Closing the Loop: Testing ChatGPT to Generate Model Explanations to Improve Human Labelling of Sponsored Content on Social Media

Closing the Loop: Testing ChatGPT to Generate Model Explanations to Improve Human Labelling of Sponsored Content on Social Media

Thales Bertaglia[1,3(✉)], Stefan Huber[1,2,3], Catalina Goanta[2], Gerasimos Spanakis[1], and Adriana Iamnitchi[1]

[1] Maastricht University, Maastricht, The Netherlands
t.costabertaglia@maastrichtuniversity.nl
[2] Utrecht University, Utrecht, The Netherlands
[3] Studio Europa, Maastricht, The Netherlands

Abstract. Regulatory bodies worldwide are intensifying their efforts to ensure transparency in influencer marketing on social media through instruments like the Unfair Commercial Practices Directive (UCPD) in the European Union, or Section 5 of the Federal Trade Commission Act. Yet enforcing these obligations has proven to be highly problematic due to the sheer scale of the influencer market. The task of automatically detecting sponsored content aims to enable the monitoring and enforcement of such regulations at scale. Current research in this field primarily frames this problem as a machine learning task, focusing on developing models that achieve high classification performance in detecting ads. These machine learning tasks rely on human data annotation to provide ground truth information. However, agreement between annotators is often low, leading to inconsistent labels that hinder the reliability of models. To improve annotation accuracy and, thus, the detection of sponsored content, we propose using chatGPT to augment the annotation process with phrases identified as relevant features and brief explanations. Our experiments show that this approach consistently improves inter-annotator agreement and annotation accuracy. Additionally, our survey of user experience in the annotation task indicates that the explanations improve the annotators' confidence and streamline the process. Our proposed methods can ultimately lead to more transparency and alignment with regulatory requirements in sponsored content detection.

Keywords: sponsored content detection · human-AI collaboration · legal compliance · social media

1 Introduction

The rise of influencers, content creators monetising online content through native advertising, has drastically changed the landscape of advertising on social

L. Longo (Ed.): xAI 2023, CCIS 1902, pp. 198–213, 2023.
https://doi.org/10.1007/978-3-031-44067-0_11

media [8,13]. This shift has increased concern about hidden advertising practices that might harm social media users. For decades, advertising rules have been applied to legacy media in such a way as to separate commercial communication from other types of content. The primary rationale behind rules relating to mandated disclosures has been that hidden advertising leads to consumer deception. Despite the increasing legal certainty that native advertising, such as influencer marketing, must be clearly disclosed, monitoring and enforcing compliance remains a significant challenge [25].

The task of automatically detecting sponsored content aims to enable the monitoring and enforcement of such regulations at scale. For instance, in the United Kingdom, the Competition and Markets Authority is one of the enforcement agencies tasked with monitoring influencer disclosures on social media, which is done using some automated techniques developed by their internal data unit[1]. In published scholarship, most existing methods frame the problem as a machine learning task, focusing on developing models with high classification performance. The success of these models depends on the quality and consistency of human-annotated data, which often suffer from low inter-annotator agreement, compromising the reliability and performance of the models [9,27]. Moreover, fully-automated approaches are insufficient for regulatory compliance, where human decision-makers are ultimately responsible for imposing fines or pursuing further investigations.

To bridge this gap, we propose a novel annotation framework that augments the annotation process with AI-generated explanations, which, to our knowledge, is the first attempt in this domain. These explanations, presented as text and tokens or phrases identified as relevant features, aim to improve annotation accuracy and inter-annotator agreement. Our experiments show that our proposed framework consistently increases agreement metrics and annotation accuracy, thus leading to higher data quality and more reliable and accurate models for detecting sponsored content. Critically, our work tackles the need for explainability in AI tools used for regulatory compliance, ensuring that human decision-makers can better understand and trust the outputs of these models. This is particularly important for market surveillance activities, which have not yet caught up with the transparency and accountability issues at the core of discussions around individual surveillance [19].

2 Related Work

Sponsored content detection has primarily been studied as a text classification problem. Works in this field generally train models in a semi-supervised setting, using posts disclosed as ads with specific hashtags as weak labels. Generally, there is a lack of focus on evaluating model performance with labelled data. Most works collect their own datasets and do not describe whether (and how) data is annotated. Since social media platforms typically do not allow data sharing, there are no standardised datasets for evaluating the task; thus, comparing results is

[1] https://www.gov.uk/cma-cases/social-media-endorsements.

challenging. Furthermore, the absence of labelled data for evaluation affects the reliability of results, as models are often not tested on undisclosed ads.

From a technical perspective, previous studies have employed traditional machine learning models with basic text features [7,29], neural networks with text embeddings [32], and multimodal deep learning architectures combining text, image, and network features [16,17]. In this paper we experiment with some of these models in addition to chatGPT and GPT-4 for classification. Although peer-reviewed research is limited due to chatGPT's recent release, some technical reports have found chatGPT to achieve state-of-the-art performance in several text classification tasks [12,23,30].

Interdisciplinary research combining computational methods with fields such as communication and media studies and law has focused on identifying influencers, describing their characteristics, and mapping the prevalence of their disclosures [2,4,21]. In the context of using explanations to improve data labelling or decision-making, research has explored AI-human collaboration and investigated the optimal integration of explanations for human interaction [6,18,22,28]. To the best of our knowledge, our paper is the first to propose using AI-generated explanations to improve the detection of sponsored content, bridging the gap between explainable AI and regulatory compliance in the context of sponsored content on social media.

3 Experimental Setup

This section describes the dataset we use, how we selected the model for sponsored content detection, generated explanations to augment the annotation process, and designed the annotation task and the user-experience survey.

3.1 Data Collection

We collected and curated our own dataset of Instagram posts for this study. We manually selected 100 influencers based in the United States using the influencer discovery platform Heepsy[2]. We selected 50 micro-influencers (between $100k$ and $600k$ followers) and 50 mega-influencers (over $600k$ followers). Then, we collected all available data and metadata from all posts for each account using CrowdTangle[3], the Meta platform that provides access to social media data for (among others) academic purposes. Our dataset includes $294.6k$ posts, 66.1% from mega-influencers and 33.9% from micro-influencers. CrowdTangle's Terms of Service do not allow (re)sharing datasets that include user-generated content; thus, we cannot share the full dataset. However, the list of the *ids* of accounts and posts is publicly available on https://github.com/thalesbertaglia/chatgpt-explanations-sponsored-content/

[2] https://heepsy.com.
[3] https://www.crowdtangle.com/.

3.2 Detecting Sponsored Content

In the first step of our experimental setup, we aim to select the most suitable sponsored content classifier for generating explanations. We evaluate three previously proposed models: (1) a logistic regression classifier with term frequency inverse-document frequency (TF-IDF) features, analogous to the approach used by [7,29], (2) a pre-trained BERT model fine-tuned for our task, comparable to [17,32], and (3) OpenAI's chatGPT (GPT-3.5-turbo as of March 2022), which achieves state-of-the-art results in various text classification tasks [12,23]. We generate GPT predictions using OpenAI's API.

To evaluate the models' performance, we select a sample from our original dataset and split the data into training and test sets by year, using 2022 for testing and all prior posts for training. This division simulates a real-world scenario where a model is deployed and used to classify unseen data for regulatory compliance. By ensuring no temporal overlap between the sets, we prevent the model from learning features correlated with a specific period. Given the high imbalance in the data (only 1.72% of posts are disclosed as sponsored), we apply the random undersampling approach proposed by Zarei et al. (2020) [32] to balance the data. We include all disclosed posts (n) and randomly sample ($2 * n$) posts without disclosures as negative examples. We allocate 90% of the balanced data before 2022 to training and the remaining 10% to validation. We use all data in 2022 as the test set.

Additionally, we labelled a sample of the test set to evaluate the model's performance in detecting undisclosed ads. Four annotators labelled 1283 posts in total, with a sample of 50 posts labelled by all annotators for calculating agreement metrics. The inter-annotator agreement was 52% in absolute agreement and 53.37 in α, indicating moderate agreement. 654 posts were labelled as sponsored (50.97%) and 629 as non-sponsored (49.03%). 91.59% of the sponsored posts did not have disclosures – i.e., they were identified as undisclosed ads.

We employ a semi-supervised approach to train the models, treating disclosed sponsored posts as positive labels for the *sponsored* class. We consider *#ad*, *#advertisement*, *#spons*, and *#sponsored* as ad disclosures. We then remove disclosures from the posts to prevent models from learning a direct mapping between disclosure and sponsorship. We train the logistic regression model using TF-IDF features extracted from word-level n-grams from the captions (unigrams, bigrams, and trigrams). For the BERT-based model, we use the *bert-base-multilingual-uncased* pre-trained model weights from HuggingFace [31]. We fine-tuned the BERT-based model for three epochs using the default hyperparameters (specified in Devlin et al. (2019) [5]).

We apply various prompt-engineering techniques to enhance GPT's predictions. As we use the same methodology for generating explanations, we provide a detailed description in the following subsection. We evaluate all models using F1 for the positive and negative classes, Macro F1 (the simple average of both classes) and Accuracy in detecting undisclosed ads – a critical metric for determining the models' effectiveness in detecting sponsored posts without explicit

disclosures, which is ultimately our goal. Table 1 presents the classification metrics for the three models, calculated based on the labelled test set.

Table 1. Performance of the different models on the labelled test set. Acc represents the models' accuracy in detecting undisclosed ads.

Model	Pos F1	Neg F1	Macro F1	Acc
Log Reg	45.33	66.50	55.92	28.71
BERT	29.30	68.84	49.07	10.85
GPT-3.5	76.09	63.93	70.01	88.98

GPT-3.5 outperforms the other models in Macro F1 and accuracy in detecting undisclosed ads. Logistic regression (Log Reg) and BERT achieve significantly low accuracy, suggesting their inability to identify undisclosed sponsored posts effectively. The difference in Macro F1 is smaller, highlighting that relying solely on this metric for evaluating models may not accurately reflect their actual performance. Therefore, having high-quality labelled data, including undisclosed ads, is crucial for proper evaluation.

BERT's inferior performance compared to Log Reg could be due to a few factors. Being pre-trained on longer texts, BERT might struggle to extract sufficient contextual information from short Instagram captions. In addition, Log Reg, when combined with TF-IDF features, effectively captures word-level n-grams that may be more effective at identifying sponsored content patterns. In contrast, BERT uses subword tokenisation, which could result in less efficient pattern recognition. Given GPT-3.5's superior performance, particularly in detecting undisclosed sponsored posts, we selected it as the model for generating explanations to augment the annotation task.

3.3 Generating Explanations with GPT

We investigated various prompts for all publicly accessible models from the GPT-3 series and GPT-4. We observed that even the smallest GPT-3 model, Ada (*text-ada-001*), performed well in sponsored content detection and identifying relevant words. Nevertheless, we noted significant performance improvements for larger models especially when employing chain-of-thought reasoning [30] and generating explanations – particularly for more ambiguous posts. Consequently, we focused on *GPT-3.5-turbo* (the default ChatGPT version as of March 2022) and GPT-4.

We found a conservative bias for both models, with a strong preference for predicting the *not sponsored* class or other negative labels over positive ones. This phenomenon appeared consistent across all *Davinci*- and *Curie*-based models, with the inverse being true for smaller *Babbage* and *Ada*-based models. We employed several prompt engineering techniques to mitigate this bias and calibrate the labels. First, we instructed the model to highlight relevant words and

generate explanations before classifying a post. This chain-of-thought prompting approach, inspired by [30], significantly reduced bias and improved prediction interpretability. Second, we used few-shot learning to refine explanation calibration, address known failure modes, and further alleviate bias [3]. Third, we experimented with different label phrasings, such as "Likely (not) sponsored", to enhance the model's ability to make less confident predictions. Finally, we directly instructed the model to favour positive labels in cases of uncertainty, aiming to identify a higher proportion of undisclosed ads. The final prompt is available on the project's GitHub repository[4].

Upon qualitative evaluation, we found that GPT-4 outperformed GPT-3.5-turbo in explanation quality and classification accuracy, especially for ambiguous posts. However, for this study, we chose GPT-3.5-turbo (hereafter referred to as "GPT") due to its advantages in speed, cost, and public accessibility. Following this approach, we obtained the most important words in a post and generated explanations for why a post may or may not be sponsored to assist annotators. The following is an illustrative example of such an explanation; we omitted the actual brand name to ensure the post's anonymity:

```
Key indicators: '@BRAND', 'LTK'.
The post promotes a fashion brand and features a discount code,
indicating a partnership. Additionally, it features a @shop.LTK
link, a platform for paid partnerships.
```

3.4 Annotation Task

We conducted a user study to evaluate how explanations can help detect sponsored content. The study consisted of an annotation task in which participants labelled 200 Instagram posts from our dataset as *Sponsored* or *Non-Sponsored*. Our objective with the task was two-fold: i) Analyse explanations as a tool for improving annotation as a resource for ML tasks – i.e., to measure their impact on data quality, which, in turn, allows for the development of better models and evaluation methods. ii) Simulate regulatory compliance with sponsored content disclosure regulations – i.e., how a decision-maker would flag posts as sponsored.

We framed the annotation as a text classification task in which annotators had to determine whether an Instagram post was sponsored based on its caption. Generally, we followed the data annotation pipeline proposed by Hovy and Lavid (2010) [15]. We instructed annotators to consider a post as sponsored if the influencer who posted it was, directly or indirectly, promoting products or services for which they received any form of benefits in return. These benefits included direct financial compensation and non-monetary benefits, such as free products or services. Self-promotion was an exception: we considered posts promoting the influencer's content (e.g. YouTube channel or podcast) non-sponsored. However, posts advertising merchandise with their brand or directly selling other goods still fall under sponsored content. We explained these guidelines to each

[4] https://github.com/thalesbertaglia/chatgpt-explanations-sponsored-content/.

annotator and provided examples of sponsored and non-sponsored posts to help reinforce the definitions.

Eleven volunteer annotators with varying levels of expertise participated in the study. All were between 20 and 30 years old, active social media users, and familiar with influencer marketing practices on Instagram. Additionally, all annotators had or were working towards a high-education degree in a European university. Demographically, the participants came from various countries. We did not specifically collect country-level information, but at a continent level, participants were from Asia, Europe, and South America. While all participants were fluent in English, none were native speakers.

We split annotators into three groups according to their level of expertise in annotating sponsored content on social media. The first group, with three people, consisted of participants with no prior experience in data annotation. The second group included four participants who previously participated in annotation tasks but had no formal training. The third group, consisting of four legal experts, had specific legal expertise in social media advertisement regulations and had participated in annotations before. We further split the subgroups of annotators into two groups regarding annotation setup: one without explanations, in which annotators only had access to the captions, and one augmented with the generated explanations. One group of four annotators labelled the posts in both setups: with and without explanations. To summarise, our study includes three distinctive groups: novices with no prior annotation experience, intermediate annotators with previous experience but no formal training, and legal experts knowledgeable in social media regulations.

To select the 200 Instagram posts for our user study, we turned to a sample previously labelled by law students in another annotation task. Although the labels and definitions used in that task differed from ours, they provided a way to identify which posts were undisclosed ads, allowing us to include them in our study. We selected posts published between 2017 and 2020 by 66 different influencers based in the United States, with 62% being mega-influencers and 38% being micro-influencers. We also included 15% of posts with clear ad disclosures (such as the hashtag #ad) as an attention check to ensure annotators noticed the disclosures. Based on the labels from the previous annotations, we estimate that 65% (130) of the posts were likely sponsored, and 50% (100) were likely undisclosed ads.

We set up the study using the open-source annotation platform Doccano[5]. Each participant had a unique project, and although all annotators labelled the same 200 posts, the labels were not shared, and each participant only had access to their annotations. The annotation interface displayed the caption of the post and the two possible labels (Sponsored and Non-Sponsored) as buttons. After the post caption, we added the generated explanations with an explicit delimitation.

Accurately measuring inter-annotator agreement is crucial in data annotation tasks, as it allows us to estimate the annotated data's quality and the decision-making process's reliability. To assess inter-annotator agreement in our study,

[5] https://github.com/doccano/doccano.

we used three main metrics: Krippendorff's Alpha (α), absolute agreement, and accuracy in detecting disclosed posts. Krippendorff's Alpha measures the degree of agreement among annotators, considering the level of agreement expected by chance alone [14,20]. The absolute agreement indicates the proportion of annotations where all annotators agreed on the same label. We also used accuracy in detecting disclosed posts as an attention check mechanism, as it measures annotators' ability to correctly identify posts with clear disclosures as sponsored. This metric is crucial because disclosures may not always be easily visible in posts [21]. We also analysed additional metrics in some experiments, which we will introduce when describing the specific experiments.

3.5 User-Experience Survey

After the annotation, we conducted a user-experience survey to gather feedback from annotators on their experience using the explanations to assist with their decision-making process. The survey consisted of seven questions, with five closed-ended and two open-ended questions. We describe all questions and the rating scale used below:

- "On a scale of 1 (not helpful) to 5 (extremely helpful), how helpful were the explanations in identifying undisclosed advertisement partnerships?"
- "How accurate, from 1 (extremely inaccurate) to 5 (extremely accurate), did you think the explanations were?"
- "How often, from 1 (0% of the time) to 5 (100% of the time), did you agree with the AI explanations?"
- "Did the AI explanations help you feel more confident in your decision-making (Yes/No)?"
- "What aspects of the AI explanations were most helpful for your decision-making process?" This was a multiple-choice question with five options: *Reasoning*, *Identifying specific words or phrases*, *Clear examples*, *Other (specify)*, and *None*.
- "In what ways did the AI explanations improve your understanding of what constitutes an undisclosed advertisement partnership?" Open-ended.
- "How could the AI explanations be further improved to better support your decision-making process? Did you find anything noticeable you want us to know?" Open-ended.

The participants who received annotations augmented with explanations all completed the questionnaires, and we ensured their anonymity by not collecting any identifiable information. Additionally, we made it clear to the annotators that their responses would be entirely anonymous.

4 Experimental Results

This section presents the main findings from the annotation task and user-experience survey. Table 2 shows the metrics comparing the agreement between

annotators who labelled the posts with and without explanations. Seven participants were in the **No Explanations** group (one with *no experience*, four with *some experience*, and two *legal experts*). The **With Explanations** group had eight people (three with *no experience*, three with *some experience*, and two *legal experts*) – one participant from the *no experience* group and three from *some experience* labelled in both settings. In addition to the metrics presented in Subsect. 3.4, we also evaluate the proportion of posts with at most one disagreement (*1-Disag*) and show the percentage of posts labelled as sponsored (*Sponsored*). The last two rows present the absolute and relative (normalised) differences in metrics between the groups. The relative differences in metrics indicate the proportional change (in percentage). Positive differences represent an increase in agreement.

Table 2. Agreement metrics comparing annotations with and without explanations.

	α	Abs	1-Disag	Acc	Sponsored
No Explanations	54.98	46.50	69.50	90.62	54.64
With Explanations	63.58	54.50	75.00	93.75	59.81
Absolute Diff	8.61	8.00	5.50	3.12	5.17
Relative Diff	15.65	17.20	7.91	3.45	9.46

Using explanations to enhance the annotations resulted in a consistent improvement across all inter-annotator agreement metrics. Specifically, there was a 15.65% increase in α and a 17.20% increase in absolute agreement. However, the final values were still relatively low, typical of annotations in complex decision-making tasks [9,10,27]. Accuracy in detecting disclosed posts also improved by 3.45%, but the final result was not perfect, suggesting that annotators still fail to identify all disclosure hashtags, even with explanations highlighting them. Additionally, the proportion of posts labelled as sponsored increased by 9.46%, indicating that explanations led annotators to identify more as sponsored. We also analyse the agreement between all pairs of annotators to measure the variation in agreement and ensure the reliability of the annotations. Table 3 summarises the pairwise agreement metrics. The *Min* and *Max* columns represent the lowest and highest agreement metric values among the annotator pairs, respectively, and the $\pm$ column denotes the standard deviation.

The pairwise metrics reveal considerable variation in the agreement between annotator pairs. For the *No Explanation* group, there was a substantial difference of 46.23 in α between the pair with the lowest and highest agreement, with a standard deviation of 10.83. This difference indicates that some annotators are significantly less reliable than others. However, the group *With Explanations* showed a consistent improvement, with less variation between pairs. The standard deviation decreased by 14.98% for absolute agreement and 7.62% for α, indicating more reliable annotations. Even the lowest-agreement pair showed

Table 3. Pairwise agreement comparing annotations with and without explanations.

	Min Abs	Max Abs	±	Min α	Max α	±
No Explanations	66.00	88.50	5.28	30.81	77.04	10.83
With Explanations	73.00	90.00	4.49	43.13	79.53	10.00
Absolute Diff	7.00	1.50	−0.79	12.31	2.48	−0.82
Relative Diff	10.61	1.69	−14.98	39.96	3.22	−7.62

significant improvement, with an increase of 10.61% for absolute agreement and 39.96% for α. These results suggest that using explanations to augment annotations led to a higher inter-annotator agreement overall, improved consistency between pairs, and even increased agreement among the least reliable annotators. To better understand the impact of augmenting the annotation with explanations, we also investigated how it affects different subgroups of annotators. We divided the subgroups into three categories: legal experts, non-experts, and annotators who labelled in both settings (with and without explanations) – this category does not include legal experts. Table 4 presents the agreement metrics for each category in both subgroups of annotators, as well as the relative difference between them. # indicates the number of participants within the subgroup. For clarity, we did not report the proportion of annotations with at most one disagreement because some subgroups contain a single pair of annotators.

Table 4. Agreement metrics for different subgroups of annotators, aggregated according to their expertise level.

	α	Abs	Acc	Sponsored	#
Legal Experts No Explanations	52.11	76.50	96.88	57.25	2
Legal Experts With Explanations	61.94	83.00	100.00	66.50	2
Relative Diff	18.86	8.50	3.23	16.16	–
Non-Experts No Explanations	62.04	62.50	93.75	53.60	5
Non-Experts With Explanations	64.89	59.50	93.75	57.58	6
Relative Diff	4.59	−4.80	0.00	7.43	–
Labelled Both No Explanations	66.74	70.00	96.88	53.12	4
Labelled Both With Explanations	73.15	74.50	100.00	54.50	4
Relative Diff	9.60	6.43	3.23	2.59	–

The annotations augmented with explanations showed consistent improvements in all subgroups, except for absolute agreement within the non-expert group. Legal experts had the most significant improvement in α (18.86%). Additionally, the proportion of posts labelled as sponsored increased significantly (16.16%), with the subgroup *Legal Experts With Explanations* having the highest value (66.5%). This subgroup and *Labelled Both With Explanations* achieved

100% accuracy in detecting disclosed sponsored posts. *Labelled Both* also had the highest α in both settings. It is important to note that higher agreement does not necessarily imply higher accuracy in correctly identifying sponsored posts. The metrics measure how much a subgroup of annotators agree on the definitions they are applying to label; they could be wrongly applying a consistent judgement. Therefore, we cannot reliably conclude which group had the best performance. Moreover, the high agreement within the subgroup *Labelled Both* could be influenced by the annotators labelling the same posts twice in both settings. Although we randomly shuffled the posts to reduce the likelihood of memorisation, repetition could still affect agreement. Nevertheless, the high proportion of sponsored content and absolute agreement for the annotation within *Legal Experts With Explanations* indicate that experts agree that there are more sponsored posts than non-experts tend to identify.

While explanations can improve the quality of annotations, they may also introduce bias by influencing annotators to rely on specific cues presented in the explanation; annotator bias is a common challenge in text annotation tasks [1,11]. To investigate potential bias introduced by explanations in our study, we examine whether annotators tended to use the same label predicted by GPT. Although we did not explicitly provide GPT's prediction as part of the explanation, the model's reasoning and highlighted words and phrases might imply the predicted label, leading to over-reliance on the model and decreasing the accuracy of annotations. Thus, it is essential to analyse the impact of GPT's predictions on annotator behaviour to ensure the reliability and fairness of the annotations. Specifically, we calculate two metrics – the distribution of posts labelled as sponsored and the majority agreement with GPT predictions – to compare the agreement between annotators who received explanations and those who did not. We use majority agreement instead of absolute to reduce the impact of low-agreement pairs and fairly compare all groups. If the agreement with GPT predictions increased in the group with explanations, it could indicate that annotators followed the model's predictions. We hypothesise that, for the *Labelled Both* group, an increase in agreement with GPT predictions proportionally more than the percentage of sponsored posts would suggest that annotators changed their judgements based on the model's cues. Table 5 summarises the results of this analysis.

The majority agreement with GPT predictions is consistently high across all subgroups, ranging from 77.5% to 92%. All subgroups that received explanations had an increase in agreement with GPT predictions compared to the corresponding No Explanations subgroup. Specifically, except for *Labelled Both*, all subgroups showed proportional increases in both metrics, indicating no clear bias for GPT predictions. However, the *Labelled Both* subgroup demonstrated a significant increase in agreement with GPT predictions compared to the proportion of sponsored posts, suggesting that the annotators changed their decision-making process after having access to explanations. While this result indicates a bias towards the model's predictions, more experiments are needed to determine its impact on data quality. Given the generally high accuracy of GPT

Table 5. Proportion of posts labelled as sponsored and majority agreement with GPT predictions across subgroups of annotators.

	Sponsored	Agreement
No Explanations	54.64	85.50
With Explanations	59.81	90.50
Relative Diff	9.46	5.85
Legal Experts No Explanations	57.25	77.50
Legal Experts With Explanations	66.50	92.00
Relative Diff	16.16	18.71
Non-Experts No Explanations	53.60	81.00
Non-Experts With Explanations	57.58	88.50
Relative Diff	7.43	9.26
Labelled Both No Explanations	53.12	78.50
Labelled Both With Explanations	54.50	87.00
Relative Diff	2.59	10.83

demonstrated in our classification experiments, relying on them could improve annotation accuracy.

On the other hand, the difference in agreement with the predictions between the *Legal Experts* subgroups adds uncertainty about the model's accuracy. The subgroup of legal experts with no explanations had the lowest agreement with GPT predictions; in contrast, those with explanations had the highest. The groups include different annotators, and *Legal Experts No Explanations* had low inter-annotator agreement; therefore, we cannot effectively measure the model's accuracy. Although we found evidence of explanations biasing the annotators, further research is needed to investigate how this result impacts data quality.

Finally, we conducted a user-experience survey to gather feedback from annotators on their experience using the explanations to assist with their annotation process. All the responses are available online on https://tinyurl.com/sponsored-annotation-survey. We ensured that the document preserves the anonymity of all parties involved in the study.

The survey results showed that 87.5% of annotators felt more confident in their decision-making with the help of explanations. Additionally, 62.5% rated the explanations highly helpful and accurate (4 out of 5). Only one participant rated them as unhelpful (2 out of 5). The average estimate of agreement with the explanations was close to the agreement with GPT predictions, with 62.5% of annotators estimating that they agreed with the explanations between 80% and 100% of the time. Notably, all annotators selected the words and phrases highlighted by the model explanations as a helpful feature, while only 37.5% selected the reasoning behind the predictions. This result indicates a preference for precise explanations. Comparable explanations could be generated from any classifier using local-explainability methods such as LIME [24]. This shows that

the methodology proposed and evaluated in our study does not rely on GPT's capability of generating longer text-based explanations and could be reproduced with simpler models.

The open-ended questions revealed two clear trends among participants. First, most participants found the highlighted words and phrases helpful in identifying brands and context-relevant hashtags in the posts. Second, participants suggested that adding the likelihood of a post being sponsored as a feature would be a useful improvement to the explanations. Overall, these results indicate that participants had a positive experience with the explanations, found them helpful and accurate, and felt they improved their decision-making.

5 Summary

Our experiments show that inter-annotator agreement metrics consistently improve when augmenting the annotation process with explanations. We observed a 15.65% increase in α and a 17.20% increase in absolute agreement among the general population of annotators. The accuracy in detecting disclosed sponsored posts improved by 3.45%, and the proportion of posts labelled as sponsored increased by 9.46%. These findings indicate that explanations not only help annotators identify more sponsored content but also enhance the reliability of annotations and reduce variation between annotator pairs. Our user-experience survey shows that most annotators found the explanations helpful and accurate, increasing their trust in decision-making. Therefore, our proposed annotation framework could lead to higher-quality data labelling and improve decision-makers' experience in regulatory compliance contexts. We made the *ids* of posts in our dataset, along with all the labels annotated by annotators and the GPT predictions, publicly available[6], offering a valuable resource that could benefit research in the field.

Nevertheless, our study has some limitations. One potential issue is the bias introduced by explanations, as annotators may rely on specific cues presented in the explanation. While we found no clear bias for most subgroups, we note that the group that labelled posts in both settings showed a significant increase in agreement with GPT predictions compared to the proportion of sponsored posts. Another area for improvement is the small sample size of legal experts and the variation in agreement metrics among different subgroups, which may impact the generalisability of our results.

Future research should investigate the impact of explanations on annotator bias and data quality and explore open-source models with greater transparency, such as LLaMA [26], instead of OpenAI's GPT – which is a privately-owned model with limited information regarding its training data. Moreover, conducting experiments with larger and more diverse samples of annotators, including more legal experts, could shed light on the role of expertise in the annotation process. Expanding the study to other annotation tasks and domains would also provide

[6] https://github.com/thalesbertaglia/chatgpt-explanations-sponsored-content/.

insights into the generalisability of our findings, potentially benefiting a broader range of applications.

Despite these limitations, it is important to consider that digital enforcement and market monitoring by authorities such as consumer agencies will exponentially grow in the coming years. Thus, monitoring techniques must consider transparency and explainability to avoid accuracy issues when applying legal sanctions.

References

1. Al Kuwatly, H., Wich, M., Groh, G.: Identifying and measuring annotator bias based on annotators' demographic characteristics. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 184–190. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.alw-1.21. https://aclanthology.org/2020.alw-1.21
2. Arriagada, A., Ibáñez, F.: "You need at least one picture daily, if not, you're dead": content creators and platform evolution in the social media ecology. Soc. Media + Soc. **6**(3), 2056305120944624 (2020). https://doi.org/10.1177/2056305120944624
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
4. Christin, A., Lewis, R.: The drama of metrics: status, spectacle, and resistance among YouTube drama creators. Soc. Media + Soc. **7**(1), 2056305121999660 (2021). https://doi.org/10.1177/2056305121999660
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
6. van Diggelen, J., et al.: Pluggable social artificial intelligence for enabling human-agent teaming. arXiv preprint arXiv:1909.04492 (2019)
7. Ershov, D., Mitchell, M.: The effects of influencer advertising disclosure regulations: evidence from instagram. In: Proceedings of the 21st ACM Conference on Economics and Computation, EC 2020, pp. 73–74. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3391403.3399477
8. Frithjof, M., et al.: The impact of influencers on advertising and consumer protection in the single market (2022). https://www.europarl.europa.eu/RegData/etudes/STUD/2022/703350/IPOL_STU(2022)703350_EN.pdf. Accessed 13 Oct 2022
9. Geiger, R.S., et al.: "Garbage in, garbage out" revisited: what do machine learning application papers report about human-labeled training data? CoRR abs/2107.02278 (2021). https://arxiv.org/abs/2107.02278
10. Geiger, R.S., et al.: Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 325–336 (2020)
11. Geva, M., Goldberg, Y., Berant, J.: Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets (2019)

12. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd-workers for text-annotation tasks (2023)
13. Goanta, C., Ranchordás, S.: The Regulation of Social Media Influencers. Edward Elgar Publishing (2020)
14. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Measures **1**(1), 77–89 (2007)
15. Hovy, E., Lavid, J.: Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. Int. J. Transl. **22**(1), 13–36 (2010)
16. Kim, S., Jiang, J.Y., Nakada, M., Han, J., Wang, W.: Multimodal post attentive profiling for influencer marketing. In: Proceedings of the Web Conference 2020, WWW 2020, pp. 2878–2884. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3366423.3380052
17. Kim, S., Jiang, J.Y., Wang, W.: Discovering undisclosed paid partnership on social media via aspect-attentive sponsored post learning. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM 2021, pp. 319–327. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3437963.3441803
18. Kim, S.S.Y., Watkins, E.A., Russakovsky, O., Fong, R., Monroy-Hernández, A.: "Help me help the AI": understanding how explainability can support human-AI interaction. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM (2023). https://doi.org/10.1145/3544548.3581001
19. Kossow, N., Windwehr, S., Jenkins, M.: Algorithmic transparency and accountability. JSTOR (2021)
20. Krippendorff, K.: Computing Krippendorff's alpha-reliability. Annenberg School for Communication Departmental Papers, Philadelphia (2011)
21. Mathur, A., Narayanan, A., Chetty, M.: Endorsements on social media: an empirical study of affiliate marketing disclosures on youtube and pinterest. Proc. ACM Hum.-Comput. Interact. **2**(CSCW), 1–26 (2018). https://doi.org/10.1145/3274388
22. Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 204–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_18
23. Pikuliak, M.: ChatGPT survey: performance on NLP datasets (2023). https://www.opensamizdat.com/posts/chatgpt_survey
24. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 1135–1144 (2016)
25. Said, Z.K.: Mandated disclosure in literary hybrid speech. Wash. L. Rev. **88**, 419 (2013)
26. Touvron, H., et al.: LLaMA: open and efficient foundation language models (2023)
27. Vidgen, B., Derczynski, L.: Directions in abusive language training data, a systematic review: garbage in, garbage out. PLoS ONE **15**(12), e0243300 (2020)
28. van der Waa, J., van Diggelen, J., Cavalcante Siebert, L., Neerincx, M., Jonker, C.: Allocation of moral decision-making in human-agent teams: a pattern approach. In: Harris, D., Li, W.-C. (eds.) HCII 2020. LNCS (LNAI), vol. 12187, pp. 203–220. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49183-3_16
29. Waltenrath, A.: Empirical evidence on the impact of disclosed vs. undisclosed advertising in context of influencer marketing on Instagram. In: ECIS 2021 Research Papers, p. 17 (2021)

30. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models (2023)
31. Wolf, T., Debut, L., Sanh, V., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6
32. Zarei, K., et al.: Characterising and detecting sponsored influencer posts on Instagram. arXiv:2011.05757 (2020)