"I was diagnosed with...": sensitivity detection and rephrasing of Amazon reviews with ChatGPT

Costanza Alfieri
DISIM
University of L'Aquila
L'Aquila, Italy
costanza.alfieri@student.univaq.it

Suriya Ganesh
School of Computer Science
Carnegie Mellon University
Pittsburgh, USA
sayyampe@andrew.cmu.edu

Limin Ge
School of Computer Science
Carnegie Mellon University
Pittsburgh, USA
liminge@andrew.cmu.edu

Jingxin Shi
School of Computer Science
Carnegie Mellon University
Pittsburgh, USA
jingxins@andrew.cmu.edu

Norman Sadeh
School of Computer Science
Carnegie Mellon University
Pittsburgh, USA
sadeh@cs.cmu.edu

2024 21st Annual International Conference on Privacy, Security and Trust (PST) | 979-8-3503-6709-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/PST62714.2024.10788076

***Abstract*—The proliferation of platforms such as e-commerce and social networks has led to an increasing amount of personal health information being disclosed in user-generated content. This study investigates the use of Large Language Models (LLMs) to detect and sanitize sensitive health data disclosures in reviews posted on Amazon. Specifically, we present an approach that uses ChatGPT to evaluate both the sensitivity and informativeness of Amazon reviews. The approach uses prompt engineering to identify sensitive content and rephrase reviews to reduce sensitive disclosures while maintaining informativeness. Empirical results indicate that ChatGPT is capable of reliably assigning sensitivity scores and informativeness scores to user-generated reviews and can be used to generate sanitized reviews that remain informative.**

***Index Terms*—LLMs, sanitizing, self-disclosure, sensitivity detection, privacy**

I. Introduction

Platforms such as e-commerce portals and social networks have contributed to the proliferation of sensitive user-generated [1]. On one hand, the disclosure of health information has had a positive impact on people's lives, diminishing stigma, spreading information, e-word-of-mouth for health product, and so on, on the other hand the disclosure of sensitive information such as sensitive health-data can have negative consequences on people who disclose it. This can range from seeing one's application for a health insurance policy denied to losing one's job or being discriminated against in various social contexts [2]–[4].

Self-disclosure represents a significant source of users' personal information, with health-related information often being identifiable in online communities such as Reddit, Twitter, Facebook, and patient.info [5]. For these reasons, the challenge of sanitizing sensitive data from texts has been widely explored [6]–[9]. Recent advancements in Large Language Models (LLMs) have fostered the exploration of these technologies for sanitizing sensitive information in texts [10].

978-1-5386-5541-2/18/$31.00 ©2018 IEEE

In this study, we delve deeper into the task of detecting sensitivity and sanitizing text with disclosed health information generated by users. Due to the significant consequences of health information disclosure, we choose to focus exclusively on this domain to further explore the potential of LLMs in this field. Furthermore, since most self-disclosure studies are conducted on well-known online communities like Reddit, Facebook, and Twitter, we aim to investigate the e-commerce domain of reviews, considered a "field of gold" for marketing purposes due to the abundance of available information [11]. For what concerns the privacy risks of online reviews, Huang *et al.* [12] show how Amazon reviews can be used to delineate the "online lifestyle" of customers to significantly improve preference prediction. Their study demonstrates that this method surpasses the widely used Big Five personality in prediction power. Indeed, when writing reviews, individuals may unintentionally reveal sensitive details about their personal relationships, visited locations, home addresses, and temporal information such as specific times of day, dates, or special occasions, that can expose intimate aspects of a user's private life [13]. In addition, Amazon collects a vast array of information from its users through various channels, including browsing behavior on the Amazon website, interactions with Amazon services like Prime Video, and use of Amazon devices such as Kindle and Alexa [14]. This data includes IP addresses, search queries, order history, and even detailed insights from devices like Kindle and Alexa, revealing not just what users are interested in but also their habits and preferences. Through the techniques of the "online lifestyles" described in [12], by leveraging customers' reviews, Amazon can retrieve sensitive information of customers that can lead to a very detailed profile of the user. This would enable a large private company to amass an even more detailed profile of users, potentially including self-disclosed health information [15]. For this reasons, we selected Amazon as a case study, specifically focusing on reviews of a health product since our focus is on health-data. Hence, we selected a product readily purchasable on the platform: Turmeric Curcumin Supplement by Natures Nutrition, a turmeric supplement often used for inflammation.

We present a methodology for leveraging LLMs to sanitize sensitive health information. Initially, we introduce the model to the concept of health-sensitive data, asking to directly identify such information. Following this, we request the model to evaluate the sensitivity and informativeness of the text. A rephrased version of the text is then generated, and the model is tasked with reassessing its sensitivity and informativeness scores. We only used ChatGPT as the large language model (LLM) since it is currently considered by many as the most advanced and versatile tool available [16].

The Research Questions (RQs) in this study are:

RQ1: Using prompts that define what sensitive information is, is it possible to reliably use ChatGPT to identify sensitive health information disclosures in user-generated content?

RQ2: Is it possible to design prompts that can be used to have ChatGPT reliably score the sensitivity and informativeness of user reviews? Are the scores stable/reliable?

RQ3: Is it possible, using prompt engineering, to use ChatGPT to automatically identify and sanitize sensitive content in user-generated reviews? In particular, is it possible to reduce the sensitivity of information disclosed in user-generated reviews while retaining the informativeness of these reviews?

The remainder of the paper is structured as follows: Section II delves into the related work; in Section III we outline the methodology employed in this study; Section IV presents and analyzes the obtained results. The limitations of this study are discussed in Section V. Finally, Section VI provides conclusions and outlines potential future works.

II. Related work

Privacy-enhancing techniques have become an increasingly important topic in current research areas, particularly in contexts involving sensitive data such as users' health information [17]. One such technique is text sanitization, which is utilized to conceal personal information within textual data [18]. In this section, we discuss works related to our study, with particular attention to LLMs applications, the task of sanitizing texts, and detecting sensitivity.

A. Large language models and privacy sanitizing

Transformer models, which are the base of LLMs, have been widely studied by [19], as they improve the efficiency and effectiveness of sequence transformation models through their revolutionary multi-head attention mechanism. This attention mechanism is capable of selectively focusing on specific parts of the data and is theoretically suitable for identifying and filtering sensitive information in a dataset. For their great adaptability, different researchers are experimenting LLMs for various applications. In their work, Kamalloo *et al.* [20] develop "HAGRID", a dataset that serves as a catalyst for the development of information-seeking models. By integrating the efforts of human annotators and LLMs, such as GPT-3.5, "HAGRID" emphasizes the capability to generate explanations with attributability and informativeness. In the same vein, Wang *et al.* [21] have developed an open-source dataset, "Do-Not-Answer", which comprises instructions that *responsible* language models should avoid following. The dataset was created by combining safety taxonomy, human evaluation, and ChatGPT, and was applied to evaluate "dangerous capabilities" of several popular LLMs to the instructions contained in the dataset. Through the analysis of these responses, they trained several BERT-like classifiers for safety evaluation. They discovered that these smaller classifiers could achieve results comparable to GPT-4 in automatic safety evaluations.

With the spreading of LLMs, different means to perform also privacy filtering and text sanitizing are emerging. In their study, Ishibashi *et al.* [22] explored knowledge sanitization techniques across different LLMs, including negative gradient methods and Rank-One Model Editing (ROME), which aims to selectively forget or modify specific information in a model. These methods can preserve the performance of LLMs while protecting privacy and are crucial for handling sensitive queries. In the work of Kim *et al.* [23], by combining wearable sensor technology with LLMs, a transformative approach to personal health monitoring can be realized, enabling LLM to accurately process information about privacy by continuously tracking important physiological data. Their work suggests that LLM can recognize what information is privacy-relevant and protect it. Tang *et al.* [24] explore the role of ChatGPT for generating synthetic data in clinical text mining. Through the use of this LLM, they created a large amount of training data without sensitive patient information.

In the realm of LLMs for privacy purposes, a pertinent case study illustrates the application of ChatGPT in addressing developers' privacy concerns. In this study, Delile *et al.* compare the responses provided by ChatGPT to privacy-related questions extracted from Stack Overflow against the accepted answers from the platform [25]. The analysis revealed that most privacy questions pertain to issues of choice/consent, and identification. It was found that for approximately 56% of the questions, ChatGPT's responses were similarly accurate compared to the accepted answers on Stack Overflow, while for the remainder, the platform's responses were slightly more accurate. This finding was achieved with an exploration based on a relatively small dataset of 92 pairs of multi-labeled questions and answers, suggesting that even modest datasets can support an assessment of ChatGPT's capability.

In summary, from the initial applications of Transformer models, to the advances in LLMs for knowledge sanitization, health monitoring, and synthetic data generation, to information abstraction, we can see the potential of LLMs for handling privacy-sensitive information and their evolving applications.

B. Sensitivity detection

Sensitivity detection is the process of identifying sensitive information within data, a task that has long been studied [26]. Sweeney elaborated the Scrub system [27] that help with basic detection for names, locations, and so on. More sophisticated, it is the approach of Architha and Gopinath [28], that employed Natural Language Processing (NLP) technologies, such as NLTK and spaCy models, for sensitivity detection and anonymization of texts. These methods enable machines to assign meanings to each word in a document, automatically identify and remove any patient identifiers (PI), and replace them with a generic PI [28]. Other approaches, such as [29], [30], offer more specific patterns for removing sensitive terms from medical records, showcasing the variety of strategies that have been proposed in different contexts.

Building upon this, Korytkowski *et al.* introduced an advanced approach that goes beyond identifying specific words to understanding sentences in context. This method aims to detect and process sensitive information, such as "mayor of Paris" or "son of the CEO," by optimizing the network structure for efficient operation even on devices with limited computational power. Hassan *et al.* [31], proposed a novel approach to text anonymization using word embeddings to represent entities within a document. Unlike traditional methods that rely on named entity recognition and manually tagged training data, this approach identifies sensitive information based on semantic relationships between words. This brings an automatic protection of an individual or organization by detecting and removing entities with similar word vectors, which indicate a close semantic relationship to the target entity.

Dou *et al.* [10] developed a taxonomy of 19 self-disclosure categories and curated a large corpus consisting of 4.8K annotated disclosure spans. Moreover, they propose an abstraction task, where the model is tasked with rephrasing detected sensitive information from a text to achieve a more generalized level of information. For instance, they demonstrate how a sentence like "I just turned 32 last month and have been really..." can be abstracted to "I recently entered my early 30s." While this approach is valuable and consistent with the original text, a stricter approach may be necessary when dealing with health information disclosure.

In this paper, our objective is to leverage LLMs and their contextual understanding capabilities to enhance sensitivity detection and sanitize texts. Specifically, we optimize prompts to detect the sensitivity and informativeness of text, then "rephrasing" the text to mitigate user privacy concerns while retaining the informative value. This method diverges from the traditional techniques employed for sensitivity detection [7], [28] and it does not require any fine-tuning for the rephrasing task [10] that can be computationally expensive.

III. Methodology

In this section, we delineate the methodology to test the sensitivity detection and the rephrasing task by using ChatGPT-4 (gpt-4-0125-preview). Our dataset and code are available in this repository.

A. Dataset

The dataset used is the "Amazon Customer Review" dataset [32] which consists of product reviews from Amazon.com, starting from the year 2008 to 2020, spanning across seven different domains, namely, book, electronics, pharmaceutical, healthcare, etc. We focused on the pharmaceutical section, which contained reviews of the "Turmeric Curcumin Supplement by Natures Nutrition", as our analysis centered around health-related information. To ensure alignment with current regulations, we exclusively considered reviews from 2018, coinciding with the enforcement of the GDPR [33]. We opted for a minimum review length of 30 words, as identified in prior studies to enhance informativeness without leading to information overload for readers [34], [35]. Consequently, our dataset comprised 1193 reviews. In this dataset, we empirically identified keywords that often appeared in reviews containing health-related information. The chosen keywords included:

> [mental, cancer, surgery, liver, kidneys, kidney, oncology, oncologist, tumor, chemo, brain, was diagnosed, allergy, allergies, depression, anxiety, heart, heart attack, stroke].

We used these keywords to select a subset of reviews likely to disclose customers' health information. This process yielded a dataset of 100 reviews (8% of the entire dataset). From this subset, we randomly sampled 20 reviews, and we added 8 non-sensitive reviews selected by the authors of this paper.

B. Prompt engineering

Once we constructed our ad-hoc dataset, we proceed with the prompt-engineering phase. Prompt engineering is the process of designing and refining text-based inputs (prompts) to optimize the output of LLMs. It involves strategically structuring prompts to elicit desired responses, improving accuracy, relevance, and the quality of LLM-generated content [36].

We experimented with different prompting strategies such as Zero-shot, Few-shot, and Chain-of-thought prompting. These strategies come under the super-set of In-Context Learning, where the model is prompted with all the information and context required to answer the question effectively.

1) Zero-shot prompting: This is the simplest strategy, where the model is prompted with a description of the task to complete. This provides no extra information about the task and expects the model to approximate accordingly. Here is the template for zero-shot with a placeholder in the input area.

> Input: {reviews}
> Sensitivity Score (out of 10):
> Redacted Version:
> Reason:

2) Few-shot prompting: This prompting strategy consists in providing a context to the LLM about the type of task. This makes it easy for the LLM to understand the expectations of the user on the given task. Prior work has shown Few-shot to perform better than zero-shot in many cases [37]. Here is the template for Few-shot with a placeholder in the input area.

> Input: My daughter has cancer and we are using this in her treatment
> Redacted Version: My daughter uses this as part of her treatment.
> Sensitivity Score (out of 10):
> Reason: "Cancer" might be a sensitive word and exposing that is not ideal.
> Input: {reviews}
> Redacted Version:
> Sensitivity Score(out of 10):
> Reason:

3) Chain-of-thought prompting: Chain-of-thought (CoT) consists in asking the LLM to enunciate the subtasks required to perform the bigger task [38]. It has been shown to provide more accurate results in complex or long running tasks. We have used two different CoT prompts across our research: **CoT 1** without a definition of sensitivity, **CoT 2** with a definition of sensitivity inspired by the definition of health-data provided by the GDPR [33]. Below are the two templates we use with a placeholder in the reviews area.

CoT 1
The following is a review of an individual's experience with a health product on Amazon. Please start by identifying any sensitive information included in this review.
Propose a sanitized version of the text that maintains its informativeness while removing sensitive details. Your output should have the following JSON format:
{{ "sens_info": Identify each piece of sensitive information included in the review. This field should be left blank if the text is non-sensitive.
"rephrased": Propose a way to sanitize the original text from sensitive information while leaving unrelated words unchanged. This field should be left blank if the text is non-sensitive.
"explanation": Explain why you rephrased different parts of the review and how this makes it less sensitive. This field should be left blank if the text is non-sensitive.
"sens_score": On a scale of 0 to 10, where 0 is no sensitive information, and 10 is a lot of sensitive information, rate the sensitivity of the original review.
"sens_reph": On the same scale, rate the sensitivity of the original review. This field should be left blank if the text is non-sensitive.
"info_score": On a scale of 0 to 10, where 0 is useless to people who read the review, and 10 is highly informative to people who read the review, rate the original review.
"info_reph": On the same scale, rate the informativeness of the rephrased review. This field should be left blank if the text is non-sensitive. }}
{reviews}

CoT 2
The following is a review of an individual's experience with a health product on Amazon. Please start by identifying any health-sensitive information about the author included in this review.
Propose a sanitized version of the text, in the voice of the reviewer, that maintains its informativeness while removing sensitive details.
Health-sensitive information includes, but is not limited to, details regarding medical conditions, treatments, or personal health experiences.
Your output should have the following JSON format:
{{ "sens_info": Identify each piece of sensitive information included in the review. This field should be left blank if the text is non-sensitive.
"rephrased": Propose a way to sanitize the original text from sensitive information while leaving unrelated words unchanged. This field should be left blank if the text is non-sensitive.
"explanation": Explain why you rephrased different parts of the review and how this makes it less sensitive. This field should be left blank if the text is non-sensitive.
"sens_score": On a scale of 0 to 10, where 0 is no sensitive information, and 10 is a lot of sensitive information, rate the sensitivity of the original review.
"sens_reph": On the same scale, rate the sensitivity of the original review. This field should be left blank if the text is non-sensitive.
"info_score": On a scale of 0 to 10, where 0 is useless to people who read the review, and 10 is highly informative to people who read the review, rate the original review.
"info_reph": On the same scale, rate the informativeness of the rephrased review. This field should be left blank if the text is non-sensitive. }}
{reviews}

4) Prompt Experiments: Our initial experiments employed a zero-shot approach, where the task for the LLM was to identify sensitive information within textual data. While we observed a reasonable degree of consistency in the model's ability to measure sensitivity, the zero-shot setup exhibited limitations. In particular, the model often overlooked less sensitive conditions, such as knee pain, failing to classify them as sensitive information. Furthermore, we encountered challenges with the fidelity of rephrased text; the rephrased versions significantly diverged from the original, occasionally altering the meaning of sensitive information.

For what concerns the experiments with the Few-shot techniques, we observed some improvements in classifying and rephrasing but still below expectation. Therefore, we compared these techniques with the results from CoT prompts. Some examples of these are reported in Table IV, in which we compared the performance across different techniques. These comparisons confirmed the limitations of zero-shot and few-shot approaches when contrasted with CoT prompting techniques, as highlighted by Wei *et al.* [38]. Our analysis suggests that CoT prompting more effectively preserves the desensitization intent, whereas zero-shot and few-shot methods may not consistently sanitize sensitive information, which could be advantageous in certain contexts.

Eventually, we experimented with the **CoT 1** on the 28 reviews selected as described in Section III-A, for 10 times. However, we still observed some flaws, such as false negative and bad rephrasing. For example, the sensitivity scores for *Review 19* were all 0. The review text is below, which mentions a general yet health-related issue in her shoulder:

Original My new favorite product! It's helped alliviate most of the pain in my shoulder. It also has the added benefit of boosting my metabolism, making it possible to maintain a steady weight.

On the other side, *Review 7* states:

Original Unfortunately, taking this for 2 days upset my stomach so bad and caused me to have diarrhea. The last straw was me having joint pain in my shoulder/neck on 2nd day. I guess this dosage is too strong for me. I am taking other supplements and have never had these symptoms so it's got to be this supplement causing all the disruption in my body. Great it worked for so many others, but I can't say the same for me.

was rephrased as follows for a sensitivity score equal to 3:

Rephrased The product didn't suit my body well; after two days of use, it upset my stomach. I also experienced joint pain. While it may have worked for others, I didn't have the same experience.

These limitations brought us to conduct a final experiment on the dataset described in Section III-A, utilizing the GPT-4 model with the **CoT 2**. We use default temperature of 1. The model scored and rephrased each review 10 times, resulting in a final dataset comprising 280 texts and their scores. Each original review was rephrased 10 times. Additionally, both the original and rephrased texts were evaluated 10 times by the model on the dimensions of informativeness and sensitivity. This process aimed to assess the stability of the scores and the quality of the rephrasing across multiple iterations. Table I shows an overview of the dataset used for the final experiment, the number of trials per review, the outcome dataset with the rephrasing and the scores from the model, the total of failed attempt of the model, and the percentage of success.

TABLE I: Overview on the final experiment

Total texts	Trials	Sampled Reviews	Non-sensitive Reviews	Outcome dataset	%Failed	%Success
28	10	20	8	280	0.017857	0.982143

C. Metrics

To address RQ1, RQ2, and RQ3, we combined statistical analysis and well-established metrics on the outcome dataset.

We assessed the quality of rephrased texts through both quantitative and qualitative methods. Quantitatively, we measured recall using ROUGE-L and precision using BLEU scores [39]. It is important to note that these metrics are not direct indicators of rephrasing quality, as they are typically used for assessing sentence similarity. To address this limitation, we incorporated human evaluation, wherein human graders assess the sensitivity and informativeness of each response, providing a more nuanced understanding of rephrasing quality.

1) Human evaluation: To address RQ1 (sensitivity detection), we gathered four annotators with diverse backgrounds, academic experiences, and nationalities. All annotators have been working and studying in the field of privacy for over two years. The participants were instructed to read the review_text (original text) and identify any health-sensitive information mentioned according to the definition embedded in the prompt:

> *Definition of Sensitivity*: "Health-sensitive information includes, but is not limited to, details regarding medical conditions, treatments, or personal health experiences."

This definition drew inspiration from Article 4(15) of the GDPR [33] to mitigate ambiguity stemming from the subjective nature of the task. Subsequently, each annotator assessed whether the sensitive information was accurately detected by reviewing the model's explanation and responding with "Y" (yes), "N" (no), or "IDK" (I don't know) in the detection column. This allowed us to set a baseline for the sensitivity detection capability of the model. We divided the dataset such that each text underwent review by two annotators [40], [41].

To explore RQ2 (score analysis), annotators were tasked with rating the sensitivity of both the review_text and rephrased_text on a scale of 0 to 10. Here, 0 signified no sensitive information, while 10 denoted a significant presence of sensitive information, according to the provided definition. For what concerns the informativeness of the review in the human evaluation, we adopted a definition provided by Sun *et al.* [35] for "an experience product" as it is the curcumin supplement. According to the author, "A review containing detailed descriptions for attributes and personal experiences is helpful to consumers of experience products", as well as platform attributes "the services provided by the platform, such as shipping, marketing promotions, influence consumer purchase decisions". Annotators were instructed to rate both the review_text (original) and rephrased_text (rephrased) on a scale of 0 to 10, where 0 indicated the review was useless to readers, and 10 implied high informativeness. Annotators were encouraged to provide reasoning behind their ratings for each answer, with a guideline provided (see Appendix B).

2) Statistical analysis: In assessing the reliability and accuracy of the model's score (RQ2), we conducted an extensive bootstrap analysis to validate the consistency and precision of the mean scores derived from our model evaluations. This analysis involved generating $n_{\text{bootstraps}} = 1000$ bootstrap samples, a rigorous methodological approach that allows for the robust estimation of 95% confidence intervals (CIs) for the mean scores across various data columns. Such a substantial number of samples enhances the statistical power of our analysis, providing a foundation for evaluating the reliability of the mean as a central tendency measure.

3) Rouge-L: Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [42] is a widely used metric in NLP for evaluating the quality of machine-generated text summaries. It is based on the Longest Common Subsequence between the model output and reference. A longer shared sequence should indicate more similarity between two sequences. This metric allowed us to measure the distance between the original and the rephrased sentences. In an ideal rephrasing, the impact on the original sentence is minimal, hence we expect high Rouge-L in cases where the rephrasing can be simple.

4) BLEU: BiLingual Evaluation Understudy (BLEU) [43] is a metric for automatically evaluating machine-translated text. It calculates similarity between a machine-generated text (e.g., the rephrased version) and one or more human-written reference texts. BLEU focuses on n-gram precision. An n-gram is a sequence of 'n' words. It checks how many n-grams from the rephrased text also exist in the reference texts [44]. Both BLEU and ROUGE were used to further evaluate the sanitizing task with ChatGPT (RQ3).

IV. Results

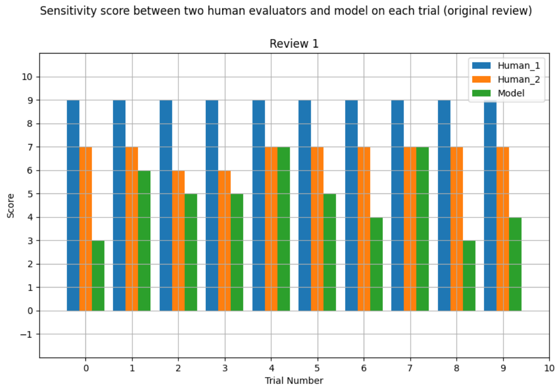
In this section, we show empirical results and their analysis.

A. Sensitivity detection

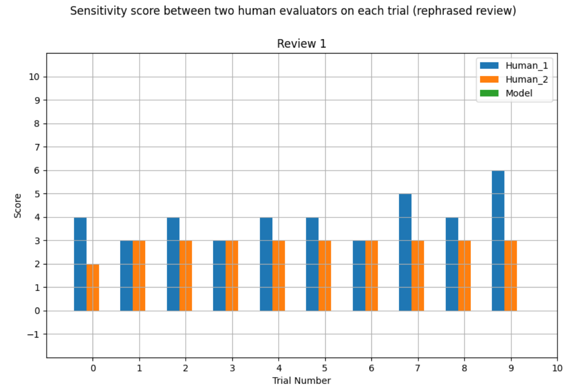
To address **RQ1**, that is, to verify whether the sensitivity detection was accurately identified by the model, we first computed the kappa score to measure agreement on the "detection" column [45], [46]. The result revealed perfect agreement between the annotators, with only one instance marked as a bad detection ("No"). Consequently, we conducted a more refined analysis of the scores provided by human evaluators. We restricted our analysis to the text instances where the model did not fail. For each review, we compared the results from each human evaluator with those from the model, focusing on both sensitivity and informativeness. Figure 1 represents an example of this analysis for the *Review 1*:

> **Original** I turned to Amazon to get this when I couldn't find it in the store in capsule form. I've been taking it since my first cancer battle over ten years ago. I was pleasantly surprised with this, no stomach upset, easy to swallow, very happy with it and it's a good value for the amount paid vs how many you get. Very please.

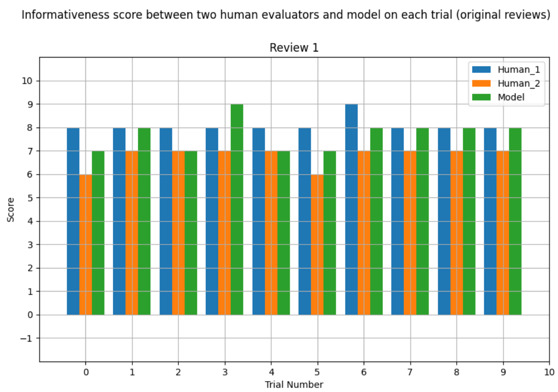
Figure 1a and Figure 1b show the sensitivity scores for the original reviews and rephrased reviews given by two human



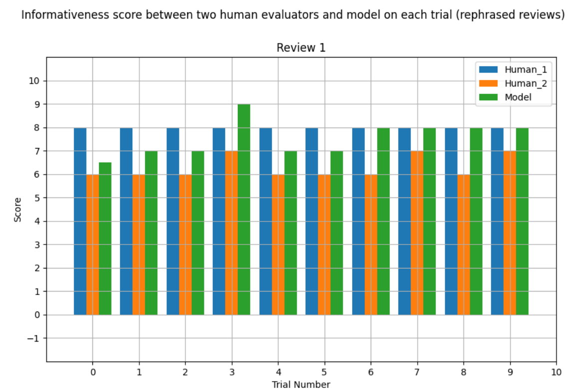
(a) Sensitivity score for the original reviews



(b) Sensitivity score for the rephrased reviews



(c) Informativeness score for the original reviews



(d) Informativeness score for the rephrased reviews

Fig. 1: Analysis of the sensitivity and informativeness score provided by the model and the human evaluators.

evaluators and the model in 10 trials. Figure 1c and Figure 1d show the informativeness scores for the original reviews and rephrased reviews given by two human evaluators and the model in 10 trials. As depicted in Figure 1a, both human evaluators identified the review as sensitive, with slightly different scores. The model's scores, however, fluctuate between 3 and 7. Regarding informativeness of the original text (Figure 1c), there is a noticeable agreement between the annotators, and the model's scores remain fairly consistent.

Turning to the sensitivity of the rephrasing of this review, illustrated in Figure 1b, there are points where the annotators' evaluations overlap, indicating a similar trend. However, the model consistently assigns lower scores than the annotators, suggesting a potential bias in its self-evaluation. An example of a rephrasing version of *Review 1* is:

> **Rephrased** I turned to Amazon to get this when I couldn't find it in store in capsule form. I've been using this product for quite some time now. I was pleasantly surprised with this, no stomach upset, easy to swallow, very happy with it and it's a good value for the amount paid vs how many you get. Very pleased.

Regarding the informativeness score of this review, Figures 1c and 1d indicate that there hasn't been a significant change, with the scores of both humans and the model being similar.

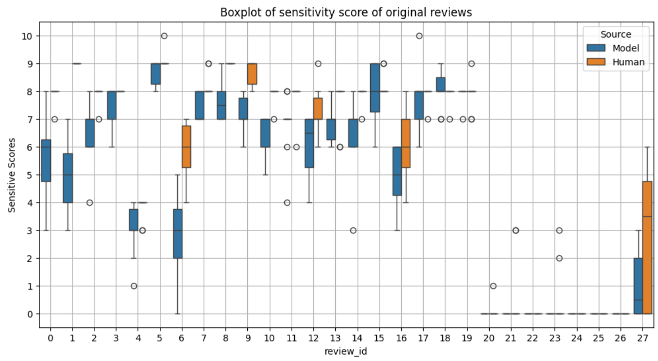
The analysis of the entire dataset is provided in the Appendix available in the data repository. Figures A1, A2, A3, and A4 show the scores for each experiment on each review.

Figure 2 shows the boxplots of the average humans' score compared with the model score. The analysis from Figure 2 revealed that 98 out of 112 boxes from human annotators have a 50th percentile range of 1 or smaller. We observed more variability in reviews containing ambiguity, such as *Review 6*:

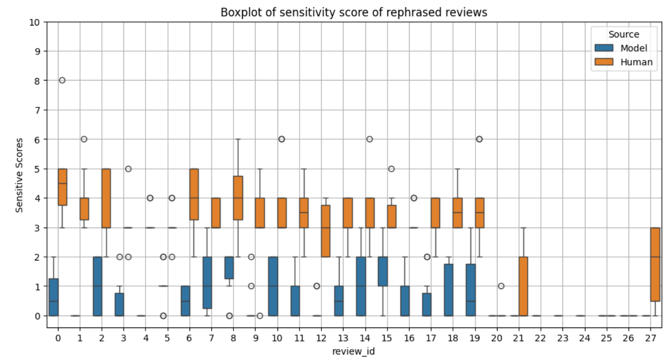
> **Original** I use it mostly for memory and overall Brain health (some evidence that it may prevent Alzheimer's), since I don't use it to reduce inflammation due to arthritis, I can't vouch for its efficacy in that regard. But I can say that it is probably one of the higher quality Turmeric Supplements, due to the addition of ingredients to improve its Bioavailability.

In this review, the author mentions an interest in brain health, although the information provided is somewhat vague.

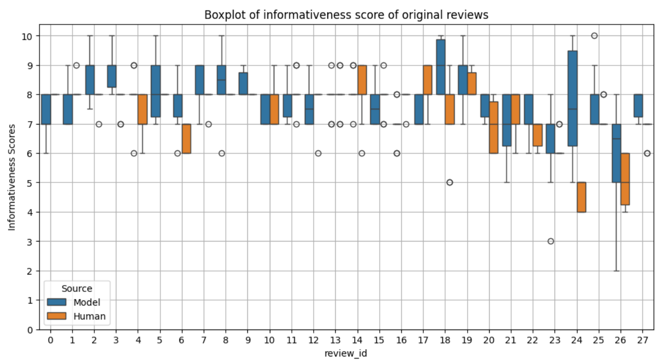
Comparing with the **CoT 1** (Section III-B), achieving self-consistency as discussed in [47] is challenging, with a broad variance rendering it impractical. This is shown in Figures 5a and 5b, which analyze the variability in sensitivity and informativeness scores of the old prompt, respectively. These findings led us to conclude that a clear definition of sensitivity is crucial for consistently enhancing model performance.



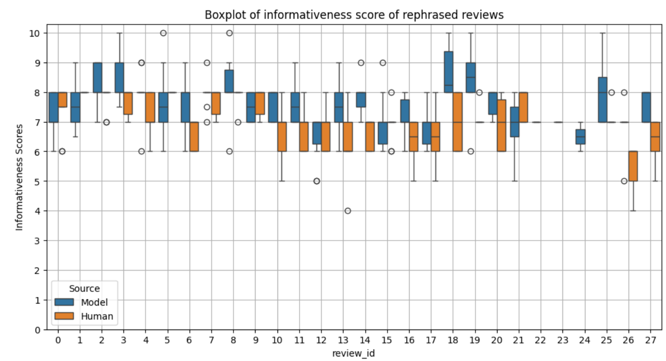
(a) Sensitivity score of original review



(b) Sensitivity score of rephrased review



(c) Informativeness score of original review



(d) Informativeness score of rephrased review

Fig. 2: Boxplot of scores of original and rephrased reviews

Answer to RQ1. Overall, we noted consistent trends between human annotators and the model regarding the detection of sensitivity in the provided text. Particularly, the model demonstrates proficiency in detecting sensitivity when prompted with a clear definition of sensitivity.

B. Score stability

Regarding **RQ2**—our goal to determine the stability of the model's scores and their alignment with human evaluations—we carried out a statistical analysis that focused on comparing the scores for each review, both original and rephrased, across two dimensions: sensitivity and informativeness. For the human evaluation, we adopted the mean value between two evaluators. Figure 2 illustrates this comparison, presenting the mean scores and their variability (as captured by the whiskers) for both the original and rephrased reviews, as assessed by human evaluators and the model.

Figure 2a represents the sensitivity scores of the original review given by the model and the humans. Regarding the model's scoring, it appears fairly consistent for the majority of the reviews, as the boxplots are not excessively wide. Exceptions are observed for *Reviews 0* and *6*. *Review 0* states:

> **Original** My husbands ortho doctor suggested him using Turmeric Curcumin to help with his knee issues, well I would say this product is working because he does not have to use the knee brace as much. So hopefully it will keep knee replacement surgery off for a bit longer.

The sensitivity of this review is debatable and somewhat hard to establish. It primarily discusses the author's husband health issues rather than the author, which may be considered less sensitive, particularly considering the prompt that asks for "any health-sensitive information about the author." Regarding *Review 6*, as discussed in Section IV-A, the author mentions interests in some health issues without clearly stating that they are personal issues. We believe these reasons contribute to the instability in the sensitivity scores. For what concerns human evaluation, Figure 2a depicts a rather stable situation, with boxplots not being excessively wide. An exception is noted for review 27, for which the same considerations of *Review 6* hold. For the sensitivity scoring of rephrased reviews, Figure 2b illustrates the model and the human scores respectively. As previously mentioned, the model scored lower than the human in rating the sensitivity of the reviews that have been rephrased by the model. This could be a potential bias of the model that requires further investigation. However, the results indicate that the scoring in this case is fairly reliable. Moreover, both human and model scores decreased, even though the values given by humans tend to be higher than those of the model.

Regarding the informativeness of the original reviews (Figure 2c), the scores given by both the model and the humans are consistent and fairly similar. An exception for human evaluation is observed for *Reviews 24* and *26* (below):

> **Original** This is my second bottle of this supplement and I

> happened to have one capsule remaining from the previous one. The difference in colour was striking. The capsules in the new bottle were of a lighter colour! So, in my disappointment my question is, do these new ones have more filler in them? I just don't know!

The informativeness of this review is arguable as it only provides information about the color of the supplement. Similar observations apply to *Review 24*, which solely focuses on the reviewer's personal delivery experience.

On the other hand, the results for the informativeness of rephrased reviews are consistent for both humans and the model. Figure 2d depicts the boxplots regarding these results.

The consistency observed in the scores, as indicated by the boxplots, led us to conclude that the mean values for each trial offer a fair representation of the model scores. The robustness of the mean as a metric was reinforced through a bootstrap analysis, with *n_bootstraps* = 1000 bootstrap samples. This substantial number of samples underpins the reliability of our original means and their 95% confidence intervals (CIs) across various data columns. The results of this analysis are shown in Table II. Specifically, columns like "info_orig_model" and "info_reph_model" show particularly narrow CIs, highlighting the mean's precision and stability as a central tendency measure for these data points. This level of precision validates the mean's role as an effective representative value for these key metrics. Conversely, columns including "sens_original_model" and "sens_original_human" exhibited broader CIs, signaling increased variability. Despite this, the mean remains a crucial summary statistic, albeit one that requires careful interpretation. Such interpretation should consider the overall data distribution and the potential influence of outliers.

TABLE II: Bootstrap results

Column	Original Mean	95% CI Lower	95% CI Upper
sens_original_model	4.66	4.27	5.07
sens_rephrased_model	0.58	0.47	0.69
info_original_model	7.77	7.65	7.92
info_rephrased_model	7.63	7.51	7.75
sens_original_human	5.59	5.16	6.00
sens_rephrased_human	3.09	2.91	3.27
info_original_human	7.42	7.29	7.55
info_rephrased_human	7.10	6.98	7.21

Therefore, we compare these mean values directly with those obtained from human evaluations. As illustrated in Figure 3, this comparison shows a close alignment between the sensitivity scores for the original reviews as determined by both humans and the model, with further details provided in Figure 3a. In the case of rephrased reviews (see Figure 3b), the model's sensitivity scores were consistently lower or equal to those of human evaluators. However, a consistent trend is evident: despite the higher scores from human evaluators, the difference between human and model evaluations remains constant, indicating a systematic pattern in the assessment process. This systematic approach, bolstered by the extensive use of $n_{\text{bootstraps}} = 1000$ bootstrap samples, ensures the reliability and validity of our findings.

Regarding the informativeness scores (see Figure 3c), the results for humans and the model are fairly close for the original reviews, with a maximum difference of 1 for the range of 50% percentile, except *Review 26* acting as an outlier here. This is also the case for the informativeness score of the rephrased reviews (see Figure 3d), although we observe some hallucination [48] of the model, as reviews 22, 23, and 24 were not rephrased, therefore the informativeness score cannot be established (-1 given by humans as the text is missing). These results for the informativeness score are interesting since a definition of this score was not provided in the prompt.

Answer to RQ2. The scores provided by the model align fairly well with the human scores. Additionally, the model's scores demonstrate a consistent stability across trials, with minimal observed variations.

C. Rephrasing evaluation

To tackle **RQ3** – evaluating the rephrasing – we conducted a comparative analysis of results obtained from various metrics. Rephrasing a sensitive sentence to render it non-sensitive while preserving its informativeness is a challenging task. This is because sometimes informativeness and sensitivity are tied to each other. Nevertheless, the model shown an ability to maintain a reasonable balance between these two dimensions. For example, the following text is *Review 13*:

> **Original** I have arthritis in my hands and my left shoulder. I had surgery 2 yrs ago on my shoulder and these pills were recommended by a friend and they really help!!! I will b ordering more soon. They r a great value!!

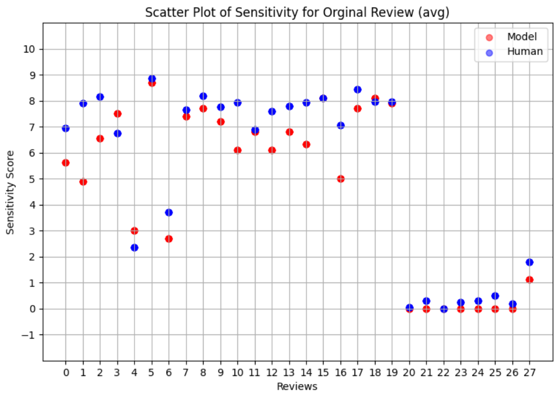
Preserving the essence of the review without mentioning medical information is difficult, as it is integral to the review's informativeness. The model's rephrased version is:

> **Rephrased** I've dealt with joint discomfort and was recommended these supplements by a friend. I found that they truly make a difference! I plan on purchasing more soon; they provide great value!

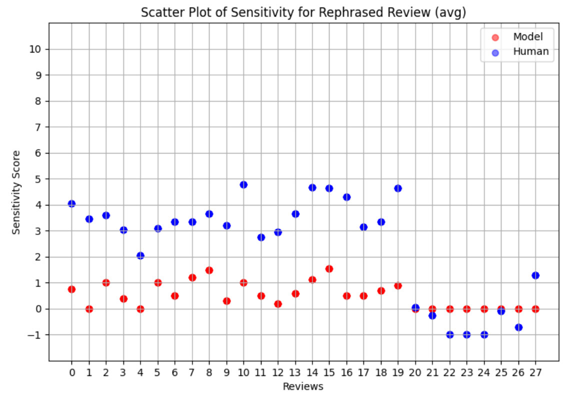
which attains a reasonable balance, while still retaining the essence of the review.

1) ROUGE-L: Table III reports the values for ROUGE-L and BLEU metrics. The mean ROUGE-L precision score is 0.4 showing similarity between the original and the rephrased text. The degree of modification isn't significantly higher because, in certain reviews, the model produces extensively altered rephrasings. This occurs because the model tends to correct all grammar and spelling errors in the rephrased text. The low Standard Deviation indicates that the model rephrased output is consistent in its similarity with the original review produced by the user. However, it is important to note that ROUGE-L, while very useful in evaluating summarization tasks, is not an exact representation for the rephrasing task. For example, it is not possible to evaluate whether the sensitive information was actually removed in the sentence.

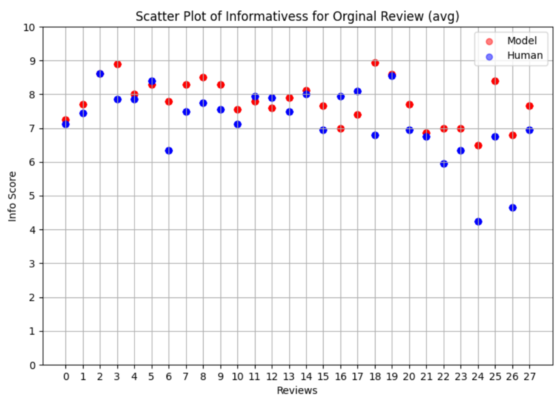
2) BLEU: Generally speaking, BLEU Score is considered good between 0.4 - 0.6 [49]. Table III shows that the average BLEU score is 0.34 and the median value is 0.15. These



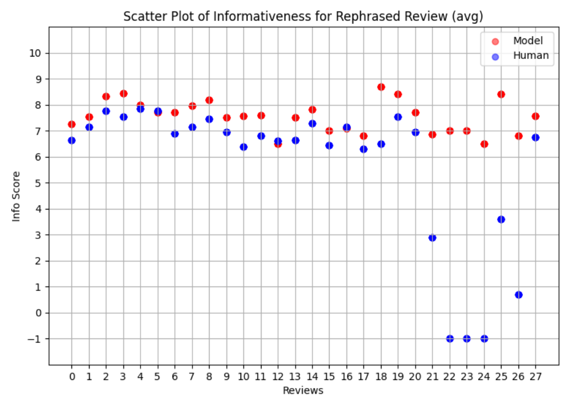
(a) Sensitivity for original reviews



(b) Sensitivity for rephrased reviews



(c) Informativeness for original reviews



(d) Informativeness for rephrased reviews

Fig. 3: Scatter plot of average scores for original and rephrased reviews

TABLE III: ROUGE-L and BLEU evaluation

Metrics	Mean	Median	St Dev
RougeL precision	0.42	0.41	0.27
RougeL recall	0.36	0.32	0.27
RougeL fmeasure	0.38	0.34	0.27
BLEU	0.34	0.15	0.38

scores show that the number of n-grams that match between the reviews is not very close, i.e., the original review and the rephrased review are not very close to each other for what concerns the word similarity. This can be explained considering the LLMs are good at correcting grammar and replacing words with their synonyms [50].

3) Human Evaluation: Figure 3 represents distances between the mean values given by humans and the mean values given by the model for each different review (x-axis). Each plot represents a different score for either original or rephrased reviews. Based on the results represented in Figure 3, there appears to be a consensus between human evaluators and the model regarding the sensitivity level of the text. While the model tends to assign slightly lower sensitivity scores to rephrased texts compared to human evaluators, the overall trend, as illustrated in Figure 3b, demonstrates alignment between human and model evaluations on average.

The model demonstrates consistency in preserving the informational content of the reviews, as evidenced by the minimal change observed between Figure 3c and Figure 3d. Furthermore, for what concerns the informativeness, the obtained *p-value* of 0.16 from the comparison of model evaluation scores before and after rephrasing suggests that there is no significant difference between the two sets of scores, as it exceeds the conventional threshold of 0.05 for statistical significance. On the contrary, it effectively reduces the sensitivity of the text, transitioning from the original sensitivity levels depicted in Figure 3a to the rephrased versions shown in Figure 3b. In this case, we obtained a *p-value* of 1.02e-8 from the comparison of model evaluation scores before and after rephrasing, which indicates a significant decrease in sensitivity from the original to the rephrased text. On the other hand, from the comparison of human evaluation before and after rephrasing, the *p-value* for sensitivity is equal to 0.00032, which confirms the results for the decrease in sensitivity. On the contrary, the *p-value* from the comparison of informativeness before and after rephrasing is equal to 0.0089, indicating a statistically significant change in the content of the rephrased sample. This could be attributed to the relatively small size of the dataset or the nature of the data used.

Answer to RQ3. The sanitized text (rephrasing) proposed by ChatGPT is overall less sensitive than the original text. The level of informativeness seems unchanged only from the model perspective. This result needs to be further investigated.

V. Threats to Validity

a) External Validity: Human evaluation was carried out by the authors of this paper. We ensured that the reviews were randomized and that the annotators had clear instructions to proceed. However, an external annotation through platforms like Mechanical Turk could ensure a more unbiased evaluation.

The dataset used is relatively small, and expanding it could significantly enhance the validity of our results. Furthermore, diversifying our sources of health-sensitive could provide valuable insights for assessing the proposed methodology. For this purpose, we are working on the dataset produced by [5] to test our methodology on a larger and diverse dataset. Similarly, a more varied product analysis could strengthen our results.

b) Construct validity: The precision of the proposed scoring scales may influence the overall scoring stability outcomes. Changing scales and evaluation of the scoring could help verifying our hypothesis. The observations about the sensitivity are generally applicable across the population, as health information (such as cancer diagnosis) is considered very sensitive by patients with and without diagnosis [51].

c) Internal validity: We found instances of LLM hallucination in ChatGPT's outputs (see Figure A.4, *Review 22* and *23*). Although these anomalies were present in a minority of responses, their potential to affect the accuracy of our findings warrants attention. Future research should focus on developing and integrating mechanisms to deal with hallucinations.

VI. Discussion and Future Works

In this study, we outlined and tested a methodology for sanitizing user-generated texts using ChatGPT. Our focus was on Amazon reviews within the health domain, emphasizing the need to maintain the usefulness of the original text while reducing sensitivity. To achieve this, we tasked the model with preserving informativeness while minimizing sensitivity. In doing so, we asked the model to provide scores for the sensitivity and the informativeness of both the original and sanitized text. Sensitivity was defined according to GDPR guidelines to ensure objectivity [33], while informativeness criteria were based on domain-specific attributes such as detailed product descriptions and personal experiences [35]. The methodology we propose is illustrated in Figure 4.

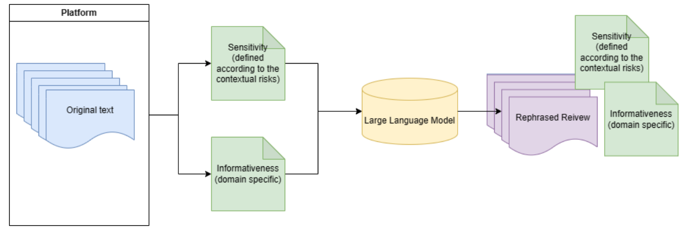


Fig. 4: Methodology proposed in this paper to sanitize sensitive text while maintaining its informativeness.

Our findings suggest that ChatGPT effectively detects sensitivity in text when provided with a clear definition (RQ1). This implies potential for developing customized sensitivity detection tools tailored to specific tasks, bypassing the limitations of generic taxonomies or dictionaries. Additionally, we found the model's sensitivity and informativeness scores to be reliable (RQ2); this result together with established metrics and human evaluation allowed us to conclude that the rephrasing was fairly good (RQ3). The dimension of informativeness needs further investigation, ideally with domain experts.

The stability of the scoring mechanism is particularly interesting, as it allows for dynamic adjustments to text sensitivity based on user needs and context. This provides a more flexible and personalized approach to sensitivity detection.

Future works should explore how our methodology can work with other LLMs such as LLAMA2, to test if the score stability stays consistent when sensitivity is clearly defined. Moreover, systematic evaluation and optimization of prompt templates used in the sanitization process represent a significant area for future research. This includes experimental studies to compare the effectiveness of various prompt designs, their context-specific performance, and the identification of prompt characteristics leading to optimal outcomes. This line of research could establish best practices for prompt creation in the context of privacy-filtering.

Acknowledgment

This research has been partially supported by the US National Science Foundation under its Security and Trustworthy Computing program (Grant CNS-1914486).

References

[1] W. Yuchao, Z. Ying, and Z. Liao, "Health privacy information self-disclosure in online health community," *Frontiers in public health*, vol. 8, p. 602792, 2021.

[2] I. Ajunwa, K. Crawford, and J. S. Ford, "Health and big data: An ethical framework for health information collection by corporate wellness programs," *The Journal of Law, Medicine & Ethics*, vol. 44, no. 3, pp. 474–480, 2016, pMID: 27587451. [Online]. Available: https://doi.org/10.1177/1073110516667943

[3] H. Chade and E. E. Schlee, "Insurance as a lemons market: Coverage denials and pooling," *Journal of Economic Theory*, vol. 189, p. 105085, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022053120300806

[4] J. Li, "Open medical big data and open consent and their impact on privacy," in *2017 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2017, pp. 511–514.

[5] M. Valizadeh, P. Ranjbar-Noiey, C. Caragea, and N. Parde, "Identifying medical self-disclosure in online communities," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 4398–4408.
[6] G. Bernardini, H. Chen, A. Conte, R. Grossi, G. Loukides, N. Pisanti, S. P. Pissis, and G. Rosone, "String sanitization: A combinatorial approach," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer, 2020, pp. 627–644.
[7] M. Korytkowski, J. Nowak, R. Scherer, and W. Wei, "Privacy preserving by removing sensitive data from documents with fully convolutional networks," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2022, pp. 277–285.
[8] W. B. Tesfay, J. Serna, and K. Rannenberg, "Privacybot: Detecting privacy sensitive information in unstructured texts," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 53–60.
[9] G. Bernardini, H. Chen, G. Loukides, N. Pisanti, S. P. Pissis, L. Stougie, and M. Sweering, "String sanitization under edit distance," 2020.
[10] Y. Dou, I. Krsek, T. Naous, A. Kabra, S. Das, A. Ritter, and W. Xu, "Reducing privacy risks in online self-disclosures with language models," 2023.
[11] J. Boegershausen, H. Datta, A. Borah, and A. T. Stephen, "Fields of gold: Scraping web data for marketing insights," *Journal of Marketing*, vol. 86, no. 5, pp. 1–20, 2022.
[12] Y. Huang, H. Liu, W. Li, Z. Wang, X. Hu, and W. Wang, "Lifestyles in amazon: Evidence from online reviews enhanced recommender system," *International Journal of Market Research*, vol. 62, no. 6, pp. 689–706, 2020.
[13] M. Burkholder and R. Greenstadt, "Privacy in online review sites," in *2012 IEEE Symposium on Security and Privacy Workshops*, 2012, pp. 45–52.
[14] Amazon. (2023) Amazon privacy notice. Accessed on: 20/02/2024. [Online]. Available: https://www.amazon.com/gp/help/customer/display.html/ref=s9_acss_bw_cg_usccpa_4a1_cta_w?nodeId=468496&pf_rd_m=ATVPDKIKX0DER&pf_rd_s=merchandised-search-1&pf_rd_r=3PEMTZ5JBZ6GFFJ084G1&pf_rd_t=101&pf_rd_p=8db4eb03-6e6e-46d1-987d-a9f0640e18ea&pf_rd_i=20907011011
[15] F. Fukuyama, B. Richman, and A. Goel, "How to save democracy from technology: ending big tech's information monopoly," *Foreign Aff.*, vol. 100, p. 98, 2021.
[16] F. Farhat, B. M. Chaudhry, M. Nadeem, S. S. Sohail, and D. Ø. Madsen, "Evaluating large language models for the national premedical exam in india: Comparative analysis of gpt-3.5, gpt-4, and bard," *JMIR Medical Education*, vol. 10, p. e51523, 2024.
[17] P. Ram Mohan Rao, S. Murali Krishna, and A. Siva Kumar, "Privacy preservation techniques in big data analytics: a survey," *Journal of Big Data*, vol. 5, pp. 1–12, 2018.
[18] A. Papadopoulou, Y. Yu, P. Lison, and L. Øvrelid, "Neural text sanitization with explicit measures of privacy risk," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 2022, pp. 217–229.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[20] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin, "Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution," 2023.
[21] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: A dataset for evaluating safeguards in llms," 2023.
[22] Y. Ishibashi and H. Shimodaira, "Knowledge sanitization of large language models," 2023.
[23] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," 2024.
[24] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?" 2023.
[25] Z. Delile, S. Radel, J. Godinez, G. Engstrom, T. Brucker, K. Young, and S. Ghanavati, "Evaluating privacy questions from stack overflow: Can chatgpt compete?" *arXiv preprint arXiv:2306.11174*, 2023.
[26] G. Xu, C. Qi, H. Yu, S. Xu, C. Zhao, and J. Yuan, "Detecting sensitive information of unstructured text using convolutional neural network," in *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2019, pp. 474–479.
[27] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system," in *Proc AMIA Annu Fall Symp*, 1996, pp. 333–337.
[28] G. Architha, "Removing sensitive part of a text," 2019.
[29] A. Tveit, O. Edsberg, T. Rost, A. Faxvaag, O. Nytro, T. Nordgard, M. T. Ranang, and A. Grimsmo, "Anonymization of general practioner medical records," in *Proceedings of the second HelsIT Conference*, 2004.
[30] M. Douglass, G. Cliffford, A. Reisner, W. Long, G. Moody, and R. Mark, "De-identification algorithm for free-text nursing notes," in *Computers in Cardiology, 2005*, 2005, pp. 331–334.
[31] F. Hassan, D. Sánchez, J. Soria-Comas, and J. Domingo-Ferrer, "Automatic anonymization of textual documents: Detecting sensitive information via word embeddings," in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 2019, pp. 358–365.
[32] I. Chatterjee, "Amazon Customer Review," 2021. [Online]. Available: https://doi.org/10.7910/DVN/W96OFO
[33] European Parliament and of the Council. (2016) Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
[34] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," *Computers in Human Behavior*, vol. 48, pp. 17–27, 2015.
[35] X. Sun, M. Han, and J. Feng, "Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products," *Decision Support Systems*, vol. 124, p. 113099, 2019.
[36] G. M. Muktadir, "A brief history of prompt: Leveraging language models. (through advanced prompting)," 2023.
[37] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv e-prints*, pp. arXiv–2302, 2023.
[38] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
[39] A. Yang, K. Liu, J. Liu, Y. Lyu, and S. Li, "Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task," *CoRR*, vol. abs/1806.03578, 2018. [Online]. Available: http://arxiv.org/abs/1806.03578
[40] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin, "Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution," *arXiv e-prints*, pp. arXiv–2307, 2023.
[41] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: A dataset for evaluating safeguards in llms," *arXiv preprint arXiv:2308.13387*, 2023.
[42] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
[44] E. Reiter, "A structured review of the validity of bleu," *Computational Linguistics*, vol. 44, no. 3, pp. 393–401, 2018.
[45] M. J. Warrens, "Five ways to look at cohen's kappa," *Journal of Psychology & Psychotherapy*, vol. 5, 2015.
[46] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[47] M. J. Min, Y. Ding, L. Buratti, S. Pujar, G. Kaiser, S. Jana, and B. Ray, "Beyond accuracy: Evaluating self-consistency of code llms," in *The Twelfth International Conference on Learning Representations*, 2023.
[48] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan, "Llm lies: Hallucinations are not bugs, but features as adversarial examples," 2023.

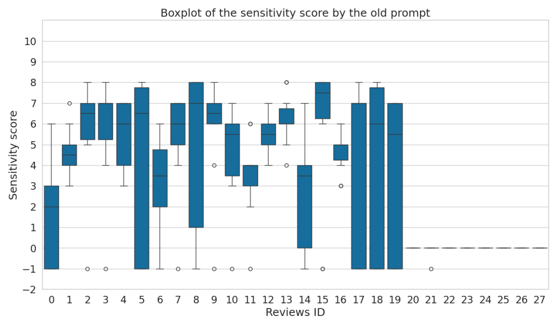
[49] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[50] M. Hosseini and S. P. J. M. Horbach, "Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review," 2023.

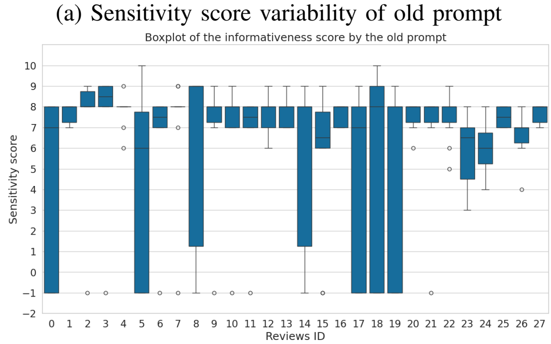
[51] D. Grande, D. A. Asch, F. Wan, A. R. Bradbury, R. Jagsi, and N. Mitra, "Are patients with cancer less willing to share their health information? privacy, sensitivity, and social purpose," *Journal of Oncology Practice*, vol. 11, no. 5, pp. 378–383, 2015.

APPENDIX

A. Analysis of the old prompt



(a) Sensitivity score variability of old prompt



(b) Informativeness score variability of old prompt

Fig. 5: Variability of scores with the prompt **CoT 1**

B. Guidelines for human evaluation

Definition of Sensitivity: "Health-sensitive information includes, but is not limited to, details regarding medical conditions, treatments, or personal health experiences." Read the review_text (original text) and identify the health-sensitive information mentioned according to the definition. Read the explanation provided by the model and establish if the sensitive information has been detected correctly by answering Y (yes), N (no), IDK (I don't know) in the column "detection".

Definition of informativeness: "A review containing detailed descriptions for attributes and personal experiences is helpful to consumers of experience products" as well as platform attributes"the services provided by the platform, such as shipping, marketing promotions, influence consumer purchase decisions". On a scale of 0 to 10 where 0 is no sensitive information, and 10 is a lot of sensitive information, rate the sensitivity of the review_text (sens_or) and rephrased_text (sens_reph). On a scale of 0 to 10, where 0 is useless to people who read the review, and 10 is highly informative to people who read the review, rate the review_text (info_or) and rephrased_text (info_reph). Please, fill the column comments with your thoughts regarding your score and anything interesting to highlight.

TABLE IV: Experiment on *Review 1*

columns	**content**
Original	I have been using it for a while now. It is the reason I have not had knee replacement surgery yet. Wife is using it to supplement her cancer treatment. Was originally taken a more expensive brand. This is more economical with the same results.
reph_zero	I've been using this product for some time and it's been beneficial for my knee condition, possibly delaying the need for surgery. My wife also finds it helpful alongside her medical treatment. We switched to this from a pricier brand, finding it equally effective and more affordable.
reph_few	I have been using it for some time now, helping me avoid more invasive procedures. My wife also finds it beneficial for her health condition. Initially, we used a costlier brand, but this one offers similar benefits at a more economical price.
reph_COT	I have been using it for a while now and have found it beneficial. My spouse also finds it helpful. Initially, we were using a more expensive brand, but this one is more economical and offers the same benefits.
Comments	COT and few entirely redacted the comments; zero kept some information to the detriment of privacy.