

# Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children’s Literature

Simone Rebora<sup>1,\*†</sup>, Marina Lehmann<sup>2,†</sup>, Anne Heumann<sup>2,†</sup>, Wei Ding<sup>2</sup> and Gerhard Lauer<sup>2</sup>

<sup>1</sup>Department of Foreign Languages and Literatures, University of Verona, Italy

<sup>2</sup>Department of Book and Reading Studies, Johannes Gutenberg-University Mainz, Germany

## Abstract

In this paper, we apply the ChatGPT Large Language Model (gpt-3.5-turbo) to the 4books dataset, a German language collection of children’s and young adult novels comprising a total of 22,860 sentences annotated for valence by 80 human raters. We verify if ChatGPT can (a) compare to the behaviour of human raters and/or (b) outperform state of the art sentiment analysis tools. Results show that, while inter-rater agreement with human readers is low (independently from the inclusion/exclusion of context), efficiency scores are comparable to the most advanced sentiment analysis tools.

## Keywords

Large Language Models, ChatGPT, 4books dataset, sentiment analysis, inter-rater agreement

## 1. Introduction

Among the many discussions stimulated by the recent success of Large Language Models (LLMs), two perspectives seem to be dominant. One that highlights the human-like behaviour of agents like ChatGPT, suggesting how they could be considered like the very first example of artificial intelligence ever realized in the history of humankind [1, 10]. Another, more cautious perspective sees them as one of the most advanced tools currently available to perform common tasks in natural language processing [22, 28].

With this paper, we explore both perspectives by profiting from 4books, a large dataset originally developed in the field of psychology and reading studies, with the goal of analysing the emotional reactions of readers towards literary texts. The peculiarity of such a dataset—composed by children’s and young adult novels in German language—offers the opportunity to study the behaviour of LLMs when confronted with a narrative genre that contributes to shaping the cognitive and emotional skills of human beings [21, 16]. Given its focus on emotions,

---


*CHR 2023: Computational Humanities Research Conference, December 6–8, 2023, Paris, France*


\*Corresponding author.

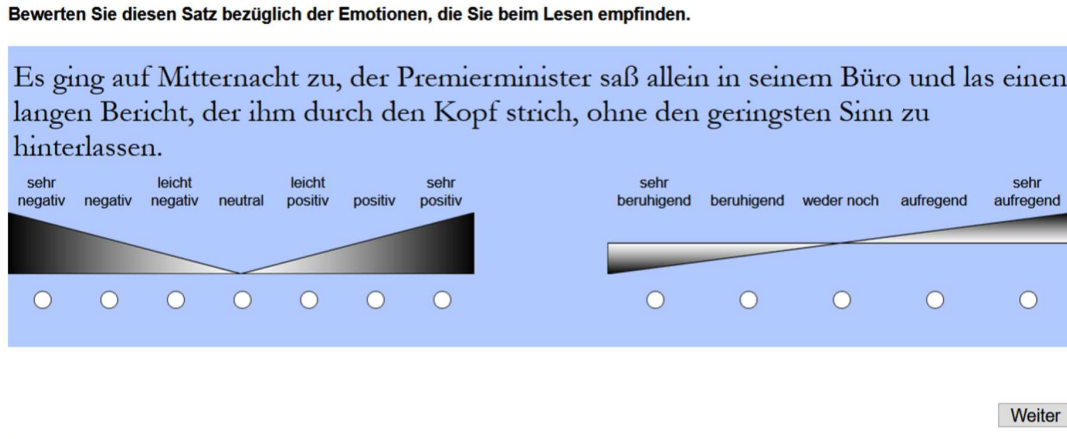
†These authors contributed equally.

✉ simone.rebora@univr.it (S. Rebora); marina.lehmann@uni-mainz.de (M. Lehmann); aheumann@students.uni-mainz.de (A. Heumann); wei.ding@uni-mainz.de (W. Ding); gerlauer@uni-mainz.de (G. Lauer)

📄 <https://orcid.org/0000-0002-1501-3774> (S. Rebora); <https://orcid.org/0000-0002-6818-6169> (M. Lehmann); <https://orcid.org/0009-0000-5791-6982> (A. Heumann); <https://orcid.org/0000-0001-5401-879X> (W. Ding); <https://orcid.org/0000-0003-0230-2574> (G. Lauer)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Sample screenshot of the 4books rating interface.

then, it could also be used to test LLMs in one of the most common tasks in natural language processing, namely Sentiment Analysis, in a still under-researched language for LLMs (German).

## 2. The 4books Dataset

The 4books dataset [19] was developed in the context of the CHYLSA project<sup>1</sup> and consists of diverse reader responses towards two children’s novels (“*Oma!*”, *schreit der Frieder* by G. Mebs and *Jim Knopf und Lukas der Lokomotivführer* by M. Ende [20, 5]) and two young adult novels (*Das Schicksal ist ein mieser Verräter* by J. Green and *Harry Potter und der Halbblutprinz* by J.K. Rowling [8, 25]).<sup>2</sup> Overall, the books contain 22,860 sentences and each book was read by 20 readers, resulting in a total of 80 participants, all native speakers of German (mean age = 23.33, sd = 6.4).<sup>3</sup>

The rating process was as follows: First, the participants evaluated the emotional impact in terms of valence (on a scale from  $-3$  to  $3$ ) and arousal (on a scale from  $1$  to  $5$ ) on sentence-level (for an example of the rating interface, see Figure 1). Sentences were shown one by one, without the possibility of reading following sentences or revising previous ratings. This was decided in order to collect the most immediate and direct reactions from readers, by disrupting the least possible their reading experience. To avoid cognitive overload, participants were asked to rate one chapter per day. At the end of each chapter, they answered a series of comprehension questions, aimed at verifying their understanding of the reading material. Finally, they were

<sup>1</sup><https://chylsa.pages.gitlab.rlp.net/chylsa-website/>.

<sup>2</sup>Due to copyright reasons, the dataset cannot be published.

<sup>3</sup>Due to the heaviness and complexity of the rating task, it was not possible to match participants’ age with books’ target audience.

asked to produce multiple ratings on the chapter and book level (including valence and arousal, but also suspense and transportation).

In this paper, we focus on the sentence ratings for valence, as it is the most frequently studied dimension in Sentiment Analysis research. We had also to exclude working on larger textual segments (like chapters and full books) due to the current limitations of context window in the most widely used LLMs.

### 3. Research Background

In the field of NLP, much attention has been given to the recent LLMs. Multiple studies compare ChatGPT to the best performing current Transformer models, so-called state-of-the-art (SOTA) models. Overview studies use ChatGPT for numerous NLP tasks [22, 14]. More specifically targeted experiments assess, e.g., ChatGPT’s understanding ability [32] or its performance on automatic genre identification [17]. Overall, ChatGPT has been evaluated in a wide range of settings: 1) comparisons between different GPT-models [14]; 2) prompting strategies [32, 31]; or 3) the effect of the prompts’ language [17]. Still, only a few studies have been dedicated to German language tasks (see, e.g., Friederichs et al. and Wang et al. [6, 30])—and none specifically to German Sentiment Analysis.

#### 3.1. Annotation

With regards to ChatGPT various ethical considerations have been discussed. Recent research has approached the question of whether LLMs could be considered as intelligent agents by using frameworks such as the Turing test [4], theory of mind [15], and world-model building [9], frequently presenting contradictory results.

However, one strong stance on the subject has been advanced in a few studies which present LLMs as able to substitute human annotators. Such a substitution could constitute one of the main confirmations of the final “emergence” of the artificial intelligence, if we consider that manual annotation is the very groundwork for any machine learning task—where the machine tries to imitate the human. Simultaneously, automatic machine annotation may dramatically decrease the financial resources needed for human annotators.

Gilardi et al. [7] compare the annotations done by ChatGPT and by crowd-workers on MTurk against a gold standard of trained human annotators. They find that ChatGPT outperforms the crowd-sourced annotators while being at the same time much cheaper to fund. Even more, the intercoder agreement exceeds not only that of MTurk annotators but also that of the trained experts, which is especially remarkable due to the frequently reported low agreement rates among human annotators [2].

In a similar experiment, Törnberg [29] shows that the LLM’s accuracy exceeds all other human annotations. Intercoder reliability is reported to be much higher when compared to human annotators. Huang et al. [11] use ChatGPT for implicit hate speech detection and generation and report an accuracy of 80% for the detection task.

While these experiments generally show promising results, others find only average performance: In the study by Ding et al. [3] the best approach with ChatGPT achieves high accuracy scores, while still being outperformed by human annotators. Reiss [24] evaluates ChatGPT’s

performance on News text classification and stresses out that the results of ChatGPT are limited in terms of scientific reliability because the scores for Krippendorff’s Alpha are not reaching the necessary threshold of 0.8.

### 3.2. Sentiment Analysis

In Sentiment Analysis, dictionary-based approaches are continuously outperformed by Transformer-based models [26]. At the same time, it is conceivable that LLMs may lead to a shift in common Sentiment Analysis techniques because promising results have been reported in other NLP applications.

The studies of Wang et al. [31] and Rathje et al. [23] are presently the most targeted experiments on Sentiment Analysis with ChatGPT. Wang et al. [31] evaluate different Sentiment scores (e.g., standard polarity values as well as an aspect-based approach) and compare ChatGPT to a fine-tuned BERT model and other SOTA models. Results show that ChatGPT can compete with a fine-tuned BERT model, but falls slightly behind other domain-specific SOTA models. Rathje et al. [23] compare ChatGPT to a dictionary- and Transformer-based approach, but for different languages and versions of ChatGPT (gpt-3.5-turbo and gpt-4). ChatGPT achieves higher performance when compared to common dictionaries, whereas results vary depending on language for the comparison with SOTA models.

Altogether, the main observation—the performance of ChatGPT in Sentiment Analysis, while generally being high, is lower than that of fine-tuned SOTA models—is confirmed by most other studies that use Sentiment Analysis next to other NLP tasks [22, 32, 18].

## 4. Experimental Setup

Overall, our experimental setup consists of two parallel studies: one focused on verifying if ChatGPT behaves like a human rater (using inter-rater agreement to measure similarity of behaviour); another comparing ChatGPT to Sentiment Analysis tools (defining a ground truth to evaluate tool efficiency). The 4books dataset is especially suited for the first study because the ratings are done by readers and not by trained experts—an important distinction as there was neither a guidebook for the raters nor the possibility of discussing possible disagreements and thus increasing inter-rater agreement.

To perform all analyses,<sup>4</sup> we used the OpenAI API. All experiments were carried out with the “gpt-3.5-turbo” model (corresponding to the default version of ChatGPT and therefore referred to as such in the following pages), the most advanced among the ones available for our newly-created account at the time of writing. Table 1 presents an overview of the prompts sent to the API.

### 4.1. ChatGPT as a Human Rater

One possibility for our first study was that of considering ChatGPT as a substitute of multiple human raters. Such a possibility could be supported by the fact that ChatGPT is able to produce

---

<sup>4</sup>All scripts available at the following link: <https://gitlab.rlp.net/srebora/chr23-sentiment-chatgpt>.

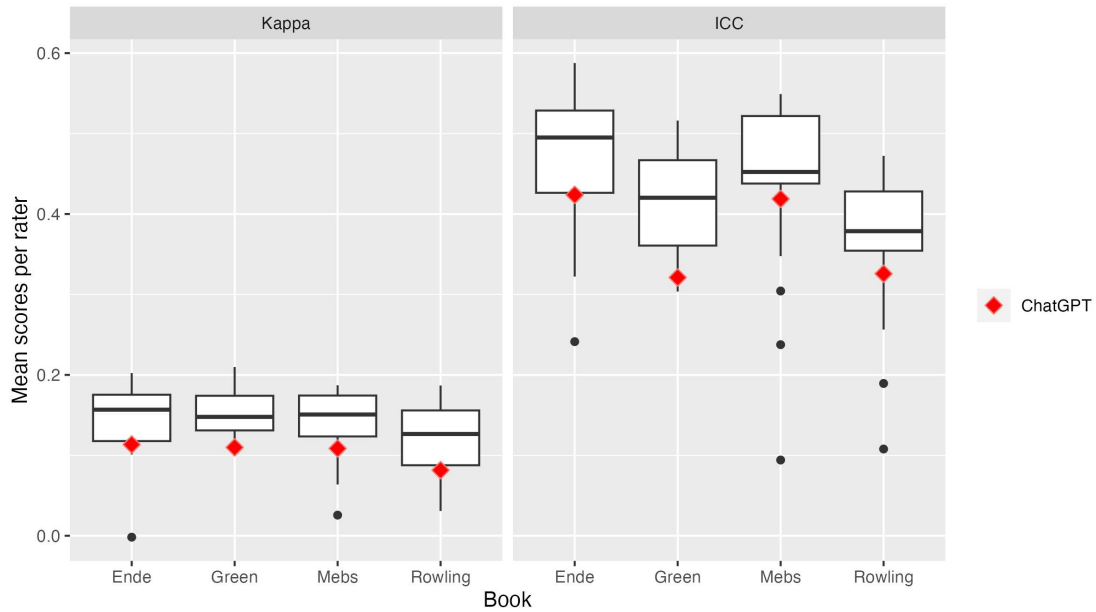
**Table 1**  
ChatGPT Prompt Overview

	P1: Sentence Rating	P2: Context Window	P3: Context Window Verification	P4: Sentiment Analysis
System Prompt	How negative or positive is this German sentence on a 1-7 scale? Answer only with an integer number (1 = very negative; 7 = very positive)	You will analyse a sequence of sentences in German language, separated by the newline symbol. You will have to evaluate the sentiment of the very last sentence on a 1-7 scale, using the previous ones just to determine its context. Answer only with an integer number (1 = very negative; 7 = very positive).	You will analyse a sequence of sentences in German language, separated by the newline symbol. You will have to evaluate the sentiment of the very last sentence on a 1-7 scale, using the previous ones just to determine its context. Answer with a json object having this key-value structure: sentence: the analysed sentence, valence: an integer number (1 = very negative; 7 = very positive)	Your role is to output the probability of a German sentence being positive, negative or neutral. Answer with a json object having this key-value structure: positive: probability of being positive, neutral: probability of being neutral, negative: probability of being negative
User Prompt (Sample)	„Sir - wie haben Sie sich die Hand verletzt?“, fragte Harry erneut und blickte mit einer Mischung aus Abscheu und Mitleid auf die geschwärtzten Finger.	„Ich wünschte einfach, das Ganze wäre nie passiert.“ Die Krebsgeschichte.“	„Sie heißt Hogwarts“, sagte Dumbledore. \n „Und wie kommt es, dass Sie an Tom interessiert sind?“	„Nun ja“, sagte sie unsicher, „ich weiß nicht ... es würde Malfoy ähnlich sehen, sich wichtiger zu machen, als er eigentlich ist ... Aber so was zu behaupten ist schon eine dicke Lüge ...“
Response (Sample)	3	3	{sentence: „Und wie kommt es, dass Sie an Tom interessiert sind?“, valence: 4}	{positive: 0.2, neutral: 0.6, negative: 0.2}

multiple answers to the same question by increasing the “temperature” of the model (thus ideally generating the answers of 20—or even hundreds—of different raters [24]). However, such a reasoning does not hold when considering the repetition of the same task on multiple samples (as done by the human raters on the 4books dataset). The “temperature” parameter, in fact, adds an element of randomness to each single answer (which can therefore be considered as new and original in itself), thus not allowing to link a group of them to a single individual. We therefore focused on the possibility of considering ChatGPT as an individual rater, by reducing its temperature to the lowest level (zero), which makes its behaviour the most deterministic [31]—and, by consequence, the most consistent [24, 23].

We verified this assumption by working on a ~10% sample of the dataset (2,000 sentences, randomly selected). Each sentence of this sample was processed by sending an API request with the system prompt<sup>5</sup> corresponding to “P1: Sentence Rating” in Table 1. This prompt was conceived to imitate as closely as possible the guidelines followed by human raters in the creation of the 4books dataset and taking as an example the prompts described by Rathje et al. [23].

<sup>5</sup>In the ChatGPT API, “system” is the type of prompt that determines the behaviour of the agent.



**Figure 2:** Inter-rater agreement scores overview.

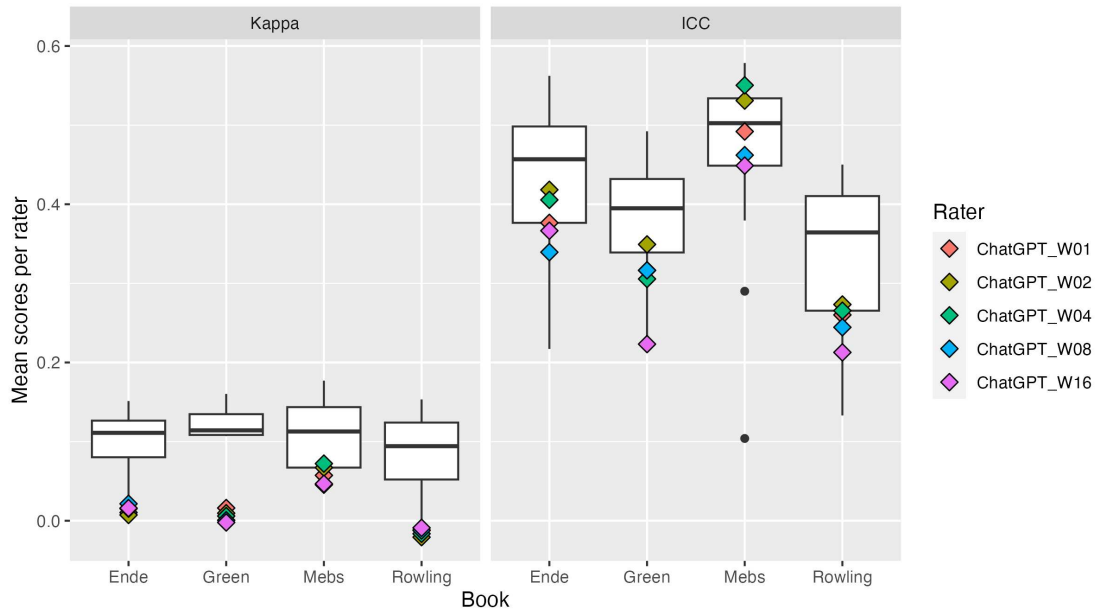
For each request, 20 different completion choices were collected and compared with each other, verifying that in 98% of the cases (1,965 out of 2,000) completions were identical. While a margin of inconsistency remained present, ChatGPT can still be considered as corresponding to a single agent (if we recognize the impossibility also for a human rater to reach a 100% consistency when repeating a task 20 times).

We then extended the analysis to the entire 4books dataset, by collecting only one completion choice with the same setup as described above. Out of 22,860 sentences, ChatGPT refused to produce such a score in 44 cases, requesting instead more context information. A qualitative analysis of these 44 sentences showed that they were mostly brief or broken clauses (a consequence of the automated text segmentation at the beginning of the rating collection). We therefore decided to exclude them from the analysis.

A comparison between ChatGPT and human ratings is shown by Figure 2<sup>6</sup>. The scores are in line with frequently reported agreement rates in sentiment annotation of literary texts [13]. In this context, ChatGPT always placed itself below the level of agreement between most human raters.

However, one of the main issues in comparing the above-described setup with human annotation is that ChatGPT was given just a single sentence at each API call, while human raters had the opportunity to contextualize it while reading the entire text. Thus, to confirm the validity

<sup>6</sup>Scores in Figure 2 are obtained by first calculating inter-rater agreement scores between each pair of raters using two different methods: Cohen’s Kappa and Intraclass Correlation Coefficient of type  $ICC(3, 1)$  [27]. Scores were then grouped based on the presence or absence of a specific rater. This produced mean scores for each rater, indicative of how much s/he agreed or disagreed with the others.



**Figure 3:** Inter-rater agreement scores overview (with context windows).

of these results we devised another experiment on a subsection of the dataset (2,000 sentences), providing context together with the sentences to be rated.

The simplest way to mimic human experience was to add to the prompt the sentences preceding the one to be annotated. We tested five different configurations, with context windows of respectively 1, 2, 4, 8, and 16 preceding sentences. The corresponding system prompt “P2: Context Window” can be found in Table 1.

Beforehand, we again performed a consistency check (with the same modality as described above) on 10% of the dataset (200 sentences), obtaining 96.6% of consistent answers.

Results of the full analysis are presented in Figure 3<sup>7</sup> and show the substantial inefficiency of context windows in increasing ChatGPT agreement. In most of the cases, agreement does not improve substantially, with just one exception for the book by Mebs.<sup>8</sup> Overall, increasing the context window does not correlate with higher inter-rater agreement, with worse results produced for longer windows (see ICC scores for 16-sentence windows).

Among the 78 cases when ChatGPT produced an invalid answer (proportionally higher than in the previous experiment), there were 13 cases when ChatGPT misunderstood the task, generating a continuation of the story instead of its sentiment evaluation.

To verify that there were no problems in distinguishing target sentence from context, we repeated the operation on 10% of the dataset (200 sentences) and changed the final part of the previous system prompt (refer to “P3: Context Window Verification” in Table 1). We verified

<sup>7</sup>Scores are calculated as in Figure 2, by excluding inter-rater agreement between the five different configurations.

<sup>8</sup>This difference seems to be mainly due to text difficulty, being Mebs addressed to 6-8 year olds.



**Table 2**  
R-squared scores overview

Book	ChatGPT	SentiArt	Pre-Trained Transformers	Fine-Tuned Transformers
Ende	0.354	0.140	0.180	0.351 (sd=0.060)
Green	0.251	0.096	0.182	0.308 (sd=0.044)
Mebis	0.330	0.145	0.217	0.390 (sd=0.062)
Rowling	0.247	0.097	0.138	0.328 (sd=0.026)

that in 72% of the cases the analysed sentence was the same as the target sentence.<sup>9</sup> A qualitative analysis of the remaining 28% revealed how in most of the cases ChatGPT simply extracted key sections from the target sentence (e.g., direct speech), but never mixed it with the previous ones.

## 4.2. ChatGPT as a Sentiment Analysis Tool

For our second hypothesis (if ChatGPT could substitute state of the art Sentiment Analysis tools) we chose a phrasing to mimic the output of Transformer models for Sentiment Analysis, which generally produce probabilities over three sentiment classes (for the prompt, refer to “P4: Sentiment Analysis” in Table 1).

The adopted procedure was the same as described above, starting from the verification of output consistency (96% ) on a subsample of 2,000 sentences.

Table 2 presents results of the analysis on the full corpus, compared to three other approaches (pre-trained and fine-tuned Transformer models and a dictionary-based approach).<sup>10</sup> To get a more fine-grained reference point, we took as ground truth the mean of all human ratings per sentence (normalized between  $-1$  and  $+1$ ). Such a value could be directly compared with the output of the dictionary-based approach (SentiArt, developed by Jacobs [12] for German language) and was used to fine-tune the Transformer model via linear regression. To make the output of ChatGPT and pre-trained Transformer models comparable, we first converted classes to numbers (e.g., “positive”:  $+1$ ; “neutral”:  $0$ ; “negative”:  $-1$ ), we then multiplied each number by the probability of the corresponding class and finally summed the obtained values (this can be synthetically described as a “weighted mean”).<sup>11</sup>

ChatGPT clearly outperforms all approaches, apart from the fine-tuned Transformer model. This result is in line with previous studies [23, 31].

<sup>9</sup>Edit distance was  $<5\%$  of its total number of characters.

<sup>10</sup>For the pre-trained Transformer model, we report the results of *cardiffnlp/twitter-xlm-roberta-base-sentiment*, which performed the best on our dataset when compared to six other models tagged for German Sentiment Analysis in *huggingface.co*—see our [GitLab repository](#) for more details. Fine-tuned Transformer model is *deepset/gbert-base*, for which we report both means and standard deviations, because r-squared scores were obtained by performing a 10-fold cross validation.

<sup>11</sup>Note how the scores obtained with this procedure for ChatGPT were strongly correlated with the ones obtained with the prompt P1 in Table 1 (Pearson’s correlation = 0.752).



## 5. Conclusions

While ChatGPT seems to be not yet fully comparable to the behaviour of human raters, it can already work as a valid substitute to the most widely used Sentiment Analysis tools in German language.

However, due to intrinsic limitations of our dataset and possible improvements of our study design, such conclusions should still be considered provisional. Among the limitations, it should be noted how the human ratings we collected were not curated, so they cannot be considered as professional annotations. There is in fact the possibility that ChatGPT’s disagreement might be due to its choosing the most fitting answer, while the majority of the raters gave a less fitting one (a possibility that is suggested by studies like the one by Gilardi et al. [7]). Due to the structure of the 4books dataset, we cannot unfortunately clarify this doubt. However, our final goal was not to check if ChatGPT behaves like a professional annotator (who indeed annotates a document as mechanically as possible), but if its behaviour is comparable to that of humans (which inevitably implies an element of disagreement).

Among the improvements, there is the possibility of repeating our analyses with other and more advanced LLMs (like GPT-4, but also Llama2, Bard, Claude2, and many others). With our paper, we just set up the groundwork for such an experimentation. Other possible improvements relate to the prompting technique, which could have approached the problem from a few-shot or chain-of-thought perspective (especially when studying the effects of context). Also, it has been noted how the formulation of the prompt itself can have an impact on LLMs performance [14], to the point that even a slightly different prompt could lead to substantially different results. While having acknowledged all this, we decided to choose the easiest possible design, as it could be more efficiently tested.

One main problem which connects all these issues, still, is that of economic resources. The OpenAI API usage for this paper cost in total \$17.52, but such a limited cost was due to the strategies we implemented to reduce API usage (such as random sampling). If we wanted to apply all analyses to the full dataset and test different prompting techniques, the cost would have easily increased to hundreds of dollars (while still using the much cheaper gpt-3.5-turbo model).

This problem connects to the issue of transparency in LLMs, which are still frequently studied as black boxes, understandable only via direct stimulation/observation, like some kind of biological system. In this regard, our research has not only provided new evidence to such a study, but has also allowed recognizing how the perspective from which LLMs are viewed—being intended just as computational tools or as possible substitutes of human agents—inevitably shapes the way in which they are studied and understood.

## References

- [1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*. arXiv. 2023. DOI: 10.48550/arXiv.2303.12712.

- [2] K. Dennerlein, T. Schmidt, and C. Wolff. “Emotionen im kulturellen Gedächtnis bewahren”. In: *Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbandes ”Digital Humanities im deutschsprachigen Raum” (DHd 2022)*. Trier, 2022, pp. 93–98. doi: 10.5283/e-pub.54152. URL: <https://epub.uni-regensburg.de/54152/>.
- [3] B. Ding, C. Qin, L. Liu, L. Bing, S. Joty, and B. Li. *Is gpt-3 a Good Data Annotator?* arXiv. 2022. doi: 10.48550/arXiv.2212.10450.
- [4] K. Elkins and J. Chun. “Can GPT-3 Pass a Writer’s Turing Test?” In: *Journal of Cultural Analytics* 5.2 (2020), pp. 1–16. doi: 10.22148/001c.17212.
- [5] M. Ende. *Jim Knopf und Lukas der Lokomotivführer*. Stuttgart: Thienemann, 2020.
- [6] H. Friederichs, W. J. Friederichs, and M. März. “ChatGPT in Medical School: How Successful Is AI in Progress Testing?” In: *Medical Education Online* 28.1 (2023). doi: 10.1080/10872981.2023.2220920.
- [7] F. Gilardi, M. Alizadeh, and M. Kubli. *Chatgpt Outperforms Crowd-Workers for Text-Annotation Tasks*. arXiv. 2023. doi: 10.48550/arXiv.2303.15056.
- [8] J. Green. *Das Schicksal ist ein mieser Verräter*. München: Hanser, 2012.
- [9] S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. *Reasoning with Language Model is Planning with World Model*. arXiv. 2023. doi: 10.48550/arXiv.2305.14992.
- [10] A. Hintze. *ChatGPT Believes It Is Conscious*. arXiv. 2023. doi: 10.48550/arXiv.2304.12898.
- [11] F. Huang, H. Kwak, and J. An. *Is chatgpt Better than Human Aannotators? Potential and Limitations of chatgpt in Explaining Implicit Hate Speech*. arXiv. 2023. doi: 10.48550/arXiv.2302.07736.
- [12] A. M. Jacobs. “Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics”. In: *Frontiers in Robotics and AI* 6 (2019), pp. 1–13. doi: 10.3389/frobt.2019.00053.
- [13] E. Kim and R. Klinger. “Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1345–1359. URL: <http://aclweb.org/anthology/C18-1114>.
- [14] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, et al. “ChatGPT: Jack of All Trades, Master of None”. In: *Information Fusion* (2023), p. 101861. doi: 10.1016/j.inffus.2023.101861.
- [15] M. Kosinski. *Theory of Mind May Have Spontaneously Emerged in Large Language Models*. arXiv. 2023. doi: 10.48550/arXiv.2302.02083.
- [16] N. Kucirkova. “How Could Children’s Storybooks Promote Empathy? A Conceptual Framework Based on Developmental Psychology and Literary Theory”. In: *Frontiers in Psychology* 10 (2019), p. 121. doi: 10.3389/fpsyg.2019.00121.
- [17] T. Kuzman, N. Ljubešić, and I. Mozetič. *Chatgpt: Beginning of an End of Manual Annotation? Use Case of Automatic Genre Identification*. arXiv. 2023. doi: 10.48550/arXiv.2303.03953.

- [18] X. Li, X. Zhu, Z. Ma, X. Liu, and S. Shah. *Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? An Examination on Several Typical Tasks*. arXiv. 2023. DOI: 10.48550/arXiv.2305.05862.
- [19] J. Lüdtke and A. M. Jacobs. “On a Rollercoaster with Frieder, Jim, Hazel and Harry: Identifying Emotional Arcs in Reader Responses to Children and Youth Books”. In: *Igel 2023*. Monopoli, 2023.
- [20] G. Mebs. “*Oma!*”, *schreit der Frieder*. Düsseldorf: Sauerländer, 2005.
- [21] M. Nikolajeva. *Reading for Learning: Cognitive Approaches to Children’s Literature*. Vol. 3. Children’s Literature, Culture, and Cognition. Amsterdam: John Benjamins Publishing Company, 2014. DOI: 10.1075/clcc.3.
- [22] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?* arXiv. 2023. DOI: 10.48550/arXiv.2302.06476.
- [23] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjeh, C. Robertson, and J. J. Van Bavel. *GPT Is an Effective Tool for Multilingual Psychological Text Analysis*. PsyArXiv. 2023. URL: <https://psyarxiv.com/sekf5/>.
- [24] M. V. Reiss. *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark*. arXiv. 2023. DOI: 10.48550/arXiv.2304.11085.
- [25] J. K. Rowling. *Harry Potter und der Halbblutprinz*. Hamburg: Carlsen, 2018.
- [26] T. Schmidt, J. Fehle, M. Weissenbacher, J. Richter, P. Gottschalk, and C. Wolff. “Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election”. In: *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. Potsdam, 2022, pp. 74–87.
- [27] P. E. Shrout and J. L. Fleiss. “Intraclass correlations: Uses in assessing rater reliability.” In: *Psychological Bulletin* 86.2 (1979), pp. 420–428. DOI: 10.1037/0033-2909.86.2.420.
- [28] X. Sun, L. Dong, X. Li, Z. Wan, S. Wang, T. Zhang, J. Li, F. Cheng, L. Lyu, F. Wu, and G. Wang. *Pushing the Limits of ChatGPT on NLP Tasks*. arXiv. 2023. DOI: 10.48550/arXiv.2306.09719.
- [29] P. Törnberg. *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. arXiv. 2023. DOI: 10.48550/arXiv.2304.06588.
- [30] J. Wang, Y. Liang, F. Meng, B. Zou, Z. Li, J. Qu, and J. Zhou. *Zero-Shot Cross-Lingual Summarization via Large Language Models*. arXiv. 2023. DOI: 10.48550/arXiv.2302.14229.
- [31] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia. *Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study*. arXiv. 2023. DOI: 10.48550/arXiv.2304.04339.
- [32] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao. *Can chatgpt Understand Too? A Comparative Study on chatgpt and Fine-Tuned Bert*. arXiv. 2023. DOI: 10.48550/arXiv.2302.10198.