Oops! An LLM Just Plagiarized Johnny!

Eric Wang
Carnegie Mellon University

Jack Lenga
Carnegie Mellon University

Prahaladh Chandrahasan
Carnegie Mellon University

Isabel Agagdaba
Carnegie Mellon University

Abstract

The increasing deployment of large language models (LLMs) in artificial intelligence systems raises concerns about the unauthorized use of publicly available text published online, including privacy violations, exposure of sensitive attributes, and misuse of intellectual property. In response, a range of defenses have emerged, including tools employing adversarial machine learning (AML) techniques to subtly perturb text and resist machine inference. This study explores the perceptions of authors regarding LLM inference and investigates the usability of one such tool, Bamboozle. We conducted $n = 6$ semi-structured interviews with online content creators to understand their mental models of AI inference, concerns about LLM usage, and preferences for protective technologies. Our findings reveal that while participants expressed concern about unauthorized usage of their text by LLMs, few take steps to mitigate this, due to limited awareness, lack of interest, or a desire to preserve reader experience. Usability testing of AML tool Bamboozle also uncovered important barriers to adoption and highlighted the need for defenses that are easy to use, minimally disruptive, and aligned with authors' expectations. We contribute actionable insights for developing "subversive AI" tools that support user agency and privacy, and describe avenues of future research.

1 Introduction

The proliferation of artificial intelligence (AI) technologies has raised important questions about how large language models (LLMs) access, analyze, and potentially misuse text published online without explicit creator consent. These AI systems can extract personal information [12], infer private attributes [11], and memorize copyrighted content [5], creating significant implications for privacy, security, and intellectual property. In response to these growing concerns, various defensive mechanisms have emerged, including anonymization tools [12], data poisoning techniques [9], and adversarial approaches that perturb text to resist LLM inference [4]. Our study builds upon the adversarial machine-learning technique developed by Agnew et al. [1] by investigating its usability among text content creators. Through a series of semi-structured interviews, we aim to understand content creators' mental models of AI inference impacts on privacy, their experiences with unwanted LLM inference, and their attitudes toward defensive tools.

We also seek to identify potential usability issues in our prototype defense tool and determine how best to align it with the concept of "subversive AI" – defensive techniques that remain imperceptible to humans while effectively thwarting machine learning models. This research ultimately aims to empower end users with greater control over their privacy and intellectual property in an increasingly AI-driven digital landscape.

We seek to answer three research questions addressing general perceptions about unwanted text inference, perceptions of defenses against text inference, and perceptions of Bamboozle.

- **RQ1**: When and why do people want protections against LLMs using their data?
- **RQ2**: What kinds of defenses against unwanted LLM inference do text content creators prefer?
- **RQ3**: What attitudes and perceptions do text content creators hold towards Bamboozle?

We reached out to text content creators who post content publicly online and conducted $n = 6$ semi-structured interviews, which also incorporated a usability test of Bamboozle. The interview focused on capturing concerns and behaviors regarding LLMs; expectations and desires for defensive technologies; and usability successes and failures of Bamboozle.

All of our participants indicated that they were concerned about LLMs to some extent. However, very few participants actually self-implemented mitigation strategies to defend against unwanted LLM inference. Reasons cited include a lack of knowledge, interest, or available resources. Additionally, we find that ease of use and reader experience are most important to authors; a tool that does not satisfy such conditions will not be used regardless of concern. We identified areas for improvement and future research.

2 Related Work

In this section, we summarize previous research about the impact of artificial intelligence (AI) on privacy, existing defense mechanisms, and usability metrics within this domain.

2.1 AI and Privacy

As AI becomes increasingly prevalent across disciplines, research into the intersection of AI and privacy has increased in prevalence. Kelley et al. conducted a survey across ten countries evaluating public perceptions of artificial intelligence [7]. Crucially, Kelley et al. found that concerns about privacy were the second most prevalent concern about AI. Furthermore, users of systems with AI integrations prefer lower consent levels for processing sensitive data and inference of sensitive attributes [2], indicating distrust in AI. These concerns are well-founded. Staab et al. demonstrated that pre-trained large language models (LLMs) can infer personal information, such as age and place of residence, from Reddit posts at a fraction of the time and cost required by humans [11].

Privacy is not the only concern regarding LLMs. These models are trained on vast amounts of publicly available text, often including copyrighted material. Usage of copyrighted material without consent has potential legal ramifications. One such issue with LLMs is training data memorization. Karamolegkou et al. conducted a study that analyzed verbatim memorization across six language model families on best-selling books and found that larger models have increased memorization capabilities, and the popularity of the books' content directly correlated with the language model's memorization capabilities [5]. This results in the partial or complete plagiarization of copyrighted content in the LLM's output.

2.2 Defensive Mechanisms

The growing concern for sensitive data inference and memorization and its effects on privacy highlight the need for methods to prevent unauthorized use of user data. Indeed, the respondents to the survey conducted by Kelley et al. desired "advanced protective technology" that would thwart such inference [7]. Many solutions that guard against unwanted inferences already exist. One of these solutions includes Adanonymizer, an LLM-based plug-in that filters out text for personally identifiable information [12]. Although it does not necessarily prevent LLMs from reading textual data, Adanonymizer prevents LLMs from extracting the most harmful, privacy-invasive content from online text.

Adversarial machine learning is a field of research that offers other defense mechanisms, including techniques such as data poisoning and prompt injection to modify user data. This induces AI models to perform inference incorrectly, reducing their ability to extract and infer sensitive attributes. Agnew et al. developed an adversarial machine-learning technique that transforms text to resist LLM inference [1]. Agnew et al. managed to significantly reduce LLM inference accuracy, achieving over 90% success in blocking PII extraction and copyright violations across multiple datasets and models.

Similar techniques have been more popular within the domain of inference on images. Nightshade, a tool designed to protect images from LLM analysis [9], uses a process known as "data poisoning" to inject random pieces of noise into an image, interfering with the LLM's ability to comprehend the image's content. Prior to Nightshade, Shan et al. designed Glaze which applies cloaks to an artist's original art by introducing perturbations to the image, thereby preventing generative AI models from mimicking their style [10].

Notably, the authors also evaluated the usability of Glaze by conducting user studies with artists, focusing on assessing their understanding of AI as it applies to inference on art [10]. The use of user studies highlights an important challenge seldom addressed when developing techniques to defend against LLM inference: for protective technologies to ultimately be effective, they must also be usable and understandable by even non-technical end users.

2.3 Usability of Defensive Mechanisms

For many adversarial machine learning techniques, an important usability concern involves the perturbations introduced to the original data [4]. Ideally, a system employing adversarial machine learning techniques should only introduce perturbations that are imperceptible to humans, and thus remain usable. Das defines adversarial machine learning that fulfills these requirements while still thwarting machine learning models as "subversive AI." Das emphasizes the importance of human-centered design of adversarial machine learning techniques to shift power away from surveillance institutions employing AI models on private data and towards end users. Unusable defenses ultimately do not empower end users to have any more control over their privacy.

One aspect of usability is comprehension. End-users may have trouble understanding domain-specific terminology and concepts to a level that would allow them to make an informed decision. Karegar et al. looked into using metaphors as one strategy to improve end-users' comprehension of the functionality of privacy-preserving machine learning [6]. However, when the use of metaphors was tested to improve users' understanding of differential privacy, the metaphors often induced additional misunderstandings and incorrect assumptions due to unexpected connections to prior knowledge, or "conceptual baggage." How best to guide user comprehension remains an open question.

Another aspect of usability is the level of satisfaction with any perturbations introduced. If a tool introduces extreme perturbations to the original work, the creator may be disincentivized to employ the tool. The SAIA-8 benchmark was developed by Logas to use as a benchmark for measuring the acceptability of perturbations introduced to images [8]. Adapting SAIA-8 to incorporate inference and defense on text is a promising direction to measure satisfaction with a tool intending to defend textual data.

2.4 Research Goals

In this paper, we aimed to expand upon the adversarial machine-learning technique developed by Agnew et al. [1] by investigating its usability. First, we investigated users' mental models of the impacts of AI inference on privacy and Agnew et al.'s adversarial machine learning technique. As in Logas's work on SAIA-8 [8], we also measured the acceptability of the perturbations introduced by our technique. We used this information to identify any potential usability issues and determine how best to correct them to align our technique with subversive AI.

3 Study Methods

Our study method consisted of a combination of a semi-structured interview and an interactive usability test, conducted with $n = 6$ participants. By combining the semi-structured interview and the usability test, we can effectively gauge both the a priori beliefs and conceptions regarding unwanted LLM inference, as well as gauge usability successes and flaws with respect to Bamboozle. For simplicity, the combined interview and usability test will be referred to as an interview in this paper. In this section we describe our recruitment process, interview design, and methods for data analysis, followed by the limitations of our study and ethical considerations.

3.1 Recruitment

We reached out to the admins of 20 subreddits, or pages with a thematic focus on Reddit, focused on posts containing creative writing or potentially private or identifiable information (PII). If we were given permission, we also posted our recruitment message directly on the subreddit itself. We also reached out to staff, moderators, and members of other communities involving such text content, including fanfiction sites like Archive of Our Own (AO3) and Facebook groups. In our recruitment message, we briefly describe the motivation and purpose of our research, then ask those interested in participating to fill out a short screening survey, implemented via Qualtrics. A comprehensive list of communities we reached out to can be found in Appendix A.

The screening survey asked interested individuals to provide contact information, describe their relationship with publishing text content, provide a sample of their work, and indicate their willingness to participate in an interview. We also asked for contact information of anyone that the respondent thought would be interested in our study, serving as additional snowball sampling. We then directly reached out to these participants to schedule interviews.

Our screening methods involved checking that the contact information provided was unique across all responses and checking that samples of work provided, if any, were valid links to content we determined were likely to be human written. All responses passed these checks. Out of the 14 respondents who filled out our survey, we reached out to 14 participants for interviews, of which 6 participants successfully completed their interview.

3.2 Interview

We conducted remote, semi-structured interviews with 6 participants over the course of April 2025. The semi-structured nature of our interview allowed us to delve more deeply into new perspectives, as well as improve the clarity and wording of our questions as we conducted more interviews. We video-recorded and transcribed interviews using Zoom, after obtaining explicit consent. Our final interview script is included in Appendix B. Participants were compensated with $25, either as an Amazon or a VISA gift card or through Paypal.

The interview began by discussing the purpose of our study and asking for consent. We then provided background information; asked about concerns and experiences relating to unwanted LLM inference; and asked about perceptions and preferences regarding tools defending against unwanted inference. Then, the participant was directed to interact with our tool while thinking aloud, which was followed by questions about their experience. Finally, we concluded with demographic questions.

Introduction When introducing the purpose of the study to the participants, we did not explicitly mention the existence of a tool we were seeking to evaluate, but rather suggested that we were broadly exploring defensive techniques as a whole. This was done to avoid priming participants to give answers specific to a tool before we introduced its existence. However, no deception was done, and all information was available in the consent form.

Background information To ensure that all participants had a consistent understanding of LLMs and text generation, we began with a definition of LLMs and associated terminology. Some terminology was not introduced until later sections that the terminology was relevant to.

AI Impacts This section of the interview investigated the participant's personal experiences with unwanted AI inference, as well as the concerns and attitudes they held towards

AI inference. Our goal was to gain a comprehensive understanding of text content creators' mental representations of AI inference as well as what modes of unwanted AI inference are prevalent.

Tool Unspecific This section of the interview investigated what types of defensive mechanisms participants desire, how participants expect such a defensive mechanism to function; how participants would interact with such a mechanism; and what properties the mechanism is expected to have. This is done before interaction or reference to our tool to avoid priming.

Tool Specific This section involves directly interacting with our tool to collect direct and indirect feedback on the efficacy and usability of our technique. We first start by asking participants to read the instructions and asking questions about its comprehensibility. Following clarifications, we ask participants to interact with the technique to secure a real piece of text content they have created, then test the result against Llama2, with the task chosen by the participant, but related to a concern they have about unwanted LLM inference. We then conclude this section by asking questions about their experience with Bamboozle, attitudes about the perturbations introduced, suggestions for improvement, and whether or not they would be willing to use this tool in the future.

Demographics We finish by asking about the extent to which participants are involved in creating and posting text content. This includes whether their work or field of education involves creating text content and how often they publish text content.

3.3 Data Analysis

We employed a grounded theory based coding approach. Rather than working independently and calculating inter-rater reliability, we employed consensus coding on all data. Two researchers conducted initial open coding in collaboration, resolving disagreements through discussion. This collaborative coding process allowed us to accurately identify recurring keywords and concepts across interviews. These codes were then further refined and synthesized into overarching themes in collaboration with a third researcher, again resolving disagreements through discussion. We discuss the themes we discovered further in Section 4. The full codebook can be found in Appendix C.

3.4 Ethics & Data Privacy

Our study includes the collection of responses that contain PII and information that could be linked back to individual participants. To mitigate this issue, we informed the participants about the study procedure and their rights prior to the interview through the consent form. All participants gave their agreement. We collected the minimal possible PII, emphasizing that interviewees may choose to not answer any question for any reason. If the participant gave their consent, we recorded the meetings, and as soon as transcriptions were complete and verified, recordings were deleted. All transcripts were stored in a private folder accessible only by the researchers. We removed personal identifiers, like names of individuals and other terms that are linkable to the participants' identity. Data presented in this paper uses only anonymized quotes and aggregate statistics.

In order to test the efficacy of Bamboozle, our study asked participants to use an LLM tool to scan text content produced by the participant. This introduces associated risks, including extraction of PII or plagiarism, if the LLM tool were to retain provided content for training. We inform the interviewees on the risks of LLM inference on their text, and only ask them to provide text they are comfortable providing. We also employ LLMs whose privacy policies indicate that they do not retain user inputted LLM prompts, though we acknowledge the risk that these policies may not be followed.

Our study methodology was approved by our institution's IRB.

3.5 Limitations

Our study is limited by several factors which should be taken into consideration when evaluating our results.

Self-reported data All data was self-reported on the survey. This introduces the possibility of inauthentic responses, fabricated responses, and repeat submissions. To deter such submissions, we ask for a sample of their work, which we screened for authenticity. We also check during the interview to make sure that individuals do not show up for an interview twice through visual verification of the individual. The interview itself also consists of self-reported data. We check for consistency across their responses and note that the chance of a participant intentionally providing inauthentic or fabricated responses is low.

Demographics Due to our low sample size, we are not able to collect a representative sample of authors that post text content online. Because of this, we also do not collect demographic information such as race, gender, age, and other related statistics, since we are unable to make meaningful conclusions about the effect of these distributions on participant responses. We did, however, collect the educational background information, and we note that our participants all completed a graduate degree program. This could result in poor generalizability towards the average text content creator.

4 Results

We first provide details about our study participants in Section 4.1. We then present the common themes we identified with relation to each of the three research questions in Sections 4.2, 4.3, and 4.4 respectively. Statements are attributed to participants as P1-P6.

4.1 Demographics

We did not collect socio-demographic data like race, gender identity, or age. We did collect educational background, with 3 participants having completed a master's degree and 3 participants having completed a PhD. However, we collected information about the frequency that participants posted text content online, which is aggregated and summarized in Table 1. The interviews lasted between 37 minutes and 68 minutes, with an average of 54 minutes.

4.2 Concerns and Behaviors

In this section, we present the common themes about preferences for defensive techniques against unwanted LLM inference expressed by participants.

Concerns Concerns regarding unauthorized use of creative content without consent or compensation by plagiarising or adapting text content via an LLM were the most common concerns. Four participants (P2, P3, P4, P5) expressed concern about AI using content without consent: "I'm putting an original idea on the Internet. How long before this, you know, makes its way without credit into something else." - P5. All four participants expressed additional concern about large entities using this for profit: "Most of the time, those are commercial products that bring in a lot of money to the companies that release them [...] if somebody like Openai does it, or Meta, [...] they earn a lot of money and profit off of creators that they don't give any credit to." - P4. All four participants also expressed concern about style appropriation: "I would rather a tech company not be able to just imitate me whenever they want." - P2.

Concerns regarding privacy invasions were also common. Four participants (P2, P4, P5, P6) were concerned about AI systems inferring or revealing personal information: "I think it would be really easy to [...] backtrace my personal information, such as 'Where is my home?'" - P6. One participant was further concerned about the increased risk of cyberstalking: "Certainly, if [stalkers are] able to access more things about me that I don't actively put out there, but that can be inferred with high accuracy, with something like a large language model... that is concerning for sure." Three participants (P2, P3, P5) also expressed worry about the disproportionate impact against vulnerable identity groups: "You look at who's making these tools, and you look at historically and currently, where their political alignments lie, and you can quickly see how this could be weaponized against certain groups of people [...] that it could be used maliciously by a given regime" - P3

Around half the participants were also concerned about the risk to their reputation: "If it's talking about how to create a malicious botnet, and it's referring to my article [about how a malicious botnet works], and I don't want to be associated with that." - P1. Two other concerns of interest were raised by P4, including the environmental impacts of AI systems: "it's bad for our environment"; and the loss of human connection and creativity as AI generated works obscure human generated works: "[AI is] essentially [...] isolating people and turning them into consumers rather than members of an active community."

Impact of LLMs on Posting Half the participants indicated that the perceived impacts of AI caused them to decrease their frequency of posting on social media platforms: "I would say I have stepped away from some social media platforms for multiple reasons, but one of those reasons that fed into it was the fact that I don't feel comfortable with perhaps my personal posts just kind of automatically being scraped to train a model" - P3. The other half stated that LLMs did not change their posting habits.

Mitigation Two participants (P1, P3) indicated that their method of mitigation against unwanted LLM inference involved keeping up with current literature against defenses, though did not describe implementing it. P3 further expressed frustration over the lack of available mitigation techniques that fit their needs: "we lack anything like that for text [...] I think it's something I unfortunately have to live with [...] they're not there yet [...] I think most of the things that I would like to happen are me just wishfully thinking, like, I wish there were stronger rules against this." - P3. P6 stated that they decrease the visibility of their Facebook posts to only their friends. The other three participants did not perform any form of mitigation. P4 was unaware of what defense mechanisms exist, resulting in their lack of mitigation: "I'm not sure what I would be able to do other than just telling people [...] I would find it [...] unpleasant if somebody used it on my work without consent." P2 and P5 indicated that they were not worried about LLM inference enough to implement mitigation strategies: "I think for me personally, like using stuff on my website that I authored to like describe me is fair game."- P5

4.3 Preferences for Defensive Tools

In this section, we present the common themes about preferences for defensive techniques against unwanted LLM inference expressed by participants.

Features Half of the participants expressed a desire for protections that do not interfere with human readability of the text: "If it's not visible, I don't care much. [But] if it actually changes my blog for a reader, then I would not at all use this." - P1. Furthermore, two participants (P3, P5) expressed that

	Counts	Frequency			
		>1/week	>1/month	>1/year	<1/year
Occupation/Field of Study	5	0	1	1	1
Academic Research	3	-	-	-	-
Non-academic Research	1	-	-	-	-
Articles	1	-	-	-	-
Other Work Related	1	-	-	-	-
Non-occupational/non-educational	6	2	2	2	0
Articles	1	-	-	-	-
Blog	3	-	-	-	-
Fan content	1	-	-	-	-
Social Media	3	-	-	-	-
Other	1	-	-	-	-

Table 1: **Participant Posting Habits:** We display the posting habits of the $n = 6$ participants in our study. We divide posts into posts made for work or academic purposes and other posts. We further divide each of these categories above into type subcategories. Frequencies are displayed for only the categories and not subcategories due to granularity differences in responses. Only the highest frequency is kept for each category per participant.

preserving accessibility was important to them: "I care a lot about accessibility on my website. And so if this were to impact accessibility in some way, I probably wouldn't do it." - P5.

Two participants (P4, P5) indicated that these issues could be mitigated if perturbations were visually marked for their reader: "Maybe some kind of framing here [...] like visual highlighting of this as not part of the original text [...] would help the reader, the human reader, to understand that this is not part of the text." - P4.

Half the participants also desired something that was easily integratable into their workflow: "if it's a program that you need to download before you do anything. Then that's a higher bar for usage" - P4. P5 suggested "something that you run command line in my, like, blog setup [...] that's pretty simple, and I run it every time.

Security and Privacy Guarantees Three participants (P2, P3, P4) expressed a desire for some guarantee that the protection tool does not collect or store data itself: "I would hope that this data wouldn't be [...] stored for use later on." - P2. P1 desired a stricter guarantee, that the tool would run completely offline. P6 wanted a guarantee that the protection tool was legitimate and not a malicious LLM pretending to be a defensive tool: "it should come from a [...] verified source that has been very vigorously stress tested with multiple versions and with multiple security researchers." P5 did not express a desire for security or privacy guarantees.

Alternative Solutions Here we note solutions suggested by participants that do not fall under inference-time adversarial perturbations. Half the participants expressed a preference for industry regulations preventing unwanted LLM inference over necessitating tools: "I think just stronger, maybe regulation [...] would just prevent this from being an option to companies." - P3. Two participants (P1, P4) expressed interest in a solution that could also prevent training on the data, whether through copyright or some method of preventing scraping respectively.

4.4 Perceptions about Bamboozle

Ease of Use All six participants indicated that Bamboozle was easy to use: "[using Bamboozle] was very easy. I mean, you just copy a text, and then you copy the [protected] text." - P4. However, participants often chose short excerpts or samples of text content to use in the study. Two participants (P4, P5) raised concerns about whether or not Bamboozle would still be easy to use when protecting extremely long texts: "If you have a text that has 30,000 characters, or maybe even more, how ergonomic it would be to [...] copy and then recopy it from the tool" - P4. Furthermore, two participants (P1, P4) were confused or dissatisfied with the explanations given by the tool's description. P4 believed that Bamboozle was meant to be used by the user of an LLM to altruistically protect the original text from unwanted inference, rather than used by the author to protect against users of LLMs. P1 was concerned that the perturbations introduced by the tool were not clearly marked out and needed to spend extra effort identifying the effect of Bamboozle.

Efficacy In all of our interviews, Bamboozle did not successfully defend against at least one prompt provided to Llama2. Most participants raised concerns about this failure. Additionally, two participants (P1, P5) were concerned that the tool was incompatible with certain media platforms, especially those with character limits: "I think the issue with [Blue Sky] is, it's quite long [...] I feel like I would just get 5 words in, and then... it would run out the text limit." - P5

Modifications Most participants were not satisfied with the modifications made to the text by Bamboozle. Some cited concerns about how the perturbations would affect reading flow: "It just looks like it's another sentence that contradicts the information in the text, and it's difficult to disambiguate." - P5 some cited concerns about how visible the modifications were: "that might also confuse, like the actual human readers, a bit, especially for creative texts." - P4.

UI P5 expressed dissatisfaction with the readability of the font color against the background.

5 Discussion

In the following section, we present further conclusions that can be drawn from our results, and use them to posit potential avenues of future research and recommendations for improvements to Bamboozle and other defensive technologies.

5.1 AI Privacy Paradox

A well-documented phenomenon in the space of online privacy behaviors is known as the "privacy paradox," the phenomenon where "while users claim to be very concerned about their privacy, they nevertheless undertake very little to protect their personal data." [3] This paradox appears to apply to concerns over AI inference on their text content as well. All six participants indicated some level of concern over unauthorized usage of their text content by AI, yet only one participant described an actual action undertaken to prevent this usage from occurring.

Several conclusions could be drawn from this phenomenon. First, any tool that defends against unwanted LLM inference must be as easy to use and integrate into existing workflows as possible. This is emphasized by half of the participants in our study desiring defenses to exhibit such a feature. If the tool is too unwieldy to use, authors will not utilize the tool even if concerned about AI inference.

Second, education efforts about the risks of AI and existing defensive techniques are extremely important. Our participants were fairly well-educated, yet only two participants mentioned keeping up with current developments on defensive technologies. Furthermore, our participants were all aware of some risks associated with AI inference; it can be reasonably inferred that individuals who are not aware of the risks would be even less equipped to defend themselves effectively.

5.2 Recommendations for Defensive Tools

Below, we present design considerations that should be made when creating a defensive tool, of which Bamboozle also needs to improve upon for future iterations. These changes need to be made in order to achieve the ultimate goal of usable and effective defenses against LLM inference that satisfy "subversive AI."

- **Perceptibility of Perturbations**: Nearly all participants agree that any amount of visible perturbation that interrupts the flow of text is unacceptable. Some possible solutions to investigate are the usage of whitespace, invisible text alterations, special characters, CSS styling, and other imperceptible or non-interfering perturbations. However, these solutions could have issues with compatibility with platforms. Other solutions include visual indicators (i.e. ascii-art style plaintext boxes around perturbations) or providing a notice to readers that perturbations have been introduced. What mechanisms are most acceptable by authors and readers versus how easy it is for attackers to mitigate is a potential future avenue of research.
- **Ease of Use**: Most participants wanted minimal interruption to their workflow, while maintaining maximum compatibility. A defensive tool may be required to be integrated directly into other applications, like text editors, browsers, or text-content-hosting platforms themselves. Such tools could also avoid compatibility issues by customizing the defense mechanism to the publishing platform. Investigations into these avenues may also be of interest for both Bamboozle and other works.
- **Efficacy**: The tool must work for participants to use the tool. Further refinement to the adversarial perturbations generated by Bamboozle is necessary. This refinement may also need to be an ongoing process: as LLMs become more robust to such adversarial inputs, further iteration on current techniques is necessary.

6 Reflection

6.1 Lessons Learned

One of our biggest recurring issues was our inconsistent meeting schedule with our project mentor. For a large part of the semester, our mentor would schedule meetings for a time during which some members were not available. Even when we communicated this issue to him, our old meeting date would persist. After our 3rd to 4th meeting, we had successfully changed the meeting time. However, during our last meeting, our mentor requested to change our meeting time again. This pattern, despite becoming less severe over time, has hindered some of our progress, as some members faced schedule conflicts throughout the semester. We have mitigated this somewhat by reporting meeting highlights to the rest of the group.

We also had some other conflicts of interest with our mentor. Our mentor was primarily focused on collecting data for a study over a longer period of time than this class. Often, his timeline did not completely line up with ours. We learned to be more proactive in resolving these differences, but we still ended up scrambling to schedule interviews in time, as our mentor wanted to save some interviews for later, when he had more availability. While we reached out to enough participants to meet the requirements, only six people responded.

Another recurring challenge was time management for certain assignments. Our meetings used to occur later in the week (every Friday), which often delayed our start on weekly tasks. To address this, we began holding team meetings earlier in the week, which helped us start assignments sooner and improve overall productivity. Despite this, we also had some differences in standard of work and expectations for timelines that hindered our work productivity. In the future, group expectations should be clearly delineated from the start, and resolutions of issues should occur earlier.

7 Future Directions

Some future studies are described in Section 5. In addition to those, we would like to expand our study to a more representative sample. Another study of interest would also be to hone in on the vulnerable populations described by some participants as being at more risk of being targeted by unwanted inference.

Future studies could directly compare Bamboozle to other defense techniques. Many users felt dissatisfaction with the current state of Bamboozle. This necessitates a better understanding of how user perception of Bamboozle compares with other tools on the market.

Future studies could also assess the usability of Bamboozle's obfuscation technique in combination with those from other tools. Despite our sample's lack of confidence for Bamboozle, pairing the tool with other techniques could improve user perception.

8 Conclusion

In this paper, we presented the beliefs and perceptions of (n=6) online text content creators with regards to unauthorized LLM inference and defenses against it. While these creators are concerned about risks associated with LLM inference, they do not actively take steps to defend their text. Furthermore, their expectations for defensive tools are strict, requiring minimal perceivability of perturbations and ease of use, while preserving efficacy. We conclude that efforts need to be made into resolving the privacy paradox as it applies to unauthorized LLM inference, and that the current techniques employed by Bamboozle are insufficient to match user needs. Identifying ways to improve upon existing techniques and make them more accessible to authors is an important step in achieving the goal of “subversive AI.”

Acknowledgments

We extend our gratitude to Professors Lorrie Cranor and Yuvraj Agarwal for their steadfast support, feedback, and encouragement throughout the semester. We also wish to express our appreciation to our project advisor, William Agnew, and his peers Harry Jiang and Cella Sum for helping us brainstorm interview questions, guide us through interviews, and help us reflect on our interviews. We would also like to thank Sauvik Das for giving valuable suggestions on our interview questions, as his suggestions helped us break down complex questions into simple, interview-friendly language.

References

[1] William Agnew, Harry H. Jiang, Cella Sum, Maarten Sap, and Sauvik Das. Data defenses against large language models, 2024.

[2] Sumit Asthana, Jane Im, Zhe Chen, and Nikola Banovic. "i know even if you don't tell me": Understanding users' privacy preferences regarding ai-based inferences of sensitive information for personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Susanne Barth and Menno D.T. de Jong. The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. *Telematics and Informatics*, 34(7):1038–1058, 2017.

[4] Sauvik Das. Subversive ai: Resisting automated algorithmic surveillance with human-centered adversarial machine learning. In *Resistance AI workshop at NeurIPS*, volume 4, 2020.

[5] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models, 2023.

[6] Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. Exploring user-suitable metaphors for differentially private data analyses. In *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security*, SOUPS'22, USA, 2022. USENIX Association.

[7] Patrick Gage Kelley, Celestina Cornejo, Lisa Hayes, Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff. "there will be less privacy, of course": how and why people in 10 countries expect ai will affect privacy in the future. In *Proceedings of the Nineteenth USENIX Conference on Usable Privacy and Security*, SOUPS '23, USA, 2023. USENIX Association.

[8] Jacob Logas, Poojita Garg, Rosa I. Arriaga, and Sauvik Das. The subversive ai acceptance scale (saia-8): A scale to measure user acceptance of ai-generated, privacy-enhancing image modifications. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–43, April 2024.

[9] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 807–825. IEEE, May 2024.

[10] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, Anaheim, CA, August 2023. USENIX Association.

[11] R. Staab, M. Vero, M. Balunović, and M. Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

[12] S. Zhang, X. Yi, H. Xing, L. Ye, Y. Hu, and H. Li. Adanonymizer: Interactively navigating and balancing the duality of privacy and output performance in human-llm interaction. *arXiv preprint arXiv:2410.15044*, 2024.

Appendix

A Communities we reached out to

A.1 Reddit Communities

Creative Writing

- https://www.reddit.com/r/creativewriting/
- https://www.reddit.com/r/AO3/
- https://www.reddit.com/r/FanFiction/
- https://www.reddit.com/r/creativecoding/about/
- https://www.reddit.com/r/selfpublish/about/
- https://www.reddit.com/r/writers/
- https://www.reddit.com/r/fantasywriters/
- https://www.reddit.com/r/copywriting/
- https://www.reddit.com/r/printSF/

Personal Info and Experiences

- https://www.reddit.com/r/AmItheAsshole/
- https://www.reddit.com/r/AmIOverreacting/
- https://www.reddit.com/r/TwoXChromosomes/
- https://www.reddit.com/r/mildlyinfuriating/
- https://www.reddit.com/r/antiwork/
- https://www.reddit.com/r/pettyrevenge/
- https://www.reddit.com/r/MaliciousCompliance/

Other Topics

- https://old.reddit.com/r/privacy/
- https://www.reddit.com/r/aiwars/

A.2 Journalist Sources

Sourced mostly through William, including journalists from: Buzzfeed and Wired

A.3 Other Relevant Sites

Transformative Works (owner of Archive of Our Own)

A.4 CMU Slack Channels and Discords

S3D, Cylab, CMU JSA, CMU Swim Club

B Interview Script

B.1 Introduction

Hi, thanks for agreeing to participate in this interview study. My name is [name], and I'll be your interviewer for today.

I'll start with a brief overview of why we're conducting this study. Increasingly, a type of artificial intelligence called large language models is being used on text posted online without permission. We'd like to investigate the impacts of this phenomenon on text content and their creators, as well as determine potential ways to defend against it. That's why we've invited you today, as someone who posts text content online.

We'll be asking you a series of questions regarding this topic. We want to stress that there are no right or wrong answers; we are interested in what you think. We'd like for you to be as honest as possible. Our goal is to publish our findings. We may use answers you give in anonymized quotes or aggregated statistics. All personal information that could identify you will be removed. You may also choose to not answer any questions for any reason. Do you have any questions so far?

Answer any questions they have

Would it be OK with you if we record this interview for note-taking purposes?

Answer any questions they have

We've prepared a consent form with more details. Please review it carefully before signing. Let us know if you have any questions about the content of the consent form.

Thanks again for agreeing to participate! If at any point you'd like to take a breather, or if you'd like to end the interview, please let us know. Any questions before we begin?

Alright, let's get started!

B.2 Definitions

To ensure your understanding of the questions that follow, I will define a few concepts that are relevant to our study.

Artificial Intelligence (AI) are computer systems that can perform tasks that usually require human intelligence, like recognizing speech, making decisions, or learning from data. Large Language Models (LLM) are Artificial Intelligence algorithms designed to process human language inputs and produce outputs in a process called inference. LLMs can be trained, or taught, to perform many tasks. These tasks can include, but are not limited to, summarization or question-answering. When asking an LLM to perform a task, the content given to the LLM is called a prompt.

Before we continue, do you have any questions about these topics?

B.3 Demographics

Now I have a few questions about your experience in writing content online:

1. Does your occupation or field of study involve creating text content?
    (a) If yes, tell me what you do?
2. Would you say written content online impacts your job function? In other words, do you depend on reading other people's online content for parts of your job?
    (a) If yes, tell me how?
3. How often do you publish or post text content, whether online or in print?
4. Do you publish written content outside of your job?
    (a) If yes, what do you publish?
    (b) How often do you do this?
5. How do you publish this form of content? In other words, by what media do you release your content online?

B.4 AI Impacts

Thank you for your input thus far. I would now like to ask you questions about your thoughts and opinions on AI and how you may use it.

1. Can you recall a time when you changed what you planned to post or decided not to post due to the influence of AI
    (a) If yes, can you share a particular example of content you posted, changed or didn't post because of AI? This can include your work or social life.
2. What is your opinion on text AI tools such as chatbots?
3. Do you or do you not have concerns about AI having access to and using your or others' writing posted online?
    (a) If yes, would you mind sharing them?
4. Could you provide some examples of how AI could use your or others' text? Describe what you would give to the LLM and what you expect to receive back.
5. Have you ever experienced unwanted LLM usage of text that you've written and posted online?
    (a) If yes,
        i. I'd like you to think about any one of these experiences that you would feel comfortable sharing with us. Could you describe this in more detail?
        ii. After the experience, did you take any steps to prevent LLM inference in the future?
            A. If yes, can you describe what steps you took and how you found that solution?
            B. If no, why did you NOT take any steps?
            C. Were there any steps you didn't take that you wished to take or thought about taking?
            D. If yes,what were the steps you thought about taking?
            E. If yes, what prevented you from taking these steps?
    (b) If no,
        i. Imagine that text you've written has become the target of unwanted LLM inference. What goals do you think the attacker would be able to achieve on your text specifically?
        ii. Do you take any steps to prevent LLM inference?
            A. If yes, can you describe what steps you take and how you found that solution?
            B. If no, why do you NOT take any steps?

C. Are there any steps you don't take that you wish to take or have thought about taking?
D. If yes, What were the steps you thought about taking?
E. If yes, What prevented you from taking these steps?

Now we'll discuss the purpose of our project in a little more detail. LLMs are capable of processing large amounts of data and finding patterns that are difficult for humans to identify. This capability can result in intentional or inadvertent negative effects. For instance, large language models can be used to process the text you've published online, and guess private information about you with high accuracy. Large language models can also be used to plagiarize your work or imitate your writing style.

Depending on whether or not the participant mentioned these parts:

1. How concerned are you about LLMs being used to plagiarize or imitate text content you've produced?
   (a) What are your specific concerns?
2. How concerned are you about LLMs being used to summarize or predict private information about you by analyzing text content you've produced?
   (a) What are your specific concerns?

B.5 Tool Specific

Now I would like to get your feedback on a prototype tool for protecting written contentdata from large language models. As you complete the following tasks, please speak your thoughts out loud as much as you are able to. First, please visit https://wagnew3.github.io/LLM-Data-Defenses/ and read the description of this tool. This includes the red box at the bottom of the page, as well as the about page accessible through a link near the top of the page.

1. How easy or difficult was the description of the tool to understand?
2. To what extent (do you think) does this tool impact your ability to control how LLMs may use your data

Now I would like you to test the tool out. In a moment, we'll provide you with a link to the tool. We'd like you to follow the instructions to protect any text of interest to you. Then, take the protected text and test it against the LLM Llama3. We will also provide you with the link to this LLM. Throughout this process, we would like you to share your screen if you are comfortable.

Links:

1. https://wagnew3.github.io/LLM-Data-Defenses
2. https://build.nvidia.com/meta/llama3-70b
3. Backup : https://www.llama2.space/

Questions after they use the tool

1. Could you share any first impressions you have of the tool?
2. How easy or difficult was the tool to use?
   (a) Why?
3. How easy or difficult was the tool to test?
   (a) Why?
4. How effective or ineffective was the tool during your testing?
5. How much or little did you understand how the tool modified your text?
6. How acceptable or unacceptable were the modifications the tool made to your text?
   (a) Would the modifications be more or less acceptable if they included false information?
   (b) Would the modifications be more or less acceptable if they included true information that is unrelated to the rest of the text?
7. What do you think people reading the protected text you post might think about the modifications?
8. Are there any situations where you would use this tool, or are there no situations where you would use this tool?
9. Could you see this being used in your everyday routine as a writer?
   (a) How could you imagine using this tool in your writing and/or work? How could it be integrated more seamlessly into your practice?
10. What changes would you like to see in this tool? This can be functionality, user interface, etc.

B.6 Tool Not Specific

1. If someone were to propose a tool to defend against unwanted AI inference on your text, what capabilities would you expect this tool to have?
   (a) Describe how you would interact with this tool and how you would use the result
2. What security/privacy guarantees would you like or need to use a tool to protect your text?

   (a) Individual writing sample, individual's writing, communal

3. What imperceptibility about the perturbation would like or need?

   (a) Friendly imperceptibility, hostile imperceptibility

4. How long is it acceptable for the tool to take to make a protection?

5. What control do you want to have over how AI uses your work?

6. Are there other approaches to preventing unwanted LLM inference you are interested in?

C Codebook

A link to our codebook can be found here.

D Recruitment Materials

D.1 Recruitment Post

Large Language Models have rapidly grown in use, with many companies, states, and other entities seeking to apply them. Large Language Models (LLMs) depend on human written text not only for training, but also for generating text. During text generation, human written text, such as news articles, interview transcripts, or social media posts may be input into large language models to provide facts, context, or enable them to answer questions about inputted human written text.

We, a group of researchers at Carnegie Mellon University, are conducting interviews to understand the privacy, copyright, and intellectual property concerns of people who publish or post text online (blogs, news sites, messaging apps, social media platforms, discord) regarding the proliferation and data sourcing practices of large language models. In addition, we are seeking to understand what tools people who publish or post text online may want to give them control over how LLMs use their text.

Please fill out this screening survey if you are interested in being interviewed. Interviews will last 60-90 minutes, and you will be compensated with a $25 via PayPal, Venmo, or a Visa gift card for participating. Filling out this screening form does not guarantee your participation in this interview.

Contact: William Agnew, wagnew@andrew.cmu.edu

D.2 Screener Survey

A link to our survey can be found here.

E IRB Forms

The consent form and IRB approval can be found here.