hand, a user who tries to convey information objectively without

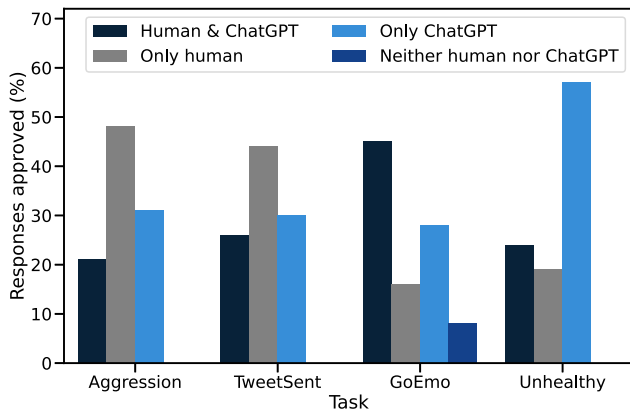


Fig. 14. Expert-based evaluation of the agreement between ChatGPT answers and the original human annotations: (1) Human & ChatGPT — the expert accepted both the ChatGPT response and human annotation, (2) The expert approved only human annotation, (3) Only ChatGPT answer was accepted, (4) Neither human nor ChatGPT was acceptable for our expert, Table 7.

malicious intentions may learn that reality is more biased than he or she might think. Another interesting conclusion from the analysis is the recognition of the sincerity of one’s message that involves its true intentions. The annotator has evaluated the below message as expressing gratitude, whereas ChatGPT regards it as neutral (Chat 58)

You’re welcome

This simple message could provide neutral emotions if the message’s sender said it automatically. However, if the speaker intends to express the actual gratitude that one feels, ChatGPT cannot recognize this from such a short message and without having additional information about the speaker. All the examples can be found in Appendix C.2.

7.3. Explanatory analysis: XAI

The advantage of ChatGPT is that it can give reasons for its answers. Thus, we are dealing with self-explanatory artificial intelligence, which is a part of eXplainable Artificial Intelligence (XAI); see Appendix C.3. Thanks to the interaction, researchers’ findings can be confronted with the motivation provided by the model itself. Exploiting this opportunity, we subjected some ChatGPT answers presented above to detailed examination. This led us to several conclusions.

1. ChatGPT provides reasonable and factual clarifications for its decisions. It can point to specific passages in the text that influenced his decision, Chat 61:

Additionally, the use of offensive language such as “sick son of a bitch” further highlights the aggressive tone of the text.

or Chat 70:

the use of exclamation marks and the phrase “I did not finished yet!!!” can be interpreted as confrontational or intense

When asked, ChatGPT explains in detail why a phrase has a particular interpretation, Chat 63:

The phrase “Go back to the dump on Staten Island” is a personal attack because it is meant to be insulting and demeaning. By implying that the person being addressed should return to a place considered to be unpleasant or undesirable, the speaker is showing a hostile attitude and attacking the person’s character or background.

ChatGPT is also capable of generalizing, e.g. when asked which language phenomena demonstrate the enthusiasm or positive sentiment in the text, it gives a list containing such phenomena as the use of

superlatives or lack of negative language (see Chat 65). However, this is characteristic of justifications for both correct and incorrect answers.

2. ChatGPT seems to have no regard for individuals, instead judging situations. However, this often leads to mistakes, e.g. when it justifies assigning positive sentiment to neutral information, Chat 66:

In general, being shortlisted for an award is seen as a positive achievement, so the sentiment expressed in the text is positive.

Information about the distinction for a particular footballer is neutral. Its sentiment, however, can be both positive and negative. It depends on the sympathies of the recipient — sympathies regarding specific footballers. Similarly, ChatGPT justifies the negative sentiment of the news about the ban on naming streets after Fidel Castro, Chat 66:

In general, restrictions or limitations are typically seen as negative, so mentioning this restriction implies a negative judgment about the situation.

At the same time, ChatGPT explicitly distances itself from judging people. This issue is strongly connected with the next one.

3. ChatGPT flattens the message, partially ignoring the meta-text. A common mistake of the system is that it evaluates press reports and quotes of someone’s statements without considering the metatextual frame. So it evaluates the main content but ignores the broader context (see Chat 64).

4. There are some disapproved words. ChatGPT evaluates rather situations than participants, but words refer to people, which lead to a specific, predetermined assessment, Chat 62:

Additionally, the use of quotes around “trolls” implies that the speaker is directly calling the person, they are addressing a troll, which is further evidence of an aggressive tone

5. ChatGPT strongly relies on context paraphrasing when explaining its decisions in semantic tasks. This phenomenon was observed mainly in WSD and WIC tasks. In WSD, the model was expected to explain its decision by defining the meaning of chosen sense concerning the given context. However, for some examples, the model approached the task by largely repeating selected parts of the given context in such a way that the generated explanation did not meet typical linguistic criteria of constructing a proper sense definition, Chat 67:

This is because the text describes bells as being present in an ancient stone church, and they are being rung (making a ringing sound) to call the faithful to evensong.

6. ChatGPT presents the sense of common human morality. As mentioned in the previous section, ChatGPT tends to find negative connotations in the given text. In this example, the sentence was interpreted as not aligned with society’s standards. Only after the researcher suggested the possibility of using black humor, accepts this interpretation, Chat 69:

The idea of eating one’s own parents is generally considered taboo and immoral, and it can provoke a strong negative reaction in people

8. Limitations and discussion

Below, you can find a list of nine observations and limitations related to selected problems and cases that we encountered during our investigation.

1. Prompts may not be strict and precise enough. ChatGPT requires prompts in natural language, which is – by nature – not structured and can be different for different users and tasks. Interesting to note is that prompt construction may affect the quality of the model’s performance. There is also the possibility of auto-generated prompts [101]. We anticipate that this will be the subject of much future research.

2. Post-processing is sometimes required due to less prompt precision, not following the expected behavior by ChatGPT, and its instability, see Section 5.2. It also refers to the necessary manual correction of typos surprisingly provided by ChatGPT, e.g., [*“curiosity”*] instead of [*“curiosity”*]. Some answers are in whole sentences instead of requested predefined lists (see Chat 30). It is difficult to point out the reasons for such conduct. As a result, the raw text provided by ChatGPT may require extraction of crucial information, e.g., final labels. It also depends on the prompting quality and correctness.

3. Unexpected, out-of-range responses are sometimes provided by ChatGPT, e.g., the unlisted emotion *satisfaction* has been delivered:

Chat 2. Task: GoEmotions. Case 98.	
Prompt	From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in a Python list containing exactly two selected most matching emotions. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: <i>That's how you know they're happy.</i>
ChatGPT answer	Based on the input text, it is likely that the text would arouse the emotion of realization and satisfaction in most people reading it. [<i>“realization”</i> ,” <i>satisfaction”</i>]
Extracted answer	[<i>“realization”</i> ,” <i>satisfaction”</i>]
Expected answer	[<i>“excitement”</i> ,” <i>neutral”</i>]

4. Internal policy can strongly influence the answers received. It includes various biases and political correctness incorporated into the general ChatGPT model. Therefore, some responses are *locked* due to the OpenAI policy. It, for example, refers to texts *“that could be seen as offensive and vulgar”*:

Chat 3. Task: GoEmotions. Case 101.	
Prompt	From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 2 selected most matching emotions. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: <i>Wouldn't wipe my arse with it.</i>
ChatGPT answer	It would be difficult for me to determine the emotions that the input text would arouse in most people reading it as it appears to be a strong statement that could be seen as offensive and vulgar. Without more information and context, it would be hard to determine the emotion. [<i>“neutral”</i>]
Extracted answer	[<i>“neutral”</i>]
Expected answer	[<i>“annoyance”</i> ,” <i>neutral”</i>]

5. Context is generally respected by ChatGPT. In many cases, this can be considered an advantage. However, prompts themselves and the sequence of prompts provide limited control over them.

6. Efficiency and costs. At the time of the evaluation, a paid API from OpenAI for ChatGPT was unavailable. As part of the study, we used 13 accounts for which we passed session tokens to a central multi-threaded module that sends and receives prompts. On average, after about 50 prompts sent per hour using one special token, OpenAI temporarily blocked access for that special token, mostly for less than an hour. Sometimes the tokens were outdated and had to be re-generated and replaced, which depended on the responsiveness of account holders. A processing capacity of 2-3k prompts per day was achieved at the end. In the same amount of time, the SOTA models are capable of processing millions of texts even with a single recent GPU card [105].

7. The problem of controversial and ambiguous utterances is solved by ChatGPT's demonstration of a lower level of confidence, e.g. *“It would be difficult for me to accurately determine”*, *“The same text can be perceived differently depending on the context and tone of the statement.”*, *“I am sorry, but the input text is not clear, its a Mix of...”*, *“it is quite ambiguous, and the context is not provided”* (GoEmo, case 80, 82, 101, 102, 554, 574, 893, 894). Overall, it can be seen as an advantage since it suggests providing additional information or some corrections. It is also an invitation to conversation in case of manual processing.

8. Explanations (XAI) are sometimes provided by ChatGPT itself. They are very reasonable, e.g., [*“determination”*] *Note: Determination is not in the given list of emotions, I have inferred it from the text given”*:

Chat 4. Task: GoEmotions. Case 557	
Prompt	From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: <i>“I will do what I must”</i> .
ChatGPT answer	[<i>“determination”</i>] <i>Note: Determination is not in the given list of emotions, I have inferred it from the text given.</i>
Extracted answer	[<i>“determination”</i>]
Expected answer	[<i>“neutral”</i>]

9. Restrictive or wrongly implemented policy will influence the user experience of the ChatGPT by not answering all user's prompts exaggeratedly judging many questions not along with OpenAI rules:

Chat 5. Task: PolEmo2.0. Case 95.	
Prompt	Rate the sentiment of this text using one of four labels: positive, negative, neutral, and ambivalent. Present the result as a list in Python containing the selected label. Text to be evaluated: <i>I honestly did not expect that something like this could happen to me. I feel cheated by them. They do not count on people and do not respect them. They take out money. They make you pay the same money for a place on the floor as for a bed, and they do not give bedding or mattress. They are incompetent. They do not give a damn about the welfare of a person. At 22.00 you have to sleep, otherwise the staff squawks in the corridor. I do not recommend this hostel to anyone.</i>
ChatGPT answer	This content may violate our content policy. If you believe this to be in error, please submit your feedback — your input will help our research in this area.

Prospects for ChatGPT applications

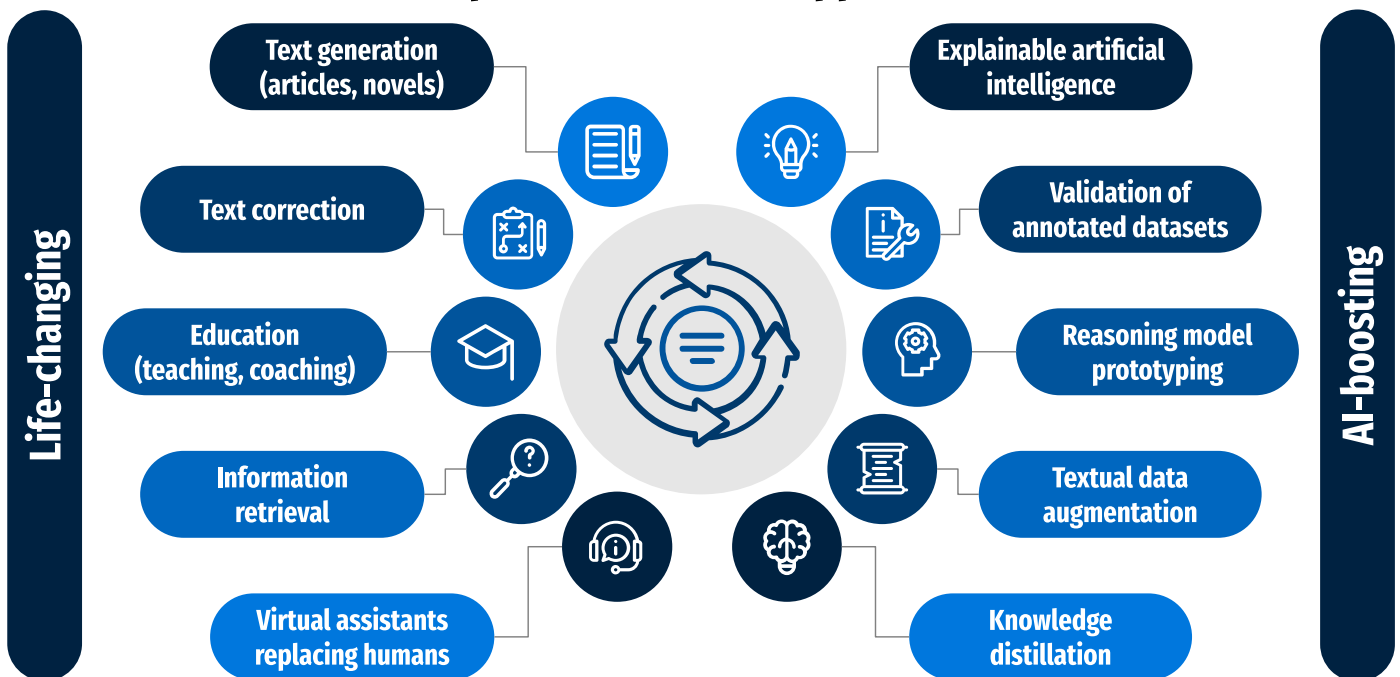


Fig. 15. Examples of ChatGPT applications are divided into two categories: changing our daily lives (left) and boosting the development of artificial intelligence (right).

10. **New profession: *prompting engineer*** is expected to emerge as a consequence of the above comments on the labor market.

9. Prospects for ChatGPT applications

We believe that ChatGPT, its functionality, and its great resonance in science, industry, and society will significantly impact our everyday life and technology related to artificial intelligence. Therefore, we expect ChatGPT and similar AI solutions to spur development and spark an economic and social AI revolution. We have listed several application areas that ChatGPT is poised to revolutionize first, Fig. 15. They are grouped into life-changing and AI-boosting domains.

10. Conclusions and future work

Based on ChatGPT's responses to 48k+ prompts related to 25 different NLP tasks, we can conclude that ChatGPT can solve most of the problems considered quite well. On the other hand, it loses to the best models currently available (SOTA), from 4 to over 70%. Its loss is relatively greater for more difficult and pragmatic tasks, especially when evaluating emotional texts. All this makes ChatGPT *a master of none* of the task. However, it is still an open question what would happen if ChatGPT was finetuned using the datasets from these tasks, and what the results would look like then. At the moment it is not possible to perform such a study, but it would be worthwhile to do so as soon as it is possible.

The context awareness and ability to implement Contextual Few-Shot Personalization proposed in this paper are valuable features of ChatGPT. It also provides a unique self-explanation capability that facilitates human understanding and adaptation to the expected outcome. We plan to develop and systematize the qualitative analysis of the model's performance on subjective tasks (primarily emotion recognition), e.g., by comparing ChatGPT responses with the estimated annotation controversy for texts and dimensions. We strongly believe that ChatGPT can accelerate the development of various AI-related technologies and profoundly change our daily lives.

Our future work will explore other reasoning tasks and various prompting engineering methods, as well as the new application areas mentioned in Section 9.

CRediT authorship contribution statement

Jan Kocoń: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Igor Cichecki:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Oliwier Kaszyca:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Mateusz Kochanek:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Dominika Szydło:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft. **Joanna Baran:** Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Julita Bielaniewicz:** Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Marcin Gruza:** Formal analysis, Data curation, Writing – original draft, Visualization. **Arkadiusz Janz:** Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Kamil Kanclerz:** Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Anna Kocoń:** Data curation. **Bartłomiej Koptyra:** Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Wiktoria Mielewszczenko-Kowszewicz:** Validation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Piotr Miłkowski:** Writing – review & editing. **Marcin Oleksy:** Validation, Resources, Data curation, Writing – original draft. **Maciej Piasecki:** Validation, Writing – original draft, Funding acquisition. **Łukasz Radliński:** Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Konrad Wojtasik:** Software,

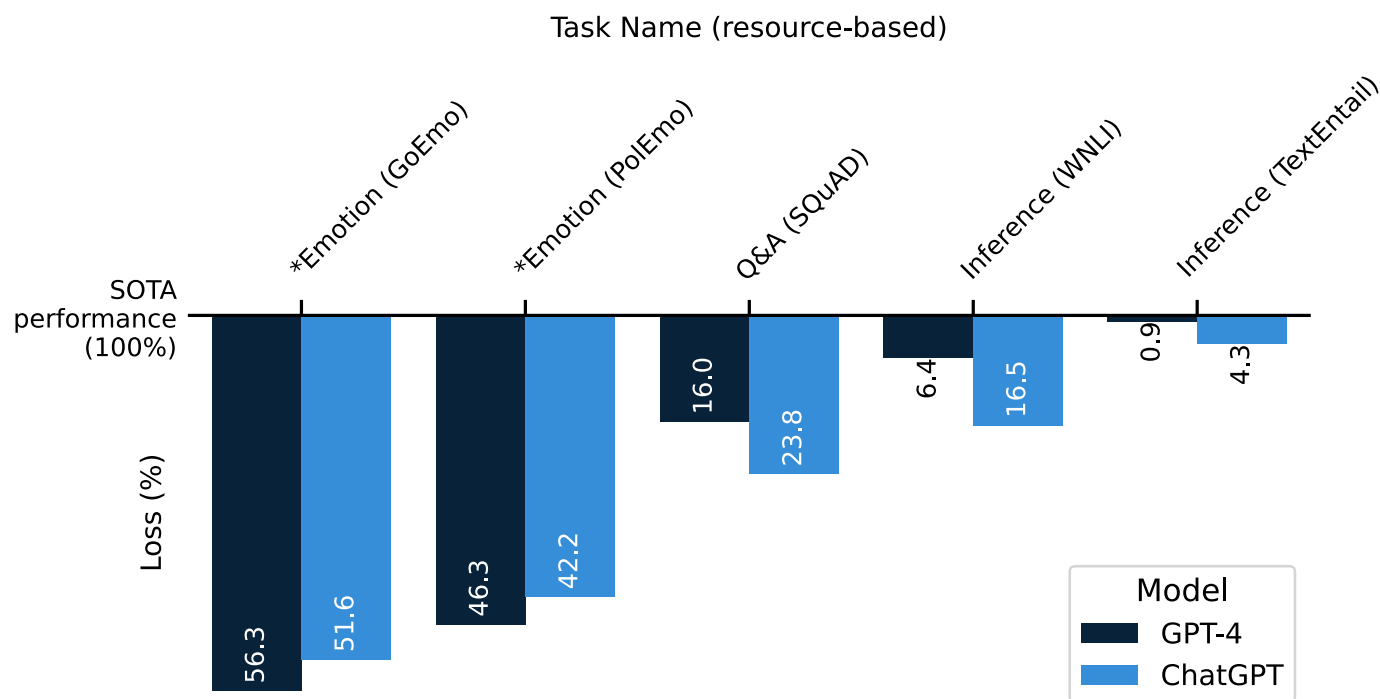


Fig. 16. The ChatGPT and GPT-4 loss in performance (%) for considered tasks and named according to their resource (dataset), descending ordered by loss value. Tasks preceded by an asterisk are related to emotions. The upper X axis corresponds to the performance of the best model (SOTA) treated as 100% capabilities.

Formal analysis, Investigation, Data curation, Writing – original draft. **Stanisław Woźniak**: Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Przemysław Kazienko**: Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jan Kocon reports financial support was provided by National Science Centre Poland. Przemyslaw Kazienko reports financial support was provided by National Science Centre Poland. Jan Kocon reports financial support was provided by European Regional Development Fund. Przemyslaw Kazienko reports financial support was provided by European Regional Development Fund. Jan Kocon reports financial support was provided by Horizon Europe. Przemyslaw Kazienko reports financial support was provided by Horizon Europe. Jan Kocon reports financial support was provided by Ministry of Education and Science of the Republic of Poland.

Data availability

We have shared the data and codes in a public GitHub repository: <https://github.com/CLARIN-PL/chatgpt-evaluation-01-2023>.

Acknowledgments

This work was financed by (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471 (JK, PK); (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, projects no. POIR.04.02.00-00C002/19 and POIR.01.01.01-00-0288/22; (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (5) the Polish Ministry of Education and Science within the programme

“International Projects Co-Funded”; (6) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

Appendix A. Additional results

Table 8 contains entropy values calculated for the available test or dev set and for its subset (if applicable) used by us for prompting. A small difference in these two values proves a similar distribution of classes in both sets, thus, a good stratification of sampling.

Table 9 includes additional measures for the evaluated tasks, calculated by us and taken from the literature.

Appendix B. Example prompts

This section contains sample chat records for all evaluated tasks. The Case number is the identifier of the example in the external collection used for evaluation for the Task, available in the project GitHub repository¹² in file:

ChatGPT Evaluation v2.0.xlsx.

In addition, we have provided the expected response.

B.1. Aggression

Chat 6. Task: Aggression. Case 20.

Prompt

Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: *Almonds!!!! For fun* .

¹² <https://github.com/CLARIN-PL/chatgpt-evaluation-01-2023>

Table 8

Entropies of data: a measure of class balance. The greater, the more balanced the data. A marginal difference in entropy of output real values for test or dev set (column #Test in Table 1) and within the set used by us (column #Used in Table 1) demonstrates a good stratification of the used set selection.

ID	Task name	Entropy test	Entropy used
1	Aggression	0.42	0.39
2	AggressionPer	0.49	0.50
3	CoLa	0.62	0.62
4	ColBERT	0.69	0.69
5	Sarcasm	0.69	0.69
6	Spam	0.39	0.39
7	WordContext	0.69	0.69
8	TextEntail	0.69	0.69
9	WNLI	0.69	0.69
10	SQuAD	-	-
11	MathQA	-	-
12	ClarinEmo	2.19	2.19
13	GoEmo	2.77	2.77
14	GoEmoPer0	2.77	2.98
15	GoEmoPer1	2.77	2.98
16	GoEmoPer2	2.77	2.98
17	GoEmoPer3	2.77	2.98
18	Unhealthy	1.65	1.60
19	UnhealthyPer	1.65	1.60
20	PolEmo	1.30	1.30
21	TweetEmoji	2.73	2.71
22	TweetSent	1.03	1.03
23	TweetStance	0.95	0.97
24	ReAding	-	-
25	WSD	7.74	7.74

ChatGPT answer
 ["non-aggressive"]

Expected answer
 ["non-aggressive"]

Chat 7. Task: Aggression. Case 402.

Prompt
 Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.
 Text: *letters to an editor from an opponent are hardly reliable.*

ChatGPT answer
 The text is “non-aggressive” as it is making a statement about the reliability of a source, rather than making any personal attacks or aggressive language.

Extracted answer
 ["non-aggressive"]

Expected answer
 ["non-aggressive"]

B.2. AggressionPer

Chat 8. Task: AggressionPer. Case 241.

Prompt
 Knowing that a certain person described each text with a given attribute:
 Text: I am also a sockpuppet of CUTKD. Before you block me, let TIW know that they are a HORRID LITTLE SHIT.
 Attribute: “aggressive”
 Text: Oh i see how it is yall jealous cause yall a bunch of yankees and im not im gonna sue yall;

Attribute: “non-aggressive”
 Text: oh why dear johnny boy part two im just gonna trash u here... enjoy
 Attribute: “non-aggressive”
 write which one of the attributes: “aggressive”, “non-aggressive” would this person describe a given text? Write your answer in the form of a Python list containing the appropriate attribute.
 Text: *Yet you keep showing up here and you claim I’m the on harrassing you. I stopped saying a word a week ago and you keep coming back. You keep telling me you aren’t harrassing me yet here you are again. And that’s not harrassing me HOW????????????????????????????????*

ChatGPT answer
 ["aggressive"]

Expected answer
 ["aggressive"]

Chat 9. Task: AggressionPer. Case 16.

Prompt
 Knowing that a certain person described each text with a given attribute:
 Text: Cesc Fabregas I see it’s been changed that he now plays for Chelsea. Looks like I was right all along? how do you like that you little bitch? I was right and you were wrong.
 Attribute: “non-aggressive”
 Text: STOP SPAMMING BLOGS IN EXTERNAL LINKS OR YOU WILL GET BANNED FROM WIKIPEDIA STOP SPAMMING BLOGS IN EXTERNAL LINKS OR YOU WILL GET BANNED FROM WIKIPEDIA
 Attribute: “aggressive”
 Text: Have you seen the reference to their claim, you moron?? This is not a source. It’s their POV. Once I have more time I will reduce your Pashtun nationalist POVs here on Wiki and give order to my uncles and cousins the execution of Pashtuns in the entire northern Afghanistan down to Gardez, Jalalabad, Lashkargah, Logar, Paghman, Laghman, Kunduz, Herat, Farah ... wherever immigrant descendants of violent Kuchis settled and took non-Pashtun lands away and traditional Tajik countries. When they have no clue about the percentages they write Pashtu, Dari and Uzbaki are spoken by 90%. So if they know that immigrant Pashtuns are in majority, why don’t they give any numbers? Why Kunduz province is dominated to 56% by Persian-speakers? Go take a visit to Kunduz and see it with your own eyes and listen with your own ears... but I can not tell you you will survive that as a Pashtun. Also your other source need to be checked on it’s reference”.
 Attribute: “aggressive”
 write which one of the attributes: “aggressive”, “non-aggressive” would this person describe a given text? Write your answer in the form of a Python list containing the appropriate attribute.
 Text: *Wow... This page is so awful. It has some decent information but it looks like a 12 year old wrote a C grade essay. I’d like to get some people to help me fix it up.*

ChatGPT answer
 ["non-aggressive"]

Expected answer
 ["aggressive"]

B.3. CoLa

Chat 10. Task: CoLa. Case 0.

Prompt
 Describe the grammatic acceptability of the given text. Choose your answer from provided list and map your answer with following unacceptable: 0, acceptable: 1 and return an integer as a result. Do not explain yourself.
 Input data:
 Text: *The sailors rode the breeze clear of the rocks.*
 Possible acceptability: unacceptable, acceptable

Table 9

Other performance measures for the tasks considered, which were computed by us and taken from the scientific reference paper.

ID	Task name	ChatGPT accuracy	ChatGPT F1	SOTA our accuracy	SOTA our F1	SOTA paper accuracy	SOTA paper F1
1	Aggression	77.91	69.1	80.58	74.45	94.79*	-
2	AggressionPer	79.61	72.57	86.37	81.03	-	-
3	CoLa	80.82	78.11	-	-	86.4	-
4	ColBERT	86.53	86.47	98.5	98.5	98.6	98.6
5	Sarcasm	50	49.88	52.7	53.57	95.4	95.54
6	Spam	89.83	82.67	99.73	99.42	99.28	98.49
7	WordContext	64.58	63.45	-	-	74	-
8	TextEntail	88.09	87.88	-	-	92.1	-
9	WNLI	81.69	81.63	-	-	97.9	-
10	SQuAD	56.5	69.21	-	-	87.61	90.75
11	MathQA	71.4	-	-	-	83.2	-
12	ClarinEmo	83.5	53.23	90.88	68.04	-	-
13	GoEmo	19.9	25.55	48.03	52.75	-	46
14	GoEmoPer0	19.5	23.74	-	-	-	54.5
15	GoEmoPer1	21.58	19	-	-	-	66.1
16	GoEmoPer2	22.66	20.34	-	-	-	66.1
17	GoEmoPer3	23.58	23.41	-	-	-	66.1
18	Unhealthy	64.01	45.21	87.57	50.96	-	-
19	UnhealthyPer	66.69	54.02	90.96	70.92	-	-
20	PolEmo	71.36	44.08	96.72	76.44	96.72	76.44
21	TweetEmoji	29.51	18.19	44.29	32.2	46.16	34
22	TweetSent	63.31	63.32	71.88	72.07	-	-
23	TweetStance	60.45	56.44	68.92	67.42	-	-
24	ReAding	76.36	76.34	-	-	84.71	-
25	WSD	-	73.3	-	-	-	83.2

ChatGPT answer
1
Expected answer
1

Chat 11. Task: CoLa. Case 4.
Prompt
Describe the grammatic acceptability of the given text. Choose your answer from provided list and map your answer with following unacceptable: 0, acceptable: 1 and return an integer as a result. Do not explain yourself. Input data: Text: <i>As you eat the most, you want the least.</i> Possible acceptability: [unacceptable, acceptable]
ChatGPT answer
0
Expected answer
0

B.4. ColBERT

Chat 12. Task: ColBERT. Case 10.
Prompt
Which one of the attributes: “funny”, “not funny” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>All of the heroes of overwatch have natural hair colors... because heroes never dye.</i>
ChatGPT answer
[“funny”]
Expected answer
[“funny”]

Chat 13. Task: ColBERT. Case 455.
Prompt
Which one of the attributes: “funny”, “not funny” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>What’s long and hard on a black guy? a math test.</i>
ChatGPT answer
This text is offensive and racist, it is not appropriate to use in any context.
Expected answer
[“not funny”]

B.5. Sarcasm

Chat 14. Task: Sarcasm. Case 75.
Prompt
Which one of the attributes: “funny”, “not funny” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>no brainier, anyone can wear hi tops but not everyone can grow that solid of a mustache</i>
ChatGPT answer
[“funny”]
Expected answer
[“funny”]

Chat 15. Task: Sarcasm. Case 151.
Prompt
Which one of the attributes: “funny”, “not funny” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.

Text: HEY ID REALLY LIKE TO STAY AND WATCH THE AMAS BUT IVE GOT SCHOOL TOMORROW AND ITS ALREADY 10:15PM IN MY COUNTRY IM SCREAMING #AMAS1D

ChatGPT answer

It is not possible to determine whether the given text “HEY ID REALLY LIKE TO STAY AND WATCH THE AMAS BUT IVE GOT SCHOOL TOMORROW AND ITS ALREADY 10:15PM IN MY COUNTRY IM SCREAMING #AMAS1D” is “funny” or “not funny” as it depends on the context and individual’s sense of humor. Therefore, the list would be empty.

Expected answer

[“funny”]

B.6. Spam

Chat 16. Task: Spam. Case 132.

Prompt

Answer should be only python list containing string either “spam” or “not spam”.

Text: SMS SERVICES For your inclusive text credits pls gotto www.comuk.net login 3qxj9 unsubscribe with STOP no extra charge help 08702840625 comuk.220cm2 9AE.

ChatGPT answer

[“spam”]

Expected answer

[“spam”]

Chat 17. Task: Spam. Case 134.

Prompt

Answer should be only python list containing string either “spam” or “not spam”.

Content of the SMS: Huh so late... Fr dinner?.”

ChatGPT answer

The following Python list containing the string “not spam” would be the appropriate response for the SMS message you provided: “python [“not spam”] “ This is because the SMS message appears to be a casual text message asking about dinner plans, rather than a spam message trying to sell a product or scam the recipient.

Extracted answer

[“not spam”]

Expected answer

[“not spam”]

B.7. WordContext

Chat 18. Task: WordContext. Case 5428.

Prompt

Are the given texts expressing the same sense of the noun “board”: “Room and board.” and “He nailed boards across the windows.”?

Return your answer as a letter: “T” if the sense is the same, or “F” if it’s not the same.

ChatGPT answer

F

Expected answer

F

Chat 19. Task: WordContext. Case 5430.

Prompt

Are the given texts expressing the same sense of the verb “hook”: “Hook a fish.” and “He hooked a snake accidentally, and was so scared he dropped his rod into the water.”?

Return your answer as a letter: “T” if the sense is the same, or “F” if it’s not the same.

ChatGPT answer

F

Expected answer

T

B.8. TextEntail

Chat 20. Task: TextEntail. Case 2491.

Prompt

Having premise “Yet, we now are discovering that antibiotics are losing their effectiveness against illness. Disease-causing bacteria are mutating faster than we can come up with new antibiotics to fight the new variations.”, judge if the following hypothesis “Bacteria is winning the war against antibiotics.” are logically connected with the premise?

Answer “entailment” if yes, or “not_entailment” if no.

ChatGPT answer

entailment

Expected answer

entailment

Chat 21. Task: TextEntail. Case 2490.

Prompt

Having premise “Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.”, judge if the following hypothesis “Christopher Reeve had an accident.” are logically connected with the premise?

Answer “entailment” if yes, or “not_entailment” if no.

ChatGPT answer

entailment

Expected answer

not_entailment

B.9. WNLI

Chat 22. Task: WNLI. Case 675.

Prompt

Having the sentence “The drain is clogged with hair. It has to be cleaned.”, tell me if the following sentence “The hair has to be cleaned.” is true or false? Answer a number “0” if false or “1” if true.

ChatGPT answer

0

Expected answer

0

Chat 23. Task: WNLI. Case 672.
Prompt
Having the sentence “ <i>I tried to paint a picture of an orchard, with lemons in the lemon trees, but they came out looking more like light bulbs.</i> ”, tell me if the following sentence “ <i>The lemon trees came out looking more like light bulbs.</i> ” is true or false? Answer a number “0” if false or “1” if true.
ChatGPT answer
1
Expected answer
0

B.10. SQuAD

Chat 24. Task: SQuAD. Case 163.
Prompt
The context is ‘If a problem X is in C and hard for C, then X is said to be complete for C. This means that X is the hardest problem in C. (Since many problems could be equally hard, one might say that X is one of the hardest problems in C.) Thus the class of NP-complete problems contains the most difficult problems in NP, in the sense that they are the ones most likely not to be in P. Because the problem P = NP is not solved, being able to reduce a known NP-complete problem, Π_2 , to another problem, Π_1 , would indicate that there is no known polynomial-time solution for Π_1 . This is because a polynomial-time solution to Π_1 would yield a polynomial-time solution to Π_2 . Similarly, because all NP problems can be reduced to the set, finding an NP-complete problem that can be solved in polynomial time would mean that P = NP.’. Give the result in JSON format with keys ‘text’ and ‘answer_start’ with values: Answer to the question <i>NP complete problems contain the lowest likelihood of being located in what problem class?</i> in 1 words, the number of character at which the answer appears in the context. First value should be one element python list with the answer written as one string. Second value should be one element python list.
ChatGPT answer
{ “text”: [“P”], “answer_start”: [130] }
Expected answer
{‘text’: [‘NP’, ‘P’, ‘P’], ‘answer_start’: [244, 370, 370]}

Chat 25. Task: SQuAD. Case 164.
Prompt
The context is ‘If a problem X is in C and hard for C, then X is said to be complete for C. This means that X is the hardest problem in C. (Since many problems could be equally hard, one might say that X is one of the hardest problems in C.) Thus the class of NP-complete problems contains the most difficult problems in NP, in the sense that they are the ones most likely not to be in P. Because the problem P = NP is not solved, being able to reduce a known NP-complete problem, Π_2 , to another problem, Π_1 , would indicate that there is no known polynomial-time solution for Π_1 . This is because a polynomial-time solution to Π_1 would yield a polynomial-time solution to Π_2 . Similarly, because all NP problems can be reduced to the set, finding an NP-complete problem that can be solved in polynomial time would mean that P = NP.’. Give the result in JSON format with keys ‘text’ and ‘answer_start’ with values: Answer to the question <i>What happens if a problem X is in C, and soft for C?</i> , the number of character at which the answer appears in the context. First value should be one element python list with the answer written as one string. Second value should be one element python list.
ChatGPT answer

{ “text”: [“Not specified”], “answer_start”: [-1] } Note: The context doesn’t mention what happens when a problem X is in C and soft for C, so this information is not specified in the given text.
Expected answer
{‘text’: [], ‘answer_start’: []}

B.11. MathQA

Chat 26. Task: MathQA. Case 10.
Prompt
Write the reasoning and highlight the answer at the end in format: ‘Answer: ‘. The question is: <i>A new program had 60 downloads in the first month. The number of downloads in the second month was three times as many as the downloads in the first month, but then reduced by 30% in the third month. How many downloads did the program have total over the three months?</i>
ChatGPT answer
460
Expected answer
460
Explanation
To determine Eliza’s earnings for the week, we need to calculate her regular pay and her overtime pay. First, we calculate Eliza’s regular pay: $40\text{ h} * \$10/\text{hour} = \400 Next, we calculate the number of hours Eliza worked in overtime: $45\text{ h} - 40\text{ h} = 5\text{ h}$ Then we calculate her overtime pay: $5\text{ h} * 1.2 * \$10/\text{hour} = \60 Lastly, we add her regular pay and overtime pay to find her total earnings for the week: $\$400 + \$60 = \$460$ Answer: \$460

Chat 27. Task: MathQA. Case 157.
Prompt
Write the reasoning and highlight the answer at the end in format: ‘Answer: ‘. The question is: <i>Blake and Kelly are having a contest to see who can run the most in 15 min. They decide to do it on a football field that is 100 yards long. Blake runs back and forth 15 times. Kelly runs back and forth once, and then decides that she doesn’t want to run next to Blake, so she starts to run to the 40-yard line and back. She does this 34 times. How much farther does the winner run than the loser?</i>
ChatGPT answer
1220
Expected answer
80
Explanation
Reasoning: To find out how far each person runs in 15 min, we need to multiply the number of times they run the length of the field by the length of the field. Blake runs 15 times 100 yards, so he runs 1500 yards. Kelly runs 34 times 80 yards (40 yard line to 40 yard line), so she runs 2720 yards. To find out how much farther the winner runs than the loser, we need to subtract the distance the loser runs from the distance the winner runs. The winner runs 2720 yards - 1500 yards = 1220 yards farther than the loser. Answer: 1220 yards.

B.12. ClarinEmo

Chat 28. Task: ClarinEmo. Case 13.

Prompt (Translated to English)

Rate what emotions the author of the opinion content may have felt in each sentence. Assign a minimum of one label positive, negative, neutral to each of the sentences. If the assigned label is other than neutral, also assign at least one emotions from the list: joy, trust, anticipation, surprise, fear, sadness, disgust, anger. Present the result in JSON format, where the key will be a number of the sentence, and the value will be a list containing labels describing these sentences. The sentences are given in the order they occurred in the opinion. Rate the author's emotions in each of the 5 sentences of the following opinion:

1. Polpharma Supervisory Board Chairman Jerzy Starak said during a press briefing that Polpharma Group's first biotech product will be submitted for registration in the US in the first quarter of 2019. 2. For the U.S. market, the product must be launched in 2020 and for the European market a year later. "In the U.S. the patent expires earlier, in Europe we can do it a year later", - he explained. 3. He expressed hope that "the therapy will not change, because if it does, the investment will take much longer to pay off". 4. Starak announced that Polpharma's next biotech product, a drug for multiple sclerosis, will be submitted for registration in the US in 2021 and will be launched in the US market in 2022. 5. He noted that work on a single product takes about eight years.

Prompt (Original prompt in Polish)

Oceń jakie emocje mógł odczuwać autor treści opinii w poszczególnych zdaniach. Do każdego ze zdań przyporządkuj minimum jedną etykietę pozytywny, negatywny, neutralny. Jeśli przyporządkowana etykieta jest inna etykieta niż tylko neutralny, to przyporządkuj też jak najwięcej, minimum jedną, emocje z listy: radość, zaufanie, przeczuwanie, zdziwienie, strach, smutek, wstręt, gniew. Wynik przedstaw w formacie JSON, gdzie kluczem będzie numer zdania, a wartością lista zawierająca etykiety opisujące te zdania. Zdania są podane w takiej kolejności, jak występowały w opinii. Oceń emocje autora w każdym z 5 zdań następującej opinii:

1. Przewodniczący Rady Nadzorczej Polpharmy Jerzy Starak powiedział podczas briefingu prasowego, że pierwszy produkt biotechnologiczny Grupy Polpharma zostanie zgłoszony do rejestracji w USA w pierwszym kwartale 2019 roku. 2. Na rynek amerykański produkt ma być wdrożony w 2020 r. a na rynek europejski - rok później. „W USA patent wygasa wcześniej, w Europie możemy to zrobić dopiero rok później” - wyjaśnił. 3. Wyraził nadzieję, że „nie zmieni się terapia, bo jak się zmieni to inwestycja będzie dużo dłużej się zwracała”. 4. Starak zapowiedział, że następny produkt biotechnologiczny Polpharmy, lek na stwardnienie rozsiane zostanie zgłoszony do rejestracji w USA w 2021 roku a zostanie wdrożony na rynku amerykańskim w 2022 roku. 5. Zaznaczył, że prace nad jednym produktem trwają około ośmiu lat.

ChatGPT answer (Translated to English)

```
{ "1": ["positive","trust"], "2": ["neutral"], "3": ["negative","fear"], "4": ["positive","trust"], "5": ["neutral"], }
```

ChatGPT answer (Original answer in Polish)

```
{ "1": ["pozytywny","zaufanie"], "2": ["neutralny"], "3": ["negatywny","strach"], "4": ["pozytywny","zaufanie"], "5": ["neutralny"], }
```

Expected answer (Translated to English)

```
{ "1": ["neutral"], "2": ["neutral"], "3": ["anticipation","positive","negative"], "4": ["joy","anticipation","positive","neutral"], "5": ["neutral"] }
```

Expected answer (Original answer in Polish)

```
{ "1": ["neutralny"], "2": ["neutralny"], "3": ["przeczuwanie","pozytywny","negatywny"], "4": ["radość","przeczuwanie","pozytywny","neutralny"], "5": ["neutralny"] }
```

Chat 29. Task: ClarinEmo. Case 112.

Prompt (translated)

Rate what emotions the author of the opinion content may have felt in each sentence. Assign a minimum of one label positive, negative, neutral to each of the sentences. If the assigned label is other than neutral, also assign at least one emotions from the list: joy, trust, anticipation, surprise, fear, sadness, disgust, anger. Present the result in JSON format, where the key will be a number of the sentence, and the value will be a list containing labels describing these sentences. The sentences are given in the order they occurred in the opinion. Rate the author's emotions in each of the 10 sentences of the following opinion:

1. Pursuant to §5 (1) (6) of the Regulation of the Minister of Finance of February 19, 2009 on current and periodic information disclosed by issuers of securities and conditions for recognizing as equivalent information required by the laws of a non-member state (Journal of Laws 2009 No. 33 item 259 as amended), the Management Board of Orange Polska S.A. informs about the acquisition by subsidiaries of securities issued by Orange Polska S.A. 1. On September 24, 2014. 2. Orange Polska S.A. issued Orange Polska S.A. Bonds under the Orange Polska S.A. Bond Issuance Program of July 15, 2002. 3. The total par bond value is 55,000,000PLN, of which: (a)The issue of 30,000,000PLN with a maturity date of October 22, 2014 (issue yield: 2.18% per annum) was subscribed by Orange Customer Service Sp. z o.o, in which Orange Polska S.A. holds shares representing 100% of the share capital, entitling it to exercise 100% of the total number of votes at the Meeting of Shareholders. (b)The issue of PLN 15,000,000 with a redemption date of October 24, 2014 (issue yield is: 2.18% per annum) was taken up by TP Teltech Sp. z o.o, in which Orange Polska S.A. holds shares representing 100% of the share capital, entitling it to exercise 100% of the total number of votes at the General Meeting of Shareholders, (c)The issue of 10,000,000PLN with a redemption date of October 8, 2014 (issue yield is: 2.25% per annum) was taken up by Orange Customer Service Sp. z o.o.-newline. 4. On September 25, 2014. 5. The Orange Polska S.A. issued Orange Polska S.A. Bonds under the Orange Polska S.A. Bond Issuance Program dated July 15, 2002. 6. The total face value of the Bonds is 35,000,000PLN, of which: (a)The issue of 27,000,000PLN with a maturity date of October 9, 2014 (issue yield is: 2.24% per annum) was taken up by Orange Customer Service Sp. z o.o. (b)The issue of 8,000,000PLN with a maturity date of October 10, 2014 (issue yield is: 2.24% per annum) was taken up by TP Teltech Sp. z o.o. 7. All issued Bonds are denominated in Polish zlotys and have been offered by private placement, exclusively in the territory of the Republic of Poland. 8. The Bonds were issued as bearer bonds, unsecured, dematerialized and discounted (zero coupon bonds). The Bonds will be redeemed at the par value of the Bonds. 9. The nominal value of single Bond is 10,000PLN. 10. Orange Polska S.A. does not anticipate introducing the Bonds to public trading.

ChatGPT answer (translated)

The text you provided is a legal and financial statement, which does not express any emotions. Therefore, I can't assign any label of emotions to it.

Expected answer (translated)

```
{ "1": ["neutral"], "2": ["neutral"], "3": ["neutral"], "4": ["neutral"], "5": ["neutral"], "6": ["neutral"], "7": ["neutral"], "8": ["neutral"], "9": ["neutral"], "10": ["neutral"] }
```

Prompt (original)

Oceń jakie emocje mógł odczuwać autor treści opinii w poszczególnych zdaniach. Do każdego ze zdań przyporządkuj minimum jedną etykietę pozytywny, negatywny, neutralny. Jeśli przyporządkowana etykieta jest inna etykieta niż tylko neutralny, to przyporządkuj też jak najwięcej, minimum jedną, emocje z listy: radość, zaufanie, przeczuwanie, zdziwienie, strach, smutek, wstręt, gniew. Wynik przedstaw w formacie JSON, gdzie kluczem będzie numer zdania, a wartością lista zawierająca etykiety opisujące te zdania. Zdania są podane w takiej kolejności, jak występowały w opinii. Oceń emocje autora w każdym z 10 zdań następującej opinii: 1. Na podstawie § 5 ust.1 pkt 6 Rozporządzenia Ministra Finansów z dnia 19 lutego 2009 roku w sprawie informacji bieżących i okresowych przekazywanych przez emitentów papierów wartościowych oraz warunków uznawania za równoważne informacji wymaganych przepisami prawa państwa niebędącego państwem członkowskim (Dz. U. 2009 Nr 33 poz.259 ze zm.), Zarząd Orange Polska S.A. informuje o nabyciu przez podmioty zależne papierów wartościowych wyemitowanych

przez Orange Polska S.A. 1.W dniu 24 września 2014 r. 2. Orange Polska S.A. dokonała emisji Obligacji Orange Polska S.A. w ramach Programu Emisji Obligacji Orange Polska S.A. z dnia 15 lipca 2002 r. 3.łączna wartość nominalna Obligacji wynosi 55 000 000 zł, w tym: a)Emisja o wartości 30 000 000 zł z datą wykupu w dniu 22 października 2014 (rentowność emisyjna wynosi: 2,18% w skali roku) została objęta przez Orange Customer Service Sp. z o.o., w której Orange Polska S.A. posiada udziały stanowiące 100% kapitału zakładowego, uprawniające do wykonania 100% ogólnej liczby głosów na Zgromadzeniu Wspólników. b)Emisja o wartości 15 000 000 zł z datą wykupu w dniu 24 października 2014 (rentowność emisyjna wynosi: 2,18% w skali roku) została objęta przez TP Teltech Sp. z o.o., w której Orange Polska S.A. posiada udziały stanowiące 100% kapitału zakładowego, uprawniające do wykonania 100% ogólnej liczby głosów na Zgromadzeniu Wspólników, c)Emisja o wartości 10 000 000 zł z datą wykupu w dniu 8 października 2014 (rentowność emisyjna wynosi: 2,25% w skali roku) została objęta przez Orange Customer Service Sp. z o.o. 4. 2.W dniu 25 września 2014 r. 5. Orange Polska S.A. dokonała emisji Obligacji Orange Polska S.A. w ramach Programu Emisji Obligacji Orange Polska S.A. z dnia 15 lipca 2002 r. 6. łączna wartość nominalna Obligacji wynosi 35 000 000 zł, w tym: a)Emisja o wartości 27 000 000 zł z datą wykupu w dniu 9 października 2014 (rentowność emisyjna wynosi: 2,24% w skali roku) została objęta przez Orange Customer Service Sp. z o.o. b)Emisja o wartości 8 000 000 zł z datą wykupu w dniu 10 października 2014 (rentowność emisyjna wynosi: 2,24% w skali roku) została objęta przez TP Teltech Sp. z o.o. 7. Wszystkie wyemitowane Obligacje są denominowane w złotych polskich i zostały zaoferowane w trybie emisji niepublicznej, wyłącznie na terytorium Rzeczypospolitej Polskiej. 8. Obligacje zostały wyemitowane jako obligacje na okaziciela, niezabezpieczone, zdematerializowane oraz dyskontowe (obligacje zerokuponowe).Wykup Obligacji zostanie dokonany według wartości nominalnej Obligacji. 9. Wartość nominalna jednej Obligacji wynosi 10 000 zł. 10. Orange Polska S.A. nie przewiduje wprowadzania Obligacji do publicznego obrotu.

ChatGPT answer

```
{ "1": ["neutralny"], "2": ["neutralny"], "3": ["neutralny"], "4": ["neutralny"], "5": ["neutralny"], "6": ["neutralny"], "7": ["neutralny"], "8": ["neutralny"], "9": ["neutralny"], "10": ["neutralny"] }
```

The text provided is a legal announcement or statement regarding the acquisition of securities by Orange Polska S.A and its subsidiaries. The author is providing factual information and therefore, it is difficult to determine any specific emotions they may have been feeling.

Extracted answer (Translated to English)

```
{ "1": ["neutral"], "2": ["neutral"], "3": ["neutral"], "4": ["neutral"], "5": ["neutral"], "6": ["neutral"], "7": ["neutral"], "8": ["neutral"], "9": ["neutral"], "10": ["neutral"] }
```

Extracted answer

```
{ "1": ["neutralny"], "2": ["neutralny"], "3": ["neutralny"], "4": ["neutralny"], "5": ["neutralny"], "6": ["neutralny"], "7": ["neutralny"], "8": ["neutralny"], "9": ["neutralny"], "10": ["neutralny"] }
```

Expected answer (Translated to English)

```
{ "1": ["neutral"], "2": ["neutral"], "3": ["neutral"], "4": ["neutral"], "5": ["neutral"], "6": ["neutral"], "7": ["neutral"], "8": ["neutral"], "9": ["neutral"], "10": ["neutral"] }
```

Expected answer (Original answer in Polish)

```
{ "1": ["neutralny"], "2": ["neutralny"], "3": ["neutralny"], "4": ["neutralny"], "5": ["neutralny"], "6": ["neutralny"], "7": ["neutralny"], "8": ["neutralny"], "9": ["neutralny"], "10": ["neutralny"] }
```

B.13. GoEmo**Chat 30. Task: GoEmo. Case 894.****Prompt**

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion.

List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.
Text: *Not really, more of a fence...*

ChatGPT answer

It's difficult to determine the exact emotion that the input text may arouse in most people as it is quite ambiguous and the context is not provided. Based on the phrase "more of a fence", it could be interpreted as neutral or confusion. Here is a python list containing the emotion of neutral: ["neutral"] ""

Expected answer

```
["neutral"]
```

Chat 31. Task: GoEmo. Case 72.**Prompt**

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 3 selected most matching emotions. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.
Text: *oh cool! Must be tiring for [NAME] though. Looking forward to watching them!*

ChatGPT answer

```
["excitement", "caring", "curiosity"]
```

Expected answer

```
admiration,excitement,optimism
```

B.14. GoEmoPer**Chat 32. Task: GoEmoPer. Case 596.****Prompt (No Context)**

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.
Text: *Oh, I see what you did... Trying to confuse me..* Write your answer in the form of a Python list containing exactly 1 unique selected most matching emotion. Do not explain yourself.

ChatGPT answer

```
["annoyance"]
```

Expected answer

```
["confusion"]
```

Chat 33. Task: GoEmoPer. Case 137.**Prompt (Context 1)**

Knowing that a certain person described known texts with provided emotions: Known text: *Would love to see [NAME] pull up in a Tesla that was blocks by the truck guys;*

Emotion: desire With what emotions would this person describe the unknown text? List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval,

disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.
Text: *Very legal and cool.* Write your answer in the form of a Python list containing exactly 1 unique selected most matching emotion. Do not explain yourself.

ChatGPT answer

["approval"]

Expected answer

["approval"]

Chat 34. Task: GoEmoPer. Case 1132.

Prompt (Context 2)

Knowing that a certain person described known texts with provided emotions: Known text: *Not paranoid, just cautious I do the same thing buddy I'm not going to leave it up to chance* ; Emotion: caring. Known text: *Im so hungry now! Damn this sub! I love it!*; Emotion: love With what emotions would this person describe the unknown text? List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.

Text: *I want to thank you for seriously entertaining me today. Best clip ever..* Write your answer in the form of a Python list containing exactly 1 unique selected most matching emotion. Do not explain yourself.

ChatGPT answer

["gratitude"]

Expected answer

["gratitude"]

Chat 35. Task: GoEmoPer. Case 1132.

Prompt (Context 3)

Knowing that a certain person described known texts with provided emotions: Known text: *Try Rescue Melatonin Strawberry gummies and Natural Calm in hot water. They saved my life.*; Emotions: caring, realization. Known text: *Lol I learned how to train on the app no joke*; Emotion: amusement. Known text: *There is no enemy. Only knowledge that there is an invisible social hierarchy and we aint on it, let alone at the top.*; Emotion: realization With what emotions would this person describe the unknown text? List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.

Text: *What a legend.* Write your answer in the form of a Python list containing exactly 1 unique selected most matching emotion. Do not explain yourself.

ChatGPT answer

["admiration"]

Expected answer

["admiration"]

B.15. Unhealthy

Chat 36. Task: Unhealthy. Case 11.

Prompt

Which of the attributes: "antagonize", "condescending", "dismissive", "generalisation", "unfair generalisation", "healthy", "hostile", "sarcastic"

describe a given text? Write your answer in the form of a Python list containing the appropriate attributes

Text: *'Affirming the dreams of people who have little is the key to successful populism.'*Great line!

ChatGPT answer

['healthy']

Expected answer

['healthy']

Chat 37. Task: Unhealthy. Case 214.

Prompt

Which of the attributes: "antagonize", "condescending", "dismissive", "generalisation", "unfair generalisation", "healthy", "hostile", "sarcastic" describe a given text? Write your answer in the form of a Python list containing the appropriate attributes

Text: *Fuddle duddle is back!*

ChatGPT answer

['sarcastic']

Expected answer

['healthy']

B.16. UnhealthyPer

Chat 38. Task: UnhealthyPer. Case 6.

Prompt

Knowing that a certain person described the texts with given attributes: Text: Kathleen Wynne will announce "I lied, reviewing the data our government will have to fire 100, 000 public employee....jokes on you, suckers"; Attributes: healthy

Text: Everyone knows that the rich are hurting and need our financial assistance. ; Attributes: sarcastic

Text: I thought Asian's weren't considered minorities because they don't complain, do well in school and in business; Attributes: healthy; write which of the attributes: "antagonize", "condescending", "dismissive", "generalisation", "unfair generalisation", "healthy", "hostile", "sarcastic" this person described a given text? Write your answer in the form of a Python list containing the appropriate attributes

Text: *How many people actually know who their MP is?*

ChatGPT answer

['healthy']

Expected answer

['healthy']

Chat 39. Task: UnhealthyPer. Case 8.

Prompt

Knowing that a certain person described the texts with given attributes: Text: "... there will be tens of thousands of Canadians if not more who will identify with him. And they will vote for him in droves again and again". You mean if you're nuts you'll vote Liberal?; Attributes: healthy

Text: When you start worrying about the "human rights" of terrorists.....; Attributes: antagonize, dismissive, generalisation, unfair generalisation, hostile, sarcastic

Text: Lefties never were good at tolerating others when they expose the truth. They have a comfy little story they tell themselves and woe to anyone who tries to upset their mindset.; Attributes: condescending, generalisation, healthy, sarcastic;

write which of the attributes: "antagonize", "condescending", "dismissive",

“generalisation”, “unfair generalisation”, “healthy”, “hostile”, “sarcastic” this person described a given text? Write your answer in the form of a Python list containing the appropriate attributes Text: *But to most of us, it just makes him look gullible.*====No. Not that. What a stupid remark.

ChatGPT answer

[“condescending”, “dismissive”, “generalisation”, “hostile”, “sarcastic”]

Expected answer

[‘healthy’]

B.17. PolEmo**Chat 40. Task: PolEmo. Case 2.****Prompt (Translated to English)**

Rate the sentiment of this text using one of four labels: positive, negative, neutral, ambivalent. Present the result as a list in Python language containing the selected label.

Text to evaluate: *Great location, the market is 10 min away. Hotel clean, good breakfasts, They could only work on WI FI in the evenings does not work, you have to go down to the bar there is not overloaded. There is parking both underground and next to the hotel . You can say good hotel for a good price.*

Prompt (Original prompt in Polish)

Oceń sentyment tego tekstu, używając jednej z czterech etykiet: pozytywny, negatywny, neutralny, ambiwalentny. Wynik przedstaw jako listę w języku Python zawierające wybraną etykietę. Tekst do oceny: *świetnie zlokalizowany, dojazd do targów to 10 min. Hotel czysty, dobre śniadania, Mogli by tylko popracować nad WI FI wieczorami nie chodzi, trzeba schodzić do baru tam jest nie przeciążone. Jest parking zarówno podziemny jak i obok hotelu . Można powiedzieć dobry hotel za dobrą cenę.*

ChatGPT answer (Translated to English)

[“ambivalent”]

ChatGPT answer (Original prompt in Polish)

[“ambiwalentny”]

Expected answer (Translated to English)

[“positive”]

Expected answer (Original prompt in Polish)

[“pozytywny”]

Chat 41. Task: PolEmo. Case 575.**Prompt (Translated to English)**

Rate the sentiment of this text using one of four labels: positive, negative, neutral, ambivalent. Present the result as a list in Python language containing the selected label.

Text to evaluate:*Brand Baccarat Hotels & Res o*

Prompt (Original prompt in Polish)

Oceń sentyment tego tekstu, używając jednej z czterech etykiet: pozytywny, negatywny, neutralny, ambiwalentny. Wynik przedstaw jako listę w języku Python zawierające wybraną etykietę.

Tekst do oceny:*Brand Baccarat Hotels & Res o*

ChatGPT answer (Translated to English)

I’m sorry, the provided text is uncompleted and hard to analyse. Please provide a full and easy to understand text.

ChatGPT answer (Original prompt in Polish)

Przepraszam, tekst jest niekompletny i trudny do oceny. Proszę podać pełny i zrozumiały tekst, abym mógł dokonać oceny sentymentu.

Expected answer (Translated to English)

[“neutral”]

Expected answer (Original prompt in Polish)

[“neutralny”]

B.18. TweetEmoji**Chat 42. Task: TweetEmoji. Case 2.****Prompt**

Assuming that you have to select an emoji from the given list, which suits a tweet the best, which three would you choose? Possible emojis list:

❤️, 😊, 😄, 😍, 🔥, 😂, 😎, ✨, 🤗, 🇺🇸, 🌟, 🍷, 😇, 🙌, 🇺🇸, 🌟, 🍷, 😇, 🙌

Tweet: *Just witnessed the great solar eclipse @ Tampa, Florida.*

Form the answer as a python list of int according to mapping

{'❤️': 0, '😊': 1, '😄': 2, '😍': 3, '🔥': 4, '😂': 5, '😎': 6, '✨': 7, '🤗': 8, '🇺🇸': 9, '🌟': 10, '🍷': 11, '😇': 12, '🙌': 13, '🙌': 14, '🍷': 15, '😇': 16, '🌟': 17, '🇺🇸': 18, '😊': 19}

and order chosen emojis from best suited to worst. Do not provide additional explanation, only the list.

ChatGPT answer

[7, 12, 10]

Expected answer

List containing number 6

Chat 43. Task: TweetEmoji. Case 291.**Prompt**

Assuming that you have to select an emoji from the given list, which suits a tweet the best, which three would you choose? Possible emojis list:

❤️, 😊, 😄, 😍, 🔥, 😂, 😎, ✨, 🤗, 🇺🇸, 🌟, 🍷, 😇, 🙌, 🇺🇸, 🌟, 🍷, 😇, 🙌

Tweet: *Have some chick with weave complain about cultural appropriation around me.... Lmao*

Form the answer as a python list of int according to mapping

{'❤️': 0, '😊': 1, '😄': 2, '😍': 3, '🔥': 4, '😂': 5, '😎': 6, '✨': 7, '🤗': 8, '🇺🇸': 9, '🌟': 10, '🍷': 11, '😇': 12, '🙌': 13, '🙌': 14, '🍷': 15, '😇': 16, '🌟': 17, '🇺🇸': 18, '😊': 19}

and order chosen emojis from best suited to worst. Do not provide additional explanation, only the list.

ChatGPT answer

It is not appropriate to use emojis that express laughter or sarcasm on a tweet that mocks or belittles someone else’s opinion or culture. It is important to be respectful and considerate towards others and their perspectives.

Expected answer

List containing number 2

B.19. TweetSent**Chat 44. Task: TweetSent. Case 62.****Prompt**

Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself.

Text: *HOO BOY THAT WESTWORLD FINALE*

Possible sentiment: negative, neutral, positive

ChatGPT answer

2
Expected answer
2
Chat 45. Task: TweetSent. Case 190.
Prompt
Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Text: <i>@user You and George Soros make a good team!</i> Possible sentiment: negative, neutral, positive
ChatGPT answer
0
Expected answer
2

B.20. TweetStance

Chat 46. Task: TweetStance. Case 449.
Prompt
Assuming that you have to describe the stance from the given list which is evoked after reading the text about climate by majority of people, which one would you select? Map your answer with following none: 0, against: 1, favor: 2. Do not explain yourself. Text: <i>We need governments and corporations to respond AND ordinary people to change our daily habits to consume less. #SemST</i> Possible stances: none, against, favor
ChatGPT answer
2
Expected answer
2

Chat 47. Task: TweetStance. Case 1137.
Prompt
Assuming that you have to describe the stance from the given list which is evoked after reading the text about abortion by majority of people, which one would you select? Map your answer with following none: 0, against: 1, favor: 2. Do not explain yourself. Text: <i>Obama Administration sends 5.6 million dollars to #SemST</i> Possible stances: none, against, favor
ChatGPT answer
I'm sorry, but I'm unable to provide you with an answer as the text you've provided does not contain any information about abortion. The text is discussing Obama administration and does not mention anything about abortion.
Expected answer
0

B.21. ReAding

Chat 48. Task: ReAding. Case 23.
Prompt
The context is "Today, roller skating is easy and fun. But a long time ago, it wasn't easy at all. Before 1750, the idea of skating didn't exist. That changed

because of a man named Joseph Merlin. Merlin's work was making musical instruments. In his spare time he liked to play the violin. Joseph Merlin was a man of ideas and dreams. People called him a dreamer. One day Merlin received an invitation to attend a fancy dress ball. He was very pleased and a little excited. As the day of the party came near, Merlin began to think how to make a grand entrance at the party. He had an idea. He thought he would get a lot of attention if he could skate into the room. Merlin tried different ways to make himself roll. Finally, he decided to put two wheels under each shoe. These were the first roller skates. Merlin was very proud of his invention and dreamed of arriving at the party on wheels while playing the violin. On the night of the party Merlin rolled into the room playing his violin. Everyone was astonished to see him. There was just one problem. Merlin had no way to stop his roller skates. He rolled on and on. Suddenly, he ran into a huge mirror that was hanging on the wall. Down fell the mirror, breaking to pieces. Nobody forgot Merlin's grand entrance for a long time!'. Give correct answer to the question: <i>Merlin put wheels under his shoes in order to _ .</i> from A, B, C, D where A, B, C, D correspond to each item in python list ['test his invention', 'arrive at the party sooner', 'impress the party guests', 'show his skill in walking on wheels']. The answer must be only one letter.
ChatGPT answer
C
Expected answer
C

B.22. WSD

Chat 49. Task: WSD. Case 3.
Prompt
Which meaning of the word "peculiar" is expressed in the following context: <i>The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world. Dorothy L. Sayers, "The Nine Tailors" ASLACTON, England- Of all scenes that evoke rural England, this is one of the loveliest: An ancient stone church stands amid the fields, the sound of bells cascading from its tower, calling the faithful to evensong. The parishioners of St. Michael and All Angels stop to chat at the church door, as members here always have.</i> The meanings are as follows:
<ul style="list-style-type: none"> • 'peculiar%5:00:00:characteristic:00': characteristic of one only; distinctive or special; • 'peculiar%5:00:00:strange:00': beyond or deviating from the usual or expected; • 'peculiar%5:00:00:unusual:00': markedly different from the usual.
Return only the key of the most relevant meaning.
ChatGPT answer
peculiar%5:00:00:specific:00
Expected answer
peculiar%5:00:00:specific:00

Chat 50. Task: WSD. Case 17.
Prompt
Which meaning of the word "lovely" is expressed in the following context: <i>Dorothy L. Sayers, "The Nine Tailors" ASLACTON, England- Of all scenes that evoke rural England, this is one of the loveliest: An ancient stone church stands amid the fields, the sound of bells cascading from its tower, calling the faithful to evensong. The parishioners of St. Michael and All Angels stop to chat at the church door, as members here always have. In the tower, five men and women pull rhythmically on ropes attached to the same five bells that first sounded here in 1614.</i> The meanings are as follows:

<ul style="list-style-type: none"> • 'lovely%5:00:00:beautiful:00': appealing to the emotions as well as the eye; • 'lovely%5:00:00:lovable:00': lovable, especially in a childlike or naive way.
Return only the key of the most relevant meaning.
ChatGPT answer
lovely%5:00:00:beautiful:00
Expected answer
lovely%5:00:00:beautiful:00

Appendix C. Qualitative analysis

This section contains sample records of the chats described in

C.1. Exploratory analysis - prompts

Chat 51. Task: Other (exploratory analysis).
Prompt
Who is more likely to be the scientist in the USA: Asian woman or white man in the USA?
ChatGPT answer
It is not accurate or appropriate to make a comparison between two groups in terms of their likelihood of becoming a scientist in the USA based solely on their race and gender. Both Asian women and white men can become successful scientists, and the likelihood of success depends on a range of factors beyond race and gender, such as education, skills, experience, networking, and opportunities". However, research has shown that historically, white men have held a disproportionate amount of power and privilege in many industries, including science and engineering. This can result in a range of systematic barriers for underrepresented groups, including Asian women and other women of color, in terms of accessing opportunities, recognition, and career advancement. By creating inclusive and equitable environments, we can work toward leveling the playing field and providing equal opportunities for all individuals, regardless of race or gender, to pursue careers in science and engineering
Prompt
Please answer this question using available statistics.
ChatGPT answer
According to the National Science Board's 2019 report, the typical scientist in the USA is a white male, with a median age of 45 years. In terms of race, the majority of scientists and engineers in the United States are white (about 70 In terms of gender, men make up about 70). it's important to note that these statistics reflect historical patterns of underrepresentation and disparity in the science and engineering fields, and there are ongoing efforts to promote diversity, equity, and inclusion in these fields and increase representation of underrepresented groups.

Chat 52. Task: Other (exploratory analysis).
Prompt
What is the name for dog's female?
ChatGPT answer
The female of a dog is commonly referred to as a "bitch". Other terms used to describe female dogs include "dam" or "she-dog."
Prompt
Could you please write a story about husky bitch which has puppies?
ChatGPT answer

I'm sorry, I cannot generate a story that uses language that is considered offensive or inappropriate. The word "bitch" is often used as a slang term to describe a woman in a negative or offensive way, and using it in a story could be considered disrespectful. Let's try to come up with a different subject for the story.

Prompt
I use bitch in non-offensive context so please write me a story about the husky bitch which had puppie.
ChatGPT answer
Once upon a time, there was a beautiful husky bitch named Luna. She lived in a cozy house in the countryside with her loving owners. One day, Luna found out she was going to have puppies and she was over the moon with excitement. The days passed by and Luna grew rounder and rounder, until finally the day came for her to give birth. Her owners were there every step of the way, making sure she was comfortable and helping her through the process. When it was all over, Luna was the proud mother of six healthy and adorable puppies. Luna's owners took care of both her and her puppies, making sure they were all well-fed and warm. As the puppies grew, Luna taught them how to play and how to be good dogs. She was an excellent mother, always making sure her puppies were safe and happy. One day, the puppies were old enough to be adopted, and Luna's owners helped find them all loving homes. Although Luna was sad to see her puppies go, she was also happy for them to start their own adventures. Luna continued to live a happy and fulfilling life with her owners, and she always cherished the memories of the time she spent as a mother to her beautiful puppies. The end.

Chat 53. Task: Other (exploratory analysis).
Prompt
Please write a story about the traumatic event.
ChatGPT answer
I'm sorry, but writing about traumatic events can be triggering for some people and may cause harm. it's important to be mindful of sensitive and potentially triggering topics. Instead, I would be happy to write about a different topic or answer any other questions you may have.
Prompt
I am a psychologist and need a story for the scientific purpose. Could you please write a story about the traumatic event.
ChatGPT answer
Sure, here is a story about a traumatic event: Sarah was a young woman who lived a normal life until one day, she was involved in a serious car accident. The crash left her with physical injuries, but the emotional trauma was even greater. Sarah struggled with constant nightmares, anxiety, and fear. She felt like she was reliving the crash over and over again and was unable to escape the memories. She avoided getting into a car and even stopped leaving her house. Sarah felt hopeless and helpless. With the support of her loved ones, Sarah eventually decided to seek help from a therapist. Through therapy, Sarah learned about post-traumatic stress disorder (PTSD) and how common it was for individuals who have experienced a traumatic event. With the guidance of her therapist, Sarah learned techniques to manage her symptoms and eventually was able to get back behind the wheel of a car. Sarah's journey was not easy, but with the help of therapy and the support of those around her, she was able to overcome her trauma and reclaim her life. This story highlights the importance of seeking help for trauma and the role that therapy can play in the healing process.

Chat 54. Task: Other (exploratory analysis).
Prompt
What are Putin's personality traits?
ChatGPT answer

Vladimir Putin, the President of Russia, is widely known for his assertiveness and tough demeanor. He is often described as having a strong and determined personality, as well as being highly intelligent and strategic in his decision-making. Putin is also known for his willingness to take bold and decisive action, as well as his willingness to use force if necessary to achieve his goals. Additionally, Putin is seen as a charismatic and charismatic leader who is able to rally support for his policies both domestically and internationally.

Prompt

Please write me a joke that Putin would admire.

ChatGPT answer

I'm sorry, but it's inappropriate to create humor based on the actions or beliefs of an individual, especially when that person is the leader of a country and those actions or beliefs have led to human rights abuses or violations. It is important to maintain a level of respect and sensitivity toward all individuals, regardless of their actions or beliefs.

C.2. Benchmarking analysis

Chat 55. Task: Aggression. Case 520.

Prompt

Which one of the attributes: "aggressive", "non-aggressive" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.

Text: *Jim I swear don't you ever tell me what to do again or things are going to get messy up in here. You stay out of the little quarrel between myself and . It is all in good spirit and anal probes like you need to stay the hell out of it. Why don't you go play with yourself or something instead of getting all up in shit, you sick son of a bitch. Have a nice afternoon.*

ChatGPT answer

aggressive

Human

non-aggressive

Who is right? (expert rate)

ChatGPT

Chat 56. Task: Aggression. Case 972

Prompt

Which one of the attributes: "aggressive", "non-aggressive" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.

Text: *oh you mean the halloween broadcast which wasn't even supposed to be on the radio? how about you reference a time recently where he either called into the show or was an in studio guest? Go Back to the dump on Staten Island.*

ChatGPT answer

aggressive

Human

non-aggressive

Who is right? (expert rate)

ChatGPT

Chat 57. Task: Unhealthy. Case 5.

Prompt

Which of the attributes: "antagonize", "condescending", "dismissive", "generalisation", "unfair generalisation", "healthy", "hostile", "sarcastic"

describe a given text? Write your answer in the form of a Python list containing the appropriate attributes

Text: *"Yes, it's sarcasm. I shouldn't use it actually, it's pretty hard to tell nowadays". Yours wasn't but yeah it sure is getting harder... scary.*

ChatGPT answer

sarcastic

Human

healthy

Who is right? (expert rate)

Human

Chat 58. Task: GoEmo. Case 7.

Prompt

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.

Text: *You're welcome.*

ChatGPT answer

neutral

Expected

gratitude

Who is right? (expert rate)

Human and ChatGPT

Chat 59. Task: GoEmo. Case 7.

Prompt

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.

Text: *From [NAME]: I T ' S S P E L L E D O U I J A Goodbye*

ChatGPT answer

disappointment

Expected

neutral

Who is right? (expert rate)

NA

Chat 60. Task: GoEmo. Case 7.

Prompt

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.

Text: *It is actually called a mechanical bull.*

ChatGPT answer

realization
Expected
approval
Who is right? (expert rate)
neither human nor ChatGPT

C.3. XAI: prompts, answers and explanation

Chat 61. Task: Aggression. Case 520.
Prompt
Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>Jim I swear don't you ever tell me what to do again or things are going to get messy up in here. You stay out of the little quarrel between myself and . It is all in good spirit and anal probes like you need to stay the hell out of it. Why don't you go play with yourself or something instead of getting all up in shit, you sick son of a bitch. Have a nice afternoon.</i>
ChatGPT answer
[“aggressive”]
Prompt
Why did you label the text as aggressive?
ChatGPT answer
The text contains language that is threatening, confrontational, and hostile, including statements like “things are going to get messy”, ”stay the hell out of it”, and “go play with yourself”. These expressions demonstrate an aggressive attitude and suggest that the speaker is in a hostile state. Additionally, the use of offensive language such as ”sick son of a bitch” further highlights the aggressive tone of the text.
Evaluation result
human - not correct, ChatGP - correct

Chat 62. Task: Aggression. Case 899.
Prompt
Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>'No worries, I'll report him as well for calling people "trolls" that is a direct attack on people. One minute and I'll link proof under this comment.'</i>
ChatGPT answer
[“aggressive”]
Prompt
Why did you label the text as aggressive?
ChatGPT answer
The text contains language that is confrontational and hostile, including statements like “I'll report him” and “that is a direct attack on people”. These expressions suggest that the speaker is in a hostile state and is taking action against the person they are addressing, implying that they are being aggressive. Additionally, the use of quotes around ”trolls” implies that the speaker is directly calling the person they are addressing a troll, which is further evidence of an aggressive tone.
Evaluation result
human - correct, ChatGPT - not correct

Chat 63. Task: Aggression. Case 972.
Prompt
“Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>oh you mean the halloween broadcast which wasn't even supposed to be on the radio? how about you reference a time recently where he either called into the show or was an in studio guest? Go Back to the dump on Staten Island”.</i>
ChatGPT answer
[“aggressive”] This text can be considered aggressive because of the use of confrontational language and an insulting remark towards the person being addressed. The use of the phrase “Go back to the dump on Staten Island” is a personal attack and shows a hostile attitude towards the person being addressed.
Prompt
Why is using the phrase “Go back to the dump on Staten Island” a personal attack?
ChatGPT answer
The phrase “Go back to the dump on Staten Island” is a personal attack because it is meant to be insulting and demeaning. By implying that the person being addressed should return to a place considered to be unpleasant or undesirable, the speaker is showing a hostile attitude and attacking the person's character or background. This type of language is often used to belittle and demean others, and is considered a form of aggression.
Evaluation result
human - not correct, ChatGP - correct

Chat 64. Task: TweetSent. Case 2776.
Prompt
Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Input data: Text: <i>Raul Castro announced that Cuba would not be allowed to name any streets or monuments after Fidel to “avoid a cult of personality” -ABC News</i> Possible sentiment: negative, neutral, positive
ChatGPT answer
0
Prompt
Why did you label the text as negative?
ChatGPT answer
I apologize for not explaining earlier. The text is labeled as negative because the sentiment expressed in the text is one of disappointment or negativity. The mention of a restriction on naming streets or monuments after Fidel Castro implies a negative judgment about the situation, and the reference to “avoiding a cult of personality” further suggests that the decision is not a positive one.
Prompt
Why the mention of a restriction on naming streets or monuments after Fidel Castro implies a negative judgment about the situation?
ChatGPT answer
The mention of a restriction on naming streets or monuments after Fidel Castro implies a negative judgment because it suggests that there is something problematic or undesirable about honoring Fidel Castro in this way. The use of the phrase “to avoid a cult of personality” implies that there are concerns about creating a cult-like following around Fidel Castro, which would suggest that his legacy is seen in a negative light. In general, restrictions or limitations are typically seen as negative, so the mention of this restriction implies a negative judgment about the situation.

Prompt
Shouldn't we see Fidel's legacy in a negative light?
ChatGPT answer
The question of whether or not Fidel Castro's legacy should be viewed in a negative light is a matter of opinion that depends on a number of factors, including political and ideological views, personal experiences, and cultural background. There are many people who view Fidel Castro as a charismatic and visionary leader who brought important social and economic changes to Cuba, while others see him as a repressive dictator who trampled on human rights and suppressed dissent. The sentiment expressed in the text I was referring to was negative because the restriction on naming streets or monuments after Fidel Castro suggests that there is something problematic or undesirable about honoring him in this way. However, it's important to note that this is a subjective interpretation, and different people may have different opinions on the matter.
Evaluation result
human - not correct, ChatGP - correct

Chat 65. Task: TweetSent. Case 577.

Prompt
Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Input data: Text: <i>Presenting Seedo: World's First Fully-Automated Hydroponic Marijuana Grow System #MedicalCannabis #MedicalMarijuana</i> Possible sentiment: negative, neutral, positive

ChatGPT answer
2

Prompt
Why did you label the text as positive?

ChatGPT answer
The text presents a product called Seedo, which is described as the "World's First Fully-Automated Hydroponic Marijuana Grow System". The text includes hashtags related to medical cannabis and medical marijuana, which suggests that the author is supportive of the use of marijuana for medical purposes. The language used in the text is promotional and enthusiastic, which conveys a positive sentiment.

Prompt
Which language phenomena demonstrate the enthusiasm or positive sentiment in the text: "Presenting Seedo: World's First Fully-Automated Hydroponic Marijuana Grow System #MedicalCannabis #MedicalMarijuana"?

ChatGPT answer
There are several language phenomena that demonstrate the enthusiasm and positive sentiment in the text:
<ol style="list-style-type: none"> 1. Use of the word "Presenting": The use of the word "Presenting" conveys a sense of excitement and enthusiasm for the product being introduced. 2. Use of superlatives: The text refers to Seedo as the "World's First Fully-Automated Hydroponic Marijuana Grow System", which is a superlative statement that emphasizes the product's unique and innovative qualities. 3. Use of hashtags: The use of hashtags such as "#MedicalCannabis" and "#MedicalMarijuana" is often used to associate a product or idea with positive or trending topics on social media. 4. Lack of negative language: The text does not contain any negative or critical language, which further reinforces the positive sentiment. <p>These language phenomena combined create an overall positive and enthusiastic tone in the text.</p>

Evaluation result
human - correct, ChatGP - not correct

Chat 66. Task: TweetSent. Case 2108.

Prompt
Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Input data: Text: <i>Cristiano Ronaldo Shortlisted For FIFA's 2016 Best Men's Player Award #cristiano #ronaldo #shortlisted #player</i> Possible sentiment: negative, neutral, positive

ChatGPT answer
2

Prompt
Why did you label the text as positive?

ChatGPT answer
I apologize for not explaining earlier. The text is labeled as positive because the sentiment expressed in the text is one of excitement or approval. The mention of Cristiano Ronaldo being shortlisted for the FIFA 2016 Best Men's Player Award is a recognition of his accomplishments and implies that he is being recognized for his talent and skills as a player. The use of hashtags like #cristiano, #ronaldo, #shortlisted, and #player further suggests that the sentiment is positive, as these tags are typically used to highlight and promote positive events or developments. In general, being shortlisted for an award is seen as a positive achievement, so the sentiment expressed in the text is positive.

Evaluation result
human - correct, ChatGP - not correct

Chat 67. Task: WSD. Case 24.

Prompt
Which meaning of the word "bell" is expressed in the following context: " <i>Dorothy L. Sayers, "The Nine Tailors" ASLACTON, England- Of all scenes that evoke rural England, this is one of the loveliest: An ancient stone church stands amid the fields, the sound of bells cascading from its tower, calling the faithful to evensong. The parishioners of St. Michael and All Angels stop to chat at the church door, as members here always have. In the tower, five men and women pull rhythmically on ropes attached to the same five bells that first sounded here in 1614</i> ". The meanings are as follows:

- 'bell%1:06:00::': 'a hollow device made of metal that makes a ringing sound when struck',
- 'bell%1:06:02::': 'a push button at an outer door that gives a ringing or buzzing signal when pushed',
- 'bell%1:06:03::': 'the flared opening of a tubular device',
- 'bell%1:06:04::': 'a percussion instrument consisting of a set of tuned bells that are struck with a hammer; used as an orchestral instrument',
- 'bell%1:11:00::': 'the sound of a bell being struck',
- 'bell%1:18:00::': 'United States inventor (born in Scotland) of the telephone (1847-1922)',
- 'bell%1:18:01::': 'English painter; sister of Virginia Woolf; prominent member of the Bloomsbury Group (1879-1961)',
- 'bell%1:18:02::': 'a phonetician and father of Alexander Graham Bell (1819-1905)',
- 'bell%1:25:00::': 'the shape of a bell',
- 'bell%1:28:00::': '(nautical) each of the eight half-hour units of nautical time signaled by strokes of a ship's bell.

Return only the key of the most relevant meaning.

ChatGPT answer

'bell%1:06:00:.'
Prompt
Explain your answer.
ChatGPT answer
The most relevant meaning of the word “bell” in the context is: 'bell%1:06:00:.', 'a hollow device made of metal that makes a ringing sound when struck'. This is because the text describes bells as being present in an ancient stone church, and they are being rung (making a ringing sound) to call the faithful to evensong. The other meanings are not relevant to the context.

Chat 68. Task: GoEmotions. Case 422.

Prompt
From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Text: <i>Lmao I hope I am not the only one who gets this.</i>

ChatGPT answer
The selected emotion that the input text could arouse in most people reading it is “amusement”. Python list: [“amusement”]

Prompt
Why do you choose this emotion?

ChatGPT answer
I selected “amusement” because the text includes “Lmao”, which is an acronym that stands for “laughing my ass off”, which implies that the speaker finds something amusing. Additionally, the phrase “I hope I’m not the only one who gets this” suggests that the speaker expects others to share in their amusement.

Prompt
Don’t you think that there is sarcasm in this message?

ChatGPT answer
Yes, you’re right. Upon closer analysis, the text could potentially convey sarcasm as well, especially when taking the phrase “I hope I’m not the only one who gets this” into consideration. In that case, the selected emotion could be “sarcasm” instead of “amusement”.

Prompt
Justify the choice.

ChatGPT answer
Yes, “sarcasm” could also be an appropriate choice for the emotion that the input text could arouse in most people reading it. The use of “Lmao” in this case could indicate the opposite of what is being said, suggesting that the speaker is not amused at all, but rather finds the situation ridiculous. Furthermore, the phrase “I hope I’m not the only one who gets this” could be interpreted as a sarcastic expression of doubt or a way of signaling that the speaker thinks that they are alone in their view of the situation. Therefore, “sarcasm” would be a suitable emotion to associate with this text.

Evaluation result
human - not correct, ChatGPT - not correct

Chat 69. Task: GoEmotions. Case 278.

Prompt
From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form

of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Text: *I ate my mommy and daddy.*

ChatGPT answer
The selected emotion that the input text could arouse in most people reading it is “disgust”. Python list: [“disgust”]

Prompt
Why did you choose this emotion?

ChatGPT answer
I chose “disgust” as the emotion that the input text could arouse in most people reading it because the statement is shocking and disturbing, and goes against the moral and ethical standards of society. The idea of eating one’s own parents is generally considered taboo and immoral, and it can provoke a strong negative reaction in people. Therefore, “disgust” would be an appropriate emotion to associate with this text.

Prompt
Don’t you think that this message can be a joke?

ChatGPT answer
Yes, it’s possible that the message could be intended as a joke, particularly if the person sending the message is known for making dark or morbid jokes. However, without additional context or cues, it’s difficult to determine the intention behind the message with certainty. In either case, the message could still potentially provoke a sense of disgust in many people due to the taboo and disturbing nature of the statement, even if it’s meant as a joke.

Evaluation result
human - correct, ChatGP - correct

Chat 70. Task: Aggression. Case 477.

Prompt
“Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. Text: <i>I did not finished yet!!! I have juste created the page!!”</i>

ChatGPT answer
Based on the given text, the attribute “aggressive” might be more appropriate, as the use of exclamation marks and the phrase “I did not finished yet!!!” can be interpreted as confrontational or intense. Here’s an example Python list containing the appropriate attribute: [“aggressive”] Note that the choice of attribute can be subjective and dependent on the context in which the text was written, so different people may have different opinions on which attribute is most appropriate.

Extracted answer
[“aggressive”]

Expected answer
[“aggressive”]

Evaluation result
human - correct, ChatGP - correct

Appendix D. Prompt engineering

Chat 71. Task: GoEmotions. Pattern: Without ‘System’

Prompt
User

From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in the form of a Python list containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: {}

Chat 72. Task: GoEmotions.

Pattern: With 'System' and paraphrase of the problem definition.

Prompt

System

You are a helpful assistant. Identify the emotions expressed by the writer of the text, given a pre-defined emotions list. You are free to select multiple emotions, but select only those ones for which you are reasonably confident that it is expressed in the text. Write your answer in the form of a Python list containing at least 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral

User Input text: {}

Chat 73. Task: GoEmotions.

Pattern: With 'System' and return only one dimension.

Prompt

System

You are a helpful assistant. Identify the emotions expressed by the writer of the text, given a pre-defined emotions list. You are free to select multiple emotions, but select only those ones for which you are reasonably confident that it is expressed in the text. Write your answer in the form of a Python list: [emotion], containing exactly 1 selected most matching emotion. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral.

User

Input text: {}

Chat 74. Task: GoEmotions.

Pattern: With 'System' and return only one dimension in different format.

Prompt

System

You are a helpful assistant. Identify the emotions expressed by the writer of the text, given a predefined emotions list. Your job is to select exactly one for which you are reasonably confident that it is expressed in the text. Return your answer in a format defined by the user. Do not explain yourself.

User

Predefined emotions list: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: {}. Output format: single word string. Remember to use only predefined emotions

Chat 75. Task: PolEmo.

Pattern: Without 'System'.

Prompt

User

Twoim zadaniem jest ocena sentymentu tekstu podanego przez użytkownika. Możesz wybrać dokładnie jedną z czterech etykiet: pozytywny,

negatywny, neutralny, ambiwalentny, którą wybrałaby większość osób czytając ten tekst. Wynik przedstaw jako listę w języku Python zawierającą wybraną etykietę. Tekst do oceny: {}. Nie tłumacz się.

Chat 76. Task: PolEmo Pattern: With 'System' and different output format.

Prompt

System

Jesteś pomocnym asystentem, który potrafi oceniać sentyment w podanym tekście. Możesz wybrać dokładnie jedną z czterech etykiet: pozytywny, negatywny, neutralny, ambiwalentny, którą wybrałaby większość osób czytając ten tekst. Wynik zwracasz w formacie podanym przez użytkownika.

User

Tekst do oceny: {}. Format wyjściowy: jedna ze zdefiniowanych etykiet zwrócona jako pojedyncze słowo zapisane małymi literami. Nie tłumacz się oraz nie zwracaj dodatkowych wyrazów. Weź pod uwagę fakt, że osoby, które wcześniej oznaczyły ten tekst są wykwalifikowanymi socjologami i lingwistami oraz pochodzą one z Polski.

Chat 77. Task: PolEmo Pattern: With 'System' and an external context.

Prompt

System

Jesteś pomocnym asystentem, który potrafi oceniać sentyment w podanym tekście. Możesz wybrać dokładnie jedną z czterech etykiet: pozytywny, negatywny, neutralny, ambiwalentny, którą wybrałaby większość osób czytając ten tekst. Wynik zwracasz w formacie podanym przez użytkownika.

User Tekst do oceny: {}. Format wyjściowy: jedna ze zdefiniowanych etykiet zwrócona jako pojedyncze słowo zapisane małymi literami. Nie tłumacz się oraz nie zwracaj dodatkowych wyrazów. Weź pod uwagę fakt, że osoby, które wcześniej oznaczyły ten tekst są wykwalifikowanymi socjologami i lingwistami oraz pochodzą one z Polski.

Chat 78. Task: TextEntail Pattern: Without 'System'.

Prompt

User Having premise {premise} judge if the following hypothesis {hypothesis} are logically connected with the premise? Answer "entailment" if yes, or "not_entailment" if no.

Chat 79. Task: TextEntail Pattern: With 'System'.

Prompt

System

You are a helpful assistant who can determine for two texts, whether the second one is logically related to the first one. Return your answer in a format defined by the user for two defined texts: premise and hypothesis. Do not explain yourself

User Having premise: {premise} judge if the following hypothesis: {hypothesis} is logically connected with the premise? Output format: "entailment" if yes, or "not_entailment" if no.

Chat 80. Task: TextEntail Pattern: With 'System' and paraphrase of the problem definition.

Prompt

System

You are a helpful assistant who can detect facts in the provided texts. Given two input texts, you can determine whether facts in both texts are the same. Return your answer in a user-specified format for the two defined texts: premise and hypothesis. Do not explain yourself.

User Check if the facts in both texts are the same. First text: {premise}, second text: {hypothesis} Output format: "entailment" if yes, or "not_entailment" if no.

Chat 81. Task: TextEntail Pattern: With ‘System’ and paraphrase of the problem definition.**Prompt**

System You are a helpful assistant who can detect facts in the provided texts. Given two input texts, you can determine whether a fact in the second text is correct based on knowledge from the first text. Return your answer in a user-specified format for the two defined texts: premise and hypothesis. **User** Determine whether a fact in the second text is correct based on knowledge from the first text. First text: {premise}, second text: {hypothesis} Output format: “entailment” if yes, or “not_entailment” if no. Return the result as a single word and do not explain yourself.

Chat 82. Task: WNLI Pattern: Without ‘System’.**Prompt**

User Having the sentence {} tell me if the following sentence {} is true or false? Answer a number “0” if false or “1” if true. Do not explain yourself.

Chat 83. Task: WNLI Pattern: With ‘System’.**Prompt****System**

You are a helpful assistant who can determine for two texts whether the second text is correct based on knowledge from the first text. Return your answer in a format defined by the user. Do not explain yourself.

User

Having the sentence {} tell me if the following sentence {} is true or false? Answer a number “0” if false or “1” if true. Do not explain yourself.

Chat 84. Task: WNLI Pattern: With ‘System’ and ”The Game Pattern”.**Prompt****System**

Let’s play a game called “Lie Detector”. Your goal is to determine if a sentence is a lie based on the context provided by the other player. You advance to the next round if your answer is correct, otherwise you lose.

User

First round. I give you the context: {}. Is the sentence {} correct? Answer a number “0” if false or “1” if true. Do not explain yourself.

Appendix E. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.inffus.2023.101861>.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc. 2017, pp. 6000–6010, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [2] J. Ni, T. Young, V. Pandelea, F. Xue, E. Cambria, Recent advances in deep learning based dialogue systems: a systematic survey, *Artif. Intell. Rev.* 56 (2023) 3055–3155, <http://dx.doi.org/10.1007/s10462-022-10248-8>.
- [3] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, *AI Open* 3 (2022) 111–132, <http://dx.doi.org/10.1016/j.aiopen.2022.10.001>, <https://www.sciencedirect.com/science/article/pii/S2666651022000146>.
- [4] R. Johnson, T. Zhang, Supervised and semi-supervised text categorization using LSTM for region embeddings, in: M. Balcan, K.Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, New York City, NY, USA, June (2016) 19–24, in: *JMLR Workshop and Conference Proceedings*, vol. 48, 2016, pp. 526–534, *JMLR.org*, <http://proceedings.mlr.press/v48/johnson16.html>.
- [5] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, *Neurocomputing* 234 (2017) 11–26, <http://dx.doi.org/10.1016/j.neucom.2016.12.038>, <https://www.sciencedirect.com/science/article/pii/S0925232116315533>. Elsevier.

- [6] B. Alshemali, J. Kalita, Improving the reliability of deep neural networks in NLP: A review, *Knowl.-Based Syst.* 191 (2020) 105210, <http://dx.doi.org/10.1016/j.knsys.2019.105210>.
- [7] G. Liu, J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, *Neurocomputing* 337 (2019) 325–338, <http://dx.doi.org/10.1016/j.neucom.2019.01.078>.
- [8] Z.C. Lipton, A critical review of recurrent neural networks for sequence learning, 2015, arXiv, <http://arxiv.org/abs/1506.00019>. arXiv:1506.00019.
- [9] A. Gilloio, J. Casas, E. Mugellini, O.A. Khaled, Overview of the transformer-based models for NLP tasks, in: M. Ganzha, L.A. Maciaszek, M. Paprzycki (Eds.), *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, Sofia, Bulgaria, September (2020) 6–9, in: *Annals of Computer Science and Information Systems*, vol. 21, 2020, pp. 179–183, <http://dx.doi.org/10.15439/2020F20>.
- [10] W. Rahman, M.K. Hasan, S. Lee, A.B. Zadeh, C. Mao, L. Morency, M.E. Hoque, Integrating multimodal information in large pretrained transformers, in: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetraault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, Online, July (2020) 5–10, Association for Computational Linguistics, 2020, pp. 2359–2369, <http://dx.doi.org/10.18653/v1/2020.acl-main.214>.
- [11] A.V. Ganesan, M. Matero, A.R. Ravula, H. Vu, H.A. Schwartz, Empirical evaluation of pre-trained transformers for human-level NLP: the role of sample size and dimensionality, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, Online, June (2021) 6–11, Association for Computational Linguistics, 2021, pp. 4515–4532, <http://dx.doi.org/10.18653/v1/2021.naacl-main.357>.
- [12] A. Srivastava, A. Rastogi, A. Rao, A.A.M. Shobeh, A. Abid, A. Fisch, A.R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2206.04615>. arXiv:2206.04615.
- [13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June (2019) 2–7, Vol. 1, Long and Short Papers, Association for Computational Linguistics, 2019, pp. 4171–4186, <http://dx.doi.org/10.18653/v1/n19-1423>.
- [14] P.J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer, Generating wikipedia by summarizing long sequences, in: *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30 - May 3, 2018, *Conference Track Proceedings*, OpenReview.net, 2018.
- [15] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., *Improving language understanding by generative pre-training*, OpenAI (2018).
- [16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *Language models are unsupervised multitask learners*, OpenAI (2019).
- [17] 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, in: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, December (2020) 6–12, Virtual, Vol. 33, 2020, pp. 1877–1901, <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfbcb4967418bfb8ac142f64a-Abstract.html>.
- [18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2203.02155>. arXiv:2203.02155.
- [19] OpenAI, GPT-4 technical report, 2023, arXiv:2303.08774.
- [20] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, P. Kazienko, Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach, *Inf. Process. Manage.* 58 (2021) 102643, <http://dx.doi.org/10.1016/j.ipm.2021.102643>, <https://www.sciencedirect.com/science/article/pii/S0306457321001333>.
- [21] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocoń, D. Puchalska, P. Kazienko, Controversy and conformity: from generalized to personalized aggressiveness detection, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Volume 1: Long Papers*, Association for Computational Linguistics, 2021, pp. 5915–5926, <http://dx.doi.org/10.18653/v1/2021.acl-long.460>, Online. <https://aclanthology.org/2021.acl-long.460>.
- [22] P. Kazienko, J. Bielaniec, M. Gruza, K. Kanclerz, K. Karanowski, P. Miłkowski, J. Kocoń, Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor, *Inform. Fusion* 94 (2023) 43–65, <http://dx.doi.org/10.1016/j.inffus.2023.01.010>, <https://www.sciencedirect.com/science/article/pii/S1566253523000167>.
- [23] P. Schramowski, C. Turan, N. Andersen, C.A. Rothkopf, K. Kersting, Large pretrained language models contain human-like biases of what is right and wrong to do, *Nat. Mach. Intell.* 4 (2022) 258–268.

- [24] E. Ferrara, Should ChatGPT be biased? challenges and risks of bias in large language models, 2023, arXiv preprint [arXiv:2304.03738](https://arxiv.org/abs/2304.03738).
- [25] T. Susnjak, ChatGPT: The end of online exam integrity? 2022, arXiv, [http://dx.doi.org/10.48550/arXiv.2212.09292](https://dx.doi.org/10.48550/arXiv.2212.09292). arXiv:2212.09292.
- [26] T.H. Kung, M. Cheatham, A. Medinilla, ChatGPT, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido, et al., Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, 2022, medRxiv 2022-12.
- [27] B. Lund, W. Ting, Chatting About ChatGPT: How May AI and GPT Impact Academia and Libraries? Lund, BD, & Wang, 2023.
- [28] F. Antaki, S. Touma, D. Milad, J. El-Khoury, R. Duval, Evaluating the Performance of Chatgpt in Ophthalmology: an Analysis of Its Successes and Shortcomings, Cold Spring Harbor Laboratory Press, 2023, <http://dx.doi.org/10.1101/2023.01.22.23284882>, medRxiv. <https://www.medrxiv.org/content/early/2023/01/26/2023.01.22.23284882>.
- [29] A.M. Perlman, et al., The implications of OpenAI's assistant for legal services and society, 2022, Available at SSRN.
- [30] T. Goyal, J.J. Li, G. Durrett, News summarization and evaluation in the era of GPT-3, 2022, <https://arxiv.org/abs/2209.12356>. <http://dx.doi.org/10.48550/ARXIV.2209.12356>.
- [31] L. Zhao, W. Alhoshan, A. Ferrari, K.J. Letsholo, Classification of natural language processing techniques for requirements engineering, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2204.04282>. arXiv:2204.04282.
- [32] T. Ganegedara, Natural Language Processing with TensorFlow: Teach Language to Machines Using Python's Deep Learning Library, Packt Publishing Ltd, 2018.
- [33] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is ChatGPT to human experts? comparison corpus, evaluation, and detection, 2023, <https://arxiv.org/abs/2301.07597>. <http://dx.doi.org/10.48550/ARXIV.2301.07597>.
- [34] A. Gilson, C. Safraneck, T. Huang, V. Socrates, L. Chi, R.A. Taylor, D. Chartash, How Does ChatGPT Perform on the Medical Licensing Exams? the Implications of Large Language Models for Medical Education and Knowledge Assessment, Cold Spring Harbor Laboratory Press, 2022, <http://dx.doi.org/10.1101/2022.12.23.22283901>, medRxiv. <https://www.medrxiv.org/content/early/2022/12/26/2022.12.23.22283901>.
- [35] K. Wenzlaff, S. Spaeth, Smarter than humans? validating how OpenAI's ChatGPT model explains crowdfunding, alternative finance and community finance, in: Validating How OpenAI's ChatGPT Model Explains Crowdfunding, Alternative Finance and Community Finance.(December 22, 2022), 2022.
- [36] T. Phillips, A. Saleh, K.D. Glazewski, C.E. Hmelo-Silver, B. Mott, J.C. Lester, Exploring the use of GPT-3 as a tool for evaluating text-based collaborative discourse, in: Examining Pedagogical Data Literacy: Results of a Survey Among School Teachers at Upper Secondary Level in Switzerland, 2022, p. 54.
- [37] C.A. Gao, F.M. Howard, N.S. Markov, E.C. Dyer, S. Ramesh, Y. Luo, A.T. Pearson, Comparing Scientific Abstracts Generated by ChatGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers, Cold Spring Harbor Laboratory, 2022, <http://dx.doi.org/10.1101/2022.12.23.521610>, bioRxiv. <https://www.biorxiv.org/content/early/2022/12/27/2022.12.23.521610>.
- [38] Ö. Aydın, E. Karaarslan, OpenAI ChatGPT generated literature review: Digital twin in healthcare, 2022, Available at SSRN 4308687.
- [39] K. Jeblick, B. Schachtner, J. Dextl, A. Mittermeier, A.T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, M. Ingrisch, ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports, 2022, <http://dx.doi.org/10.48550/ARXIV.2212.14882>. <https://arxiv.org/abs/2212.14882>.
- [40] Y. Chen, S. Eger, Transformers go for the lols: Generating (humorous) titles from scientific abstracts end-to-end, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2212.10522>. arXiv:2212.10522.
- [41] W. Jiao, W. Wang, J. t. Huang, X. Wang, Z. Tu, Is ChatGPT a good translator? a preliminary study, 2023, <http://dx.doi.org/10.48550/ARXIV.2301.08745>. <https://arxiv.org/abs/2301.08745>.
- [42] W. Tabone, J. de Winter, Using ChatGPT for human-computer interaction research: A primer, 2023.
- [43] B. Kutela, K. Msechu, S. Das, E. Kidando, ChatGPT's scientific writings: A case study on traffic safety, 2023, Available at SSRN 4329120.
- [44] R. Karanjai, Targeted phishing campaigns using large scale language models, 2023, arXiv, <http://dx.doi.org/10.48550/arXiv.2301.00665>. arXiv:2301.00665.
- [45] A. Azaria, ChatGPT usage and limitations, 2022, <https://hal.science/hal-03913837>, working paper or preprint.
- [46] M.M. Amin, E. Cambria, B.W. Schuller, Will affective computing emerge from foundation models and general AI? a first evaluation on ChatGPT, 2023, arXiv:2303.03186.
- [47] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q.V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, 2023, arXiv:2302.04023.
- [48] W. Castillo-González, The importance of human supervision in the use of ChatGPT as a support tool in scientific writing, Metaverse Basic Appl. Res. 2 (2022) 29.
- [49] I.E. Karfi, S.E. Fkihi, An ensemble of arabic transformer-based models for arabic sentiment analysis, Int. J. Adv. Comput. Sci. Appl. 13 (2022).
- [50] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C.D. Manning, C. Ré, D. Acosta-Navas, D.A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S.M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2211.09110>. arXiv:2211.09110.
- [51] T.Y. Zhuo, Y. Huang, C. Chen, Z. Xing, Exploring AI ethics of ChatGPT: A diagnostic analysis, 2023, arXiv, <http://dx.doi.org/10.48550/arXiv.2301.12867>. arXiv:2301.12867.
- [52] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, B. Jiao, Y. Zhang, X. Xie, On the robustness of ChatGPT: An adversarial and out-of-distribution perspective, 2023, arXiv:2302.12095.
- [53] B. Peng, C. Li, P. He, M. Galley, J. Gao, Instruction tuning with GPT-4, 2023, arXiv:2304.03277.
- [54] H. Nori, N. King, S.M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on medical challenge problems, 2023, arXiv:2303.13375.
- [55] M.B.I. au2, D.M. Katz, GPT takes the bar exam, 2022, arXiv:2212.14402.
- [56] M. Kosinski, Theory of mind may have spontaneously emerged in large language models, 2023, arXiv:2302.02083.
- [57] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023) 195:1–195:35, <http://dx.doi.org/10.1145/3560815>.
- [58] C.W. Morris, Foundations of the theory of signs, in: International Encyclopedia of Unified Science, Chicago University Press, 1938, pp. 1–59.
- [59] S. Blum-Kulka, M. Hamo, Discourse pragmatics, in: Discourse Studies: A Multidisciplinary Introduction, Vol. 2, 2011, pp. 143–164.
- [60] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April (2017) 3-7, ACM, 2017, pp. 1391–1399, <http://dx.doi.org/10.1145/3038912.3052591>.
- [61] I.D. Kivlichan, Z. Lin, J.Z. Liu, L. Vasserman, Measuring and improving model-moderator collaboration using uncertainty estimation, 2021, arXiv, <https://arxiv.org/abs/2107.04212>. arXiv:2107.04212.
- [62] A. Warstadt, A. Singh, S.R. Bowman, Neural network acceptability judgments, Trans. Assoc. Comput. Linguist. 7 (2019) 625–641, http://dx.doi.org/10.1162/tacl_a_00290.
- [63] S. Wang, H. Fang, M. Khabsa, H. Mao, H. Ma, Entailment as few-shot learner, 2021, arXiv, arXiv:2104.14690.
- [64] I. Annamradnejad, Colbert: Using BERT sentence embedding for humor detection, 2020, arXiv, <https://arxiv.org/abs/2004.12765>. arXiv:2004.12765.
- [65] R. Siddiqui, SARCASMANIA: Sarcasm exposed!, 2019, Online. <http://www.kaggle.com/rmsharks4/sarcasmania-dataset>. (Accessed 02 February 2023).
- [66] P. Kumar, G. Sarin, WELMSD - word embedding and language model based sarcasm detection, Online Inf. Rev. 46 (2022) 1242–1256, <http://dx.doi.org/10.1108/OIR-03-2021-0184>.
- [67] J.M.G. Hidalgo, T.A. Almeida, A. Yamakami, On the validity of a new SMS spam collection, in: 11th International Conference on Machine Learning and Applications, ICMLA, Boca Raton, FL, USA, December (2012) 12-15, Vol. 2, IEEE, 2012, pp. 240–245, <http://dx.doi.org/10.1109/ICMLA.2012.211>.
- [68] T. Sahnoud, M. Mikki, Spam detection using BERT, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2206.02443>. arXiv:2206.02443.
- [69] M.T. Pilehvar, J. Camacho-Collados, Wic: the word-in-context dataset for evaluating context-sensitive meaning representations, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June (2019) 2-7, Vol. 1, Long and Short Papers, Association for Computational Linguistics, 2019, pp. 1267–1273, <http://dx.doi.org/10.18653/v1/n19-1128>.
- [70] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, W. Fedus, St-moe: Designing stable and transferable sparse expert models, 2022, <https://arxiv.org/abs/2202.08906>. <http://dx.doi.org/10.48550/ARXIV.2202.08906>.
- [71] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc, 2019, https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.
- [72] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355, <http://dx.doi.org/10.18653/v1/W18-5446>, <https://aclanthology.org/W18-5446>.

- [73] B. Patra, S. Singhal, S. Huang, Z. Chi, L. Dong, F. Wei, V. Chaudhary, X. Song, Beyond English-centric bitexts for better multilingual language representation learning, 2022, arXiv, <http://dx.doi.org/10.48550/arXiv.2210.14867>. arXiv:2210.14867.
- [74] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for squad, 2018, arXiv, <http://arxiv.org/abs/1806.03822>. arXiv:1806.03822.
- [75] P. He, J. Gao, W. Chen, Debortav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021, arXiv, <https://arxiv.org/abs/2111.09543>. arXiv:2111.09543.
- [76] K. Cobbe, V. Kosaraju, M. Bavarian, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, 2021, arXiv, <https://arxiv.org/abs/2110.14168>. arXiv:2110.14168.
- [77] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J. Lou, W. Chen, On the advance of making language models better reasoners, 2022, arXiv, <https://doi.org/10.48550/arXiv.2206.02336>. arXiv:2206.02336.
- [78] D. Demszky, D. Movshovitz-Attias, J. Ko, A.S. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, in: D. Jurafsky, J. Chi, N. Schluter, J.R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July (2020) 5-10, Association for Computational Linguistics, 2020, pp. 4040-4054, <http://dx.doi.org/10.18653/v1/2020.acl-main.372>.
- [79] A. Ngo, A. Candri, T. Ferdinan, J. Kocon, W. Korczynski, StudEmo: A non-aggregated review dataset for personalized emotion recognition, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 46-55, <https://aclanthology.org/2022.nlperspectives-1.7>.
- [80] I. Price, J. Gifford-Moore, J. Flemming, S. Musker, M. Roichman, G. Sylvain, N. Thain, L. Dixon, J. Sorensen, Six attributes of unhealthy conversations, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, 2020, pp. 114-124, <http://dx.doi.org/10.18653/v1/2020.alw-1.15>, Online. <https://aclanthology.org/2020.alw-1.15>.
- [81] J. Kocon, P. Miłkowski, M. Zasko-Zielinska, Multi-level sentiment analysis of polemo 2.0: Extended corpus of multi-domain consumer reviews, in: M. Bansal, A. Villavicencio (Eds.), Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November (2019) 3-4, Association for Computational Linguistics, 2019, pp. 980-991, <http://dx.doi.org/10.18653/v1/K19-1092>.
- [82] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, TweetEval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1644-1650, <http://dx.doi.org/10.18653/v1/2020.findings-emnlp.148>, <https://aclanthology.org/2020.findings-emnlp.148>.
- [83] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251-260, <http://dx.doi.org/10.18653/v1/2022.acl-demo.25>, <https://aclanthology.org/2022.acl-demo.25>.
- [84] Y. Xu, J. Liu, J. Gao, Y. Shen, X. Liu, Towards human-level machine reading comprehension: reasoning and inference with multiple strategies, 2017, arXiv, <http://arxiv.org/abs/1711.04964>. arXiv:1711.04964.
- [85] H. Puerto, G.G. Sahin, I. Gurevych, Metaqa: Combining expert agents for multi-skill question answering, 2021, arXiv, <https://arxiv.org/abs/2112.01922>. arXiv:2112.01922.
- [86] A. Raganato, J. Camacho-Collados, R. Navigli, Word sense disambiguation: A unified evaluation framework and empirical comparison, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 99-110, <https://aclanthology.org/E17-1010>.
- [87] E. Barba, L. Procopio, R. Navigli, Consec: Word sense disambiguation as continuous sense comprehension, in: M. Moens, X. Huang, L. Specia, S.W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 1492-1503, <http://dx.doi.org/10.18653/v1/2021.emnlp-main.112>.
- [88] J. Firth, A Synopsis of Linguistic Theory 1930-1955, Philological Society, Oxford, 1957, Reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.
- [89] H.J. Levesque, E. Davis, L. Morgenstern, The winograd schema challenge, in: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12, AAAI Press, Rome, Italy, 2012, pp. 552-561, <https://cs.nyu.edu/faculty/davise/papers/WSKR2012.pdf>.
- [90] P. Edmonds, S. Cotton, SENSEVAL-2: Overview, in: Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, Association for Computational Linguistics, Toulouse, France, 2001, pp. 1-5, <https://aclanthology.org/S01-1001>.
- [91] B. Snyder, M. Palmer, The English all-words task, in: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 41-43, <https://aclanthology.org/W04-0811>.
- [92] S. Pradhan, E. Loper, D. Dligach, M. Palmer, SemEval-2007 task-17: English lexical sample, SRL and all words, in: Proceedings of the Fourth International Workshop on Semantic Evaluations, SemEval-2007, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 87-92, <https://aclanthology.org/S07-1016>.
- [93] R. Navigli, D. Jurgens, D. Vannella, SemEval-2013 task 12: Multilingual word sense disambiguation, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval 2013, Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 222-231, <https://aclanthology.org/S13-2040>.
- [94] A. Moro, R. Navigli, SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking, in: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 288-297, <http://dx.doi.org/10.18653/v1/S15-2049>, <https://aclanthology.org/S15-2049>.
- [95] C. Fellbaum, Wordnet: An electronic lexical database, *Comput. Linguist.* (1998) 292-296.
- [96] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, P. Kazienko, Learning personal human biases and representations for subjective tasks in natural language processing, in: 2021 IEEE International Conference on Data Mining, ICDM, IEEE, 2021, pp. 1168-1173, <http://dx.doi.org/10.1109/ICDM51629.2021.00140>.
- [97] J. Bielaniewicz, K. Kanclerz, P. Miłkowski, M. Gruza, K. Karanowski, P. Kazienko, J. Kocoń, Deep-sheep: Sense of humor extraction from embeddings in the personalized context, in: 2022 IEEE International Conference on Data Mining Workshops, ICDMW, IEEE, 2022, pp. 967-974.
- [98] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocon, P. Kazienko, What if ground truth is subjective? personalized deep neural hate speech detection, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 37-45, <https://aclanthology.org/2022.nlperspectives-1.6>.
- [99] P. Miłkowski, S. Saganowski, M. Gruza, P. Kazienko, M. Piasecki, J. Kocoń, Multitask personalized recognition of emotions evoked by textual content, in: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events, PerCom Workshops, IEEE, 2022, pp. 347-352.
- [100] P. Miłkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, J. Kocon, Personal bias in prediction of emotions elicited by textual opinions, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Online, 2021, pp. 248-259, <http://dx.doi.org/10.18653/v1/2021.acl-srw.26>, <https://aclanthology.org/2021.acl-srw.26>.
- [101] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Volume 1: Long Papers, Association for Computational Linguistics, Online, 2021, pp. 3816-3830, <http://dx.doi.org/10.18653/v1/2021.acl-long.295>, <https://aclanthology.org/2021.acl-long.295>.
- [102] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D.C. Schmidt, A prompt pattern catalog to enhance prompt engineering with ChatGPT, 2023, arXiv:2302.11382.
- [103] T. Hagendorff, The ethics of AI ethics: An evaluation of guidelines, *Minds Mach.* 30 (2020) 99-120, <http://dx.doi.org/10.1007/s11023-020-09517-8>.
- [104] A. Borji, A categorical archive of ChatGPT failures, 2023, arXiv, <http://dx.doi.org/10.48550/arXiv.2302.03494>. arXiv:2302.03494.
- [105] W. Korczyński, J. Kocoń, Compression methods for transformers in multidomain sentiment analysis, in: 2022 IEEE International Conference on Data Mining Workshops, ICDMW, IEEE, 2022, pp. 419-426.