Machine Learning with Applications 16 (2024) 100541

Contents lists available at ScienceDirect

Machine Learning with Applications

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

ChatGPT: A meta-analysis after 2.5 months

Christoph Leiter *, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, Steffen Eger

Natural Language Learning Group (NLLG), Data and Web Science Group, University of Mannheim, B6 26, Mannheim, 68159, Baden-Wuerttemberg, Germany

ARTICLE INFO

Dataset link: https://github.com/NL2G/ChatGPTReview

Keywords:
ChatGPT
Sentiment analysis
Emotion analysis
Science
Large language models

ABSTRACT

ChatGPT, a chatbot developed by OpenAI, has gained widespread popularity and media attention since its release in November 2022. However, little hard evidence is available regarding its perception in various sources. In this paper, we analyze over 300,000 tweets and more than 150 scientific papers to investigate how ChatGPT is perceived and discussed. Our findings show that ChatGPT is generally viewed as of high quality, with positive sentiment and emotions of joy dominating social media. Its perception has slightly decreased since its debut, however, with joy decreasing and (negative) surprise on the rise, and it is perceived more negatively in languages other than English. In recent scientific papers, ChatGPT is characterized as a great opportunity across various fields including the medical domain, but also as a threat concerning ethics and receives mixed assessments for education. Our comprehensive meta-analysis of ChatGPT's perception after 2.5 months since its release can contribute to shaping the public debate and informing its future development. We make our data available.[1]

1. Introduction

ChatGPT[2] — a chatbot released by OpenAI in November 2022 which can answer questions, write fiction or prose, help debug code, etc. — has seemingly taken the world by storm. Over the course of just a little more than two months, it has attracted more than 100 million subscribers, and has been described as the fastest growing web platform ever, leaving behind Instagram, Facebook, Netflix and TikTok (Haque, Dharmadasa, Sworna, Rajapakse, & Ahmad, 2022).[3] Its qualities have been featured, discussed and praised by popular media,[4] laymen[5] and experts alike. On social media, it has (initially) been lauded as "Artificial General Intelligence",[6] while more recent assessment hints at limitations and weaknesses e.g. regarding its reasoning and mathematical abilities (Borji, 2023; Frieder et al., 2023) (the authors of this work point out that, as of mid-February 2023, even after 5 updates, ChatGPT can still not accurately count the number of words in a sentence – see Fig. 13 – a task primary school children would typically solve with ease.).

However, while there is plenty of anecdotal evidence regarding the perception of ChatGPT, there is little hard evidence via analysis of different sources such as social media and scientific papers published on it. In this paper, we aim to fill this gap, considering the first 2.5 months after ChatGPT's release.[7] We ask how ChatGPT is viewed from the perspectives of different actors, how its perception has changed over time and which limitations and strengths have been pointed out. Here, we focus specifically on Social Media (Twitter), collecting over 300k tweets, as well as scientific papers from Arxiv and Semantic Scholar, analyzing more than 150 papers. Because of its large user base Twitter comprises a wide range of domains and user groups making it an interesting source to study the early reception of ChatGPT. Also, the high number of daily tweets provides a reasonable data volume for our time period. Arxiv as a repository for scientific articles is not peer-reviewed and therefore reflects insights by academia from early on and

* Corresponding author.
E-mail addresses: christoph.leiter@uni-mannheim.de (C. Leiter), ran.zhang@uni-mannheim.de (R. Zhang), yanran.chen@uni-mannheim.de (Y. Chen), jonas.belouadi@uni-bielefeld.de (J. Belouadi), daniil.larionov@uni-mannheim.de (D. Larionov), V.Fresen@crif.com (V. Fresen), steffen.eger@uni-mannheim.de (S. Eger).

[1] https://github.com/NL2G/ChatGPTReview.
[2] chat.openai.com/.
[3] https://time.com/6253615/chatgpt-fastest-growing/.
[4] www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177.
[5] www.youtube.com/watch?v=OcXKiTDODFU&t=1151s.
[6] https://twitter.com/MichaelTrazzi/status/1599073962582892546.
[7] Due to the considered time period, sometimes we may use words like *recent* to refer to the beginning of 2023.

https://doi.org/10.1016/j.mlwa.2024.100541
Received 29 May 2023; Received in revised form 11 December 2023; Accepted 3 March 2024
Available online 5 March 2024

Semantic Scholar as a search engine for scientific papers allows to find early adopters throughout several other sources.

We find that ChatGPT is overall characterized in different sources as of high quality, with positive sentiment and associated emotions of *joy* dominating. In scientific papers, it is characterized predominantly as a (great) opportunity across various fields, including the medical area and various applications including (scientific) writing as well as for businesses, but also as a threat from an ethical perspective. The assessed impact in the education domain is more mixed, where ChatGPT is viewed both as an opportunity for shifting focus to teaching advanced writing skills (Bishop, 2023) and for making writing more efficient (Zhai, 2022), but also a threat to academic integrity and fostering dishonesty (Ventayen, 2023). Its perception has, however, slightly decreased in social media since its debut, with *joy* decreasing and *surprise* on the rise. In addition, in languages other than English, it is perceived with a more negative sentiment.

By providing a comprehensive assessment of its current perception, our paper can contribute to shaping the public debate and informing the future development of ChatGPT.

Our paper is structured into 8 sections. In Section 2 we discuss related work. Section 3 describes our analysis on Twitter data, Section 4 describes our analysis of scientific papers and Section 5 describes other datasources we have considered. Moreover, Section 6 discusses design choices and future work, Section 7 presents the conclusion of our work and Section 8 highlights potential limitations.

2. Related work

The two most closely related works are Borji (2023) and Haque et al. (2022). Haque et al. (2022) study Twitter reception of ChatGPT after about 2 weeks, finding that the majority of tweets have been overwhelmingly positive in this early period. They have much smaller samples which they manually annotate and use unsupervised topic modeling to determine topics. They also do not look at scientific papers, but only at social media posts. Borji (2023) presents a catalog of failure cases of ChatGPT relating to reasoning, logic, arithmetic, factuality, bias and discrimination, etc. The failure cases are based on selected examples mostly from social media. Beese, Altunbaş, Güzeler, and Eger (2022) and Bowman (2022) discuss the increase of negative papers over time using NLP tools, which is also related to our study. In contrast to their work, we only discuss very recent trends over the last months; our methodological setup is also very different.

Around the time we released the pre-print version of this paper, Taecharungroj (2023) released their Twitter analysis of ChatGPT. They analyze 233,918 tweets with Latent Dirichlet allocation (LDA) topic modeling to determine the key topics that are discussed. They find topics like "news", "technology" and "creative writing". While they also model a topic "reaction", they do not analyze sentiment, which is the core aspect of our work. More than that, we also analyze papers and other resources. Feng et al. (2023) analyze ChatGPT mentions of December and January on Twitter and Reddit with a focus on streaming applications. Besides also performing LDA topic modeling, they present a sentiment analysis of tweets with regards to streaming, but do not focus on the temporal progression. Since our analysis in mid-February, other authors have conducted similar work, varying in time frame and focus of the topic. For example, Miyazaki, Murayama, Uchiba, An, and Kwak (2023) analyze references to generative AI on Twitter from 2019 to 2023 and explore how they are connected to different occupations. Li et al. (2023), Mogavi et al. (2023) and Tlili et al. (2023) focus on a tweet analysis to determine implications and concerns of ChatGPT for education. Raman, lathabhai, Diwakar, and Nedungadi (2023) analyze social media mentions of 183 scientific articles that have been published until 08.03.2023, identify regions important for ChatGPT research and find core research topics. Liu et al. (2023) survey scientific works on Arxiv until 01.04.2023 and analyze submission counts, scientific fields, as well as word clouds. They also structure papers after their use cases. A more recent quantitative evaluation of ChatGPT analyzes over 1.25 million tweets and confirms that positive sentiment is stronger than negative.[8] Gabashvili (2023) even proposes a second-order review, i.e., they survey existing works on ChatGPT and find 11 reviews; 9 that tackle a specific research area and 2 that evaluate across fields.

With the rapid releases of further large language models (LLMs) like GPT4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023), the research continues to become more fast-paced and astonishing results have been achieved. For example, Chen and Eger (2022) use ChatGPT for humorous scientific abstract generation; language models can create style-conditioned poetry (Belouadi & Eger, 2023); GPT4 can (to some extent) perform neuron explanation of other language models (Bills et al., 2023) and generally LLM's continue to solve unexpected tasks with new prompting approaches (e.g. Yao et al., 2023). Some researchers even state that we are witnessing beginning approaches to artificial general intelligence (AGI) (Bubeck et al., 2023).

Table 1
Information of the collected dataset.

Attribute	Detail
Date range	2022-11-30 to 2023-02-09
Number of tweets	334,808
Language counts	61
English tweets	228 127
Number of users	168,111

3. Social media analysis

To acquire insights into public opinion and sentiment on ChatGPT, we conduct sentiment and emotion analysis of public attitudes as well as comparisons of sentiment distribution across topics and languages. Because of its large coverage of user opinions, we choose Twitter as our social media source and collect tweets since the publication date of ChatGPT. For this reason, our data reflects the public opinion as reflected by Twitter (see Section 6). The following will introduce the dataset, methods and results.

3.1. Dataset

We obtain data through the use of a hashtag search tool *snscrape*,[9] setting our search target as #ChatGPT. After acquiring the data, we deduplicate all the retweets and remove robots. We detect robots with two key metrics: the average time between each tweet and the number of total tweets in the examining period. Specifically, we define users as robot accounts if the average tweet interval between two consecutive tweets is less than 2 h. Thereby, we discard 15 such users.

Our final dataset contains tweets in the time period from 2022-11-30 18:10:57 to 2023-02-09 17:24:45. The information is summarized in Table 1. We collect over 330k tweets from more than 168k unique user accounts. The average "age" over all user accounts is 2807 days. On average, each user generated 1.99 tweets during the analysis time period. The dataset contains tweets across 61 languages. Over 68% of them are in English, other major languages are Japanese (6.4%), Spanish (5.3%), French (5.0%), and German (3.3%).

[8] https://hxycorn.github.io/Twitter-Sentiment-Analysis-about-ChatGPT/.
[9] https://github.com/JustAnotherArchivist/snscrape.

3.2. Methods

In this section, we introduce the methods we use for sentiment and emotion analysis. Since the first introduction of transformers (a network architecture based on attention mechanisms Vaswani et al., 2017), language modeling has witnessed a revolution, leading to massive advancements on multiple NLP tasks. These tasks include language generation tasks such as machine translation and language understanding tasks such as sentiment or emotion analysis. Additionally, pre-training transformer-based models on large corpora of texts often eliminates the need for large fine-tuning datasets. These pre-trained models can be fine-tuned on a smaller, task-specific dataset and yield state-of-the-art performance. We take advantage of these advancements and conduct our analysis using task-specific transformer-based models. For thorough background information on topics such as pre-training, fine-tuning, transformers and language models, we refer readers to recent surveys (e.g. Lin, Wang, Liu, & Qiu, 2022; Zhao et al., 2023).

Sentiment analysis. We utilize the multilingual sentiment classifier from Barbieri, Espinosa Anke, and Camacho-Collados (2022) to acquire the sentiment label. This XLM-Roberta (Conneau et al., 2020) based language model is trained on 198 million tweets, and fine-tuned on Twitter sentiment datasets in eight different languages. In their paper, the model performance on sentiment analysis varies among languages (e.g., the F1-score for Hindi is only 53% in multilingual setting), but the model yields comparable results in English with an F1-score of 71% compared to other fine-tuned models (Barbieri et al., 2022; Dashtipour et al., 2016; Nguyen, Vu, & Tuan Nguyen, 2020; Wankhade, Rao, & Kulkarni, 2022; Xu, Cao, Du, & Wang, 2022). Thus, we choose English as our sole input language and collect negative, neutral, and positive sentiments over time (represented as classes 0,1,2, respectively).

We translate all tweets into English via a multilingual machine translation model developed by Facebook (M2M) (Fan et al., 2021).[10] This model supports direct translation between any pair of 100 languages. It is suitable for our study because it yields comparable results to english-centric translation models and supports 100 languages (our dataset covers 61 languages). The study by Mohammad, Salameh, and Kiritchenko (2016) analyzes the effect of translation on sentiment (from Arabic to English) and finds that automatic sentiment analysis of English translations can lead to competitive results to sentiment analysis directly using the source language. They also point out that poor translations often occur when dealing with ambiguous words, cultural-specific sarcasm, metaphors and inappropriate word-reordering, all of which can pose challenges for humans as well. Thus, we argue that with more advanced automatic translation and classification tools (compared to those available in 2016), we can mitigate substantial drops in performance due to translation.

Emotion analysis. In addition to sentiment, we do a more fine-grained analysis based on the emotions of the tweets. We use an emotion classifier (a BERT base model) fine-tuned on the GoEmotions dataset (Demszky et al., 2020) that contains texts from Reddit and their emotion labels based on Ekman's taxonomy (Ekman, 1992) to categorize the translated English tweets into 7 dimensions: *joy, surprise, anger, sadness, fear, disgust* and *neutral*.[11]

[10] https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100.

[11] We use it to predict a single label for each tweet, despite that it being a multi-label classifier (https://huggingface.co/monologg/bert-base-cased-goemotions-ekman).

Table 2
Sentiment distribution of all tweets.

Sentiment	Number of tweets
Positive	100,163
Neutral	174,684
Negative	59,961

3.3. Results & analysis

In this section, we report the results of both sentiment analysis and emotion analysis.

3.3.1. Results of the sentiment analysis

Table 2 summarizes the sentiment distribution of all tweets. While the majority of the sentiment is neutral, there is a relatively large proportion of positive sentiment, with 100k instances, and a smaller but still notable number of tweets of negative sentiment, with 60k instances. Table 3 provides sample tweets belonging to different sentiment groups.

To examine the sentiment change over time, we plot the weekly average of sentiment and the weekly percentage of positive, neutral, and negative tweets in Fig. 1. From the upper plot, we observe an overall downward trend of sentiment (black solid line) during the course of ChatGPT's first 2.5 months: an initial rise in average sentiment was followed by a decrease from January 2023 onwards. We note, however, that the decline is mild in absolute value: the average sentiment of a tweet decreases from a maximum of about 1.15 to a minimum of 1.10 (which also indicates that the average sentiment of tweets is slightly more positive than neutral). We also report the average sentiment of English tweets (dotted line) and non-English tweets (dashed line). The absolute difference is small, but we can clearly identify the division of sentiment between English and non-English tweets. The difference in sentiment is narrowing over time, but overall tweets in English have a more positive perception of ChatGPT. This may suggest that ChatGPT is better at English, which constituted the majority of its training data. This hypothesis is based on the assumption that users tweet and interact with ChatGPT in the same language. However, it is also possible that people tweet in one language and interact with ChatGPT in another language and the difference in sentiment may also be caused by other reasons. We will discuss this in more details in our topic-based analysis below.

The bar plots in the lower part of the figure represent the count of tweets per week and the line plots show the percentage change of each sentiment class. While the percentage of negative tweets is stable over time, the percentage of positive tweets decreases and there is a clear increase in tweets with the neutral sentiment. This may indicate that the public view of ChatGPT is becoming more rational after an initial hype of this new "seemingly omnipotent" bot.

During the course of 2.5 months after ChatGPT's debut, OpenAI announced 5 new releases claiming various updates. Our data covers the period of the first three releases on the 15th of December 2022, the 9th of January, and the 3rd of January in 2023. The two latest releases on the 9th of February and the 13th of February are not included in this study.[12] The three update time points of ChatGPT are depicted as vertical dashed lines in the lower plot of Fig. 1. We can observe small short-term increases in sentiment (solid line in the upper part of the picture) after each new release.

Sentiment across language and topic. We notice from Fig. 1 that the sentiments among English and non-English tweets vary. Here we analyze sentiment based on all 5 major languages in our ChatGPT dataset, namely English (en), Japanese (ja), Spanish (es), French (fr), and

[12] https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

Table 3
Sample tweets of positive (2), neutral (1), and negative sentiment (0) along with their topic.

Tweet	Sentiment	Topic
"*Here we had yet exchanged about the power of open #KI APIs, now we are immersed in the amazing answers of #ChatGPT*".	2	Science & technology
"*I've been playing around with this for a few hours now and I can firmly say that i've never seen anything this developed before. Curious to see where this goes. #ChatGPT*"	2	Diaries & daily life
"*The U.S. company wants to add a filigrane to the texts generated by #ChatGPT. [url] via @user #tweetsrevue #cm #transfonum*"	1	Business & entrepreneurs
"*When you're trying to be productive but the memes keep calling your name.#TBT #ChatGPT #Memes*"	1	Diaries & daily life
"*@user I just tested this for myself and itŠs TRUE. The platform should be shut down IMMEDIATELY #chatgpt #rascist #woke #leftwing*"	0	News & social concern
"*I'm starting to think a student used #ChatGPT for a term paper. If that's the case, the technology isn't ready yet. #academicchatter*"	0	Learning & educational

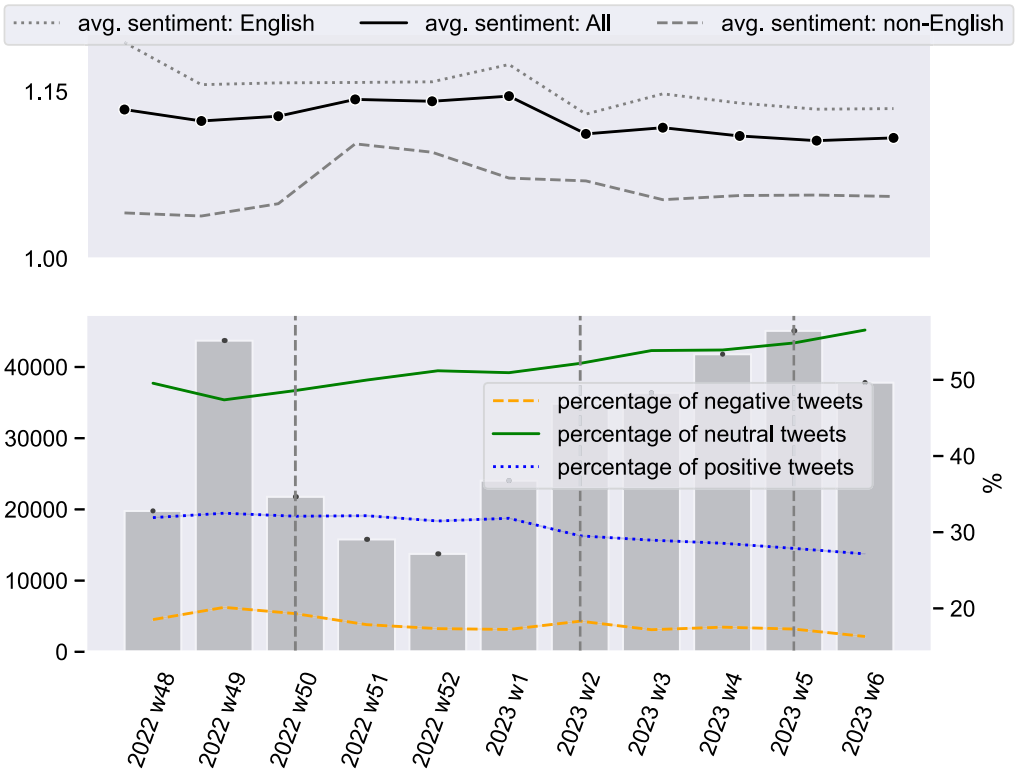


Fig. 1. Upper: weekly average of sentiment over all languages (solid line), over English tweets (dotted line) and non-English tweets (dashed line). Lower: Tweet counts distribution and sentiment percentage change at weekly level aggregation.

German (de). Fig. 2 demonstrates the weekly average sentiment of each language over time. As indicated by our previous observation in Fig. 1, tweets in English have the most positive view of ChatGPT. It is also worth noting that over the time period, the sentiment of English, German, and French tweets are trending downward while Spanish and Japanese tweets start from a low point and trend upwards.

To answer why this is the case, we introduce topic labels into our analysis. To do so, we utilize the monolingual (English) topic classification model developed by Antypas et al. (2022). This Roberta-based model is trained on 124 million tweets and fine-tuned for multi-label topic classification on a corpus of over 11k tweets. The model has 19 classes of topics. We only focus on 5 major classes, which cover 86.3% of tweets in our dataset: science & technology (38.6%), learning & educational (15.2%), news & social concern (13.0%), diaries & daily life (10.2%), and business & entrepreneurs (9.3%). The upper plot of Fig. 3 depicts the topic distribution in percentage by different languages. The share of the science & technology topic ranks the highest in all 5 languages. However, German and French tweets have a relatively higher share of learning & educational and news & social concern topics compared to English and Spanish. We report the sentiment distribution over different topics in Fig. 4. From this plot, we notice that the topic business & entrepreneurs has the lowest proportion of negative tweets while the topic news & social concern contains the highest proportion of negative tweets. For the other three topics, even though their share of positive tweets is similar, the diaries & daily life topic contains more negative tweets proportionally.

This observation may explain the differences in sentiment distribution among different languages. Compared to other languages, English

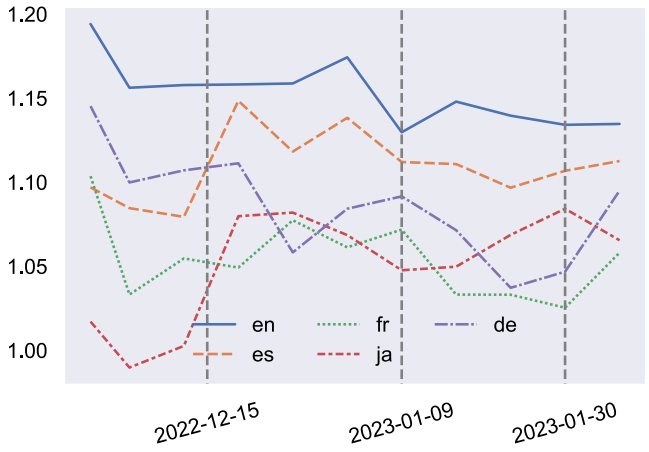


Fig. 2. Weekly sentiment distribution averaged per language.

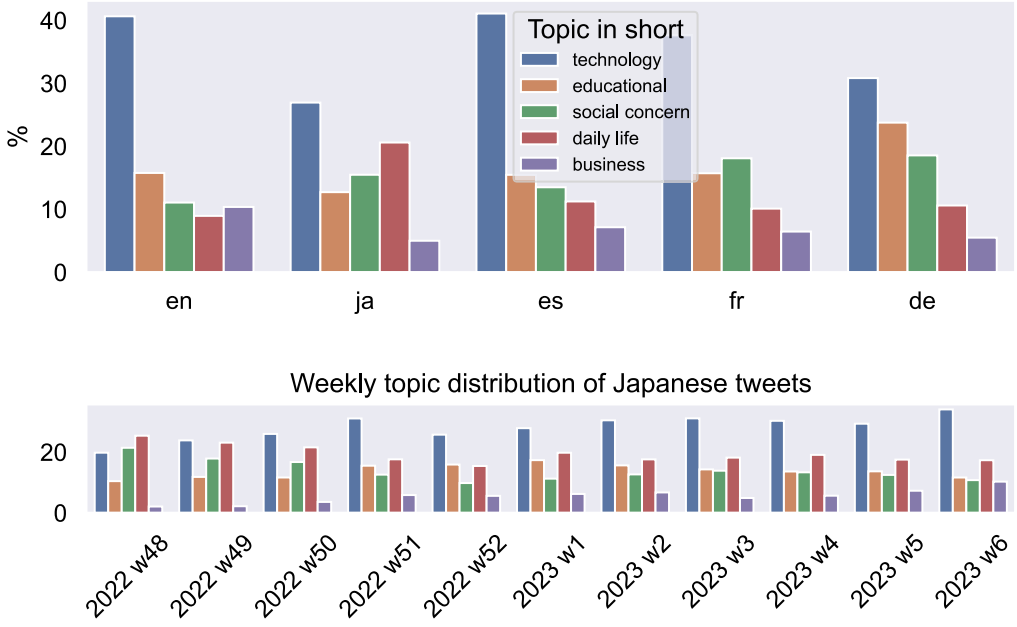


Fig. 3. Upper: topic distribution per language. Lower: topic distribution over time for Japanese tweets.

tweets have the highest proportion of business & entrepreneurs and science & technology, both of which contain the lowest share of negative views about ChatGPT. French and German tweets have a similar proportion of news & social concern topics, which may result in their slightly less positivity than English tweets, though the three of them have similar overall trends. The case for Japanese and Spanish is unique in terms of the low initial sentiment. The lower plot in Fig. 3, which shows the topic distribution change over time for Japanese tweets, may explain this phenomenon. We can observe an evident increase in topics concerning business & entrepreneurs and science & technology, which contribute more positivity, and a decrease in news & social concern, which reduces the share of negative tweets. The same explanation may apply to Spanish tweets.

Aspect of the sentiment. To further obtain an understanding of aspects of sentiments in negative and positive tweets, two co-authors of this paper (one male Ph.D. student and one female Ph.D. student) manually annotated and analyzed the sentiment expressed within 40 tweets. Their kappa agreement is 0.80. We draw 20 of these 40 tweets as random positive tweets from the period including the last two weeks of 2022 and the first week of 2023, where the general sentiment reaches the peak. We draw the other 20 tweets as random negative tweets from the second week to the fourth week of 2023, where the general sentiment declines. We are particularly interested in what users find positive/negative about ChatGPT, which in general could relate to many things, e.g., its quality, downtimes, etc.

Our analysis of the 20 positive tweets during the first period shows a prevalent positive sentiment towards ChatGPT's ability to generate human-like and concise text. Specifically, 14 out of 20 users reported evident admiration for the model and the text it produced. Users particularly noted the model's capacity to answer complex medical questions, generate rap lyrics and tailor texts to specific contexts. Notably, we also discovered instances where users published tweets that ChatGPT completely generated.

As for the randomly selected negative tweets of the second period, 13 out of 20 users expressed frustration with the model. These users voiced concerns about potential factual inaccuracies in the generated text and the detectability of the model-generated text. Additionally, a few users expressed ethical concerns, with some expressing worries about biased output or the potential increase in misinformation. Our analysis also revealed that a minority of the 20 users expressed concerns over job loss to models like ChatGPT. Overall, these findings suggest that negative sentiment towards ChatGPT was primarily

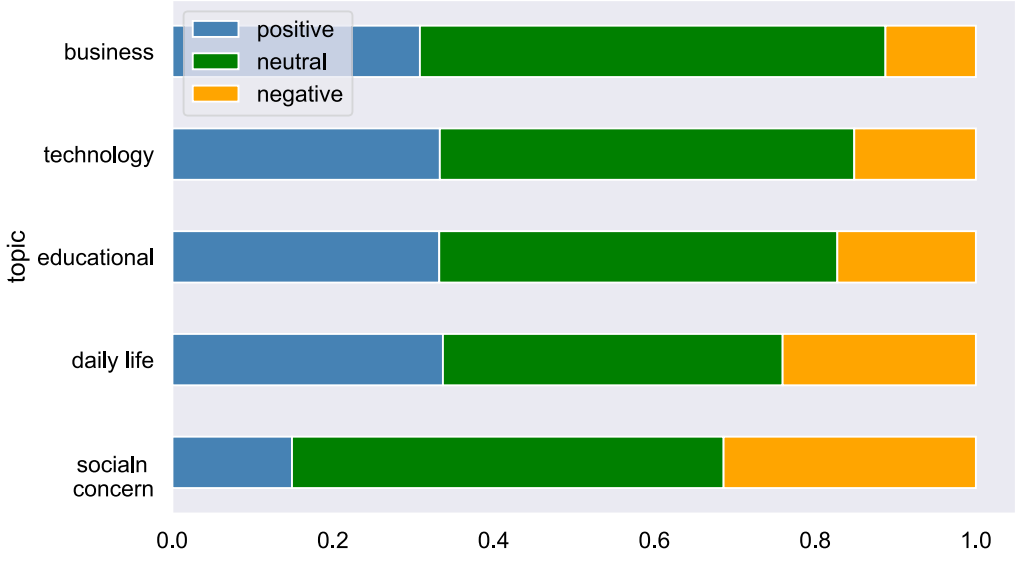


Fig. 4. Sentiment distribution per topic.

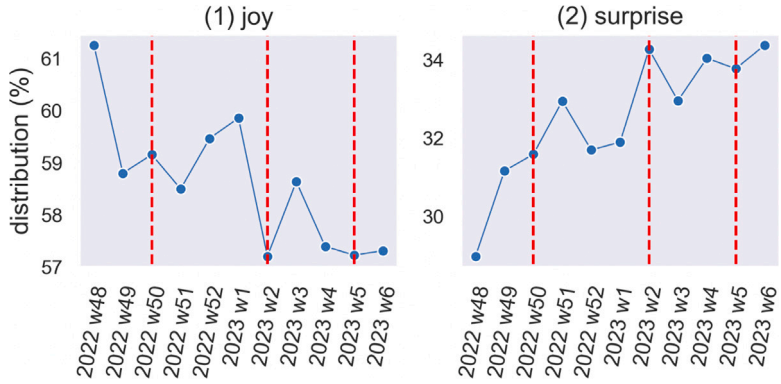


Fig. 5. Weekly emotion distribution of (1) *joy* and (2) *surprise* tweets over time. The percentages denote the ratio of *joy*/*surprise* tweets to non-neutral tweets. We mark the update time points with red dashed lines.

driven by concerns about the model's limitations and its potential impact on society, particularly in generating inaccurate or misleading information.

As part of our analysis, we manually evaluated the sentiment categories for the 40 samples analyzed. We found that 25% (5 out of 20) of the automatically classified sentiment labels were incorrect during the first period. In the second period, we found that 20% (4 out of 20) of the assigned labels were incorrect. The majority of the misclassified tweets were determined to have a neutral sentiment. Despite these misclassifications, we consider the overall error rate of 22.5% (9 out of 40) acceptable for our use case.

Especially, errors may cancel out in our aggregated analysis, and it is worth pointing out that the main confusions were with the neutral class, not the confusion of negative and positive labels, as according to the taxonomy of sentiment classification errors defined by Zimbra, Abbasi, Zeng, and Chen (2018), the majority of identified errors belongs to the categories "neutral mistaken for sentiment", "subtle positive" or "subtle negative". We consider the effect of these errors insubstantial to the aggregated analysis.

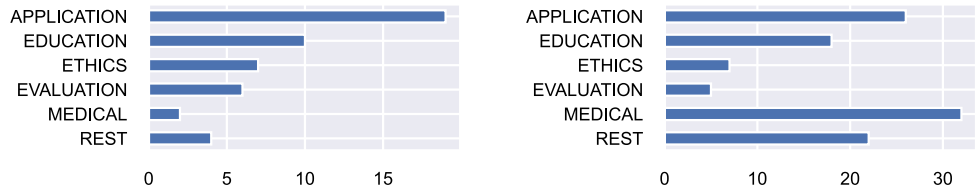
3.3.2. Results for emotion analysis

Among all 334,808 tweets, the great majority are labeled as *neutral* (~70%), followed by the ones classified as *joy* (17.6%) and *surprise* (9.8%); the tweets classified as the remaining 4 emotions compose only 2.7% of the whole dataset.

We demonstrate the weekly changes in the emotion distribution of *joy* and *surprise* tweets in Fig. 5. Here we only show the percentage distribution denoting the ratio of the tweets classified as a specific emotion to all tweets with emotions (i.e., the tweets which are not labeled as *neutral*). We observe that the percentage of *joy* tweets generally decreases after the release, though it rises to some degree after each update, indicating that the users have less fun with ChatGPT over time. On the other hand, the percentage of *surprise* tweets is overall in an uptrend with slight declines between the update time points.

To gain more insights, we manually analyze five randomly selected tweets per emotion category for the release version and each of the three update dates.We draw these random samples from the tweets posted two days after each date considering the difference in time zones. Here, we focus on the *joy* and *surprise* tweets, as they dominate in the tweets with emotions. We also include an analysis of *fear* tweets. The observed peak of the distribution trend of fear tweets appears at the first two update time points, which we believe could provide more insight into the users' concerns across different updates. To summarize, we collect a total of 60 tweets for manual analysis (5 tweets × 4 dates × 3 emotions); we show one sample for each emotion in Table 4.

Joy. Our annotation suggests that 2 out of 20 tweets were misclassified to this category. Among the 18 tweets correctly classified, 12 tweets directly expressed admiration or reported positive interactions with ChatGPT such as successfully performing generation tasks or acquiring answers, 1 tweet conveyed a positive outlook of AI in Non-Fungible



(a) Arxiv (b) Semantic Scholar

Fig. 6. Topics of papers found on Arxiv and Semantic Scholar that include ChatGPT in their title or abstract. *Ethics* also includes papers addressing AI regulations and cybersecurity. *Evaluation* denotes papers that evaluate ChatGPT with respect to biases or more than a single domain.

Table 4
Sample automatically labeled *joy*, *surprise* and *fear* tweets. We mask the user and url information in the table.

Tweet	Emotion
#ChatGPT is so excellent, so fun and I touch it every day, but IŠm looking for a way to use it every day.	*joy*
Wow just wow, Just asked #ChatGPT to write a vision statement for #precisiononcology [emoji] #AI is [emoji] @user @user @user @user [url]	*surprise*
ChatGPT is taking over the internet, and I am afraid, the world for good! #ChatGPT	*fear*

Token (NFT) and game production, and 5 tweets expressed joy which is, however, not (directly) related to ChatGPT. Interestingly, even though 3 tweets did not pertain directly to ChatGPT, they expressed delight in a talk, post, or interview about ChatGPT, and all of them were posted after the second or third updates.

Fear. 1 out of 20 tweets was found to be misclassified, and 1 tweet expressed fear but was unrelated to ChatGPT. Among the remaining 18 tweets, 9 expressed scariness of ChatGPT because of its strong capability, 1 user argued that Google should be scared of ChatGPT, and the rest 8 tweets reported various concerns including providing wrong/malicious information, job loss and the unethical use of ChatGPT. It is noteworthy that 7 out of the 8 tweets demonstrating concerns were published after the second or third updates.

Surprise. Among the sampled tweets, we found that more misclassified tweets may exist compared to the other two categories, as the model tends to classify sentences with question marks as surprise. It is also a challenging task for humans to identify "surprise" in a short sentence, as this emotion may involve different cognitive and perceptual processes. Moreover, "surprise" could have both negative and positive connotations. Hence, we do a four-way manual sentiment annotation for the *surprise* tweets: positive, negative, mixed and unrelated. 12 out of 20 tweets were found to be correctly classified as surprise, among which 6 tweets conveyed positive surprise due to ChatGPT's impressive performance, 2 tweets expressed negative surprise in terms of providing inaccurate information and prejudice against AI, and 4 tweets expressed positive surprise about ChatGPT but with negative concerns regarding unethical uses. We further consider all tweets expressing surprise before and after the 2nd update. Before the 2nd update, there were 1.13 times more positive sentiment surprise tweets than negative ones (4942 vs. 4372); after the second update, the ratio is roughly equal (5065 vs. 5093).

The decrease in *joy* tweets and the increase in negative *surprise* tweets over time – even though on relatively small levels – indicates a more nuanced and rational assessment of ChatGPT over time, similar to the overall decline of positive sentiment over time found in our initial sentiment analysis. We still notice that, apart from *neutral*, *joy* is the most frequently expressed emotion for tweets relating to #ChatGPT in our dataset.

Table 5
Number and source of scientific papers examined. Here, Semantic Scholar comprises only non-Arxiv papers.

Source	Number of instances
Arxiv	48
Semantic Scholar	104

4. Analysis of Arxiv & Semantic Scholar

Besides our Twitter analysis we consider scientific papers from Arxiv and Semantic Scholar to analyze the early adoption of ChatGPT in academia.

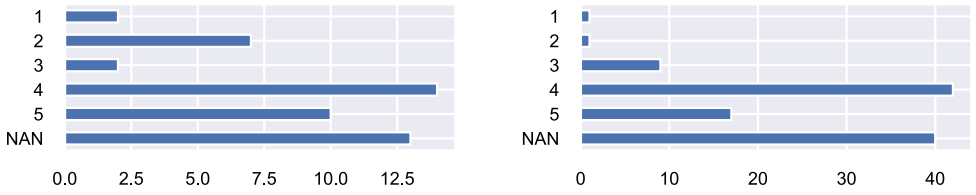
4.1. Dataset

Given the limited time frame of ChatGPT's availability (we consider papers until 09.02.2023), a substantial portion of potentially relevant papers on it are not yet available in officially published form. Thus, we focus our analysis on two sources of information: (1) preprints from Arxiv, which may or may not have already been published; and (2) non-Arxiv papers identified through Semantic Scholar. The Arxiv preprints primarily comprise computer science and similar "hard science" disciplines. Arxiv papers may represent the cutting-edge research in these fields (Eger, Li, Netzer, & Gurevych, 2018). On the other hand, non-Arxiv Semantic Scholar papers encompass a broad range of academic disciplines, including the humanities and social sciences. We do not automatically classify papers but resort to manual annotation, which is feasible given that there are only ∼150 papers in our dataset, see Table 5.

We classify papers along three dimensions; see Table 6 for a concise overview:

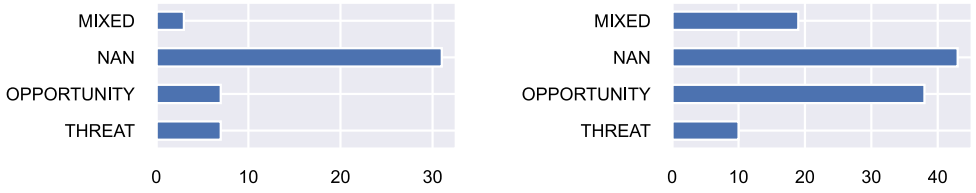
(i) their `quality` assessment of ChatGPT. That means, we annotate how the papers' authors perceive the quality of ChatGPT. This perception could be based on benchmarks, subjective biases and further sources. Future work could further annotate and explore these factors that contribute to the perception. We assign scores from 1 to 5, where 1 indicates very low and 5 very high perceived quality, and 3 is neutral. 2 and 4 offer more fine-grained annotation. NAN indicates that the paper does not discuss the quality of ChatGPT. To align the annotatorsŠ decision processes, we conducted two initial annotation rounds after which we discussed which samples should receive which score.

(ii) their `topic`. After checking the papers and some discussion, we decided on six different topics. These are *Ethics* (which includes biases, fairness, security, etc.), *Education*, *Evaluation* (which includes reasoning, arithmetic, logic problems, etc. on which ChatGPT is evaluated), *Medical*, *Application* (which includes writing assistance or using ChatGPT in downstream tasks such as argument mining, coding, etc.) and *Rest*. We note that a given paper could sometimes be classified into multiple classes, but we are interested in the dominant class. In these cases we give the



(a) Arxiv (b) Semantic Scholar

Fig. 7. Performance quality in papers found on Arxiv and Semantic Scholar that include ChatGPT in their title or abstract. On a scale of 1 (bad performance) to 5 (good performance), this indicates how the performance of ChatGPT is described in the papers' titles/abstracts. *NAN* indicates that no performance sentiment is given.



(a) Arxiv (b) Semantic Scholar

Fig. 8. Social impact in papers found on Arxiv and Semantic Scholar that include ChatGPT in their title or abstract. The labels indicate (based on abstract and title) which effect the authors believe ChatGPT will have on the social good. *NAN* indicates that no social sentiment is given.

following precedence: Medical > Education > Ethics > Evaluation > Rest > Application.

(iii) their `impact` on society. We distinguish *Opportunity*, *Threat*, *Mixed* (when a paper highlights both risks and opportunities) or *NAN* (when the paper does not discuss this aspect).

Example annotations are shown in Table 7.

4.2. Results

Four co-authors of this paper (three male Ph.D. students and one male faculty member) initially annotated 10 papers on all three dimensions independently without guidelines. Agreements were low across all dimensions. After a discussion of disagreements, we devised guidelines for subsequent annotation of 10 further papers. This included (among others) only looking at paper abstracts for classification, as the annotation process would otherwise be too time-consuming, and which labels to prioritize in ambiguous cases. Abstracts are a good compromise because abstracts are (highly condensed) summaries of scientific papers, containing their main message. This time, agreements were high: the kappa agreement is 0.63 on average across all pairs of annotators for `topic`, 0.70 for `impact` and 0.80 Spearman for `quality`, averaged across annotators. In total, we annotated 48 papers from Arxiv and 104 additional papers from Semantic Scholar.

4.3. Analysis

Fig. 6 shows the topic distributions for the Arxiv papers and the papers from Semantic Scholar. The main topics we classified for the Arxiv papers are *Education* and the *Application* in various use cases. Only a few papers were classified as *Medical*. Conversely, Semantic Scholar papers are most frequently classified as *Medical* and *Rest*.

This indicates that *Medical* is of great concern in more applied scientific fields not covered by Arxiv papers. Further, Fig. 7 shows the distributions of quality labels we annotated. The labels 4 and 5 have high numbers of occurrences, i.e., many papers report a strong performance of ChatGPT. Figs. 8(a) and 8(b) show the distributions for our annotations of the social impact. If a social impact sentiment is provided, ChatGPT is most frequently described as an opportunity. For

Table 6
Annotation categories for Arxiv and Semantic Scholar papers along with their labels. Quality refers to the assessment of papers with regard to ChatGPT; 5 indicates that the papers attest it very high quality and 1 indicates very low quality. Impact refers to whether papers describe ChatGPT as a Threat or Opportunity (for society).

Category	Labels
`Topic`	Ethics, Education, Evaluation, Medical, Application, Rest
`Quality`	0 (= NAN), 1,2,3,4,5
`Impact`	Threat, Opportunity, Mixed, NAN

Arxiv, the number of papers which see ChatGPT as an opportunity is the same number as papers that see it as a threat.

In the second part of the analysis, we consider the annotations from Arxiv and Semantic Scholar together. Fig. 9 displays the intersection of *performance quality* and *social impact*. It shows that authors who report a high performance quality for ChatGPT (4/5) in most cases also believe that it will have a positive social impact. Also, there is a high number of papers which report no performance quality or social impact (NAN). Papers that report a low performance quality (1/2) either state no social impact or perceive it as *Mixed* or a *Threat*, but not as *Opportunity*. Fig. 10 shows the intersection between *performance quality* and *topic*. For every topic, the majority of papers describe a high performance quality of ChatGPT. Also, most papers that report low quality are found for *Application* and *Education*. Lastly, Fig. 11 presents the intersection of *topic* and *social impact*. Here, papers in the categories *Application*, *Medical* and *Rest* mostly describe ChatGPT as an opportunity for society. For *Education*, the number of papers that see ChatGPT as a threat is almost equal to the number of those that view it as an opportunity. For *Evaluation*, a comparably high number of abstracts articulate mixed sentiments towards the social impact. Finally, in the *Ethics* category, ChatGPT is mostly seen as a threat.

We also consider the development of each annotated category over time, using all considered papers from Arxiv and those of Semantic Scholar that have an attached publication date. Overall, the number of papers that are published every week is increasing, highlighting the current importance of the topic. Compared to the Twitter data, the sample size of papers is small, hence, other trends are difficult to reliably describe. In Fig. 12, we show that the topics *Evaluation* and *Ethics* have not been considered as a main topic in most early

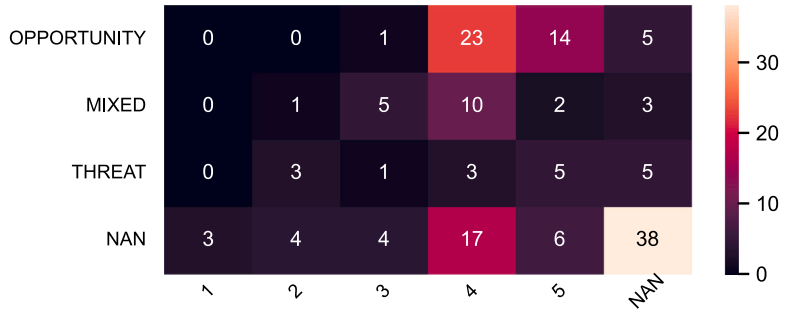


Fig. 9. Heatmap of performance quality (x-axis) and social impact (y-axis) of papers from Arxiv and Semantic Scholar. On a scale of 1 (bad performance) to 5 (good performance), the performance quality indicates how the performance of ChatGPT is described in the papers' titles/abstracts. NAN indicates that no performance quality is given. The social impact indicates (based on abstract and title), which effect the authors believe ChatGPT will have on the social good. *NAN* indicates that no social impact is given.

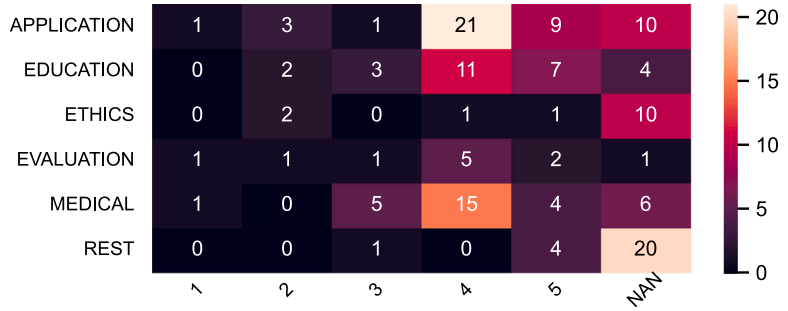


Fig. 10. Heatmap of performance quality (x-axis) and topic (y-axis) of papers from Arxiv and non-Arxiv papers retrieved from Semantic Scholar. On a scale of 1 (bad performance) to 5 (good performance), the performance quality indicates how the performance of ChatGPT is described in the papers' titles/abstracts. NAN indicates that no performance quality is given. *Ethics* also comprises papers addressing AI regulations and cybersecurity. *Evaluation* denotes papers that evaluate ChatGPT with respect to multiple aspects.

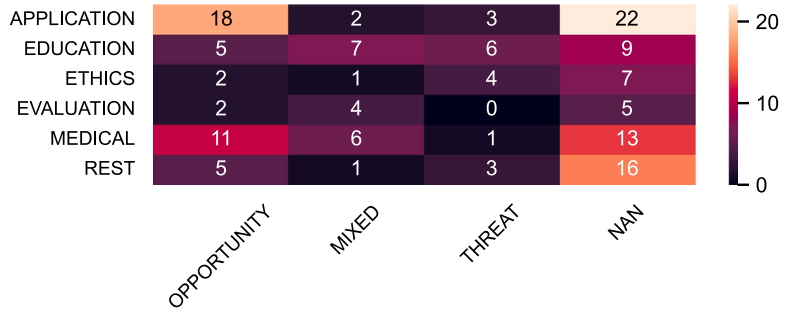


Fig. 11. Heatmap of social impact (x-axis) and topic (y-axis) of papers from Arxiv and non-Arxiv papers retrieved from Semantic Scholar. The social impact indicates (based on abstract and title), which effect the authors' belief ChatGPT will have on the social good. *NAN* indicates that no social impact is described. *Ethics* also comprises papers addressing AI regulations and cybersecurity. *Evaluation* denotes papers that evaluate ChatGPT with respect to biases or more than a single domain.

papers of December 2022. Further, the amount of papers in *Medical* and *Rest* increases, especially since the beginning of February, showing the newly gained, widespread recognition of ChatGPT in many areas outside the NLP community.

To conclude, the analysis of papers exemplifies the explosive attention ChatGPT is getting. They mostly see ChatGPT as an opportunity for society and praise its performance. Threats perceived in *Education* and *Ethics* could, for example, be linked to concerns about plagiarism (e.g. Yeadon, Inyang, Mizouri, Peach, & Testrow, 2023). As another example, Susnjak (2022) finds that ChatGPT poses a problem to online exam integrity.

5. Analysis of other sources

Up to now, we analyzed the public opinion of the ChatGPT model by analyzing Arxiv/Semantic Scholar papers and Twitter for sentiment. However, it is important to note that there are other resources that can provide valuable insights into the model. One such resource is GitHub repositories, which contain a wealth of information about the ChatGPT model. This includes third-party libraries that can be used to programmatically leverage or even enhance the functionality of the model,[13] as well as lists of prompts that can be used to test its abilities.[14]

Other valuable resource are blog posts and the discussion of failure cases,[15] which can help us understand the limitations of the model and how they can be addressed. These resources provide important

[13] https://github.com/stars/acheong08/lists/awesome-chatgpt
https://github.com/saharmor/awesome-chatgpt
https://github.com/humanloop/awesome-chatgpt.

[14] https://github.com/f/awesome-chatgpt-prompts
https://chatgpt.getlaunchlist.com
https://promptbase.com/marketplace.

[15] https://github.com/giuven95/chatgpt-failures
https://docs.google.com/spreadsheets/d/1kDSERnROv5FgHbVN8z_bXH9gak2IXRtoqz0nwhrviCw.

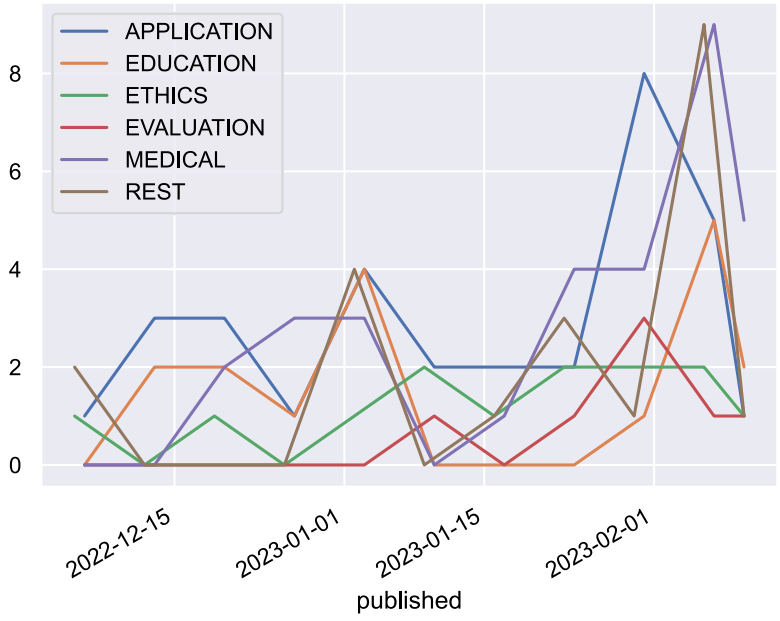


Fig. 12. Paper topics per 7 days. The downtrend in the last days is an artifact of the time aggregation.

Table 7
Sample annotated abstracts from Arxiv (top 2) and Semantic Scholar (bottom) along with annotation dimensions topic, quality (attributed to ChatGPT) and impact on society. We mark rationales for the quality labels in bold and underline rationales for the impact on society. For example, for the first text we chose a perceived quality of 5 as only highly positive descriptions are given. For the second text we chose a perceived quality of 3 due to ChatGPT being described giving competitive results as a positive view, while lagging behind on low-resource tasks is a negative viewpoint. Also, we abbreviate with *Edu(cation)*, *App(lication)*, *Eval(uation)* and *Opp(ortunity)*.

Abstract	Topic	Quality	Impact
"*This study evaluated the ability of ChatGPT, [...]. **The study found that ChatGPT is capable of exhibiting critical thinking skills and generating highly realistic text with minimal input, making it a potential threat to the integrity of online exams, particularly in tertiary education settings where such exams are becoming more prevalent.** [...]*".	Edu.	5	Threat
"*This report provides a preliminary evaluation of ChatGPT for machine translation [...]. **[W]e find that ChatGPT performs competitively with commercial translation products (e.g., Google Translate) on high-resource European languages, but lags behind significantly on low-resource or distant languages.** [...]*". (Jiao, Wang, tse Huang, Wang, & Tu, 2023)	App.	3	NAN
"*Of particular interest to educators, **an exploration of what new language-generation software does (and does not) do well**. Argues that the new language-generation models make instruction in writing mechanics irrelevant, and that educators should shift to teaching only the more advanced writing skills that reflect and advance critical thinking. [...]*". (Bishop, 2023)	Edu.	4	Opp.
"*We investigate the mathematical capabilities of ChatGPT by testing it on publicly available datasets [...]. **ChatGPT's mathematical abilities are significantly below those of an average mathematics graduate student.** [...]*". (Frieder et al., 2023)	Eval.	1	NAN

feedback to the developers and can inform future development efforts, ensuring that the ChatGPT model continues to evolve and improve.

We constructed a small dataset (50 entries) of such online resources and enlisted two coworkers to annotate their sentiment. Our analysis of these resources (see Table 8) reveals that shared prompts lists, as well as other GitHub repositories, exhibit overwhelmingly positive sentiment, while blog posts display a mix of positive, neutral, and negative sentiments in nearly equal proportions. We further observed that lists of failure cases showed the poorest overall sentiment, a finding which intuitively makes sense. Failure cases often involve math problems,[16] a domain where ChatGPT frequently provides confidently incorrect answers. However, our findings were not entirely consistent, as some positive blog posts suggest that ChatGPT performs well in symbolic execution of code,[17] indicating that the issue may lie in prompt tuning rather than ChatGPT's general capabilities, i.e., ChatpGPT can handle math problems better when they are formulated as programs, not prose. Neutral[18] and negative[19] blog posts tend to focus less on the quality of ChatGPT's outputs and more on concerns related to OpenAI's restrictions or potential negative social impacts.

Table 8
Observed sentiment found in various independent resources around the web.

Type	Positive	Neutral	Negative
Prompt sharing sites	100%	0%	0%
GitHub repositories	92%	8%	0%
Blog posts	22%	44%	33%
Lists of failure cases	0%	0%	100%

[16] https://twitter.com/GaryMarcus/status/1610793320279863297.
[17] https://www.engraved.blog/building-a-virtual-machine-inside.
[18] https://thezvi.substack.com/p/jailbreaking-the-chatgpt-on-release.
[19] https://davidgolumbia.medium.com/chatgpt-should-not-exist-aab0867abace.

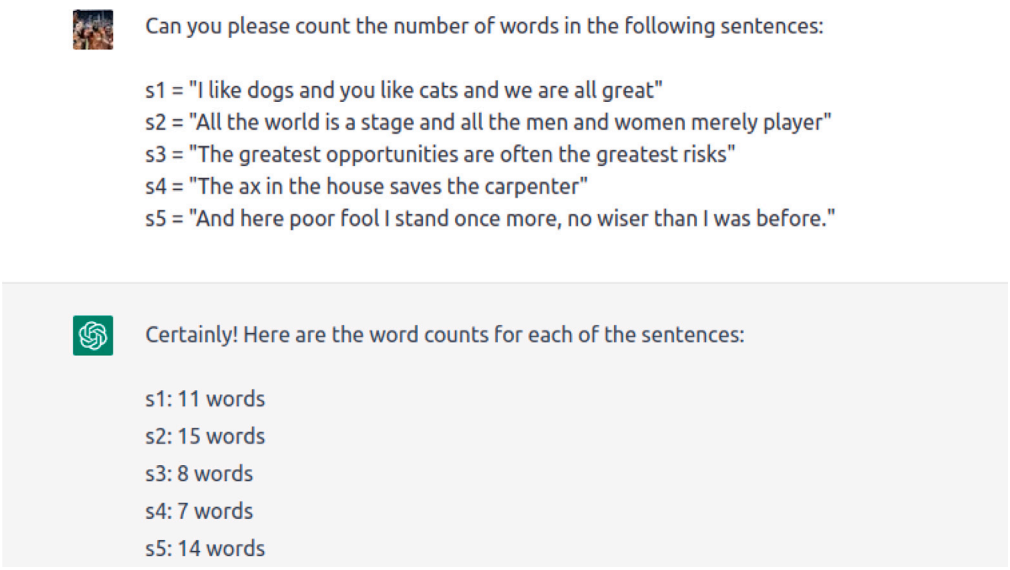


Fig. 13. Evidence of counting failures of ChatGPT. The correct answers are 12, 14, 8, 8, 14, of which ChatGPT gets false 3/5 (by an error of one). ChatGPT indicated that it ignored punctuation and quotation marks when counting and also miscounted without any quotation and punctuation marks.

6. Discussion

In this section, we discuss the chosen scope of our paper and potential future research goals that we did not or only partly address and describe their feasibility for future work.

Our work analyzes the reception of ChatGPT during 2.5 months after its release. We consider this period to be especially interesting because of the world wide recognition the new technology has received in this initial phase of "hype". During this time, there has been a widespread exploration of the tool in various fields and opinions were formed. Since the pre-print of our paper was released, other strong models, like GPT4 (OpenAI, 2023) or the Bing chatbot[20] with internet access have been released. Therefore, future work could consider changes in the perception over larger time frames and for other models. Indeed, Eger et al. (2023) find that the academic usage of ChatGPT may be declining, at least in computer science and AI related fields. Other work has found that for some tasks the performance of GPT4 and ChatGPT has been declining since their release (Chen, Zaharia, & Zou, 2023).

Because of the 2.5 month time frame, our original work was time and resource restricted. Hence, as one source we selected academia, as represented by Arxiv and Semantic Scholar to reflect a wide range of early adopters. While these two do not represent all available paper sources, they cover a broad range of domains, especially in computer science, and many relevant ChatGPT papers have been published here (Eger et al., 2023). Further, because of its large user base, we chose Twitter to consider a wide range of public opinions. Still, user dimensions like age, education or political orientation can be confounded with their perception of ChatGPT. The usage of further data sources may address this issue, however, this was beyond the scope of our work, which relied on immediate and automatic access on large user bases. The setup of a potentially unbiased dataset, e.g. via a global survey, would be very time consuming, likely missing the early adoption phase. Since the early adoption phase of ChatGPT is over now, it might be impossible to acquire such a dataset in hindsight. Another issue with running similar analyses now is that many social media platforms, e.g. Twitter and Reddit, have restricted their public data access throughout 2023.[21]

The goal of our paper is to give an overview of the reception of ChatGPT in diverse domains, instead of detailed analyses of specific domains. Future work could explore further dynamics within each of these domains, for example, how international undergraduate students use ChatGPT to understand topics in the domain of education. Generally, the usage of ChatGPT for such specific usergroups is often already considered as topic of the papers that we have analyzed. Papers after these first 2.5 months continue to explore ChatGPT in various scenarios. For example, Eysenbach (2023) used ChatGPT to write a discussion about using ChatGPT in the specific domain of medical education.

Subsequent to our analysis, the successor model GPT4 (OpenAI, 2023) was released. Also, ChatGPT received many additional updates. Both might have influenced the perspective on these models. As our goal is to capture the perception during the early release of this technology, we leave this evaluation to future work.

Future work could also use existing work that analyzes the sentiment and emotions revolving around ChatGPT to predict the public reaction for the case that a new model outperforms ChatGPT and GPT4 by a large margin.

Finally, future work might explore the usage of stronger classifiers for the sentiment and emotion classification. While the classifiers that we have used already achieve a good performance, ensembles or newer classifiers based on models like LLaMA might outperform them and lead to further improvements of our results. However, ensemble methods often do not strongly improve the performance if the individual models already deliver good results. For example, ensemble of deep learning language models from the study of Alsayat (2022) shows 2%–5% improvement in classification accuracy and the ensemble model from Minaee, Azimi, and Abdolrashidi (2019) barely show any improvement.

7. Conclusion

In this paper, we conducted a comprehensive analysis of the perception of ChatGPT, a chatbot released by OpenAI in November 2022 that has attracted over 100 million subscribers in only two months. We analyzed over 300k tweets and more than 150 scientific papers to understand how ChatGPT is viewed from different perspectives, how its perception has changed over time, and what its strengths and limitations are. We found that ChatGPT is generally perceived positively, with high quality, and associated emotions of joy dominating. However, its perception has slightly decreased since its debut, and

[20] www.bing.com.

[21] See for example https://twitter.com/XDevelopers/status/1641222788911624192.

in languages other than English, it is perceived with more negative sentiment. Moreover, while ChatGPT is viewed as a great opportunity across various scientific fields, including the medical domain, it is also seen as a threat from an ethical perspective and in the education domain. Our findings contribute to shaping the public debate and informing the future development of ChatGPT.

Future work should investigate developments over longer stretches of time, consider popularity of tweets and papers (via likes and citations), investigate more dimensions besides sentiment and emotion and look at the expertise of social media actors and their geographic and demographic distribution. Finally, as language models like ChatGPT continue to evolve and gain more capabilities, future research can assess their real (rather than anticipated) impact on society, including their potential to exacerbate and mitigate existing inequalities and biases.

8. Limitations and ethical considerations

In this work, we automatically analyzed social media posts using NLP technology like sentiment and emotion classifiers and machine translation systems, which are error-prone. Our selection of tweets was biased via the employed hashtag, i.e., #ChatGPT. Our human annotation was in some cases subjective and not without disagreements among annotators. Our selection of Semantic Scholar papers was determined by the search results of Semantic Scholar, which seem non-deterministic. Our search for papers on Arxiv was restricted to mentions in the abstracts and titles of papers and our annotations were only based on titles and abstracts. Further, we only considered the first 2.5 months since ChatGPT's release. We hope our work will be inspirational for others to consider ChatGPT's perception at regular intervals, from which short-, mid- and long-term dynamics can be assessed.

CRediT authorship contribution statement

Christoph Leiter: Writing – original draft, Formal analysis, Data curation, Writing – review & editing. **Ran Zhang:** Writing – original draft, Formal analysis, Data curation. **Yanran Chen:** Writing – original draft, Formal analysis, Data curation. **Jonas Belouadi:** Writing – original draft, Data curation, Formal analysis. **Daniil Larionov:** Writing – original draft, Data curation, Formal analysis. **Vivian Fresen:** Data curation. **Steffen Eger:** Conceptualization, Data curation, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We make our data available on GitHub: https://github.com/NL2G/ChatGPTReview.

Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work the author(s) used ChatGPT in order to aid the writing process. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

Funding

The NLLG group gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant "Metrics4NLG" and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1. This paper was written while on a retreat in the Austrian mountains in the small village of Hinterriß.

References

Alsayat, A. (2022). Improving sentiment analysis for social media applications using an ensemble deep learning language model. *Arabian Journal for Science and Engineering*, *47*(2), 2499–2511.

Antypas, D., Ushio, A., Camacho-Collados, J., Silva, V., Neves, L., & Barbieri, F. (2022). Twitter topic classification. In *Proceedings of the 29th international conference on computational linguistics* (pp. 3386–3400). Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 258–266). Marseille, France: European Language Resources Association.

Beese, D., Altunbaş, B., Güzeler, G., & Eger, S. (2022). Detecting stance in scientific papers: Did we get more negative recently? arXiv preprint arXiv:2202.13610.

Belouadi, J., & Eger, S. (2023). ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 7364–7381). Toronto, Canada: Association for Computational Linguistics.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., et al. (2023). Language models can explain neurons in language models. URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html. (Date accessed: 14.05.2023).

Bishop, L. M. (2023). A computer wrote this paper: What ChatGPT means for education, research, and writing. *SSRN Electronic Journal*.

Borji, A. (2023). A categorical archive of ChatGPT failures. ArXiv, abs/2302.03494.

Bowman, S. (2022). The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 7484–7499).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Chen, Y., & Eger, S. (2022). Transformers go for the LOLs: Generating (humourous) titles from scientific abstracts end-to-end. arXiv preprint arXiv:2212.10522.

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv preprint arXiv:2307.09009.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., et al. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive Computation*, *8*, 757–771.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4040–4054). Online: Association for Computational Linguistics.

Eger, S., Leiter, C., Belouadi, J., Zhang, R., Kostikova, A., Larionov, D., et al. (2023). NLLG quarterly arxiv report 06/23: What are the most influential current AI papers? arXiv preprint arXiv:2308.04889.

Eger, S., Li, C., Netzer, F., & Gurevych, I. (2018). Predicting research trends from arxiv. ArXiv, abs/1903.02831.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553.

Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education*, *9*, Article e46885.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., et al. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, *22*(1).

Feng, Y., Poralla, P., Dash, S., Li, K., Desai, V., & Qiu, M. (2023). The impact of ChatGPT on streaming media: A crowdsourced and data-driven analysis using Twitter and reddit. https://yunhefeng.me/material/IDS_ChatGPT_Streaming_Media__camera_ready_.pdf.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., et al. (2023). Mathematical capabilities of ChatGPT. ArXiv, abs/2301.13867.

Gabashvili, I. S. (2023). The impact and applications of ChatGTP across Industries and disciplines. URL https://doi.org/10.17605/OSF.IO/87U6Q.

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). "I think this is the most disruptive technology": Exploring sentiments of ChatGPT early adopters using Twitter data. ArXiv, abs/2212.05856.

Jiao, W., Wang, W., tse Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. ArXiv, abs/2301.08745.

Li, L., Ma, Z., Fan, L., Lee, S., Yu, H., & Hemphill, L. (2023). ChatGPT in education: A discourse analysis of worries and concerns on social media. arXiv preprint arXiv:2305.02201.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132.

Liu, Y.-H., Han, T., Ma, S., Zhang, J.-Y., Yang, Y., Tian, J., et al. (2023). Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. ArXiv, abs/2304.01852.

Minaee, S., Azimi, E., & Abdolrashidi, A. (2019). Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. arXiv preprint arXiv:1904.04206.

Miyazaki, K., Murayama, T., Uchiba, T., An, J., & Kwak, H. (2023). Public perception of generative AI on Twitter: An empirical study based on occupation and usage. arXiv preprint arXiv:2305.09537.

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., et al. (2023). Exploring user perspectives on chatgpt: Applications, perceptions, and implications for ai-integrated education. arXiv preprint arXiv:2305.13114.

Mohammad, S. M., Salameh, M., & Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research*, *55*, 95–130.

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 9–14). Online: Association for Computational Linguistics.

OpenAI (2023). GPT-4 technical report. ArXiv, abs/2303.08774.

Raman, R., lathabhai, h., Diwakar, S., & Nedungadi, P. (2023). Early research trends on ChatGPT: a review based on Altmetrics and science mapping analysis.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? ArXiv, abs/2212.09292.

Taecharungroj, V. (2023). What can ChatGPT do? Analyzing early reactions to the innovative AI chatbot on Twitter. *Big Data and Cognitive Computing*, *7*(1).

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., et al. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*(1), 15.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Ventayen, R. J. M. (2023). OpenAI ChatGPT generated results: Similarity index of artificial intelligence-based contents. *SSRN Electronic Journal*.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780.

Xu, Y., Cao, H., Du, W., & Wang, W. (2022). A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations. *Data Science and Engineering*, *7*(3), 279–299.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., et al. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, *58*(3), Article 035027.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN Electronic Journal*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, *9*(2), 1–29.