Exploring ChatGPT's ability to detect privacy violations in photo sharing

CHRISTINE CHEN*, BEN MALONE*, KENDRICK MERNITZ*, CHRISTOPHER PAGE*, and APU KAPADIA, Luddy School of Informatics, Computing, and Engineering, USA
ARUN BALAJI BUDURU, Indraprastha Institute of Information Technology - Delhi, India

The proliferation of photo sharing on Online Social Media (OSM) has vastly increased the risk of privacy violations. This work investigates Large Language Model's (LLM), specifically ChatGPT's, proficiency in identifying and explaining privacy violations within photos. Using a dataset of 68 images encompassing common privacy-sensitive scenarios, we tasked three different models of ChatGPT to rate these images on their extent of privacy violation. These ratings were then compared to those made by humans. Our findings indicate that ChatGPT is effective in detecting privacy violations, with better performance when using fine-tuned models trained with expert privacy data. Further research in this topic can contribute to the development of automated tools to enhance privacy protections in this era of rampant photo sharing.

CCS Concepts: • **Security and privacy** → **Social aspects of security and privacy**; **Privacy protections**; *Usability in security and privacy*; • **Computing methodologies** → *Scene understanding*; *Visual inspection*; *Natural language generation.*

1 INTRODUCTION

The proliferation of photo sharing on online social media platforms has given rise to critical privacy concerns [21]. Countless photos are uploaded every day, many of which may contain sensitive or personal information, inadvertently exposing individuals to privacy violations ranging from minor embarrassments to serious security risks [15]. The increased volume of digital content has made it impractical for human moderators to effectively monitor and identify potential privacy issues in shared photos [8]. Herein lies the potential of artificial intelligence, particularly large language models (LLMs), in automating the detection of privacy-sensitive content in complex media. Such systems could not only assist with flagging potentially problematic images but also in educating users about privacy concerns related to their shared content. Recently, Amon et al. [4] explored the privacy perceptions of people in the context of meme sharing and Li et al. [15] identified a taxonomy of privacy violations present in photos shared online. Utilizing the dataset from Amon et al. and informed by Li et al.'s taxonomy, our research aims to explore this possibility of utilizing LLMs to offer users real-time suggestions to mitigate privacy risks in the context of photo sharing.Our work investigates ChatGPT's ability to discern sensitive content in images. Specifically we explore the following research questions: (R1) How well does ChatGPT detect privacy violations in photos?; (R2) How does ChatGPT reason about privacy violations in photos? Our findings provide a deeper understanding of ChatGPT's capabilities in identifying privacy-sensitive content within photos in the context of online photo sharing.

2 METHODOLOGY

2.1 Experimental Design

We conducted an experimental study of ChatGPT 4.0, which was selected due to its unique photo analysis capabilities. The three Models we analyzed were: (1) Base Model of ChatGPT (Base Model); (2)Base Model of ChatGPT prompted with a definition of sensitive content from prior literature [15] (Taxonomy Prompt Model); and (3) Fine-tuned ChatGPT model, specifically trained on photo sharing and privacy research literature (Fine-tuned Model).

Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Copyright remains with the author(s).
*All authors contributed equally to this research.

2.2 Data Collection

Our photo dataset consisted of 68 "meme" photos [4], each of which had been previously ascribed an average privacy rating derived form ratings, categorized provided by 245 human raters. These ratings were originally on a 1-5 scale; we categorized these ratings into three categories (tertiles) based on percentile: high, medium, and low. Every photo in the dataset was presented to each of the three models, which would then be prompted to assess each photo's potential for privacy concern. For the Base Model and Fine-Tuned Model, the prompt was: "Give this image a score of low, medium, or high in regard to the potential for a privacy violation, and explain why in a single paragraph." For the Taxonomy Prompt Model, the prompt was: "Based on the definition of sensitive content provided, give this image a score of low, medium, or high in regard to the potential for a privacy violation, and explain why in a single paragraph."

2.3 Qualitative Analysis Techniques

To answer our second research question about how ChatGPT reasons about privacy sensitive content within photos, we compared the occurrences of specific words in the outputs of the Base Model and Fine-tuned Model, targeting words central to privacy, data protection, and ethical issues. We selected words appearing at least 10 times in the outputs and, using the Porter stemming algorithm [20], concentrated on a curated list of 20 relevant word stems. The occurrence rate of each word stem was calculated by dividing its total occurrences by total number of word occurrences (excluding stop words) in a model's output. This process, focused on key word stems, enabled a direct comparison between the Baseline Model and the Fine-tuned Model in terms of concepts used to explain privacy concerns in photos.

2.4 Search Criteria Establishment for Training Fine-tuned Model

We fine-tuned the base model by creating a custom version of ChatGPT 4.0 where we specified how it should behave based on the papers selected below. *(1) Keyword Selection and Search Strategy:* To find papers that could be used to train the Fine-tuned Model, the term "photo privacy" was used to query Google Scholar. The scope of the search was limited to the first 10 pages of results, among which the top five most cited papers [1, 5, 13, 17, 18] were selected, based on the assumption that higher citation counts correlate with greater academic impact. As a secondary search strategy, results were sorted by the recency of publication. The time frame of this search was confined to papers published from 2019 to the present, with a preference toward more recent scholarly contributions. The search was restricted to the first 10 pages of results, and the five most recent and cited papers [12, 14, 16, 22, 23] were chosen in order to capture the recent and significant developments. *(2) Supplementary literature:* Although we plan a more extensive selection of research papers from a wide array of experts in the future, we faced practical limits on the number of papers the model would accept. For this work, we chose to focus on ten papers from our research group [2, 3, 6, 7, 9–11, 19, 21] on this topic.

3 RESULTS

3.1 Quantitative Analysis

The quantitative analysis involved calculating the percent match between ChatGPT's privacy violation ratings and those provided by 245 human raters, as shown in Table 1. This involved comparing ChatGPT's response (low, medium, high) for each photo with the human rating along with quantifying the degree of agreement. For instances where ChatGPT's ratings did not match with the human ratings, the analysis further categorized whether ChatGPT overestimated or underestimated the privacy risk. This comparison provided insight into the model's tendency toward caution or leniency in privacy assessment. Given the three categories of ratings, a baseline of 33.33% match rate represents random guessing.

Analysis shows that all three models overestimated privacy risks, with the Fine-tuned Model outperforming the others in match rate accuracy, with second highest overestimation rate. This suggests that the incorporation of expert knowledge can only somewhat enhance the model's ability to interpret privacy concerns in a manner akin to human understanding.

	Base Model	Taxonomy Prompt Model	Fine-tuned Model
Match rate	51.47%	54.41%	58.82%
Overestimation rate	81.82%	74.19%	78.57%

Table 1. ChatGPT vs. Human Raters Privacy Ratings - Percent Match and Over/Underestimation Analysis[1]

	Base Model		Taxonomy Prompt Model		Fine-tuned Model	
	Precision	Recall	Precision	Recall	Precision	Recall
High	55.55%	86.96%	61.76%	91.30%	71.43%	86.96%
Medium	35.00%	31.83%	33.33%	27.27%	42.86%	54.55%
Low	66.66%	34.78%	62.50%	43.48%	66.67%	34.78%

Table 2. Precision and Recall - Separated by Models and Rating Category

To better understand the overestimation, we examined the precision and recall for the three categories (low, medium, high). As seen in Table 2, the Base Model performs poorly (or close to random) for the medium category, with stronger recall of the high category, and has poor recall of the low category. The Taxonomy Prompt Model improves this situation a little, and the Fine-tuned Model greatly improves precision for the high category. We note that recall remains poor for the low category, but the performance for differentiating between the high and low categories are improved.

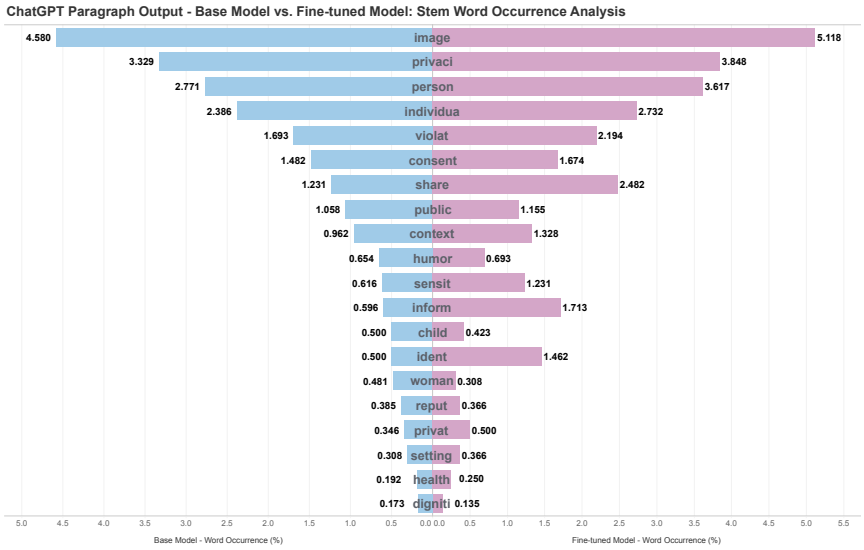

Fig. 1. Comparative Histogram of ChatGPT's Privacy Violation Terms: Base Model vs. Fine-Tuned Model

[1]Baseline: 33.33%

3.2 Qualitative Analysis

Figure 1 illustrates the differences in ChatGPT's interpretation of privacy violations between the Base Model and the Fine-tuned Model, highlighting the Fine-Tuned model's increased emphasis on privacy and data protection terms, such as "ident", "inform", "person", "share", "context". This comparison visualizes the distinct assessment tendencies of each model regarding privacy violations. On the other hand, the stability in occurrences of stems such as "woman", "reput", and "humor' across both models points to consistent thematic treatment across the dataset.

4 DISCUSSION

A critical observation across all model versions was the tendency to overestimate privacy risks. This overestimation raises important questions about risks of excessive censorship or misinterpretation. Furthermore, thematic shifts observed between the Base model and the Fine-tuned Model suggest significant implications for the application of language models in privacy risk evaluation. These trends suggest that the Fine-tuned Model could be more adept at preemptively identifying privacy violations in online content, particularly on social media. Fine-tuning of LLMs in emphasizing privacy-related terms demonstrates the potential for automation of privacy risk assessment.

5 LIMITATIONS

One notable limitation to our experimental design is how we translated the human privacy ratings to low, medium and high. Average human privacy ratings are continuous values whereas GPT ratings are discrete. Another limitation is possible inaccuracy in the human privacy ratings. The privacy ratings data was collected from regular social media users who may not know what constitutes as a privacy violation. Also, participants are located in the US and may not be a general representative sample. The Taxonomy Prompt Model and Fine-tuned Model were trained with privacy research written by privacy experts. Future work could collect human privacy ratings from a panel of experts for comparison with fine-tuned LLM models. Experiments in this work were conducted using ChatGPT 4. Note that as OpenAI continuously updates its models, these experiments may not be replicable.

6 CONCLUSIONS

In this work, we examined ChatGPT's effectiveness in identifying privacy violations in digital images as compared to human judgement. Although ChatGPT was reasonably proficient, models trained with expert privacy data had higher performance rates, indicating that optimizing ChatGPT's capabilities involves training the model with a clear understanding of what constitutes a privacy violation. At the same time, in cases where ChatGPT's output did not match the human ratings, it tended to overestimate the degree of violation, posing a risk of excessive censorship, especially with increased fine tuning. When looking at ChatGPT's reasoning for its ratings, we observe a tendency in the trained model to place more emphasis on words related to potential sharing of identifiable information. Although our work focused on ChatGPT 4, our findings support the utility of LLMs in automating privacy violations detection.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the Department of Defense via Purdue University under funding agency 13000844-031 and the National Science Foundation under grant no. CNS-2207019. We thank Bennett Bertenthal and Mary Jean Amon for their ideas and assistance as well as Susan Zvacek for her help with copy editing.

REFERENCES

[1] Shane Ahern, Dean Eckles, Nathaniel S. Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Over-exposed? privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 357–366. https://doi.org/10.1145/1240624.1240683

[2] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I. Bertenthal, and Apu Kapadia. 2020. Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. In *2020 IEEE Symposium on Security and Privacy (SP)*. 1350–1366. https://doi.org/10.1109/SP40000.2020.00006

[3] Mary Jean Amon, Nika Kartvelishvili, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2022. Sharenting and Children's Privacy in the United States: Parenting Style, Practices, and Perspectives on Sharing Young Children's Photos on Social Media. 6, CSCW1, Article 116 (apr 2022), 30 pages. https://doi.org/10.1145/3512963

[4] Mary Jean Amon, Aaron Necaise, Nika Kartvelishvili, Aneka Williams, Yan Solihin, and Apu Kapadia. 2023. Modeling User Characteristics Associated with Interdependent Privacy Perceptions on Social Media. *ACM Transactions on Computer-Human Interaction* 30, 3 (2023), 1–32. https://doi.org/10.1145/3577014

[5] Andrew Besmer and Heather Richter Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1563–1572. https://doi.org/10.1145/1753326.1753560

[6] Sanchari Das, Tousif Ahmed, Apu Kapadia, and Sameer Patil. 2021. Does This Photo Make Me Look Good? How Posters, Outsiders, and Friends Evaluate Social Media Photo Posts. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 46 (apr 2021), 32 pages. https://doi.org/10.1145/3449120

[7] Shawn E. Fagan, Lauren Wade, Kurt Hugenberg, Apu Kapadia, and Bennett I. Bertenthal. 2021. Sharing Photos on Social Media: Visual Attention Affects Real-World Decision Making. In *Advances in Human Factors in Robots, Unmanned Systems and Cybersecurity*, Matteo Zallio, Carlos Raymundo Ibañez, and Jesus Hechavarria Hernandez (Eds.). Springer International Publishing, Cham, 199–206.

[8] V. U. Gongane, M. V. Munot, and A. D. Anuse. 2022. Correction to: Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining* 12 (2022), 171. https://doi.org/10.1007/s13278-022-00991-9

[9] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. 2020. Automatically Detecting Bystanders in Photos to Reduce Privacy Risks. In *2020 IEEE Symposium on Security and Privacy (SP)*. 318–335. https://doi.org/10.1109/SP40000.2020.00097

[10] Rakibul Hasan, Yifang Li, Eman Hassan, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. 2019. Can Privacy Be Satisfying? On Improving Viewer Satisfaction for Privacy-Enhanced Photos Using Aesthetic Transforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300597

[11] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. 2020. Privacy Norms and Preferences for Photos Posted Online. *ACM Trans. Comput.-Hum. Interact.* 27, 4, Article 30 (aug 2020), 27 pages. https://doi.org/10.1145/3380960

[12] Mathias Humbert, Benjamin Trubert, and Kévin Huguenin. 2019. A Survey on Interdependent Privacy. *ACM Comput. Surv.* 52, 6, Article 122 (oct 2019), 40 pages. https://doi.org/10.1145/3360498

[13] Maritza Johnson, Serge Egelman, and Steven M. Bellovin. 2012. Facebook and privacy: it's complicated. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (Washington, D.C.) *(SOUPS '12)*. Association for Computing Machinery, New York, NY, USA, Article 9, 15 pages. https://doi.org/10.1145/2335356.2335369

[14] Tao Li and Lei Lin. 2019. AnonymousNet: Natural Face De-Identification With Measurable Privacy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[15] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards A Taxonomy of Content Sensitivity and Sharing Preferences for Photos. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14. https://doi.org/10.1145/3313831.3376498

[16] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* 54, 2, Article 31 (mar 2021), 36 pages. https://doi.org/10.1145/3436755

[17] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (Berlin, Germany) *(IMC '11)*. Association for Computing Machinery, New York, NY, USA, 61–70. https://doi.org/10.1145/2068816.2068823

[18] Michelle Madejski, Maritza Lupe Johnson, and Steven Michael Bellovin. 2011. *The Failure of Online Social Network Privacy Settings.* Technical Report. Department of Computer Science, Columbia University. https://doi.org/10.7916/D8NG4ZJ1

[19] Sabid Bin Habib Pias, Imtiaz Ahmad, Taslima Akter, Apu Kapadia, and Adam J. Lee. 2022. Decaying Photos for Enhanced Privacy: User Perceptions Towards Temporal Redactions and 'Trusted' Platforms. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 437 (nov 2022), 30 pages. https://doi.org/10.1145/3555538

[20] Martin Porter and Richard Boulton. 2001. *The English (Porter2) stemming algorithm.* http://snowball.tartarus.org/algorithms/english/stemmer.html

[21] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2018. "You don't want to be the next meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, Baltimore, MD, 143–157. https://www.usenix.org/conference/soups2018/presentation/rashidi

[22] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. 2023. A Survey on Metaverse: Fundamentals, Security, and Privacy. *IEEE Communications Surveys Tutorials* 25, 1 (2023), 319–352. https://doi.org/10.1109/COMST.2022.3202047

[23] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. 2019. Privacy Perceptions and Designs of Bystanders in Smart Homes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 59 (nov 2019), 24 pages. https://doi.org/10.1145/3359161