How can ChatGPT help humans in Dark Web content classification? Assessing GPT models reliability and effects of explanations on human decisions

Víctor-Pablo Prado-Sánchez, Adrián Domínguez-Díaz, Daniel Rodríguez, José-Javier Martínez-Herráiz, Luis de-Marcos

Department of Computer Science, University of Alcalá, Spain

Alcalá de Henares, España

{victor.prado, adrian.dominguez, daniel.rodriguezg, josej.martinez, luis.demarcos}@uah.es

***Abstract**. This study evaluated the reliability of ChatGPT for Dark Web content classification and the effect of its explanations on human classification decisions. The study objectives were to determine whether ChatGPT can be considered a reliable classifier, identify the content categories where it performs differently than humans, and assess if its explanations can enhance human understanding of these contents. Results show that while ChatGPT is outperformed by supervised models, its reliability is similar to human classifiers, outperforming them in technology-related categories. Additionally, while ChatGPT's explanations allowed humans to better understand its decisions, their effect was not consistent across different people.*

Keywords. Dark Web, Darknet, Cybersecurity, Language models, LLM, NLP, ChatGPT, GPT-3.5.

1 Introduction

The increased volume and complexity of activities conducted on the Dark Web pose challenges for the classification of its contents, especially in the context of illegal and potentially harmful activities (Avarikioti et al., 2018). Consequently, there has been a growing interest in applying machine learning techniques to address this issue (Al Nabki et al., 2017). In this regard, various approaches based on supervised machine learning models have been developed to detect and classify illegal activities on the Dark Web (Jin et al., 2022).

Given the linguistic nature of the contents to be classified, the performance of pre-trained language models in this task has been explored. For example, the use of models from the BERT (Bidirectional Encoder Representations from Transformers) family has proven effective in classifying text across various domains, including Dark Web content (Devlin et al., 2019; Jin et al., 2022). The need to evaluate the performance of different models has led to the creation of datasets with manually labeled texts from Dark Web sites, such as DUTA (Darknet Usage Text Addresses) and CoDA (Comprehensive Darkweb Annotations) (Al Nabki et al., 2017; Jin et al., 2022). Despite the achieved performance, these models require a prior phase of supervised learning or fine-tuning of the language model on a portion of the dataset to effectively classify them (Arslan et al., 2021). This limitation suggests that the classifiers' performance in real-world environments would be limited by the availability and quality of labeled data, as well as the possibility of overfitting to the training data, making it difficult to adapt to the dynamic and heterogeneous nature of the Dark Web (Al Nabki et al., 2017).

An alternative that would reduce the need for labeled data and the effort of human classifiers is the use of large language models, such as GPT-3 and its derivatives, through zero-shot or few-shot prompting (Kalyan, 2024; Roumeliotis & Tselikas, 2023). When using these techniques, the language model is not specifically fine-tunned for the required task, and just receives prompt messages with the description of the task to perform (zero-shot) or the description and a few classification examples (few-shot). The model must rely on its pre-existing knowledge and understanding, which it has acquired during training on a broad range of data, to perform the classification task (Reiss, 2023). While this approach has been tested for financial text classification (Loukas et al., 2023) or sentiment analysis tasks (L. Li et al., 2023), it performance and reliability remains to be tested for Dark Web content classification, where different categories of content have specific and somewhat obscure terminology which could make both humans and artificial classifiers struggle to determine the correct category of content. Understanding how ChatGPT performs compared to other classification models and to humans would facilitate its integration in real-work environments.

In addition to its usage as an isolated classifier, the capability of ChatGPT to explain its classification decisions in natural language could also be used to help humans better understand Dark Web contents. In other contexts, research such as (Huang et al., 2023) explores ChatGPT ability to explain implicit hate speech, offering a valuable tool for understanding and addressing this sensitive issue. Similar capabilities in the Dark Web context could help law enforcement agents to better identify and understand illegal or potentially harmful contents.

This study aims to address several key objectives: firstly, to determine whether ChatGPT can be considered reliable for Dark Web content classification tasks; secondly, to identify the types of Dark Web contents where ChatGPT classifier performance differs more from humans; and thirdly, to assess the extent to which ChatGPT explanations can help improve human understanding of Dark Web contents for classification tasks. To achieve these objectives, the study provides a comprehensive evaluation of ChatGPT's ability to classify Dark Web content by comparing its results with other models and with humans. Furthermore, it investigates how ChatGPT's explanations impact human classification of Dark Web contents. This paper provides a comprehensive view of the potential role of ChatGPT and humans in collaborative Dark Web content classification. To that end, the following research questions have been defined:

- RQ1: Can ChatGPT be considered reliable for Dark Web content classification tasks?

- RQ2: In what content categories does ChatGPT classification performance differ from classification performed by human?

- RQ3: To what extent can ChatGPT explanations help improve human understanding of Dark Web contents for classification tasks?

The remainder of this document is structured as follows: Section 2 provides an overview of related work on text content classification and on GPT models explanatory capabilities. Section 3 details the methodology employed to evaluate ChatGPT reliability and the effects of its explanations on human classification decisions. Section 4 presents the results and discusses the findings considering the research questions. Finally, Section 5 outlines the conclusions and discusses future lines of work.

2 Related Work

2.1 ChatGPT classification metrics and reliability in zero-shot prompting

The natural language processing capabilities of the latest LLMs, such as GPT models, enable tackling text classification tasks without the need for specific prior training for the classification categories.

In the financial domain, (Loukas et al., 2023) proposes using conversational GPT models, such as GPT-3.5 and GPT-4, to classify financial texts with minimal examples and compares their performance with other pre-trained and fine-tuned models. GPT-3.5 and GPT-4 outperform other models in classification with 1 or 3 examples, achieving F1 scores of 75.2% and 83.1%, respectively, but trail behind fine-tuned language models. The study highlights the ease of implementing effective classifiers through services like OpenAI, though it notes the high cost for small organizations as a drawback. Reliability is assessed using a confidence metric, where values close to 1 indicate high reliability and values close to 0 indicate low confidence. These confidence values are crucial for evaluating the certainty of model predictions across different configurations in the financial domain.

In addressing the challenge of detecting harmful content on social networks, (L. Li et al., 2023) employs ChatGPT as a detection model, comparing its results with human annotations. Achieving an 80% precision in identifying harmful content, the study emphasizes the consistency of classifications while acknowledging the impact of the prompt used on its performance. ChatGPT's reliability is assessed by comparing its results with annotations from MTurkers, reporting approximate values of around 80% for Precision, around 70% for Recall, and around 75% for F1-score. These results indicate ChatGPT's ability to accurately identify harmful comments compared to human annotations.

Finally, in (Reiss, 2023) examines the reliability of ChatGPT for annotation and text classification tasks in various application contexts. While its potential is recognized, concerns are raised about its non-deterministic nature, which can lead to variable results even with small changes in inputs. Caution is advised when using ChatGPT for these tasks without additional validation, such as comparisons with human-annotated data. In the study, a Krippendorff's Alpha value above 0.8 is considered to indicate reasonable reliability in ChatGPT outputs for text classification tasks. Values below this threshold raise questions about the reliability of the results. It is recommended to validate ChatGPT outputs with human-annotated data to ensure reliable results.

2.2 Dark Web Language and existing datasets and classifiers

In the realm of Dark Web research and online illegal activity detection, several studies have been conducted exploring different machine learning algorithms and language models for text classification through supervised learning. This is due to the need to address the inherent challenges in analyzing the specific and diverse language that characterizes the Dark Web. Furthermore, the scarcity of public data and the anonymity and cryptography techniques used in this clandestine digital environment have driven the creation of specialized datasets. For instance, (Al Nabki et al., 2017) introduces the DUTA dataset for active domains on the Dark Web, comprising 6,831 documents classified into 26 categories. It was found that the combination of TF-IDF and Logistic Regression achieves a weighted F1 score of 97% in classifying a subset of the dataset into 9 categories.

In an attempt to overcome some limitations of the DUTA dataset, (Jin et al., 2022) introduces the CoDA dataset, consisting of 10,000 web documents obtained from the Dark Web for text analysis. Linguistic differences between the Dark Web and the Surface Web are examined, and differences between the CoDA and DUTA datasets are analyzed. The performance of various Dark Web page classification methods is also evaluated, achieving a weighted F1 score of 92.49% using the pre-trained BERT language model, following fine-tuning for document classification. In (Jin et al., 2023), an enhanced classifier based on a pre-trained BERT adaptation with texts obtained from the Dark Web is presented, achieving an F1 score of 94.25% in classifying the contents of the CoDA dataset.

2.3 ChatGPT natural language explanations

The ability of ChatGPT to provide natural language explanations of its decisions is mentioned. We explore how these explanations can be beneficial to individuals by providing a clear and easily accessible understanding of the reasoning behind the responses generated by the model.

In the context of recommender systems (RSs), ChatGPT can generate explanations that justify the recommendation and contextualize it in a way that is relevant to the user (Silva et al., 2024). Research indicates that personalized explanations are more effective than generic ones, especially when the recommended item is unfamiliar to the user, highlighting the importance of tailoring explanations to each individual.

Multiple studies have examined ChatGPT's ability to provide natural language explanations in the context of sentiment analysis (Huang et al., 2023, 2024). These studies show that GPT models can generate effective explanations regarding hate speech detection and have compared its explanations with those generated by humans, evaluating their quality and alignment with human standards.

The role of ChatGPT in generating interpretable explanations in mental health analysis is also noted, with promising results suggesting its utility in this area. In (Yang et al., 2023), the limitations of previous research are addressed through a comprehensive evaluation of the models' mental health analysis and emotional reasoning abilities of ChatGPT across 11 datasets and 5 tasks. Results show that ChatGPT generates explanations that approach human performance, showing great potential in explainable mental health analysis.

Finally, (B. Li et al., 2023) assesses ChatGPT's overall capability across seven information extraction tasks. The performance, explainability, calibration, and fidelity of ChatGPT are analyzed. While the performance varies greatly depending on the information extraction setting, the study shows that ChatGPT provides high-quality and trustworthy explanations for its decisions.

3 Methodology

3.1 CoDA dataset selection

The CoDA dataset (Comprehensive Darkweb Annotations) was utilized, representing a valuable public resource for Dark Web analysis, comprising a collection of 10,000 web documents intended for text-based research in this environment. These documents cover a wide range of topics and were classified into ten distinct thematic categories. Primarily in English, the documents were sourced from onion services on Tor, i.e., the Dark Web, providing significant insight into this relatively unexplored digital space.

The documents were categorized into ten thematic categories, including Drugs, Financial, Gambling, Cryptocurrency (Crypto), Hacking, Arms/Weapons (Arms), Violence, Electronics, as well as Pornography and Others. This wide variety of categories allows for an assessment of various aspects of activity on the Dark Web. It is worth noting, however, that it was necessary to exclude web documents categorized as Pornography due to conflicts with OpenAI's content policies.

3.2 Zero-shot classification with GPT 3.5 Turbo model

For the classification of documents from the Dark Web, the decision was made to employ OpenAI's GPT-3.5 language model, particularly the variant known as GPT-3.5 Turbo. This choice was based on the demonstrated ability of this model to efficiently process text and understand the complex and varied context characteristic of the Dark Web (Ye et al., 2023).

The selection of this model variant was justified by its better balance between capabilities and cost compared to other models such as GPT-4, which was 20 times more expensive at the time of the study, and its better suitability to the problem's needs, with an input context of 16,385 tokens that is suitable for the size of documents to be classified.

The model received the necessary data to perform the classification task under a zero-shot prompting model (Chen et al., 2023). For each Dark Web site in the dataset, ChatGPT was sent a prompt that explained the classification task to perform and included the site text content. The task explanation included the name and descriptions of the possible categories, as defined in the CoDA dataset (Jin et al., 2022), and requested an answer that included the name of the category that best described the content, as well as to briefly explain the reasons why it has chosen that category. The language used for the prompt was English, as the bulk of documents included in CoDA are in this language. The parameters of the GPT model call were set by the system's defaults.

3.3 Reliability assessment and human-GPT disagreement analysis

Based on classification results, standard classification performance metrics and inter-coder reliability measures were calculated, comparing ChatGPT classification with the original human-made classification included in the dataset. Then, for those documents in which ChatGPT and the human classification disagreed, a comprehensive analysis and tie-breaking process was conducted to determine up to what point ChatGPT, or human assigned classes were right. The tie-breaking process was performed by three independent reviewers over a random sample of 180 documents, distributed in 20 documents per category, based on the original dataset classification.

To carry out the tie-breaking process, a double iteration approach was implemented. In the first iteration, the disagreements were reviewed without accessing the explanation that ChatGPT provided to justify the class it selected. In this initial phase, reviewers determined if they agreed with the class originally assigned by humans (which was included in the dataset) or with the class assigned by ChatGPT relying solely on the analysis of the content of the site in question. Reviewers could answer if they thought the original classification was right (0), the ChatGPT classification was right (1), or if they were not sure (2). Subsequently, in the second iteration, researchers were exposed to the explanation provided by ChatGPT, so the evaluators had the opportunity to consider the reason behind ChatGPT decision before making a final decision. Again, they were allowed to provide one of the three possible answers. Based on the results of the second iteration, each document was assigned to a final category, assigning the original dataset category or the one selected by ChatGPT when at least two or the three reviewers agreed with either of them, or a "Not sure" value when that condition was not met.

The average percentage of agreement with humans and with ChatGPT was calculated for the whole sample after each iteration. Average percentage of agreement was also calculated for each category, grouping documents by the final category they were assigned to at the end of the tie-breaking process. Finally, inter-coder agreement was calculated to determine the level of agreement between the three reviewers for the whole sample and for each category.

The main objective of this process was first to assess whether human or ChatGPT performed better in difficult to classify documents, and to identify in which categories human or the GPT model performed differently. In addition, the two iterations approach was adopted to determine whether the explanation provided by ChatGPT influenced evaluators' decision making, leading them to change their initial classification decision or keeping it unchanged.

4 Results and discussion

4.1 ChatGPT classification reliability

As shown in (Table 1), the performance of supervised classifiers (Jin et al., 2022) reveals several important considerations. First, it is observed that supervised classifiers such as SVM, CNN and BERT, show generally superior performance to ChatGPT with the GPT-3.5 Turbo model in terms of accuracy, recall and F1-score. This suggests that supervised approaches may benefit from a larger labelled dataset specific to the classification task compared to the zero-shot learning model used by ChatGPT. The result is in line with studies showing that GPT models without specific training offer good performances, but always below other algorithms prepared specifically for the problem to be addressed (Kocoń et al., 2023).

Table 1. Performance comparison (weighted average) with supervised classifiers

Model	Precision	Recall	F1
SVM	91.59%	91.17%	91.19%
CNN	88.08%	87.30%	87.23%
BERT	92.51%	92.50%	92.49%
ChatGPT	**85.63%**	**82.63%**	**83.14%**

It is important to consider the way supervised classifiers are trained and the possibility of overfitting. Supervised models, when trained on labelled data, can be overfitted to the specific details of the training set. Since the datasets are drawn from a limited number of sources, the performance of these classifiers in production environments may suffer when classifying documents from other sources. On the other hand, ChatGPT is based on a zero-shot approach, which means that it lacks prior training on the specific documents to be classified. This could lead to a more generalizable performance, maintaining similar performance values in real environments to those obtained on the dataset.

To evaluate the reliability of ChatGPT as a potential substitute for a human in Dark Web classification tasks, intercoder reliability values such as Cohens Kappa and Krippendorf's Alpha were examined. These values provide a measure of the consistency between ChatGPT's and the original human-made classification. Intercoder reliability metrics (Table 2) indicate a high level of agreement between ChatGPT and the original human CoDA classifiers. With a Cohen's Kappa of 0.8, a Weighted Cohen's Kappa of 0.82 and a Krippendorff's Alpha of 0.79, these metrics exceed (Cohen's Kappa) or get near (Krippendorffs' Alpha) the thresholds considered indicative of excellent agreement and high reliability.

In clinical studies, a Kappa value of 0.8 is considered indicative of excellent agreement. Furthermore, it is mentioned that when both sensitivity and specificity are less than 0.9, they can never produce a Kappa of 0.8 or higher, even with a raw agreement approaching maximum sensitivity and specificity (Feuerman & Miller, 2008). On the other hand, an Alpha value close to 1 indicates high reliability in measuring interobserver agreement (Krippendorff, s. f.). High Alpha values, such as 0.8 or higher, suggest good reliability in the data.

These metrics indicate that ChatGPT could be used effectively to complement the classification performed by humans on the dataset. Its high level of agreement with the original human classifiers suggests that the model can provide accurate and reliable classification of the data, making it a valuable tool for improving and streamlining classification processes in different contexts.

Table 2. Intercoder reliability metrics between human and ChatGPT classifiers

Cohen's Kappa	0.8
Weighted Cohen's Kappa	0.82
Krippendorff's Alpha	0.79

4.2 Human vs ChatGPT reliability

The results of the tie-breaking process between human classifiers and ChatGPT (Table 3) reveal valuable insights into the performance and reliability of ChatGPT and human classifiers across different content categories. For each category, the level of agreement with human classification and with ChatGPT classification was calculated by dividing the number of documents that reviewers agreed to assign to the category (with the agreement of 2 or more reviewers) by the total number of documents that the dataset or ChatGPT, respectively, originally assigned to the same category. In addition, the average level of agreement between the three reviewers was measured for the whole sample and for each category by calculating the Fleiss' Kappa intercoder reliability metric. Fleiss' Kappa is related to Cohen's Kappa but applicable to more than 2 raters (Powers, 2012)

Table 3. Comparison of reviewers' agreement with humans and with ChatGPT across diverse categories

Final Class	Agree with human	Agree with ChatGPT	Fleiss' Kappa
Gambling	10 / 20 (.5)	1 / 1 (1)	-0.18
Arms/Weapons	8 / 20 (.4)	2 / 4 (.5)	-0.09
Electronics	11 / 20 (.55)	4 / 4 (1)	0.61
Hacking	9 / 20 (.45)	18 / 18 (1)	0.41
Drugs	3 / 20 (.15)	8 / 12 (.67)	0.15
Crypto	5 / 20 (.25)	15 / 20 (.75)	0.74
Financial	3 / 20 (.15)	13 / 40 (.325)	0.38
Violence	1 / 20 (.05)	5 / 6 (.83)	0.32
Others	4 / 20 (.2)	44 / 66 (.66)	0.14
Total	**54 / 180 (.3)**	**110 / 180 (.61)**	**0.32**

Overall, reviewers show higher degree of agreement with ChatGPT that with humans across the sample, although the intercoder reliability, considering Cohen's Kappa reference values (Altman, 1990) is just fair (> 0.2) when considering the whole sample. This probably indicates that the sampled documents are generally difficult to classify both for ChatGPT and humans alike. There could be different reasons, such as difficult to understand language, ambiguous content that could fit multiple categories or the lack of a well stablished classification protocol that unequivocally determines a document class when multiple values are possible. An in-depth analysis of the reviewers' decisions during the tie-breaking process would be required to understand low intercoder reliability in each category properly.

Tie-breaking results show that reviewers had higher level of agreement with ChatGPT than with the original human-made classification, with most categories left with 10 or less documents assigned to their original category (from the original 20 documents per category) and an overall agreement of 30% across the sample, clearly below 50%. These values suggests that ChatGPT, with an overall agreement of 61%, could be a more reliable classifier than humans for Dark Web content classification tasks, performing better in extensive document classification and with difficult to classify documents. This conclusion must be taken with caution though, as low intercoder reliability in the tie-breaking process means that agreement percentages are not strongly grounded. However, it shows a tendency that could be confirmed by additional studies.

Focusing on specific categories, high differences in agreement percentage between humans and ChatGPT, accompanied by a moderate or good intercoder reliability (Fleiss' Kappa > 0.4) can be found in the "Electronics" (0.55 for human vs 1.0 for ChatGPT), "Hacking" (0.45 for human vs 1.0 for ChatGPT) or "Cryptocurreny" (0.25 for human vs 0.75 for ChatGPT) categories. These results suggest that the original human classifiers struggled to understand advanced technological terminology related to electronic devices, hacking and cryptocurrency products and services offered through the Dark Web, while ChatGPT may have a stronger performance due to its capabilities to process and understand complex and dynamic technical language.

Even with insufficient or fair intercoder reliability during the tie-break, high differences in agreement levels in "Drugs", "Violence" and "Others" categories should also be highlighted. According to the reviewers, the original dataset contained many documents misclassified in these categories. Although an in-depth analysis of the documents is out of the scope of this paper, as an example, reviewers reported that many documents originally classified as "Violence" where just describing historical events or contained political discussions which did not fit the category description of "Human trafficking, hitman, kidnapping, poisoning, torture, extortion, sextortion, sex slavery, blackmail, etc.", and were classified as "Others" by ChatGPT. This case could be pointing towards the importance of reporting the classification protocol followed during the creation of a dataset, so ambiguous cases could be resolved unequivocally in future classification studies. That classification protocol is not included in the CoDA dataset description, and thus it's not easy to determine classification errors for certain documents in the original dataset.

4.3 Effect of ChatGPT explanations on human classification

Before and after explanation tie-breaking results (Table 4) show that reviewer's level of agreement with ChatGPT increased 20 percentage points after the exposition to its explanations, reducing almost equally the agreement with the original dataset and the agreement uncertainty by approximately 10 percentage points each, while Fleiss' Kappa intercoder reliability sightly increased by 0.05 units. This shows that while the explanations provided by ChatGPT had an impact on the average levels of agreements, convincing some reviewers to agree with ChatGPT decisions; the effect was not consistent in a per-document basis for the three reviewers, becoming insufficient to increase intercoder reliability from low (> 0.2) to moderate (> 0.4) level. The brevity or the insufficient quality of the explanations could be limiting factors that prevent reviewers to agree in the classification. Future studies should measure the quality of explanations and acknowledge possible differences between reviewers during the classification process to better understand the influence of explanations in classification decisions.

Table 4. Before and after explanations tie-breaking results showing average level of agreement and intercoder reliability.

	Before	After	Diff
Fleiss' Kappa	0.27	0.32	+0.05
Level of agreement			
With human	73 / 180 (.4)	54 / 180 (.3)	-0.1
With ChatGPT	74 / 180 (.41)	110 / 180 (.61)	+0.2
Not sure	33 / 180 (.18)	16 / 180 (.09)	-0.09

As an example of the effect of explanations, in Table 5 we can observe a record from the CoDA dataset which is initially categorized as Others, but categorized as Drugs by ChatGPT. While one of the reviewers initially agreed with the original classification, after evaluating the reason behind ChatGPT's categorization, she reassessed her tie-break decision and changed it, finally agreeing with ChatGPT response.

Table 5. Tie-break decision altered by ChatGPT explanation.

Content	urlaub / vacation corona hat uns im griff … bis auf weiteres nur notprogramm. corona got us … no shipping for some time to abroad … unser notprogramm für euch auf darkmarket, empire und cannazon… greensupreme ID_EMAIL —–begin pgp public key block—– version: gnupg ID_VERSION gafo3nirma== =gfqo —–end pgp public key block—–
Before explanation	Others
ChatGPT Category	Drugs
ChatGPT categorization reason	The content mentions 'greensupreme' and provides a PGP public key block, which are typically associated with the sale

	of drugs on darknet markets. Additionally, phrases like 'no shipping for some time to abroad' suggest that the website is offering drug products for sale.
After explanation	Drugs

5 Conclusions

This study investigated the reliability and performance of ChatGPT, specifically the GPT-3.5 model, in classifying Dark Web content using a zero-shot learning approach. In addressing the first research question, it was found that although supervised classifiers such as SVM, CNN, and fine-tuned BERT models outperform ChatGPT in precision, recall, and F1-score, ChatGPT still shows significant reliability, comparable to human classifiers. The review of discrepancies between ChatGPT and the original human classification showed improved ChatGPT performance in technology-related categories and revealed difficulties in certain ambiguous cases both for humans and models alike. Lastly, the research demonstrated that ChatGPT's brief natural language explanations influence human understanding and decision-making, but they are not sufficient to drastically increase intercoder agreement as measured by Fleiss' Kappa metric.

The promising results suggest several future research directions. These include refining classification protocols to improve consistency and reliability, especially for ambiguous cases, and investigating the fine-tuning of GPT models on domain-specific datasets to enhance accuracy and reliability. Finally, conducting longitudinal studies to evaluate ChatGPT's performance in real-world Dark Web monitoring and law enforcement scenarios will be crucial for understanding its impact on operational efficiency and accuracy over time.

Finally, some limitations that could put the generalizability and accuracy of the study should be noted. First, the small sample size for human-GPT disagreement analysis and potential reviewer bias complicates the assessment of performance discrepancies. Additionally, the two steps iterative review process may not fully isolate the impact of ChatGPT's explanations, as reviewers could be influenced by their initial decisions or the mere presence of an AI-generated explanation. Finally, the quality and clarity of ChatGPT's explanations themselves are not systematically evaluated, which could affect their usefulness in improving human understanding. These factors underscore the need for cautious interpretation and highlight areas for future research refinement to enhance the study's robustness.

Acknowledgments

This work has been funded by the Spanish Ministry of Science and Innovation, under the "Project for the analysis and recovery of criminal evidence associated with hidden networks" (PARCHE), with reference PID2021-125645OB-I00 of the 2021 call for "Knowledge Generation Projects".

References

Al Nabki, M. W., Fidalgo, E., Alegre, E., & de Paz, I. (2017). Classifying Illegal Activities on Tor Network Based on Web Textual Contents. En M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 35-43). Association for Computational Linguistics. https://aclanthology.org/E17-1004

Altman, D. G. (1990). *Practical Statistics for Medical Research*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429258589

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., & Goujon, A. (2021). A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain. *Companion Proceedings of the Web Conference 2021*, 260-268. https://doi.org/10.1145/3442442.3451375

Avarikioti, G., Brunner, R., Kiayias, A., Wattenhofer, R., & Zindros, D. (2018). *Structure and Content of the Visible Darknet* (arXiv:1811.01348; Número arXiv:1811.01348). arXiv. https://doi.org/10.48550/arXiv.1811.01348

Buldin, I. D., & Ivanov, N. S. (2020). Text Classification of Illegal Activities on Onion Sites. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 245-247. https://doi.org/10.1109/EIConRus49466.2020.9039341

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). *Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review* (arXiv:2310.14735). arXiv. https://doi.org/10.48550/arXiv.2310.14735

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: Sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice*, *14*(5), 930-933. https://doi.org/10.1111/j.1365-2753.2008.00984.x

Graczyk, M., & Kinningham, K. (s. f.). *Automatic Product Categorization for Anonymous Marketplaces*.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *Companion Proceedings of the ACM Web Conference 2023*, 294-297. https://doi.org/10.1145/3543873.3587368

Huang, F., Kwak, H., Park, K., & An, J. (2024). *ChatGPT Rates Natural Language Explanation Quality Like Humans: But on Which Scales?* (arXiv:2403.17368). arXiv. https://doi.org/10.48550/arXiv.2403.17368

Jin, Y., Jang, E., Cui, J., Chung, J.-W., Lee, Y., & Shin, S. (2023). DarkBERT: A Language Model for the Dark Side of the Internet. En A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7515-7533). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.415

Jin, Y., Jang, E., Lee, Y., Shin, S., & Chung, J.-W. (2022). *Shedding New Light on the Language of the Dark Web* (arXiv:2204.06885; Número arXiv:2204.06885). arXiv. https://doi.org/10.48550/arXiv.2204.06885

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048. https://doi.org/10.1016/j.nlp.2023.100048

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Krippendorff, K. (s. f.). *Computing Krippendorff's Alpha-Reliability*.

Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. (2023). *Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness* (arXiv:2304.11633). arXiv. https://doi.org/10.48550/arXiv.2304.11633

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2023). *«HOT» ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media* (arXiv:2304.10619; Número arXiv:2304.10619). arXiv. https://doi.org/10.48550/arXiv.2304.10619

Loukas, L., Stogiannidis, I., Malakasiotis, P., & Vassos, S. (2023). *Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance* (arXiv:2308.14634; Número arXiv:2308.14634). arXiv. https://doi.org/10.48550/arXiv.2308.14634

Powers, D. M. W. (2012). The Problem with Kappa. En W. Daelemans (Ed.), *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 345-355). Association for Computational Linguistics. https://aclanthology.org/E12-1035

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark* (arXiv:2304.11085; Número arXiv:2304.11085). arXiv. https://doi.org/10.48550/arXiv.2304.11085

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*, *15*(6), Article 6. https://doi.org/10.3390/fi15060192

Silva, Í., Marinho, L., Said, A., & Willemsen, M. C. (2024). Leveraging ChatGPT for Automated Human-centered Explanations in Recommender Systems. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 597-608. https://doi.org/10.1145/3640543.3645171

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). *Towards Interpretable Mental Health Analysis with Large Language Models* (arXiv:2304.03347). arXiv. https://doi.org/10.48550/arXiv.2304.03347

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X. (2023). *A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models* (arXiv:2303.10420). arXiv. https://doi.org/10.48550/arXiv.2303.10420