Spread and (Mis)use of Evaluative Expressions in Human Written and LLM-Based Generated Text

Maurice Langner[a] and Ralf Klabunde[b]
Linguistic Data Science Lab, Ruhr-University Bochum, Germany

Keywords: AI-Text Detection, Evaluative Expressions, Large Language Models, Data-To-Text NLG.

Abstract: We investigate the capacity of Large Language Models (LLMs) to generate evaluative expressions in a data-driven manner. The linguistic object of investigation is the production of justified and adequate evaluative language, such that the evaluative stance of the text is motivated by the underlying data. We use the SportSett corpus for generating summaries of basketball games. The input data is converted into RDF triples that are fed into GPT-4 and GPT-3.5, prompting the models to produce game summaries using evaluative adverbs and judgemental language. We annotated the generated texts and the original summaries for their propositional content contained in the line score and box score of each game, as well as for evaluative adverbs and their polarity. The results show that the models struggle to correctly interpret the numerical data and coherently assess the quality of team-wise and player-wise performances both within games and across games, often producing contradictory evaluations and displaying the lack of global evaluative scales.

1 INTRODUCTION

Generative Large Language Models (LLMs) have reached a quality that makes the resulting generated texts almost indistinguishable from texts written by human authors. This opens the door to a wide field of abuse, for example fake news generation, plagiarism, sophisticated spam formulation, and further text-based fraud schemes. There are a number of technical solutions for detecting AI-generated text, with varying degrees of success (Sadasivan et al., 2023), but there is astonishingly little work with a strong focus on linguistic characteristics of texts, and the use of discourse-oriented features for AI-text detection.

We propose to consider the use of evaluative expressions in texts for deciding whether the text has been generated by the use of an LLM, or whether the text is based on genuine authorship. Evaluative items – adverbs like *astonishingly*, *unfortunately*, or adjectives like *fair*, *outstanding* – express an evaluation of some state of affair that is based on an estimation of an expected value and the degree of divergence from that value which is, in turn, rooted in experience or known facts.

We should emphasize that using evaluative items is not solely a proposition- and, thus, clause-related decision, but rooted in the deployment of the overall discourse. Evaluative means awaken interest in a text, making it enjoyable to read, and these "evaluation foci" are not arbitrarily set in a text but follow strategies for establishing coherence. The following two examples from the SportSett Corpus, a modified version of the RotoWire Corpus (Wiseman et al., 2017) with NBA game summaries – the data we are using for our research – demonstrates this:

(1) *The Bucks showed superior shooting, going 46 percent from the field, while the Knicks went only 41 percent from the floor*

(2) *The Grizzlies shot 50 percent from the field, led by strong performances from Courtney Lee and Mike Conley. Lee scored 22 points (9 - 14 FG, 4 - 5 3Pt), while Conley led all scorers with 24 (9 - 14 FG, 3 - 4 3Pt) and 11 assists.*

In (1), the evaluative adjective *superior* expresses a shooting quite clearly above an expected one, and the adverb *only* expresses the contrary. As a result, the clauses with the respective evaluative items are interpreted as being contrastive.

The evaluative adjective *strong* in example (2) expresses a performance that is above the performance

[a] https://orcid.org/0009-0005-2169-064X
[b] https://orcid.org/0000-0003-1103-6431

Langner, M. and Klabunde, R.
Spread and (Mis)use of Evaluative Expressions in Human Written and LLM-Based Generated Text.
DOI: 10.5220/0012590600003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 2*, pages 646-654
ISBN: 978-989-758-680-4; ISSN: 2184-433X

that would be expected. The subsequent sentence provides the explanation why Lee's and Conley's performance was above the expectations by stating their scoring.

Although the use of evaluative items is often linked to rhetorical relations like contrast and explanation (Benamara et al., 2017; Trnavac and Taboada, 2012), we do not consider this relationship in this paper. Instead we analyze the distribution of evaluation items in texts written by humans and AI-generated texts for classifying. The results can be used for a post-hoc investigation on the (im)practicality of establishing rhetorical relations between clauses containing evaluative expressions, however.

We also do not intend to outperform existing methods for AI-text detection by our approach. Learning-based approaches typically use a bundle of different features and modes for classification, but without considering linguistic aspects of textual coherence in a satisfying way. The aim of this paper is to show that integrating linguistic evidence into the classification task – here the use of evaluative items – results in strong hints for detecting LLM-generated texts, if these texts are not purely descriptive, but convey expressive meaning as well.

2 RELATED WORK

Several different approaches have been proposed for AI-text detection: watermarking techniques for generated texts in order to support their detection (Kirchenbauer et al., 2023), perplexity-based methods (Mitchell et al., 2023), expected per-token log probability of texts for detecting thresholds that separate AI-generated texts from human written ones (Solaiman et al., 2019), and combinations thereof, to name just a few. In general, detection rates decrease with short texts and for human written texts containing just segments generated by a LLM. In addition to these document-oriented approaches, AI-text detection is also possible on sentence level (Wang et al., 2023).

What these approaches have in common is their low consideration of linguistic insights on text organization. There is a long tradition in Linguistics to analyze texts as a linguistic unit with a multi-layered organisation around information structural categories and different types of meaning – propositional, expressive and evaluative (Adam, 1992; Halliday and Hasan, 1976; Beaver et al., 2017). Since LLMs are to a large extend black boxes w.r.t. textual organisation criteria, these insights could be used for AI-text detection, as we demonstrate in this paper on the distribution and adequacy of expressive items.

Further related work concerns data-to-text natural language generation (NLG), the task of generating text from tabular data, where the use of evaluative expressions is motivated by a sufficient deviation of an observed value from an expected one. As we show for a vehicle domain (Langner and Klabunde, 2022), it is possible to determine at the early stage of content determination in an NLG pipeline whether some feature combination justifies an evaluative adverb or some other linguistic marker of evaluation by using regression models. In general, the concept of 'denial of expectation' best clarifies the intuition behind the mechanism: Experts have certain expectations of feature values given the remaining values in a feature set. In the basketball domain this means, if a score seems to fall out of a series, it deviates from the experts' expectation of what the value should be, given its context. This deviation may either be positive or negative in polarity, justifying the usage of evaluative language with this respective polar stance.

In NLG, evaluative adverbs are more generally attributed to affective language generation (de Rosis and Grasso, 2000). Evaluative items are generated in order to convey information with a specific stance (Elhadad, 1991). These systems communicate sensitive data, such as exam marks (Mahamood et al., 2007) or user-specific content (Balloccu et al., 2020). Large language models are used for affective language generation (Goswamy et al., 2020; Santhanam and Shaikh, 2019), but toxicity and fact hallucination have an immense negative influence on the output quality (Ji et al., 2023; Dušek and Kasner, 2020; Shen et al., 2020). GPT models are also employed in this field (Goswamy et al., 2020).

3 DATA USED

We are using the SportSett dataset (Thomson et al., 2020), a modified version of the RotoWire dataset (Wiseman et al., 2017) for data-to-text generation.[1] The SportSett dataset contains 6150 NBA basketball game summaries from different years and seasons. At the core of the tabular data is a set of different scores that are attributed to teams or individual players for different time spans of a game, e.g. the whole game, one of the four main periods of a game or even a play, which could be paraphrased as a short sequence of turns or actions. The scores are domain-dependent and comprise information on the points (pts) made, rebounds (oreb/dreb/treb), turnovers (tov, also loss of

[1] https://github.com/nlgcat/sport_sett_basketball

the ball to the other team), steals (stl), blocks (blk) and a special feature called pm (plus minus) which can also be negative and expresses what point difference to the opponent team was achieved while the player was on the field. Furthermore, points are the sum of free throws (ft), field golds (fg) and three point (fg3) field golds, that are each further subdivided in goals attempted (-a) and goals made (-m). Scores attributed to whole teams (cumulative values of individual player scores) are listed in the line scores, scores of individual players are listed in the box score. Both, line score and box score, are also separately listed for different spans of the game, e.g. periods or plays.

In this paper, we concentrate on the general game information and scores that relate to the overall game only, mainly due to the context limitations of LLMs. Adding further input data on period-wise scores would have exceeded the context limitation of the smaller LLM.

In order to generate game summaries from the data, we constructed a set of RDF triples for each of 50 randomly chosen game summaries. We take into account the following features: The score types, whose column names in the database are self-explaining, are used directly as relations in the RDF triples. Besides the score types for line score and box score, we included `home team`, `visiting team`, `stadium name`, `venue`, `attendance` and `capacity`. The subject was either the game itself or the team or player name respectively, while the assigned object is the value from the database cell. The triples were concatenated in the same order for every text to be generated. The models we chose for generation are GPT-4 and its predecessor GPT-3.5-Turbo by openAI, which are the currently largest available models for our use case. For each of the random 50 games, both models were instructed to generate two texts, a game summary with neutral tone and one with judgemental tone using evaluative adverbs. The prompts are as follows:

neutr.: *Translate the following box and line scores into a neutral NBA basketball game summary. Use non-judgemental words.*

eval.: *Translate the following box and line scores into an evaluative NBA basketball game summary. Use evaluative adverbs and judgemental language.*

We did not prescribe the polarity of the evaluation, hence the prompt did not introduce bias. In addition to the prompts, the concatenated RDF triples were fed into the models as well. In order to reduce the context size of GPT-3.5 and at the same time preserve comparability across text groups (GPT-4, GPT-3.5 and the original summaries), we removed some of the player-related triples with 0 as value, i.e. players from the bench who did not participate in the game. The original human-written game summaries also contain period-wise information, global information from previous seasons, and overall player scores from several games in a season. Since the models do not have access to this information, they were not annotated in the summaries. We only compare the information the models have access to.

We annotated the game summaries and the evaluative texts from both models[2] according to the presence of the score types, their association with either a player or a team. Further, we annotated evaluative adverbs, their polarity (positive or negative), and evaluations in regard to the global game quality. We subdivided the team and player annotations into winner and looser.

4 APPROACH

The motivation for producing evaluative language is grounded in the tabular data as described in the previous section.

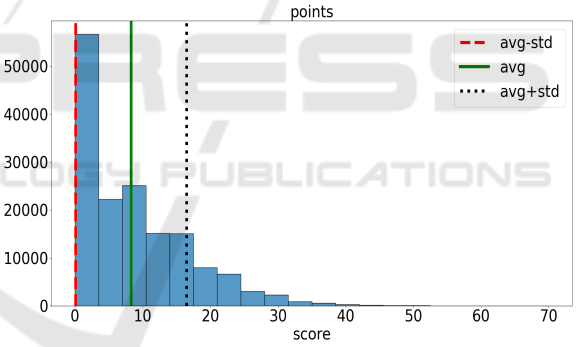

Figure 1: Player in game counts (y), score (x) and thresholds for the points score.

Let us exemplify this with expressions from the game with id 2120, which is also included in the annotated data for this paper. In this NBA game, Damian Lillard scored 50 points on his own, which is far above the average of 8.28 points across all database entries in the corpus, and 11.10 with zero score performances excluded. Figure (1) shows that the majority of points per player in a game are below 20 points, while only a very small number of player performances in some game provided more than 40 points, with a corpus-wide maximum of 70 points.

The original game summaries nearly always start with the final score of both teams in the first sentence

[2]https://github.com/MMLangner/Spread_and_Misuse_of_evaluative_language_in_LLMs/

of the summary. The score of 50 points is so remarkable that the author of the text not only decided to name it as the first piece of information in the text, but refers to it again later, to be seen in the excerpts below, describing Lillard as *playing on another level*. Although the author of this text does not use evaluative adverbs, the statement perfectly expresses the extraordinary status of this score. Both models, GPT-4 and GPT-3.5, also recognized this performance as exceptional, but GPT-3.5 seems to have the more adequate realisation of this denial of expectation, since GPT-4's generation of the adjective *solid* puts the exceptional status into perspective. The position of this adjective on the evaluative scale does not capture the extraordinariness of the described value.

In order to determine whether or not using evaluative language is legitimate, the "amount" of deviation that licenses its usage must be quantified. Assuming the average performance of a player (8.28) might be unjust, depending on talent and role of the player, the average of each individual player is much more adequate. The corpus-wide average for Damian Lillard is 24.74, with a standard deviation of 8.81. So even this threshold is exceeded by the 50 points score he achieved in game 2120.

Orig.	Despite 50 points from Portland 's Damian Lillard , the Toronto Raptors beat the visiting Trail Blazers on Friday , 117 - 115. [...] Damian Lillard and C.J . McCollum [...] combined for 74 points on Friday , with 50 coming from point guard Damian Lillard . Lillard , who has been playing on another level [...].
GPT-4	Damian Lillard exhibited an exceptional performance scoring a solid 50 points, making him undeniably the best player for his trailblazers.
GPT-3	Damian Lillard had a sensational game, leading his team with an incredible 50 points.

Thresholds for this model cannot be arbitrarily chosen, but empirical studies show that the average value plus or minus the standard deviation as an approximate threshold justifies the use of evaluative items (Langner and Klabunde, 2023).

In Figure (1), the graph shows the distribution of scores in the SportSett domain, where each score represents the performance of a specific player in a specific match. The global average (green solid line) lies at about 8 points, with two stronger thresholds being the average with standard deviation added (black dotted line) or subtracted (red dashed line). The interval between the lower and upper thresholds matches exactly the majority of scores within the domain. Given an adverb produced to express how positive Lillards performance of 50 points is, our method is to compare the score to the in-game average value and the player-related average of the `points` score, which establish the weaker thresholds. For a positive evaluation, the evaluated score must exceed these averages, since a higher score of `points` is considered positive and desirable. As stronger thresholds, we modify the averages by the respective standard deviation values. A positive polarity of the evaluative expression, given a feature where a higher score is better, therefore entails addition of the standard deviation. The evaluated score is again compared to these stronger thresholds. The distribution of real game data shown in Figure (1) is skewed, implying that the lower threshold may cause more errors, also with adverbs found in the original summaries.

In order to assess whether we can leverage evaluative language as a means to improve on AI detection methods from a linguistically motivated viewpoint, we let LLMs generate evaluative language in game summaries. On the basis of these empirically motivated thresholds for the data the models have access to, we automatically assess whether evaluative language is licensed by the underlying data or not. If the evaluative language is not licensed or even contradictory, we judge this as indication that our approach provides a good indicator for detecting AI-generated text.

5 RESULTS

First, we analysed the vocabulary of the three text groups: the summaries, the evaluative texts produced by GPT-4 and the evaluative texts produced by GPT-3.5. As null hypothesis, we assume that there is no significant difference between groups, since we expect language models of such a size to be capable of simulating the lexical choice the sports summary genre demands for. Furthermore, we assume that there is a significant difference within the group of summaries, since human-written texts tend to be more lexically variant, and that there is an insignificant difference within the groups of texts produced by the LLMs, since inference based on maximizing the probability of the output word sequence (beam search) makes the outcome lexically more deterministic than human-formulated texts.

Methodically, we lemmatize all the tokens in the union of all 50 annotated texts, remove stop words and create word counts for each text and separately

for each text group. The resulting distributions should not be assumed to be normally distributed, which is why we use the Mann-Whitney-U-test and the KS-test in order to determine whether there are significant differences between groups (assuming they are not sampled from the same distribution) and the Kruskal-Wallis-test for significance tests within the text groups. Additionally, we use the Euclidean distance as a similarity measure between and within groups in order to shed light on the distribution from a second, more common perspective. For Euclidean distance within groups, we calculate the distance for each text–pair combination and average over the distance values.

The results of the significance tests show that between groups, the texts generated by GPT-4 vary highly significantly from both the GPT-3.5 generated texts and the original summaries, while there is no significant difference between GPT-3.5 texts and the summaries (see Table 1). This is valid for both significance tests used. Within groups, all results are significant (Table 2), but the degree of significance varies strongly. The least significant differences are found in the group of GPT-4 texts, showing more homogeneity than the other one. GPT-3.5 texts vary much stronger than the GPT-4 texts (p=3.8e-11), but the largest significance is found in the original summaries (p=2.3e-40). We relate these differences to a higher lexical variance in the original summaries and a much more homogeneous lexical configuration of the GPT-4 models. Annotators also confirmed that lexicalisation and phrasal collocation were repetitive throughout GPT-4 and GPT-3.5 texts, whereas this was not the case for original summaries. This might be related to the way the GPT models are fine-tuned to match task-specific data.

In regard to Euclidean distance, we found that within groups, distance measures are quite equal for all three text groups, GPT-4 amounting to an average of 22.59, GPT-3.5 provides a value of 25.16 and the summaries a value of 24.51. Overall this means that the texts from GPT-3.5 as well as the original summaries are slightly less homogeneous, but the differences are rather minor.

More meaningful are the distance measures between groups. Between GPT-4 and GPT-3.5, the distance value is 350.86, while the models in comparison to the summaries more than double this distance with values larger than 760 (GPT-4 to summaries: 763.27; GPT3.5 to summaries: 799.62). According to this metric, GPT-3.5 is even more dissimilar from the original summaries than the GPT-4 output.

Both methods imply that there is a huge difference between the top model GPT-4 and the original summaries. Euclidean distance judges GPT-3.5 closer to GPT-4 and far away from the summaries, while significance tests see it positioned between GPT-4 and the summaries.

Table 1: Mann-Whitney-U-Test and KS-test on vocabulary distribution between groups.

group	MW-U (p)	KS (p)
4 vs 3.5	5814481.5 (p=3.8e-25)	0.157 (p=3.8e-35)
3 vs s.	5197648.0 (p=0.117)	0.024 (p=0.310)
4 vs s.	5915579.5 (p=2.7e-32)	0.182 (p=1.2e-46)

Table 2: Kruskal-Wallis-test on vocabulary distribution within groups.

group	K-W (p)
within GPT4	96.763 (p=5.653e-05)
within GPT3.5	143.060 (p=3.834e-11)
within summaries	314.109 (p=2.336e-40)

5.1 Evaluative Adverbs

We chose as thresholds for validating evaluative expressions the average of each score across all players in a game as well as the average score for each player globally. We assume that the addition or subtraction of the standard deviation to or from the average are stronger and empirically more reasonable thresholds. We group by the models and summaries respectively, as well as by separate score types and the polarity.

Before going into analytical details, it is important to mention that there is a huge class imbalance in regard to evaluative adverbs, both between groups as much as between the score types within each group. The most numerous group is the `points` score within the GPT-4 generated texts. Overall, GPT-4 generated 156 evaluative adverbs, while GPT-3.5 only produced 6. 12 occurred in the summaries. Also adjectives are used in all three text groups in order to express evaluative stance, the sparsity in GPT-3.5 and the summaries is hence due to our focus on evaluative adverbs. In future research, extending the analysis to adjectives and contrast relations will increase the amount of relevant data.

We can state for GPT-4 that there is a significant bias towards producing positive evaluations for the winning team or a member of the winning team, and negative evaluations for the defeated team and its players, as Table (3) shows.

The class imbalance between the score types such

Table 3: Polarity bias (adverb counts) in GPT-4 texts.

reference	positive	negative
player winner	**31**	2
player loser	6	**23**
team winner	**52**	3
team loser	7	**30**

as `points` or `assists` is just an inherent domain-specific issue, showing that the scored points are the most notable feature. Table (4) shows that player-related evaluative adverbs are explainable by the two weak average thresholds with about 77 and 73 percent fit. The stronger thresholds causes the fit to drop to or even below random with 50 percent and 29 percent correctness only. Surprisingly, the team-related adverbs show the opposite picture, where the evaluative adverbs are captured by the weak average thresholds only at random level ($\sim$50 percent). The stronger thresholds validate zero percent for the in-game average modified by the standard deviation and only about 20 percent for the global team average modified by standard deviation. The match for evaluations on team-level scores is therefore significantly worse than for single player scores.

Due to data sparseness, we cannot identify a polarity bias as explained above for GPT-3.5 or the summaries. With respect to the evaluative adverbs, it is hardly possible to draw any reliable conclusions about the GPT-3.5 outcomes, since the number of instances is just too small, but the impression is that it is basically random whether the expressions for teams are captured by the thresholds or not. The player-related expressions fit better, for the weak thresholds and the strong game-related data, but due to a lack of data, this is not a reliable statement. GPT-3.5 completely failed to abide by the task prescription given in the prompts. Although it used adjectives for expressing evaluative content, it failed to realize evaluation in the form of adverbs. GPT-3.5 basically failed the NLG task and hence the premise to an analysis of the occurring evaluative adverbs.

For the player-related evaluations given in the original summaries in Table 5, the agreement with the thresholds is promising with 100 percent and 81 percent for the weaker thresholds and 71% and about 40 % for the stronger ones. An important point to be mentioned here, that puts the numbers of incorrect instances in perspective, is the fact that in contrast to the LLM-generated texts, the summaries often explicitly name the background information for using the evaluative expression. For example, a player's performance in the previous 5 games or the present season has been mentioned. While the authors of the summaries have access to information permitting for further ways of grounding the evaluation, the LLMs do not have access to those and can only be evaluated with respect to the given texts. Using additional information for evaluating the adverbs in the original summaries would introduce bias, hence, we only use those thresholds we also use to interpret the LLM outputs.

5.2 Content Selection

Although a thorough analysis of hallucinations and factual correctness of the LLM output is beyond the scope of this paper, we analysed the share of propositional factoids from the database that were present in the LLM output and scrutinized these for correctness. Although there is again a huge class imbalance within the set of score types as well as between the LLMs, their correctness level is on par with about 80% correctness each. This means that about 20 percent of the database facts given in the LLMs' input are incorrectly transferred to the output (Table 6 and Table 7).

5.3 Contradictions and Overt Faults

The annotation work drew up some erroneous formulations that emphasize the dimension and momentousness of the lack in reasoning that LLMs show w.r.t. evaluative language. Examples generated by GPT-4 are given in Table (8).

The errors shown in Table (8) not only root in problems with basic maths as in items (1) to (3), but also in a basic misconception of the semantics of the score type. A good example for this is example (2), where the LLM fails to understand that having less turnovers is better, which is inconsistent with the statement in (4).

Another important source of errors is represented by (6), where GPT-4 failed to consider the information in its input that the scoreless players did not participate in the game, so cannot have scored at all, which is consequently not noteworthy.

Even more numerous, but less obvious is the lack of global evaluative scales for the scores which thus are not mirrored in the surface realisation. Examples (7) and (8) show that GPT-4 misses to correctly evaluate D'Angelo Russell's performance of 50% field goals made here, where his personal average is 40 percent and the player average overall only 35%. On the other hand in (8), percentages of 43% field goals made is judged "commendable". These outputs are related to the polarity bias already shown by the LLMs in the distribution of adverb polarity in regard to winning or loosing teams and players. The winners tend to be depicted positively and the looser negatively, indifferent to whether the evaluation is grounded in the data or not.

6 CONCLUSION

The analysis of semantic distance between the text groups and the significance tests of the word distri-

Table 4: Adverb analysis for player-related adverbs produced by GPT-4, separately listed.

ref	score	pol.	game avg	game avg +/− std	player avg	player avg +/− std
player	sec	positive	2(1.0): 0(0.0)	1(0.5): 1(0.5)	2(1.0): 0(0.0)	0(0.0): 2(1.0)
player	sec	negative	0(0.0): 1(1.0)	0(0.0): 1(1.0)	0(0.0): 1(1.0)	0(0.0): 1(1.0)
player	tov	positive	1(0.5): 1(0.5)	0(0.0): 2(1.0)	1(0.5): 1(0.5)	0(0.0): 2(1.0)
player	tov	negative	3(1.0): 0(0.0)	3(1.0): 0(0.0)	3(1.0): 0(0.0)	2(0.67): 1(0.33)
player	pf	positive	0(0.0): 1(1.0)	0(0.0): 1(1.0)	0(0.0): 1(1.0)	0(0.0): 1(1.0)
player	oreb	positive	4(1.0): 0(0.0)	3(0.75): 1(0.25)	3(0.75): 1(0.25)	1(0.25): 3(0.75)
player	treb	positive	3(0.75): 1(0.25)	2(0.5): 2(0.5)	4(1.0): 0(0.0)	0(0.0): 4(1.0)
player	blk	positive	3(1.0): 0(0.0)	3(1.0): 0(0.0)	1(0.33): 2(0.67)	1(0.33): 2(0.67)
player	stl	positive	1(1.0): 0(0.0)	1(1.0): 0(0.0)	1(1.0): 0(0.0)	0(0.0): 1(1.0)
player	ast	positive	5(1.0): 0(0.0)	5(1.0): 0(0.0)	5(1.0): 0(0.0)	3(0.6): 2(0.4)
player	fga	negative	0(0.0): 1(1.0)	0(0.0): 1(1.0)	0(0.0): 1(1.0)	0(0.0): 1(1.0)
player	fgm	negative	0(0.0): 2(1.0)	0(0.0): 2(1.0)	0(0.0): 2(1.0)	0(0.0): 2(1.0)
player	fg3a	positive	1(1.0): 0(0.0)	1(1.0): 0(0.0)	1(1.0): 0(0.0)	0(0.0): 1(1.0)
player	fg3m	positive	2(1.0): 0(0.0)	2(1.0): 0(0.0)	2(1.0): 0(0.0)	2(1.0): 0(0.0)
player	pts	positive	28(0.9): 3(0.1)	18(0.58): 13(0.42)	26(0.84): 5(0.16)	12(0.39): 19(0.61)
player	pts	negative	8(0.44): 10(0.56)	1(0.06): 17(0.94)	9(0.5): 9(0.5)	2(0.11): 16(0.89)
player	pm	negative	5(1.0): 0(0.0)	3(0.6): 2(0.4)	5(1.0): 0(0.0)	2(0.4): 3(0.6)
player	all	both	**66(0.77)**: **20(0.23)**	**43(0.5)**: **43(0.5)**	**63(0.73)**: **23(0.27)**	**25(0.29)**: **61(0.71)**
team	pts	positive	3(0.2): 12(0.8)	0(0.0): 15(1.0)	3(0.2): 12(0.8)	1(0.07): 14(0.93)
team	pts	negative	0(0.0): 2(1.0)	0(0.0): 2(1.0)	0(0.0): 2(1.0)	0(0.0): 2(1.0)
team	all	both	**21(0.47)**: **24(0.53)**	0(0.0): **45(1.0)**	**23(0.51)**: **22(0.49)**	**10(0.22)**: **35(0.78)**

Table 5: Adverbs in GPT-3.5 texts and original summaries, summed over features and polarities.

group	ref	feature	game avg	game avg +/− std	player avg	player avg +/− std
GPT-3.5	player	all	3(1.0): 0(0.0)	3(1.0): 0(0.0)	3(1.0): 0(0.0)	1(0.33): 2(0.67)
GPT-3.5	team	all	2(0.5): 2(0.5)	0(0.0): 4(1.0)	3(0.75): 1(0.25)	2(0.5): 2(0.5)
orig.	player	all	**21(1.0)**: 0(0.0)	**15(0.71)**: 6(0.29)	**17(0.81)**: 4(0.19)	8(0.38): 13(0.62)
orig.	team	all	3(0.75): 1(0.25)	0(0.0): 4(1.0)	3(0.75): 1(0.25)	0(0.0): 4(1.0)

Table 6: GPT-4 content selection: correctness of named features (in comparison to its input from the database).

feature	correct (%)	incorrect (%)
sec	4(0.4)	6(0.6)
tov	49(0.88)	7(0.12)
vio	0(0)	0(0)
pf	11(0.92)	1(0.08)
df	0(0)	0(0)
oreb	1(0.03)	36(0.97)
dreb	4(1.0)	0(0.0)
treb	33(0.92)	3(0.08)
blk	17(0.94)	1(0.06)
stl	13(1.0)	0(0.0)
fta	4(1.0)	0(0.0)
ast	28(0.85)	5(0.15)
ftm	6(0.86)	1(0.14)
fga	16(0.73)	6(0.27)
fgm	19(0.76)	6(0.24)
fg3a	8(0.89)	1(0.11)
fg3m	10(0.83)	2(0.17)
pts	**279(0.87)**	**42(0.13)**
pm	8(0.89)	1(0.11)
all	**510(0.81)**	**118(0.19)**

Table 7: GPT-3.5 content selection: correctness of named features (in comparison to its input from the database).

feature	correct (%)	incorrect (%)
sec	0(0)	0(0)
tov	20(0.91)	2(0.09)
vio	2(1.0)	0(0.0)
pf	4(0.57)	3(0.43)
df	0(0)	0(0)
oreb	3(0.09)	30(0.91)
dreb	7(1.0)	0(0.0)
treb	54(0.93)	4(0.07)
blk	34(0.92)	3(0.08)
stl	42(0.98)	1(0.02)
fta	2(1.0)	0(0.0)
ast	50(0.93)	4(0.07)
ftm	1(1.0)	0(0.0)
fga	19(0.95)	1(0.05)
fgm	18(0.56)	14(0.44)
fg3a	4(0.8)	1(0.2)
fg3m	20(0.87)	3(0.13)
pts	**292(0.87)**	**45(0.13)**
pm	8(1.0)	0(0.0)
all	**580(0.84)**	**111(0.16)**

bution have shown the large gap between the human-formulated texts on the one side, and the LLM-generated counterparts on the other side, backing the hypothesis that the generated texts are structurally and lexically far more deterministic and predictable. Although the word distribution within all groups is significant, the stronger significance for the original summaries once more underlines the more deterministic character of the GPT-4 and GPT-3.5 produced texts. In regard to Euclidean distance, GPT-3.5 is even

Table 8: Faulty examples generated by GPT-4.

1	However, their less impressive assists (28) as compared to the Celtics' 27 were noteworthy.
2	The Celtics suffered from an excessive turnover rate of 16 turnovers overall compared to Spurs' 18.
3	He had more turnovers than points scored (11 points, 3 turnovers)
4	He showed superior control with only 15 turnovers compared to the Kings' 11.
5	The Wizards demonstrated superior ball handling only committing 11 turnovers compared to the Pacers also with 11 (They displayed excellent ball control, committing only 11 turnovers compared to the Pacers' 11).
6	Regrettably, Nets' Isaiah Whitehead, Jahlil Okafor, Nik Stauskas performed poorly, with each of them failing to make a single point.
7	D'Angelo Russell struggled with his shooting, only making 50% of his attempted field goals. (40% avg, globally 35%.)
8	Moreover, their teamwork and synchronization were evident in their commendable 22 assists and a shooting rate of 43.75% from the field and 52.38% from beyond the arc.

less similar to the original summaries than the GPT-4 texts. The analysis of adverbs shows that there is a fundamental difference between evaluations of single players and the team performance. Our thresholds capture single player descriptions from the original texts nearly perfectly, validating that the thresholds are substantially useful for modeling, whereas data sparseness renders the results on team-addressed evaluative expressions in the original summaries unreliable.

Using the thresholds to assess the adequacy of the GPT-4 output shows the shortcomings of the LLMs in correctly grounding the evaluations in the data. It also shows that evaluative expressions are less adequate for team scores, where the match is sometimes lower than random, indicating structural bias, misconception of evaluative markers and the semantics of some feature names. The smaller GPT-3.5 model could not reliably be evaluated in regard to adverbs and their (in-)correct usage since the model simply failed to adhere to the task, producing only a fraction of data points needed. By a selection of failed contrast relations and evaluative adjectives, we furthermore give empirical evidence for the LLMs' inability to establish global evaluative scales and apparent issues in comparing simple numerical expressions, that permit the distinction from human-written texts. Even the proportion of evaluative language instances analysed here already shows the strong polarity bias of LLMs and their inability to produce coherent evaluations on discourse level.

We judge these findings as indication that the validity check of evaluative expressions is a promising linguistic means to complement existing methods for AI-text detection.

7 FUTURE WORK

In further research, annotation of evaluative adjectives and contrast relations is a promising measure to overcome data sparseness and extend our approach. This will also allow to assess validity and coherence of global evaluative scales across more instances of evaluative language. Furthermore, deeper analysis of the variance in discourse structure, which showed to be another substantially useful predictor for telling LLM generated texts and the original summaries apart, is a promising means to enrich and ultimately improve present approaches to AI detection with linguistic knowledge.

REFERENCES

Adam, J.-M. (1992). *Les textes: types et prototypes. Récit, description, argumentation, explication et dialogue*. Nathan, Paris.

Balloccu, S., Pauws, S., and Reiter, E. (2020). A NLG framework for user tailoring and profiling in healthcare. In Consoli, S., Recupero, D. R., and Riboni, D., editors, *Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, March 17, 2020*, volume 2596 of *CEUR Workshop Proceedings*, pages 13–32. CEUR-WS.org.

Beaver, D., Roberts, C., Simons, M., and Tonhauser, J. (2017). Questions under discussion: Where infor- mation structure meets projective content. *Annual Review of Linguistics*, 3:265–284.

Benamara, F., Taboada, M., and Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 34(1):201–264.

de Rosis, F. and Grasso, F. (2000). Affective natural language generation. In Paiva, A., editor, *Affective Interactions: Towards a New Generation of Computer Interfaces*, pages 204–218. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dušek, O. and Kasner, Z. (2020). Evaluating semantic accuracy of data-to-text generation with natural language inference. In Davis, B., Graham, Y., Kelleher, J., and Sripada, Y., editors, *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Elhadad, M. (1991). Generating adjectives to express the speaker's argumentative intent. In *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 1*, AAAI'91, page 98–103. AAAI Press.

Goswamy, T., Singh, I., Barkati, A., and Modi, A. (2020). Adapting a language model for controlled affective text generation. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A watermark for large language models.

Langner, M. and Klabunde, R. (2022). Realizing a denial of expectation in pipelined neural data-to-text generation. In Confalonieri, R. and Porello, D., editors, *Proceedings of the 6th Workshop on Advances in Argumentation in Artificial Intelligence 2022 co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022), Udine, Italy, November 28, 2022*, volume 3354 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Langner, M. and Klabunde, R. (2023). Validating predictive models of evaluative language for controllable Data2Text generation. In Keet, C. M., Lee, H.-Y., and Zarrieß, S., editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 313–322, Prague, Czechia. Association for Computational Linguistics.

Mahamood, S., Reiter, E., and Mellish, C. (2007). A comparison of hedged and non-hedged nlg texts. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, page 155–158, USA. Association for Computational Linguistics.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. (2023). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, page 24950–24962. JMLR.org.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2023). Can ai-generated text be reliably detected?

Santhanam, S. and Shaikh, S. (2019). Emotional neural language generation grounded in situational contexts. In Burtenshaw, B. and Manjavacas, E., editors, *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, pages 22–27. Association for Computational Linguistics, Tokyo, Japan.

Shen, X., Chang, E., Su, H., Niu, C., and Klakow, D. (2020). Neural data-to-text generation via jointly learning the segmentation and correspondence. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7155–7165, Online. Association for Computational Linguistics.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. (2019). Release strategies and the social impacts of language models.

Thomson, C., Reiter, E., and Sripada, S. (2020). SportSett:basketball - a robust and maintainable data-set for natural language generation. In Sánchez, D., Hervás, R., and Gatt, A., editors, *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 32–40, Santiago de Compostela, Spain. Association for Computational Lingustics.

Trnavac, R. and Taboada, M. (2012). The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, 3(34):301–318.

Wang, P., Li, L., Ren, K., Jiang, B., Zhang, D., and Qiu, X. (2023). SeqXGPT: Sentence-level AI-generated text detection. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. Association for Computational Linguistics.

Wiseman, S., Shieber, S., and Rush, A. (2017). Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.