Using ChatGPT for Annotation of Attitude within the Appraisal Theory: Lessons Learned

Mirela Imamović
University of Hildesheim
imamovic@uni-hildesheim.de

Silvana Deilen
University of Hildesheim
deilen@uni-hildesheim.de

Dylan Glynn
University of Paris 8
dsg.up8@gmail.com

Ekaterina Lapshinova-Koltunski
University of Hildesheim
lapshinovakoltun@uni-hildesheim.de

Abstract

We investigate the potential of using ChatGPT to annotate complex linguistic phenomena, such as language of evaluation, attitude and emotion. For this, we automatically annotate 11 texts in English, which represent spoken popular science, and evaluate the annotations manually. Our results show that ChatGPT has good precision in itemisation, i.e. detecting linguistic items in the text that carry evaluative meaning. However, we also find that the recall is very low. Besides that, we state that the tool fails in labeling the detected items with the correct categories on a more fine-grained level of granularity. We analyse the errors to find systematic errors related to specific categories in the annotation scheme.

1 Introduction

The present paper investigates the potential of using large language models (LLMs), specifically ChatGPT for annotating pragmatic categories. Recent advances in Artificial Intelligence (AI), propelled by LLMs such as ChatGPT, have led to substantial improvements in automating complex linguistic tasks such as coherent text generation, text simplification, machine translation, error detection, and question answering. They have produced unprecedented results in a wide range of applications including linguistic annotations.

We focus on the feasibility of using ChatGPT to annotate linguistic items expressing evaluation, attitude and emotion according to the framework based on Appraisal Theory (Martin and White, 2005). More specifically, we probe the tool for the annotation of evaluative language in spoken popular science discourse. To this end, we automatically annotate selected English texts extracted from a dataset of TED talks and evaluate the annotation results manually to see how well ChatGPT can recognise linguistic items carrying the pragmatic meaning under analysis and if these items were correctly classified according to the pre-defined categories of the Appraisal Theory. We also analyse erroneous cases to find the systematic failure of the tool in assigning specific categories.

So far, especially when investigating complex pragmatic and discourse features, most of the annotation work is still done manually. However, manual corpus annotation is time-consuming and requires specialised skills and training. Automating this task would significantly reduce the annotator's workload, save resources and costs, and would allow researchers to build and annotate larger and thus more representative corpora. Moreover, manual annotation is error-prone and is subject to inconsistencies which can be avoided by automatic or semi-automatic procedures integrated into the process of annotation.

Thus, our aim is two-fold. On the one hand, we aim to evaluate ChatGPT in a demanding annotation task of complex linguistic phenomena. On the other hand, we want to learn about the challenges present in the data at hand. This knowledge will allow us to detect phenomena that are particularly hard to annotate, which in turn will allow us to create better guidelines for human annotators.

The remainder of the paper is structured as follows. In Section 2, we outline the basic principles of the Appraisal Theory and present the categories we consider in our analysis. Section 3 describes related works that utilise LLMs for annotating. In Section 4, we provide details on the data used for this study, as well as procedures to annotate the data. We present and discuss our results in Section 5. An outlook for future work is given in Section 6. We discuss the limitations of our work in Section 7.

2 Theoretical Background

We use Appraisal Theory (Martin and White, 2005) developed under the general framework of Sys-

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 112–123

temic Functional Linguistics (SFL, Halliday and Matthiessen, 2014). Appraisal is related to the interpersonal metafunction of language and it consists of three interrelated domains: Attitude (expressions of opinions and feelings), Engagement (positioning of the writer/speaker toward other voices within the discourse) and Graduation (intensifying or down-toning attitude and engagement). We focus on the three sub-types of Attitude: Affect (emotions), Judgement (moral judgements about human behaviours) and Appreciation (evaluations of objects, products and processes). We use the original appraisal annotation scheme and consider these three sub-types of Attitude as the main semantic categories (level 1). Furthermore, we focus on the sub-categories of Attitude on a more fine-grained level (level 2) and their related subvalues that represent the finest level of granularity (level 3). The categories on levels 2 and 3 include (with the 3rd level given in brackets): for Affect DIS/INCLINATION (fear; desire), UN/HAPPINESS (misery, antipathy; cheer, affection), IN/SECURITY (disquiet, apprehension; confidence, trust), and DIS/SATISFACTION (ennui, displeasure; interest, pleasure), for Judgement SOCIAL ESTEEM (normality, capacity, tenacity), and SOCIAL SANCTION (veracity, propriety), and for Appreciation REACTION (impact, quality), COMPOSITION (balance, complexity), and VALUATION. Our detailed scheme is presented in Figure 1.

3 Related Work

3.1 Annotating Appraisal

Appraisal Theory for annotation of evaluative language in English book reviews was used by Read et al. (2007). The authors proposed a multi-step strategy and analysed the inter-annotator agreement (IAA) for both itemisation and category assignment. The agreement varied depending on the level of abstraction in the Appraisal hierarchy, with a better result (a mean F-score of 0.698) for the highest level and a much worse result (a mean F-score of 0.395) at the most fine-grained level. Interestingly, the authors reported that the agreement was dependent on the category annotated: a better agreement was achieved for Attitude if compared to Engagement or Graduation. In our work, we focus on Attitude only.

Mora and Lavid-López (2018) also used Appraisal theory to annotate English and Spanish reviews. The authors stated some problematic issues in the application of the categories in Appraisal theory for the specific dataset at hand. The reported IAA was very high in both itemisation and classification of the main categories tasks. However, the authors also stated problematic cases, e.g. in the case of long and complex sentences that need contextualisation to convey an evaluative meaning. The agreement on a more fine-grained level was much lower and contrary to the findings of Read et al. (2007), the main problems here were caused by the subtypes of Attitude.

More recent work on annotation using Appraisal Theory includes (Dong and Fang, 2023). However, they do not report any agreement scores. In our study, we analyse the agreement between the automatic annotation by ChatGPT and a human evaluator. We expect that the results for categories on the higher level of granularity will be higher.

3.2 LLMs for annotation tasks

The potential of large language models (LLMs) for data annotation has been explored in some recent studies. For instance, some authors tested the potential of LLMs for crowd-sourcing approaches. Testing LLMs (with a focus on ChatGPT) against crowd-workers, Gilardi et al. (2023) showed that for most of the annotation tasks, ChatGPT's accuracy was higher than that of crowd-workers. The models' IAA also exceeded that of both trained annotators and crowd-workers. As using ChatGPT is more cost-effective than hiring professional annotators and even crowd-workers, the authors conclude that LLMs have a significant potential to transform common text-annotation procedures and to increase their efficiency. This was also confirmed by Ostyakova et al. (2023) who compared ChatGPT with not only crowd-workers but also human experts. The authors stated that in some cases LLMs could achieve human-like performance following a multi-step pipeline on complex discourse annotation.

However, for efficient crowd-sourced annotation procedures, LLMs should be provided with sufficient guidance and demonstrated examples as it was shown by He et al. (2023). The authors proposed a two-step approach called 'explain-then-annotate'. They created prompts for every demonstrated example, which they then subsequently utilized to prompt the language model GPT-3.5 to provide an explanation for why the specific ground truth answer/label was chosen for that particular example. Following this, they constructed the

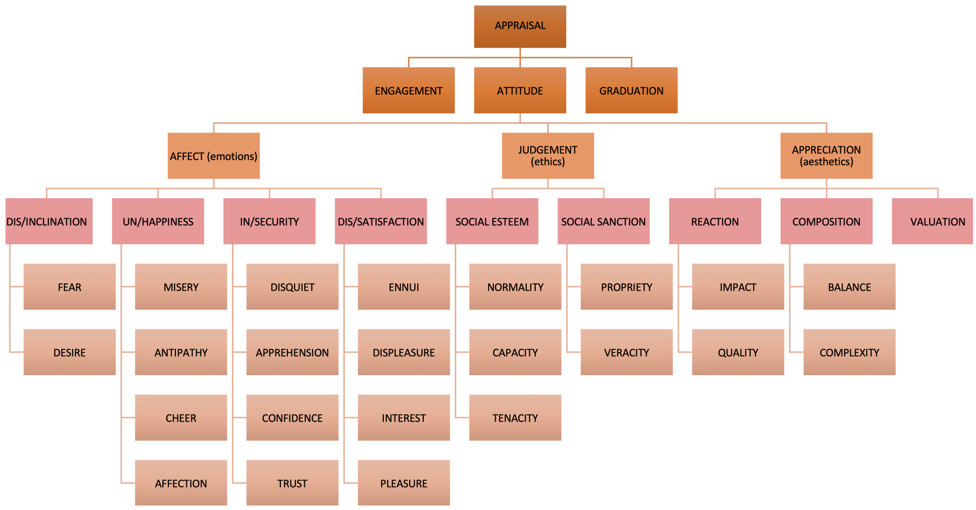

Figure 1: Analysis scheme according to Appraisal Theory, based on Martin and White (2005)

few-shot chain-of-thought prompt with the self-generated explanation and employed it to annotate the unlabeled data. The authors conducted experiments on three tasks and their results were that for the two out of the three tasks, GPT-3.5 achieved results that were comparable to those obtained through crowd-sourced annotation.

This was also confirmed by Ding et al. (2023) who found that LLMs have the potential to accurately annotate data for different NLP tasks while requiring only a fraction of the time and cost of human annotators.

The range of annotation tasks for LLMs, i.e. linguistic categories in focus, is varying from named entity recognition and relation extraction (Zhang et al., 2023), semantic verb classes (Straková et al., 2023), sentence meaning structure (Ettinger et al., 2023) to more complex task of pragmatic annotation. For instance, Yu et al. (2023) explored the feasibility of using LLMs in the annotation of apologies. Their study showed that the models were able to identify key features of apologies with high accuracy. The models seemed to achieve results comparable with human annotations.

In another study (Nedilko, 2023), ChatGPT was used for emotion detection by leveraging the prompt engineering and zero-shot as well as few-shot learning methodologies based on multiple experiments showing improvement over their baseline model. However, the authors state that although ChatGPT provides stable results, especially if asked for a specific output format, there is still an element of volatility due to the conversational nature of the model. They also note that the context window limitation does not allow for working with larger datasets. Furthermore, the authors admit that the emotion detection task remains challenging for machines in general.

In our study, we use ChatGPT for a pragmatic annotation task, which is challenging not only for machines but also for humans.

4 Research Design

4.1 Data

For this study, we selected 11 texts (25,117 words) from a dataset of TED talks collected for a bigger project on the analysis of evaluative language[1]. These texts cover talks in eleven disciplines (Arts, Business, Education, Entertainment, History, Medicine, Natural science, Philosophy, Politics and Law, Psychology, and Technology). The individual text size is given in Appendix B. The communicative aim of the talks is two-fold. On the one hand, they serve for the knowledge transfer between experts and laypeople. On the other hand, they also aim at entertaining the audience. We assume that evaluative language is used for both better knowledge transfer and entertainment purposes. The texts are transcripts of speeches available on the TED website[2]. We selected talks by both female and male native speakers of North American English. For all the selected texts, there are also translations into German available. However, their analysis remains beyond the scope of this paper.

[1] The data used for the current study including the annotation results is provided in the GitHub repository https://github.com/katjakaterina/chatgptanno.

[2] https://www.ted.com

4.2 Data Analysis

The annotation task was conducted using the large language model ChatGPT[3]. ChatGPT was used via the chat interface with default settings (temperature = 1).

In our prompt, we first included information on Appraisal Theory and then asked the tool (in the same prompt) to annotate all the evaluative linguistic instances of Affect, Judgement and Appreciation in the given text. We also requested to only assign one value to each evaluative linguistic instance. Another request (still the same prompt) was to focus on verbs, nouns, adjectives and adverbs, i.e. explicit (directly inscribed) attitudes and to only annotate the linguistic instances that correspond to Affect, Judgement and Appreciation and its sub-values instead of annotating the whole sentence. The exact prompt we used is given in the Appendix A. As we added basic information on Appraisal Theory to our prompt, our prompting can be classified as instruction-based or context-based prompting, rather than zero-shot prompting. The human annotator followed the same guidelines given in the prompt. The annotation scheme used is presented in Section 2.

5 Results

5.1 Itemisation

In the first step, we experimented with running the same prompt several times. More precisely, we ran it two times for each text and observed diverging results. In some cases, the output did not contain the actual items that were supposed to be tagged as instances of Appraisal, i.e. the output would not contain any data. In this case, we ran the prompt again. The number of items identified in subsequent queries varied, even though the prompts were run within minutes of each other. The fact that we used ChatGPT via the chat interface and with the default setting (temperature = 1) might explain why the system created very different outputs in the two runs.

Table 1 presents the numbers of items returned for each text for the first two successful prompts (with the same prompt formulation).

The differences in the number of items returned were significant (Chi^2, df = 10, p = 4.575^{-10}) and the discrepancy for the texts Art, Med, and Pol[4] is, perhaps, noteworthy. These differences may inform the results concerning the classification of items, reported below. The ensuing results on the accuracy of the itemisation, i.e. identification of evaluative items, are all based exclusively on the returns for the first query.

In the second step, we analysed the output annotated by ChatGPT in terms of item recognition precision – if all instances marked by the tool were correctly marked as evaluative. A human annotator, a trained linguist with a theoretical background in Appraisal Theory, evaluated the output of ChatGPT. The results showed that out of 381 items tagged by the tool, 21 were false positives, which means that ChatGPT achieved a high precision (94.49%) in itemisation.

Then, the human annotator analysed a subsample of 5 texts (Art, Bus, Edu, Ent, His) to assess the itemisation results in terms of recall (the number of evaluative items in the texts missed by ChatGPT). The results showed that a total of 485 evaluative instances remained undetected. The number of the automatically detected items (true positives) in these five texts comprises 177, so the tool achieves a recall of 26.74%, which is rather low.

5.2 Error analysis

To determine if some types of evaluation were more problematic than others, the items missed by ChatGPT in the five texts were manually annotated for the most coarse-grained level of evaluative distinction, namely Affect, Judgement and Appreciation. The results showed that in total, Appreciation was most frequently missed (230 items) followed by Judgement (159) and Affect (96). We also noticed that this tendency was not the same across the texts. While in Bus, the erroneous tagging was more evenly distributed across the categories (20 Affect, 16 Judgement and 19 Appreciation), in His, Judgement was much more frequently missed than the other two categories (27 Affect, 76 Judgement and 45 Appreciation) and in Edu, it was Appreciation where ChatGPT most frequently failed to detect evaluative meaning (15 Affect, 37 Judgement and 88 Appreciation). However, to be able to relate this

[3]Our study was conducted in November 2023, i.e., the results are based on GPT-4 Turbo, the latest version of ChatGPT available at the time of writing.

[4]The text IDs represent disciplines: Art=Arts, Bus=Business, Edu=Education, Ent=Entertainment, His=History, Med=Medicine, NatSci=Natural Science, Phil=Pilosophy, Pol=Politics and Law, Psy=Psychology, Tech=Technology. However, the analysis of disciplines remains beyond the scope of this study.

Prompt run/Text	Art	Bus	Edu	Ent	His	Med	NatSci	Phil	Pol	Psy	Tech
1st run	48	22	35	63	29	46	37	21	26	33	30
2nd run	20	28	21	30	50	75	16	27	50	16	16

Table 1: Number of items retrieved in two runs of the prompt per text.

Text part/Text	Art	Bus	Edu	Ent	His
Full-length	35	21	35	59	27
Two-part	59	62	61	45	50
Three-part	82	50	107	105	92

Table 2: Number of items retrieved for two-part and three-part text-splitting procedure for 5 texts versus full-length texts.

variation in the results to the domain effects, we need to analyse more texts per discipline.

An example of a missed item representing Affect is illustrated in (1) and another example illustrating several Attitude types missed by the tool is shown in (2).

(1) *They felt science had stagnated since the days of the scientific revolution that had happened in the 17th century.*

(2) *A couple of years ago, I put a video on Youtube, and in the video, I acted out every terrible [Judgement] conference call you've ever been on. It goes on for about five minutes, and it has all the things that we hate [Affect] about really bad [Appreciation] meetings.*

In the next step, we experimented further with the prompts and inputs. For instance, we noticed that ChatGPT could not deal with longer texts returning approximately the same number of items for each text (around 30). So, we split each of the 5 texts (that we manually analysed) into two and three parts. The text parts were run in separate sessions. Table 2 shows the number of items (both true and false positives) returned for the two-part and three-part text-splitting procedure in comparison to the output for full-length texts (true positives). We also slightly modified the prompt and asked the tool to include the context around the identified evaluative instances. Splitting the texts into two halves doubled the output in terms of the number of identified items (both true and false positives). Splitting the texts into three parts returned the highest amount of identified items (the highest being 107). We assume that ChatGPT can deal better with shorter texts. An example of an item which was missed in the first prompt but captured in the three-part-split text is illustrated in (3) (with the item in question marked in bold). It also correctly marked *immoral* as Judgement: PROPRIETY. However, the tool still failed to identify other evaluative items (as those underlined that were tagged by the human annotator) within the same context.

(3) *And there's a lot to be overwhelmed [Affect] about, to be fair [Judgement] – an environmental crisis [Judgement], wealth disparity [Judgement] in this country unlike we've seen since 1928, and globally, a totally **immoral** [Judgement] and ongoing wealth disparity [Judgement]. Xenophobia's [Judgement] on the rise. The trafficking [Judgement] of women and girls. It's enough to make you feel very overwhelmed [Affect].*

As an example, one text (that was also used for the analysis) in our data set had a total of 148 false negatives in the first prompt. The number of true positives in this output was 27. In the experiments with two- and three-part-split texts, the tool returned 50 and 92 items (both true and false positives) respectively. In another text, the number of detected items increased from 35 (true positives) to 59 (two-part-split) and 82 (three-part-split) (both true and false positives). However, even with the increased number of detected items, there are still many false negatives left, hence the recall remaining low. Another issue is the linguistic context, i.e. the tool coming up with evaluative items based on the contextual information as illustrated in examples (4), (5) and (6). Those phrases (*astounding, revolutionary, unequal*) do not appear in the original text. However, the tool listed them as evaluative based on the context (the sentences in quotation marks). ChatGPT also wrongly classified the items as Attitude.

(4) *astounding - "how could the word 'scien-*

tist' not have existed until 1833?" [Appreciation: IMPACT*].*

(5) *revolutionary - "pledged to bring about, and what's so amazing about these guys is, not only did they have these grandiose undergraduate dreams" [Appreciation:* VALUATION*].*

(6) *unequal - "wealth disparity in this country" [Judgement:* PROPRIETY*].*

Although the prompt we used explicitly stated that only evaluative instances that appear in the text should be listed, ChatGPT still performed this surprising action. The fact that the tool can "retrieve" an evaluative expression based on the contextual information is indicative of potentially recognising implicit evaluation/attitude. However, more studies are needed to confirm this assumption.

Interestingly, the output also contained 15 items that did not occur in the text. This is not a definite number, as checking each instance manually would be laborious. This is an indicator of hallucinations observed by other existing studies using ChatGPT for NLP applications (Zhang et al., 2023; Peng et al., 2023; Guerreiro et al., 2023).

5.3 Classification

We start with the analysis of the main categories of Attitude from the Appraisal scheme (Affect, Judgement, Appreciation) as defined in Section 2. The tool recognised 360 items in total (160 Affect, 89 Judgement and 108 Appreciation, 2 were marked as "difficulty" and 1 as "ease"). The categories assigned by the human annotator were 95 Affect, 101 Judgement and 164 Appreciation. The level of IAA between the tool and human annotator at this 1st level of abstractness was calculated using Cohen's weighted Kappa (Cohen, 1968). The human classification for these three nominal categories was done independently of the tool and resulted in a Kappa of 0.52 (lower.bound = 0.43, upper.bound = 0.61, $p < .0001$). As a general guide, values between 0.4 and 0.75 are taken to be indicative of a fair level of agreement that is above chance. However, given that we are only dealing with three categories and at a level of abstractness that is normally unproblematic for human raters, this score should not be indicative of successful categorisation on the part of ChatGPT.

At the second and intermediary level of abstractness, ChatGPT classified the following Appraisal categories: DISINCLINATION (19), INCLINATION (8), UNHAPPINESS (16), HAPPINESS (33), INSECURITY (17), SECURITY (16), DISSATISFACTION (20), SATISFACTION (31), SOCIAL ESTEEM (49), SOCIAL SANCTION (40), REACTION (65), COMPOSITION (25), VALUATION (18). The tool also marked "difficulty" (2) and "ease" (1) that were not given in the prompt and are not considered as original appraisal categories. The classifications of the human annotator are DISINCLINATION (10), INCLINATION (1), UNHAPPINESS (14), HAPPINESS (19), INSECURITY (16), SECURITY (14), DISSATISFACTION (11), SATISFACTION (10), SOCIAL ESTEEM (57), SOCIAL SANCTION (44), REACTION (76), COMPOSITION (23), VALUATION (65).

At the third and finest level of abstractness, 133 (34.91%) labels matched between ChatGPT and the human annotator. The tool incorrectly classified 248 (65.09%) labels. The second and third levels of abstractness resulted in Kappas of 0.39 (lower.bound = 0.024, upper.bound = 0.75, $p < .0001$) and 0.34 (lower.bound = -0.10, upper.bound = 0.78, $p < .0001$), respectively. Although the third level of abstractness has 24 categories[5], in contrast to the second level which has 13, the IAA was marginally better. Given that only the 1st level of abstractness produced results that show any workable degree of agreement between ChatGPT and the human annotator, we restricted further investigation to this level. Two examples of a disagreement between the tool and the human annotator are shown in (7) and (8).

(7) *And I'm talking about something far more valuable than office furniture. I'm talking about time* (tagged as Judgement: CAPACITY by ChatGPT and as Appreciation: VALUATION by the human annotator).

(8) *If you're a doctor, you can do some good things, but if you're a caring doctor, you can do some other things* (tagged as Affect: TRUST by ChatGPT and as Judgement: PROPRIETY by the human annotator).

These examples show a strong disagreement in the classification of attitude types at the highest level of granularity, i.e. the most abstract attitude categories between the tool and the human annotator. Looking at the context, it becomes clear that

[5]The third level originally contains 23 categories, as indicated in Figure 1. However the label "valuation" is used at both the second and the third level.

valuable does not judge someone's capacity to do something, but evaluates the worth of something and that being *caring* is not a direct expression of emotion, but rather a moral judgement of someone.

As expected, some categories revealed a higher rate of agreement between ChatGPT and the human annotator than others. Table 3 presents the residuals of a Pearson's Chi^2 test for independence between the human and machine classification.

Examining the residuals of a Chi^2 test for independence between the human and machine classification results show that the highest level of agreement was for Judgement, followed by Affect and then Appreciation, but the differences between them were not particularly remarkable. We can also see that none of the mismatches between machine and human were particularly noteworthy. The residuals suggest that there might be two exceptions to this. Firstly, it appears to be relatively unlikely that if ChatGPT classifies an instance as Appreciation that the human will classify it as Affect, a mismatch between Appreciation and Judgement being much more likely. Secondly, the most likely mismatches do occur when ChatGPT classifies the instance as Affect, but the human as Appreciation or when ChaptGPT classifies it as Appreciation, but the human as Judgement. These potential exceptions aside, the results overall suggest relatively random errors in classification, with none of the categories at level 1 abstractness proving substantially easier or more difficult to classify.

5.4 Individual text effects on classification

Having established that a pairwise comparison between ChatGPT and human annotator produces classification only weakly in agreement, the next step is to determine whether this is equally the case for different texts. Of particular interest are the potential effects that domain variation, often associated with different disciplines, may have on ChatGPT's ability to accurately classify certain categories. Although our sample does not allow any claims concerning specific domain effects of certain disciplines, we do observe significant differences between the texts for the level of IAA relative to the different classifications at the 1st level of abstractness.

To determine for which categories agreement or disagreement between the human and machine classifications were significantly higher or lower than chance, we ran a log-linear analysis looking at the number of classifications for each 1st-level category for each text. Figure 2 presents the results in the form of a mosaic plot. Each box represents the classification for the machine and human annotator for each text. The size of the box represents the relative frequency and, based on Pearson residuals, the colour indicates whether the the combination is significantly higher or lower than expected. When both ChatGPT and the human propose the same category and that combination is blue, especially dark blue, this represents a noteworthy and high level of agreement. If the box is blue but the two annotators classified the categories differently, it represents a high level of disagreement. Red indicates a lower number of that combination that would be expected and grey boxes simply indicate that the frequency of that combination is not noteworthy. To interpret the plot, one looks at the top row of the top line of boxes for agreement on Affect, the middle row of the middle line for Appreciation and the bottom row of the bottom line of boxes for agreement on Judgement. For instance, for Affect, we have a substantial agreement for the texts His, Med and Pol, but substantial disagreement for Ent where ChatGPT classifies instances as Affect, but the human as Appreciation and then again for Psy, where the ChatGPT classifies Affect, but the human as Judgement. Looking across the three categories for the 11 texts, we see that the best agreement is for Judgement with six texts (Edu, Ent, NatSci, Phil, Pol and Psy). In contrast, in only 3 texts is there significant agreement for Appreciation and Affect. The two red boxes are difficult to interpret but indicate that ChatGPT and the human are less likely to disagree with the combination of Affect and Appreciation. Psy and Pol are the only two texts where two of the categories are classified in a significantly similar way. In all the other texts, it is only one category or none that reveals significant agreement for a category. Art and Tech reveal no significant agreement for any of the categories. Although clear patterns are difficult to discern, it is only for Affect that we see significant disagreements, neither Appreciation nor Judgement revealing any at all. Although Appreciation has no significant agreement, it only has three texts where there is significant agreement. Judgement has six texts with agreement and none with significant disagreement. These results appear to contradict the results presented in Table 3 where Judgement is best and Appreciation is worse in terms of agreement. The difference in the results indicates that text variation does have a significant effect on the IAA of the classification.

	Human Affect	Human Appreciation	Human Judgement
ChatGPT Affect	6.639010	-2.611789	-3.187939
ChatGPT Appreciation	-4.649263	5.560443	-2.457828
ChatGPT Judgement	-3.668086	-3.668086	6.928130

Table 3: Pearson residuals for Chi^2 test of independence between raters for level 1.

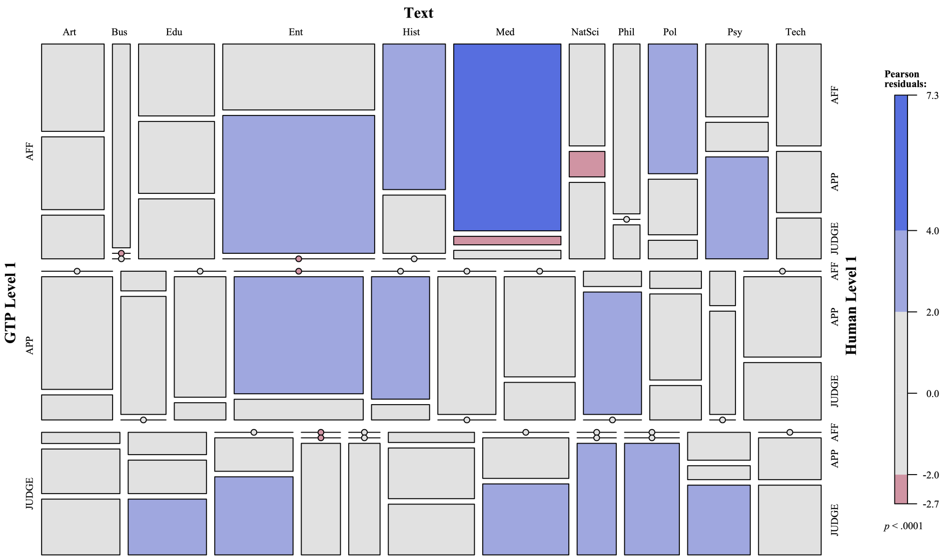

Figure 2: Text variation and inter-annotator level 1 classification

Lastly, in the previous section, we saw that ChatGPT itemisation, when run more than once, showed substantially different numbers of items for texts Art, Med and Pol. It is possible, therefore, that ChatGPT struggled more with these texts. If that is the case, we do not see it reflected at the stage of classification for level 1.

6 Conclusion and Future Work

The present paper evaluates the use of ChatGPT as an annotation tool for evaluative language in spoken popular science discourse. Our results showed that while the concordance rate between the tool and the human annotator in terms of itemisation precision was rather high (94,49%), it was rather low in terms of the sub-category label classification (35%). Although the tool succeeds in recognising evaluative items, it fails in retrieving all of them (as shown by the low recall) and in correctly classifying them.

Also, the tool produced hallucinations in the output. This is especially problematic as human annotators not only have to manually check the results in terms of recall and the assigned category, but they also have to double-check if the annotated item appears in the text at all. This problem could be eliminated by changing the form of the output (see discussion of limitations below).

Our observations also show that most of the recognised evaluative items (58,89%) were adjectives. However, a more detailed analysis of part-of-speech (pos) categories would be an asset. For this, we plan to automatically pos-tag and parse the corpus, which will also help us to define systematic morpho-syntactic patterns of explicit evaluative expressions. However, it would be also interesting to find out if ChatGPT could identify and annotate implicit (indirect) attitudes in texts which is a task demanding even for human annotators, which is also amongst the tasks for our future work.

Besides that, the tool performed better on some texts than on others. However, as we only analysed one text per discipline we cannot determine if the varying results depend on the text discipline or other factors, such as the speaker's style or the topic of the text.

Interestingly, regardless of the text length, the number of instances annotated by ChatGPT was more or less the same (approximately 30 instances per text). Our results show that splitting the texts and performing the annotation task on shorter parts improves the output. This suggests that text length could potentially have an impact on the results. With this assumption, the recall rate of future studies could be improved by adjusting the text length. This, however, does not guarantee that some items will not be missed out completely or not be false positives.

We also found that the output varied across several runs of the same prompt. This means that the results of the automatic annotation are not reliable in terms of consistency, which is in line with other studies. While other existing annotation projects found that GPT-4 was more reliable in annotation than other GPT models (e.g. Pérez et al., 2023), we conclude that even the latest model GPT-4 Turbo exhibits low consistency compared to human annotator for the annotation of evaluative language. We explain the inconsistency of ChatGPT as an annotator of complex linguistic phenomena by the non-deterministic nature of LLM: identical input with minor word alterations in prompts leads to different outputs (see Reiss, 2023, for more details). Thus, improving and designing effective prompts is important for optimal model performance. We will also explore and compare different prompting strategies (few-shot vs. zero-shot). By doing so, we will investigate whether the tool's ability of in-context learning through prompting can increase the concordance rate. Moreover, we will test and compare further existing LLM tools (e.g. Bard[6]).

We also plan to collect annotations by multiple human annotators and compare them with those of ChatGPT and other LLMs. We will use the findings of this study to improve the annotation guidelines for human annotators in terms of itemisation and classification of appraisal categories. Crucially, we will follow up on the evaluative items that ChatGPT incorrectly assigned and observe if the human annotators also tend to disagree regarding those linguistic items, explaining what kind of tasks the tool performs best on. We will also analyse the evaluation span as its length seems to have an impact on the ChatGPT's output and may also cause inconsistencies in human annotator decisions.

Besides, we will annotate several texts of the same discipline which will allow us to identify tendencies in terms of stylistic or domain-specific variation. Therefore, we plan to analyse texts from different genres and compare the results for different text types. Further, we will cross-lingually test ChatGPT's annotation of evaluation in German translations of the analysed texts by comparing originals with translations and exploring how evaluation is translated and if pragmatic meaning is altered.

[6]https://bard.google.com/chat.

7 Discussion of Limitations

Our work has several limitations. First of all, only one human annotator evaluated the output of ChatGPT and performed the manual annotation of the texts. Evaluation by multiple human annotators would provide us with information if there is a correlation between erroneous classification by ChatGPT and disagreement of Appraisal classification between human annotators.

Second, labelling evaluative items directly in the input texts (e.g. by marking the items with XML tags) would be more advantageous instead of listing out single linguistic instances such as nouns or adjectives. In this way, the context of the evaluative instances was not contained in the output. An example of the ChatGPT output generated using our prompt is given in Figure 3 in the Appendix.

Third, we did not split the texts into 2 and 3 parts from the very beginning which could have potentially given us a higher recall.

Besides that, our analysis is limited in terms of genre and language, as we analysed texts of TED talks in English only. The performance of ChatGPT may vary across different text types and also across different topics. Since our dataset contains one text per discipline, we are not able to correlate the performance of the tool with genres or disciplines.

Also, our analysis is restricted to one LLM only. Testing more LLM-based tools would give us a better idea of their usability for our annotation tasks.

Ethics Statement

The data used in this study are collected from the TED website, which is publicly available. The texts are anonymised and do not contain any personal information.

References

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Min Dong and Alex Fang. 2023. Appraisal Theory and the Annotation of Speaker-Writer Engagement. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 18–26, Nancy, France. Association for Computational Linguistics.

Allyson Ettinger, Jena D. Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "You Are An Expert Linguistic Annotator": Limits of LLMs as Analyzers of Abstract Meaning Representation. *arXiv preprint*, 2310.17793.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Nuno M Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*.

M.A.K. Halliday and C.M.I.M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*, 4 edition. Routledge.

Xingwei He, Zheng-Wen Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. *ArXiv*, abs/2303.16854.

JR Martin and Peter Robert Rupert White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave MacMillan.

Natalia Mora and Julia Lavid-López. 2018. Building an annotated dataset of app store reviews with appraisal features in English and Spanish. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 16–24, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Andrew Nedilko. 2023. Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 616–620, Toronto, Canada. Association for Computational Linguistics.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254, Prague, Czechia. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. *arXiv preprint arXiv:2303.13780*.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar, and David E Losada. 2023. DepreSym: A Depression Symptom Annotated Corpus and the Role of LLMs as Assessors of Psychological Markers. *arXiv preprint arXiv:2308.10758*.

Jonathon Read, David Hope, and John Carroll. 2007. Annotating expressions of appraisal in English. In *Proceedings of the Linguistic Annotation Workshop*, pages 93–100, Prague, Czech Republic. Association for Computational Linguistics.

Michael V Reiss. 2023. Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. *arXiv preprint*, arXiv:2304.11085.

Jana Straková, Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. Extending an event-type ontology: Adding verbs and classes using fine-tuned LLMs suggestions. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 85–95, Toronto, Canada. Association for Computational Linguistics.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023. Assessing the potential of AI-assisted pragmatic annotation: The case of apologies. *arXiv: 2305.08339*.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making Large Language Models as Active Annotators. *arXiv:2310.19596*.

A Appendix: The Prompt

Do you know Appraisal Theory developed my Martin and White (2005)? Appraisal Theory consists of Attitude, Engagement and Graduation. Attitude consists of Affect, Judgement and Appreciation. Affect consists of these values: inclination (sub-value desire), disinclination (sub-value fear), happiness (sub-values cheer and affection), unhappiness (sub-values misery and antipathy), security (sub-values confidence and trust), insecurity (sub-values disquiet and apprehension), satisfaction (sub-values interest and pleasure), dissatisfaction (sub-values ennui and displeasure). Judgement consists of these values: social esteem (sub-values normality, capacity and tenacity) and social sanction (sub-values veracity and propriety). Appreciation consists of these values: reaction (sub-values impact, quality), composition (sub-values balance and complexity) and valuation. Can you annotate all the evaluative linguistic instances of Affect, Judgement and Appreciation in this text and create a list for me? Assign one of the sub-values of Affect, Judgement and Appreciation that are in brackets to evaluative linguistic instance. Please assign only one value to each evaluative linguistic instance. Focus on verbs, nouns, adjectives and adverbs. Do not annotate the whole sentence. Please annotate only the linguistic instances that correspond to Affect, Judgement and Appreciation and its sub-values. This is the text:

B Appendix: The Dataset

Topic	Text ID	Words
Art	Art	2,268
Business	Bus	1,209
Education	Edu	3,096
Entertainment	Ent	3,215
History	His	1,809
Medicine	Med	1,906
Natural Science	NatSci	3,233
Philosophy	Phil	1,693
Politics & Law	Pol	4,501
Psychology	Psy	1,177
Technology	Tech	1,010

Table 4: Details on the dataset.

C Appendix: ChatGPT Output

Affect
1. fascinating (happiness - cheer)
2. surprised (insecurity - disquiet)
3. interesting (satisfaction - interest)
4. magical (happiness - cheer)
5. nice (happiness - cheer)
6. cool (satisfaction - pleasure)
7. impressive (satisfaction - interest)
8. fail (dissatisfaction - displeasure)
9. fun (happiness - cheer)
10. love (inclination - desire)
11. wrong (dissatisfaction - displeasure)
12. love (inclination - desire)
13. fallible (insecurity - disquiet)
14. err (dissatisfaction - displeasure)
15. mistakes (dissatisfaction - displeasure)
16. biased (dissatisfaction - displeasure)
17. genius (security - confidence)

Judgement
1. logical (social esteem - capacity)
2. right (social sanction - veracity)
3. wrong (social sanction - veracity)
4. impressive (social esteem - capacity)
5. powerful (social esteem - capacity)
6. genius (social esteem - capacity)
7. elegant (social sanction - propriety)
8. quirky (social sanction - propriety)
9. buggy (social sanction - propriety)
10. rational (social sanction - veracity)

Appreciation
1. powerful (valuation)
2. remarkable (valuation)
3. extraordinary (impact - quality)
4. staggering (valuation)
5. sophisticated (composition - complexity)
6. innovative (composition - complexity)
7. technological (composition - complexity)
8. insightful (composition - complexity)

Figure 3: Visualisation of the output by ChatGPT