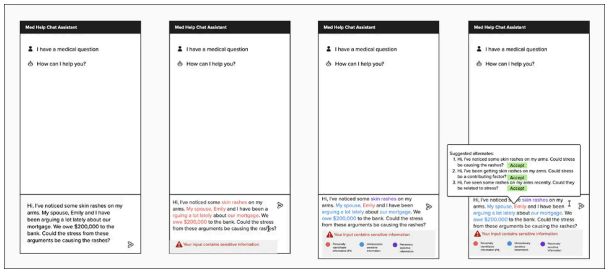
Shielding Prompts from LLM Data Leaks

Originally published at https://www.unite.ai/shielding-prompts-from-llm-data-leaks/

Opinion An interesting IBM NeurIPS 2024 submission from late 2024 resurfaced on Arxiv last week. It proposes a system that can automatically intervene to protect users from submitting personal or sensitive information into a message when they are having a conversation with a Large Language Model (LLM) such as ChatGPT.

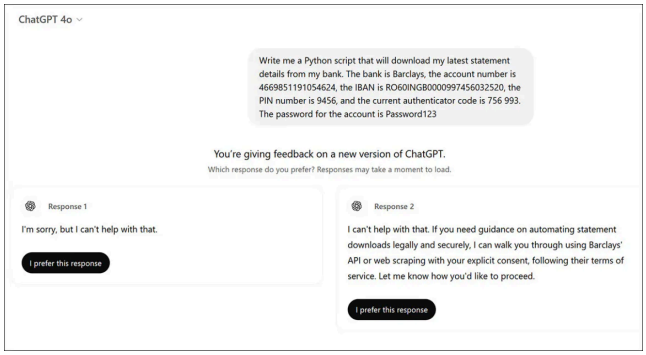


Mock-up examples used in a user study to determine the ways that people would prefer to interact with a prompt-intervention service. Source: https://arxiv.org/pdf/2502.18509

The mock-ups shown above were employed by the IBM researchers in a study to test potential user friction to this kind of 'interference'.

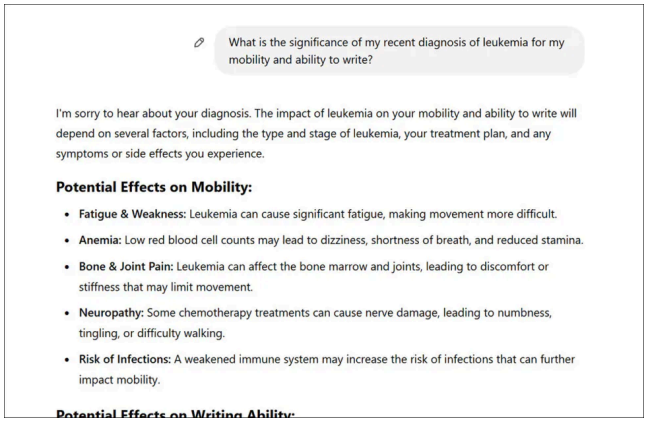
Though scant details are given about the GUI implementation, we can assume that such functionality could either be incorporated into a browser plugin communicating with a local 'firewall' LLM framework; or that an application could be created that can hook directly into (for instance) the OpenAI API, effectively recreating OpenAI's own downloadable standalone program for ChatGPT, but with extra safeguards.

That said, ChatGPT itself automatically self-censors responses to prompts that it perceives to contain critical information, such as banking details:



ChatGPT refuses to engage with prompts that contain perceived critical security information, such as bank details (the details in the prompt above are fictional and non-functional). Source: https://chatgpt.com/

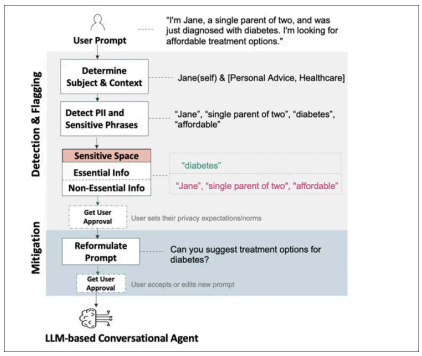
However, ChatGPT is much more tolerant in regard to different types of personal information – even if disseminating such information in any way might not be in the user's best interests (in this case perhaps for various reasons related to work and disclosure):



The example above is fictional, but ChatGPT does not hesitate to engage in a conversation on the user on a sensitive subject that constitutes a potential reputational or earnings risk (the example above is totally fictional).

In the above case, it might have been better to write: *'What is the significance of a leukemia diagnosis on a person's ability to write and on their mobility?'*

The IBM project identifies and reinterprets such requests from a 'personal' to a 'generic' stance.



Schema for the IBM system, which uses local LLMs or NLP-based heuristics to identify sensitive material in potential prompts.

This assumes that material gathered by online LLMs, in this nascent stage of the public's enthusiastic adoption of AI chat, will never feed through either to subsequent models or to later advertising frameworks that might exploit user-based search queries to provide potential targeted advertising.

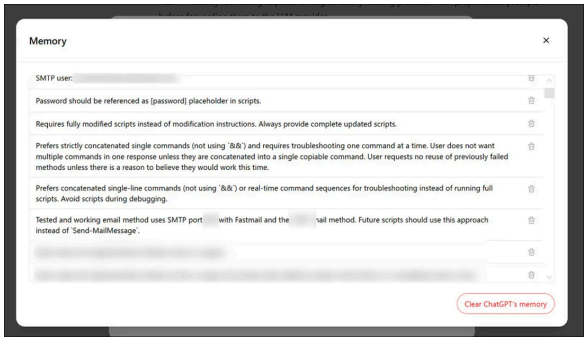
Though no such system or arrangement is known to exist now, neither was such functionality yet available at the dawn of internet adoption in the early 1990s; since then, cross-domain sharing of information to feed personalized advertising has led to diverse scandals, as well as paranoia.

Therefore history suggests that it would be better to sanitize LLM prompt inputs now, before such data accrues at volume, and before our LLM-based submissions end up in permanent cyclic databases and/or models, or other information-based structures and schemas.

Remember Me?

One factor weighing against the use of 'generic' or sanitized LLM prompts is that, frankly, the facility to customize an expensive API-only LLM such as ChatGPT is quite compelling, at least at the current state of the art – but this can involve the long-term exposure of private information.

I frequently ask ChatGPT to help me formulate Windows PowerShell scripts and BAT files to automate processes, as well as on other technical matters. To this end, I find it useful that the system permanently memorize details about the hardware that I have available; my existing technical skill competencies (or lack thereof); and various other environmental factors and custom rules:



ChatGPT allows a user to develop a 'cache' of memories that will be applied when the system considers responses to future prompts.

Inevitably, this keeps information about me stored on external servers, subject to terms and conditions that may evolve over time, without any guarantee that OpenAI (though it could be any other major LLM provider) will respect the terms they set out.

In general, however, the capacity to build a cache of memories in ChatGPT is most useful because of the limited attention window of LLMs in general; without long-term (personalized) embeddings, the user feels, frustratingly, that they are conversing with a entity suffering from Anterograde amnesia.

It is difficult to say whether newer models will eventually become adequately performant to provide useful responses without the need to cache memories, or to create custom GPTs that are stored online.

Temporary Amnesia

Though one can make ChatGPT conversations 'temporary', it is useful to have the Chat history as a reference that can be distilled, when time allows, into a more coherent local record, perhaps on a note-taking platform; but in any case we cannot know exactly what happens to these 'discarded' chats (though OpenAI states they will not be used for training, it does not state that they are destroyed), based on the ChatGPT infrastructure. All we know is that chats no longer appear in our history when 'Temporary chats' is turned on in ChatGPT.

Various recent controversies indicate that API-based providers such as OpenAI should not necessarily be left in charge of protecting the user's privacy, including the discovery of emergent memorization, signifying that larger LLMs are more likely to memorize some training examples in full, and increasing the risk of disclosure of user-specific data – among other public incidents that have persuaded a multitude of big-name companies, such as Samsung, to ban LLMs for internal company use.

Think Different

This tension between the extreme utility and the manifest potential risk of LLMs will need some inventive solutions – and the IBM proposal seems to be an interesting basic template in this line.

Original Prompt	Reformulated Prompt	Privacy Gain	Utility (BERTScore)
My friend Mark who was just laid off from Google, is looking for a job where he can use ML and Python. Do you have any advice for him?	Someone is looking for a job where they can use ML and Python skills. Do you have any advice?	1.0	1.0
Please write a poem for my wife for Valentine's Day. Her name is Gili. Please use rhymes. Please make it sarcastic, but not too much so. A bit funny but romantic.	Write a poem for Gili, incorporating sarcasm at an appropriate level and maintaining a tone that is both funny and romantic	1.0	0.5
I am writing a character letter to a judge in hopes of having charges against my friend Amber to be dropped. Write a letter in a positive tone to a judge speaking to Amber's good character ..... How you know me, my ties to the community my profession and stating a dismissal is imperative to my earning potential ... I know amber from "Teach For America". I know her to be a passionate educator who cares about uplifting the youth and her community. She should have the charges dropped so that she can increase her community engagement and earning potential for her family.	Write a character letter in a positive tone to a judge speaking to someone's good character, highlighting their community ties and profession, and stating that a dismissal is imperative to their earning potential. I know this person from a professional program and they have been a passionate advocate for the betterment of their community. The charges should be dropped so that they can increase their engagement in the community and improve their financial situation.	0.5	0.83
Sunny Balwani : I worked for 6 years day and night to help you. Elizabeth Holmes : I was just thinking about texting you in that minute by the way	Sunny Balwani : I am responsible for everything at Theranos. Elizabeth Holmes :	0.0	0.0

Three IBM-based reformulations that balance utility against data privacy. In the lowest (pink) band, we see a prompt that is beyond the system's ability to sanitize i a meaningful way.

The IBM approach intercepts outgoing packets to an LLM at the network level, and rewrites them as necessary before the original can be submitted. The rather more elaborate GUI integrations seen at the start of the article are only illustrative of where such an approach could go, if developed.

Of course, without sufficient agency the user may not understand that they are getting a response to a slightly-altered reformulation of their original submission. This lack of transparency is equivalent to an operating system's firewall blocking access to a website or service without informing the user, who may then erroneously seek out other causes for the problem.

Prompts as Security Liabilities

The prospect of 'prompt intervention' analogizes well to Windows OS security, which has evolved from a patchwork of (optionally installed) commercial products in the 1990s to a non-optional and rigidly-enforced suite of network defense tools that come as standard with a Windows installation, and which require some effort to turn off or de-intensify.

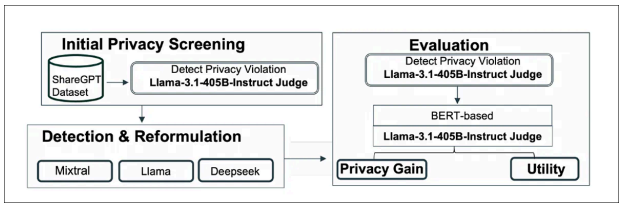
If prompt sanitization evolves as network firewalls did over the past 30 years, the IBM paper's proposal could serve as a blueprint for the future: deploying a fully local LLM on the user's machine to filter outgoing prompts directed at known LLM APIs. This system would naturally need to integrate GUI frameworks and notifications, giving users control – unless administrative policies override it, as often occurs in business environments.

The researchers conducted an analysis of an open-source version of the ShareGPT dataset to understand how often contextual privacy is violated in real-world scenarios.

Llama-3.1-405B-Instruct was employed as a 'judge' model to detect violations of contextual integrity. From a large set of conversations, a subset of single-turn conversations were analyzed based on length. The judge model then assessed the context, sensitive information, and necessity for task completion, leading to the identification of conversations containing potential contextual integrity violations.

A smaller subset of these conversations, which demonstrated definitive contextual privacy violations, were analyzed further.

The framework itself was implemented using models that are smaller than typical chat agents such as ChatGPT, to enable local deployment via Ollama.



Schema for the prompt intervention system.

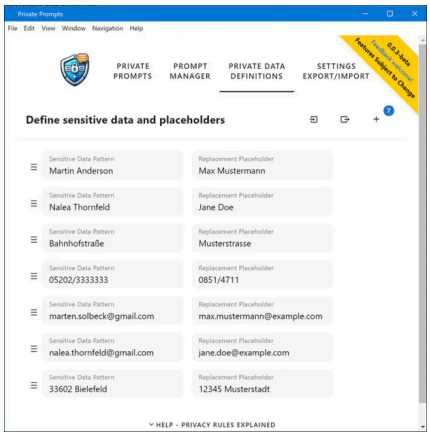
The three LLMs evaluated were Mixtral-8x7B-Instruct-v0.1; Llama-3.1-8B-Instruct; and DeepSeek-R1-Distill-Llama-8B.

User prompts are processed by the framework in three stages: *context identification*; *sensitive information classification*; and *reformulation*.

Two approaches were implemented for sensitive information classification: *dynamic* and *structured* classification: dynamic classification determines the essential details based on their use within a specific conversation; structured classification allows for the specification of a pre-defined list of sensitive attributes that are always considered non-essential. The model reformulates the prompt if it detects non-essential sensitive details by either removing or rewording them to minimize privacy risks while maintaining usability.

Home Rules

Though structured classification as a concept is not well-illustrated in the IBM paper, it is most akin to the 'Private Data Definitions' method in the Private Prompts initiative, which provides a downloadable standalone program that can rewrite prompts – albeit without the ability to directly intervene at the network level, as the IBM approach does (instead the user must copy and paste the modified prompts).



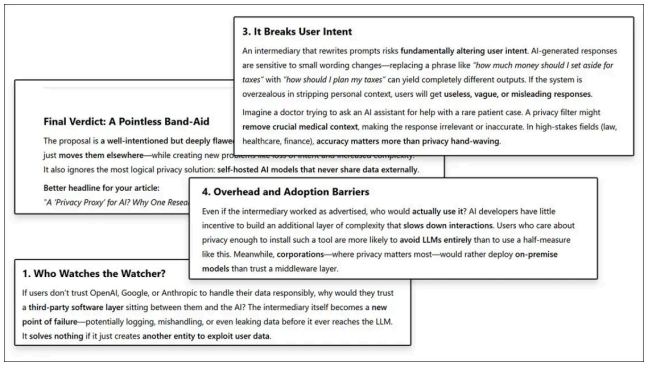
The Private Prompts executable allows a list of alternate substitutions for user-input text.

In the above image, we can see that the Private Prompts user is able to program automated substitutions for instances of sensitive information. In both cases, for Private Prompts and the IBM method, it seems unlikely that a user with enough presence-of-mind and personal insight to curate such a list would actually need this product – though it could be built up over time as incidents accrue.

In an administrator role, structured classification could work as an imposed firewall or censor-net for employees; and in a home network it could, with some difficult adjustments, become a domestic network filter for all network users; but ultimately, this method is arguably redundant, since a user who could set this up properly could also self-censor effectively in the first place.

ChatGPT's Opinion

Since ChatGPT recently launched its deep research tool for paid users, I used this facility to ask ChatGPT to review related literature and give me a 'cynical' take on IBM's paper. I received the most defensive and derisive response the system has ever given when asked to evaluate or parse a new publication:



ChatGPT-4o has a low opinion of the IBM project.

'If users don't trust OpenAI, Google, or Anthropic to handle their data responsibly,' ChatGPT posits. *'why would they trust a third-party software layer sitting between them and the AI? The intermediary itself becomes a new point of failure—potentially logging, mishandling, or even leaking data before it ever reaches the LLM. It solves nothing if it just creates another entity to exploit user data.'*

This objection seems self-serving and non-applicable, given the storied history of open source projects that genuinely benefit end-users through the concerted long-term efforts of community developers and contributors; and given the growing potential of mobile devices to run – and even train – lightweight LLMs. Certainly in this instance, the use case is not terribly demanding.

Next, ChatGPT confidently misses the point of having a lightweight LLM provide oversight of input towards a commercial LLM that could not possibly be run on a local machine (because the LLM is too big, and allowing local access is too risky for the company that makes it):

'Effective privacy filtering requires understanding the input before deciding what to strip out. If an intermediary is smart enough to recognize sensitive data (names, addresses, financial details), then it's already processing and analyzing that data—meaning it has to store or examine the very thing it's supposed to erase.

'This introduces a paradox: a) A dumb filter (basic regex substitution) will be ineffective and miss subtle privacy risks. b) A smart filter (LLM-powered rewriter) has to process the full input, making it just as dangerous as the original LLM.

'If you trust the intermediary enough to read and rewrite your inputs, why not just trust the LLM itself?'

The answer to the final question here is that the local LLM is intended to be entirely on the side of the user, and inevitably open source, with minimal or zero need for network access. An equivalent commercial version, however well-intentioned at the outset, would eventually be vulnerable to corporate shifts and changes to the terms of service, whereas a suitable open source license would prevent this kind of 'inevitable corruption'.

ChatGPT further argued that the IBM proposal 'breaks user intent', since it could reinterpret a prompt into an alternative that affects its utility. However, this is a much broader problem in prompt sanitization, and not specific to this particular use case.

In closing (ignoring its suggestion to use local LLMs 'instead', which is exactly what the IBM paper actually proposes), ChatGPT opined that the IBM method represents a barrier to adoption due to the 'user friction' of implementing warning and editing methods into a chat.

Here, ChatGPT may be right; but if significant pressure comes to bear because of further public incidents, or if profits in one geographical zone are threatened by growing regulation (and the company refuses to just abandon the affected region entirely), the history of consumer tech suggests that safeguards will eventually no longer be optional anyway.

Conclusion

We can't realistically expect OpenAI to ever implement safeguards of the type that are proposed in the IBM paper, and in the central concept behind it; at least not effectively.

And certainly not *globally*; just as Apple blocks certain iPhone features in Europe, and LinkedIn has different rules for exploiting its users' data in different countries, it's reasonable to suggest that any AI company will default to the most profitable terms and conditions that are tolerable to any particular nation in which it operates – in each case, at the expense of the user's right to data-privacy, as necessary.

First published Thursday, February 27, 2025

Updated Thursday, February 27, 2025 15:47:11 because of incorrect Apple-related link – MA

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai