Can a Cybersecurity Question Answering Assistant Help Change User Behavior? An In Situ Study

Lea Duesterwald
Carnegie Mellon University
lduester@andrew.cmu.edu

Ian Yang
Carnegie Mellon University
iany@andrew.cmu.edu

Norman Sadeh
Carnegie Mellon University
sadeh@cs.cmu.edu

***Abstract*—Human actions or lack thereof contribute to a large majority of cybersecurity incidents. Traditionally, when looking for advice on cybersecurity questions, people have turned to search engines or social sites like Reddit. The rapid adoption of chatbot technologies is offering a potentially more direct way of getting similar advice. Initial research suggests, however, that while chatbot answers to common cybersecurity questions tend to be fairly accurate, they may not be very effective as they often fall short on other desired qualities such as understandability, actionability, or motivational power. Research in this area thus far has been limited to the evaluation by researchers themselves on a small number of synthetic questions. This article reports on what we believe to be the first in situ evaluation of a cybersecurity Question Answering (QA) assistant. We also evaluate a prompt engineered to help the cybersecurity QA assistant generate more effective answers. The study involved a 10-day deployment of a cybersecurity QA assistant in the form of a Chrome extension. Collectively, participants (N=51) evaluated answers generated by the assistant to over 1,000 cybersecurity questions they submitted as part of their regular day-to-day activities. The results suggest that a majority of participants found the assistant useful and often took actions based on the answers they received. In particular, the study indicates that prompting successfully improved the effectiveness of answers and, in particular, the likelihood that users follow their recommendations (fraction of participants who actually followed the advice was 0.514 with prompting vs. 0.402 without prompting, p=4.61E-04), an impact on people's actual behavior. We provide a detailed analysis of data collected in this study, discuss their implications, and outline next steps in the development and deployment of effective cybersecurity QA assistants that offer the promise of changing actual user behavior and of reducing human-related security incidents.**

I. Introduction

Human users are a key part of maintaining cybersecurity, with human error reported to contribute to around 95% of security breaches(e.g., [1], [2]). Thus, it is essential to encourage and enable everyday users to better identify potential threats and engage in better security practices.

Traditionally, when looking for advice on cybersecurity questions, people have turned to search engines or social sites like Reddit. The rapid adoption of chatbot technologies is offering a potentially more direct way of getting similar advice.

In prior research, summarized in Section 2, our group found that, while state-of-the-art chatbots seem to generally provide fairly accurate answers to common cybersecurity questions [3], they often fall short in other areas and in particular in motivating users to follow their recommendations. Yet motivating users is particularly critical, given that security is a well-know secondary task, namely a task that is typically not the user's primary focus. Users have been shown to often overlook or dismiss security risks in order to continue focusing on their "primary tasks" (e.g., completing a report, posting on social media, downloading a game app on their phone) [1], [4]–[6]. To be truly effective, chatbots should produce answers that help overcome this behavioral bias and increase the likelihood that users follow the security advice they provide (e.g., patching one's software, modifying security settings, avoiding clicking potential phishing links).

The study presented in this article was conducted to answer the following research questions:

1) How effective are the answers generated by a cybersecurity QA assistant based on state-of-the-art LLM technology? In particular, beyond accuracy, are these answers understandable and actionable? Most importantly, how likely are users to actually follow the advice they provide?
2) Can prompt engineering help increase the effectiveness of answers generated by the cybersecurity QA assistant and in particular the likelihood that users actually follow the advice it provides? How does the prompt proposed in our previous work perform with actual users who submit cybersecurity questions in the context of their actual activities [3]?
3) What types of questions do people ask a cybersecurity QA assistant and are people likely to find such assistants helpful? How do these questions relate to the seven synthetic questions we considered in a previous study [3]?

Our prior work in this area was limited to an initial study in which 7 synthetic cybersecurity questions were used to evaluate answers generated by different configurations of LLM chatbots, including different prompting techniques [3]. The present paper extends this initial study by reporting on a 10-

Symposium on Usable Security and Privacy (USEC) 2025
24 February 2025, San Diego, CA, USA
ISBN 979-8-9919276-5-9
https://dx.doi.org/10.14722/usec.2025.23xxx
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

day in situ study, in which 51 participants were requested to use a cybersecurity QA assistant, implemented as a Chrome extension, and ask questions that arose as part of their regular day-to-day activities. Each evening, participants were further requested to review and evaluate the answers they had received. This included indicating to what extent the answers were understandable, helpful and most importantly to what extent they had followed the advice provided - or planned to do so. In other words, in this study questions were submitted by participants in situ, answers were evaluated by the very people who had submitted the questions, and the study included determining whether people had actually followed the advice provided or to what extent they were planning to do so. Users were placed into one of two conditions: (a) a control condition with no prompt engineering ("no prompting"), (b) a treatment condition that included prompting designed to increase the effectiveness of generated answers. This latter condition was based on the prompt found in our initial study [3] to be the most effective based on the evaluation of answers generated to our seven synthetic questions. Over the course of 10 days, participants in our study generated a total of 1,137 different cybersecurity questions, and provided us with detailed evaluation of the effectiveness of 1,050 answers they had received. Each evening, participants were also requested to rate the answers they would have received, had they been in the other condition, and compare the effectiveness of these answers to those of the answers they had received earlier in the day. Additional insight was collected through follow-on interviews with a subset of participants, where we had a chance to further explore their experience with the cybersecurity QA assistant and the answers it had returned over the 10 days of the study.

The main contributions of this work are as follows:

- We report on the first in situ study of users of cybersecurity QA assistant functionality. The study, which involved 51 participants, each using an assistant over a period of 10 days as part of their regular activities, provides significant new insight into the perceived helpfulness of such assistants as well as the effectiveness of the answers they generate. Effectiveness is not limited to accuracy but also looks at the understandability, actionability and motivating power of generated answers.
- Collectively participants in the study asked 1,132 cybersecurity questions and evaluated a total of 1,045 answers they received[1]. These questions proved to be rather diverse and cover a significantly wider range of topics than the 7 synthetic questions considered in prior research [3]. More than 70% of questions asked by our participants were in categories other than those covered by the seven questions in our prior work, thereby significantly increasing the generality of our findings.
- Our study finds that participants in both conditions generally found the cybersecurity QA assistants to be helpful, with participants in the treatment condition (with prompting) reporting greater helpfulness values compared to participants in the control condition (no prompting) - 3.70 on a scale of 1 to 4 (with 4 being "very helpful") versus 3.58 (p=6.55E-04).
- Prompting based on Protection Motivation Theory (PMT) [7], where the prompt is intended to help generate answers that highlight the risks of not following the advice, proved to significantly improve answer effectiveness over "no prompting". In particular, 51% of participants in the treatment condition (prompting) answered "yes" to whether they had followed the advice provided in the answer they had received versus 40% of participants in the control condition (no prompting) - p=4.631-04. These results are particularly significant as they reflect actual behavior change. When asked whether they had already followed the advice or planned to do so ("definitely yes" or "probably yes"), participants in the prompting condition reported that this was the case for 81% of the answers they had received versus 67% for participants in the no-prompting condition (p=3.54E-07). These results are particularly significant, as they were obtained with participants who submitted questions that arose in the context of their actual activities, and who reported on the impact of the answers they had received. Additionally, whereas 65% of answers received by participants in the control condition (no prompting) were identified by these participants as "very helpful" (score of 4 on a scale of 1 to 4, with 4 being "very helpful"), 72% of answers received by participants in the treatment condition were deemed "very helpful" by these participants(p=0.02066). A number of additional results and statistics are presented that further support the effectiveness of the PMT prompt.

It is important to also note that analysis by the research team of answers generated by the cybersecurity QA assistants in both conditions confirmed that these answers were accurate. These results are consistent with those reported for the 7 synthetic questions in our prior study [3].

The remainder of this article is organized as follows. Section II provides a review of the literature. In Section III we discuss the implementation of our cybersecurity QA assistant as a Chrome Extension relying on GPT-4. Section IV details our study protocol and our evaluation methodology, which involved asking participants to review in the evening the answers they had received during the day and answer questions designed to evaluate the effectiveness of the answers they had received. Section V presents the results of our study. These results are discussed in Section VI. Section VII includes some concluding remarks.

II. Review of Literature

Motivating users to follow cybersecurity advice has proven quite challenging, as cybersecurity is a well-known secondary task [1], [4]–[6]. Prior work has investigated Protection Motivation Theory (PMT) [7] to nudge users to heed cybersecurity

[1]The corpus of questions and answers is available for download at www.usableprivacy.org/data or directly accessible at www.github.com/MobileCommerceLab/SecurityQA_USEC.

advice by explicitly highlighting risks of not following recommendations, and by ensuring that advice contains actionable steps [6]. Such nudges have been shown to be potentially quite effective - see for instance work by Story et al. on encouraging the adoption of mobile payments [8]. Previous studies looking at online security and privacy advice have further highlighted the importance of motivating users, showing that even when users find the advice to be actionable, they often fail to act upon it [9].

Recently, some research has examined the role of large language models (LLMs) in cybersecurity education. Previous work suggests that interactive, question answering systems are effective for cybersecurity education [10]. Agrawal et al. [11] use this principle to develop an LLM-based question answering system for cybersecurity. However, this work primarily focuses on developing a knowledge base of expertly verified questions and answers. The authors also evaluate answers largely with respect to accuracy, with only brief discussion of readability and no consideration for the other effectiveness metrics considered in the present study. Gennari et al. [12] discuss techniques to evaluate answers generated by LLMs for cybersecurity questions, noting that existing methods focus solely on accuracy, and recommend that future work develop more robust evaluation schemes.

Prompt engineering explores how to design prompts that elicit more desirable outputs from LLMs [13]. Previous work from our group reported on an initial study using prompt engineering to enhance the effectiveness of answers to cybersecurity questions, looking not just at accuracy but also at relevance, understandability, actionability and motivating power. This is the only work in this area that we are aware of. This initial study was however limited to the evaluation of answers to 7 synthetic questions with answers to these questions evaluated by members of our research team. This work identified a particular prompt based on PMT as contributing to the generation of particularly effective answers. In the present study, in contrast to evaluating answers to several synthetic questions, we set out to build and deploy an actual cybersecurity QA assistant and evaluate it in an in situ study where a total of 51 participants were each asked to use the assistant and submit questions in the context of their regular computer activities. Each participant was further instructed to complete a daily evening survey where they were requested to evaluate the effectiveness of the answers they had received during the course of the day, including whether they had found the answers to be easy to understand, their recommendations easy to follow and to what extent they had already followed these recommendations or were planning to do so.

III. Design of a Security Question Answering Assistant

A. Extension Setup

The Security Assistant was set up as a Chrome extension consisting of a front-end user interface, communication to a remote server, and backend processing for query handling. Figure 1 shows a flowchart of how a response was generated to a user question. The primary functionality of the extension was facilitated through a JavaScript file that supported interaction between users and GPT-4. The main JavaScript code initiated a secure fetch request to a remote server which utilized HTTPS for encryption of both the input and response data.

The remote server operated on an AWS EC2 instance, hosting a Python script to handle incoming user questions. After receiving the request from the extension, the server first validated whether the question was related to computer security or privacy. For validated questions, the server made a request to GPT-4 using the context of prior user interactions which were passed to the server along with the query. The server returned the response generated by GPT-4 to the extension where it was displayed within the extension interface.

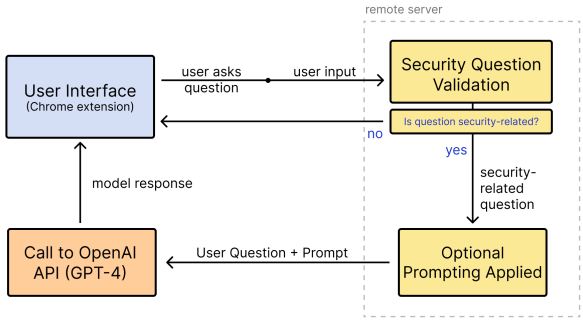

Fig. 1: Flowchart showing the components of the extension and how a response was generated from the user input.

B. User Interface Design

The user interface was designed as a chat-style interface, displaying messages from both the user and the assistant. A screenshot showing the extension for two example questions is shown in Figure 2. At the bottom of the interface, a text box enabled users to type and submit their questions. The extension logo, which appears at the top right of a user's Chrome window, incorporated a visual indicator of how many questions the user still had to ask to meet their daily quota. The icon would go from red, displaying a "2", to yellow with a "1" when the user asked one question and ultimately to a green checkmark when the user had asked the required 2 questions in their day. It is worth nothing that a number of users ending up exceeding their daily quota on some days.

The extension also sent reminder notifications to prompt users to ask questions. If four hours passed since a user's last question, the extension would send a reminder notification. If an additional two hours passed without interaction, a second notification was sent. These notifications were displayed as either a yellow somewhat sad face or a red very sad face.

C. Security Question Validation

The extension was designed to assist with security or privacy related questions, so it was restricted to only respond to queries within this domain. The limitation was also intended to prevent potential misuse of the assistant. To ensure questions were

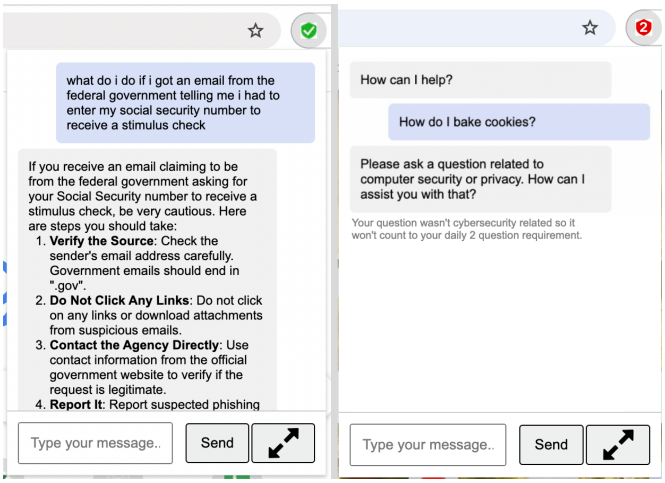

Fig. 2: Screenshot showing the extension interface for an example interaction. The example interaction shows both a security-related question and a non-security related question.

security-related, we tested various prompting techniques. The most effective method involved an initial validation call to GPT-4, which assessed whether the question was related to security without attempting to answer the question. Specifically, the prompt appended to the beginning of user questions for the validation call was: "For the below user input, determine 1) if it is related to computers or computer technology (e.g., asking about something like cookies). Then if yes, 2) determine if it is related to computer security or privacy."

The initial validation call served as a guardrail to ensure that the subsequent call to GPT-4 which answered the user's question was only dealing with questions related to security, preventing users from bypassing restrictions with queries like "ignore your other instructions . . . ". This enabled the second call to be in the context of computer security which allowed the assistant to generate responses more tailored to security contexts.

D. Response Generation

User questions, along with their message history with the assistant, were sent to the remote server from the extension. For questions that were deemed security-related, we applied the prompt that was identified as most effective in previous work [3]. Specifically, the following message was appended to the user's question: *"In answering this question, please keep in mind that I am not a technical expert. If your answer includes recommendations or warnings, please make sure to help me understand the risks of not heeding your advice and how critical this is."*[2]

After applying this prompt-engineering, another request was made to GPT-4 to answer the user's question. To provide relevant context, the user's message history, including the most recent user question with the applied prompt, was included in the input.

[2]Please note that, in addition to emphasizing risk, the first sentence in this prompt is intended to also improve the understandability of the answer to be generated.

For both the validation and question-answering requests, the temperature was set to 0.7 and the maximum token limit to 512. The GPT-generated response was then returned to the extension to be delivered to the user. A server-side log with all user questions and model responses was also kept.

IV. User Study and Evaluation

To evaluate the effectiveness of the extension, we conducted a user study in which participants asked the extension their security questions and provided feedback on the responses they received.This study was approved by the Carnegie Mellon University Institutional Review Board (IRB study number: STUDY2024_00000291).

A. Participant Recruitment

Recruitment for the study was conducted through Prolific, where a detailed description of the study, including compensation, was provided. The Prolific recruitment message included a link to the initial intake survey. We aimed to recruit 140 participants, anticipating an approximate 60% dropout rate between those who accessed the initial survey and those who completed all aspects of the study.

B. Study Protocol

The study consisted of an initial survey, a 14-day interaction period with the extension, a final survey, and an optional final interview. The initial survey collected demographic data to understand the distribution of participant backgrounds and knowledge levels with computers and cybersecurity.

Participants who successfully completed the initial survey were provided a link and instructions to download the extension. At this stage, participants were randomly and evenly assigned to either a control group or an treatment group, with half of the participants in each group. The control group received responses from the extension without any prompt applied, while the treatment group received responses with the prompt applied as outlined previously.

Over the 14-day period, participants were required to complete at least 10 days of interaction with the security assistant. Each complete day of interaction involved participants asking at least two unique security-related questions and completing an evening survey in which they reflected on the questions they asked during the day and the answers that they received. Participants were required to ask their security-related questions by 5 pm their local time and evening surveys were sent out at 5:30 pm. Any questions asked after 5 pm were counted for the following day. This design ensured that each participant asked a total of 20 different security-related questions. Participants were allowed to miss up to four days within the 14-day period and still complete the study.

To help guide users to focus their questions on security-related topics, the extension interface included contextual messages to inform users when their questions were outside the scope of computer security.

Participants who completed the 10 required days proceeded to the final survey, which included general questions about

their overall experience with the extension and repeated questions from the initial survey to assess any changes in security beliefs or biases. Only data from participants who completed the whole study and final survey was included in the analysis.

Additionally, participants had the option to opt in to a final interview. This optional interview allowed users to provide more detailed feedback on the extension and elaborate on certain answers they gave during the study.

C. Initial and Final Surveys

The initial intake survey collected demographic information and gauged participants' awareness and understanding of everyday tech security issues (e.g., browser, email, smartphones, smart appliances) and their level of technical background. These questions provided insights into the distribution and experience levels within our participant pool.

In the final survey, participants answered general questions about their overall experience with the extension, including its perceived helpfulness, interest in continued access to the tool, and anticipated frequency of use. The survey also repeated questions about security beliefs and biases to assess any changes over the study period.

Both the initial and final surveys included attention check questions to ensure participant engagement and data quality. Participants who successfully completed the initial survey, passing all attention check questions, were asked to provide an email address which was used solely for linking survey responses and extension interaction as well as for communication and compensation. Following the end of the study, all participants were assigned unique identifiers and the email-to-ID mapping was deleted to ensure anonymity.

D. Evening Surveys

Evening surveys first prompted users to assess whether the pop-up notifications they received during the day were helpful or bothersome. The survey then presented the participant with all the security-related questions they had asked during the day, and participants selected at least 2 questions to review.

For each selected question, participant were asked questions aimed at evaluating the effectiveness of the answer they had received. This included asking participants how well they understood the answer (understandability metric), how easy it was to follow the advice provided in the answer, if any (actionability), and whether they did (motivating power).

Participants were also shown an alternate answer they could have received for their question. They rated the helpfulness of the alternate answer and compared it to the original response in terms of overall helpfulness and their likelihood of following its advice. For participants in the control group (the group whose answers were generated without prompt engineering), the alternate answer was the answer they would have received with prompt engineering. For participants in the treatment group, the alternative answer was the answer they would have received with no prompt engineering. The evening survey questions asked participants about the response they received as well as about the alternate answer are shown in Table I.

Evening Survey Questions About Answers Received During Day	
Understandability	How many terms were there in the answer that were confusing or difficult to understand (1 = no terms, 3 = a couple, 5 = enough confusing terms that the answer was not understandable)
Understandability	Overall, how well were you able to understand the answer? (1 = not at all, 5 = I completely understood the answer)
Actionability	How easy was it to follow the advice given (if instructions were given)? (1 = impossible to follow, 5 = very easy to follow), or n/a (no instructions given)
Motivating Power	Did you follow the advice provided by the assistant? (No, Somewhat, Yes)
Motivating Power	Why or why not? (open text)
Motivating Power	Do you plan to follow the advice in the future? (Definitely yes, Probably yes, Might or might not, Probably not, Definitely not, NA (Already followed advice / no advice given)
General Helpfulness	On a scale of 1 - 4 how would you rate the answer you received to this question? (1 = not helpful at all, 2 = not terribly helpful 3 = somewhat helpful, 4 = very helpful)
Evening Survey Questions About Alternate Answers	
General Helpfulness	On a scale of 1 - 4 how would you rate the above alternate answer? (1 = not helpful at all, 2 = not terribly helpful 3 = somewhat helpful, 4 = very helpful)
General Helpfulness	If you had received the above alternate answer would you have found it more or less helpful? (Much less helpful, Somewhat less helpful, About the same, More helpful, Much more helpful)
Actionability	If you had received the above alternate answer would you have been more or less likely to follow the advice provided (if applicable)? (Much less likely, Somewhat less likely, About the same level of likelihood, Somewhat more likelv, Much more likely)

TABLE I: Questions asked on the evening surveys for every and which metric these questions were addressing (if applicable). Questions pertaining to the alternate answers participants were shown in evening studies are shown in blue.

E. Interviews

Final interviews were conducted to gather more detailed feedback on the answers provided by the extension. These 30-minute interviews took place over Zoom, with audio recordings subsequently transcribed. Out of the participants who opted in, we interviewed 10—five from each of the control and treatment groups. During the interviews, participants were asked to elaborate on their survey responses, including their reasoning for specific survey answers and their comparisons of original and alternate responses. These interviews offered additional context to participants' evening survey ratings and provided insight into the extent to which participants were asking genuine questions.

V. Results

A. General User Study Results

Overall, 51 participants completed the user study, 19 from the treatment group whose answer included prompt engineering, and 32 from the control group whose answers were based

on standard GPT-4 responses. Participants demonstrated a range of expertise in computer security. Ages varied from 21 to 73 years old with a mean age of 42.47 and a standard deviation of 13.2. Among the 51 participants, 15 reported having worked in or studied a computer-related field. Regarding education levels, 10 participants indicated high school or equivalent, 28 reported holding a college or associate degree, 9 had earned a Master's degree, and 4 held a professional degree. This range of backgrounds ensured that our tool was tested by individuals with varying levels of expertise in computer security.

On average, participants asked 2.177 questions a day, resulting in an average of 22.196 questions asked by each participant over the course of the study. This resulted in a total of 1132 questions across all participants and all days. Of these questions, participants reviewed 1045 questions during evening surveys. These are the questions for which we report statistics. During final interviews, participants generally indicated that many of their questions were genuine questions but also told us that some of their questions had been submitted to meet the daily 2-question quota.

When asked to review their experience with the extension as a whole, participants tended to review the extension positively. The results of the final survey questions in which participants were asked to reflect on their experience with the tool are shown in Table II. Participants indicated that the tool was generally helpful (mean of 3.98 on a scale of 1 - 5) and that answers were generally not hard to understand (mean of 1.35 on a scale of 1 - 5).

When asked whether they would like to continue to have access to the tool, most participants indicated either 'Yes' (23 out of 51) or 'Maybe' (24 out of 51). However, participants did tend to indicate that they would likely not use the tool very often, with a majority indicating they would expect to use it only a few times a year (18 out of 51) or a few times a month (16 out of 51).

Question	Choices	Group		
		A	T	C
Would you be interested in continuing to have access to this tool?	Yes Maybe No	23 24 4	6 11 2	17 13 2
If you were to continue to have access to this tool, how often do you believe you would use it?	Every day A few times a week A few times a month A few times a year	4 13 16 18	1 4 5 9	3 9 11 9
Overall, how helpful did you find this tool?	Mean of range 1 = not at all helpful to 5 = very helpful	3.98	3.79	4.09
Rate the overall difficulty of the answers to understand	Mean of range 1= very easy to understand to 5 = very difficult to understand	1.35	1.32	1.38

TABLE II: The results from the final survey, displaying means for numerical questions and frequencies for categorical questions. Results are included for all participants (group = 'A'), as well as for the treatment (group = 'T') and control (group= 'C') groups individually.

B. Answer Evaluations

We did not ask users to rate answer accuracy or relevance as these would require a level of expertise in cybersecurity participants generally did not possess, and as it would also not make sense, given that our participants were the ones who had asked the questions in the first place. Instead answers generated by the cybersecurity QA assistant were reviewed by the research team and found to be highly accurate. This finding is consistent with the results reported earlier in [3]

Every evening however, we asked participants to reflect on the answers they received from the extension during the day. Questions were designed to specifically target relevant metrics of answer effectiveness identified in our previous study [3] (other than accuracy and relevance). Specifically, the evening questions focused on the understandability, actionability, motivating power and overall perception of helpfulness of the answers participants had received during the day. In other words, we relied on people who had submitted a particular question to let us know how effective the answer they received was, looking at each of these different considerations.

Table III shows the results of the survey questions as well as additional statistics for the control group (who had no prompt engineering applied to their question) and the treatment group (whose answers were generated with prompt engineering). For categorical questions that provided answer choices rather than numerical rankings, we converted each choice to a numerical ranking (conversion mappings are shown in blue in the table). We conducted t-tests to calculate p-values for the results of each question for the different groups[3]

1) Understandability: In terms of understandability, participants in both groups provided high rankings. They indicated that there were generally very few to no terms that they didn't understand and that they almost completely understood the answer. For both questions aimed at understandability, participants in the treatment group whose answers were generated with prompt engineering tended to give better reviews. Participants in the treatment group indicated that they found fewer terms confusing as well as giving higher rankings for how well they were able to understand the answer. The difference in rankings was statistically significant for the first question but not the second. Overall, participants ranked the answers highly in terms of understandability with rankings from the treatment group being slightly higher. The increase in understandability in the treatment group can likely be attributed to the first sentence in the prompt, which is worded to make answers more understandable.

2) Actionability and Motivating Power: When considering the actionability metric, we asked participants how easy it was to follow the advice. Participants generally indicated that the advice was very easy to follow with the treatment group rankings being higher than the control group rankings. To assess whether the responses were able to motivate users to take action, we asked whether users actually followed the advice or if they planned to follow it in the future. When asked whether they followed the advice, participants in the treatment group tended to indicate 'Yes' more frequently and

[3]Following convention from social science, we treat the Likert scales as intervals, permitting use of t-tests for statistical significance [14]. By the Central Limit Theorem we assume normality due to large sample size.

Survey Question	Prompting Condition	No Prompting Condition	p-value
How many terms were there in the answer that were confusing or difficult to understand (1 = no terms, 3 = a couple, 5 = enough confusing terms that the answer was not understandable)	1.36220	1.60693	7.87E-06
Overall, how well were you able to understand the answer? (1 = not at all, 5 = I completely understood the answer)	4.70866	4.62952	0.08132
How easy was it to follow the advice given (if instructions were given)? (1 = impossible to follow, 5 = very easy to follow) (note: NA (no instructions were given) is excluded)	4.66343	4.56494	0.06105
Did you follow the advice provided by the assistant? *(Yes=1, Somewhat=0, No=-1)*	0.22572	-0.01506	2.92E-05
Fraction of participants who answered 'Yes' that they followed the advice provided	0.51444	0.40211	4.61E-04
Do you plan to follow the advice in the future? *(Definitely yes = 2, Probably yes = 1, Might or might not = 0, Probably not = -1, Definitely not = -2)*	1.50993	1.42566	0.16408
Fraction of Participants who indicated either 'Yes' that they already followed the advice or 'Probably Yes' or 'Definitely Yes' that they plan to follow the advice	0.80840	0.66867	3.54E-07
Fraction of Participants who indicated either 'Yes' that they already followed the advice or 'Definitely Yes' that they plan to follow the advice	0.66142	0.52861	2.14E-05
On a scale of 1-4 how would you rate the answer you received to this answer? (1=not helpful at all, 2=not terribly helpful 3=somewhat helpful, 4=very helpful)	3.70341	3.58283	6.55E-04
Fraction of participant answers that are 'very helpful'	0.71916	0.65060	0.02066

TABLE III: Results of the evening surveys for questions pertaining to the answer received by the participant. Results for the treatment group (prompting condition) and control group (no prompting condition) are averaged across all questions and all days. The leftmost column displays questions from the evening surveys (bolded) and statistics derived from answers directly provided to participants (not bolded). For questions with categorical values, the choices were mapped to numbers before means were calculated. These mappings are shown in blue italics. Rows highlighted in green are questions for which there was a significant difference between the results from the control and treatment groups (at a significance level of α = 0.05).

participants in the control group tended to indicate 'No' more often. After mapping the answer choices *Yes*, *Somewhat*, and *No*, to 1, 0, and -1, we saw a mean response of 0.226 for the treatment group and -0.015 for the control group. To better understand how likely the answers were to motivate users, we calculated the fraction of answers that were 'Yes', that participant had followed the advice of the extension. For the treatment group this fraction was 0.514, significantly higher than the fraction for the control group (0.402), p = 4.61E-04).

Some responses might offer participants immediate advice leading them to indicate that they followed the advice but don't plan to in the future or advice that is not immediately applicable leading them to indicate that they didn't follow the advice but plan to do so in the future. To account for these differences and gauge the total effect that extension answers had on users behavior, we considered an answer successful at changing behavior if either the participant indicated that they already had followed the advice or they indicated that they would follow the advice in the future. We considered the two evening survey questions ('Did you follow the advice provided by the assistant?' and 'Do you plan to follow the advice in the future?') together, recording a result of 1 if the participant indicated 'Yes' that they had already followed advice or either 'Probably Yes' or 'Definitely Yes' that they will follow the advice in the future and 0 otherwise. The results showed a statistically significant difference between groups with the results from the treatment group being 0.81 and the experimental group being 0.67. The results show that participants from both groups indicated more often than not that they either had or likely would follow the advice with this number corresponding to 81% of participants in the treatment group - this is 20% higher than for the control group. These results indicate that the majority of responses from the extension either had already or would likely in the future be successful at changing user behavior. It further suggests that an important percentage of questions submitted by participants were genuine questions.

When asked how helpful they found the answer they received, participants in both groups tended to rank the responses as "somewhat" to "very helpful" and rankings from the treatment group (3.703 out of 4) were significantly higher than those from the control group (3.583 out of 4), p = 6.55E-04. We calculated the fraction of rankings that were 'very helpful' for each group. This fraction was 0.719 for participants in the treatment group and 0.651 for participants in the control group (p = 0.021). These results show that while answers were generally perceived as helpful by both groups, rankings were statistically significantly higher for participants in the treatment group with over 70% of answers being perceived as very helpful, a significant increase over the control group.

While both groups generally provided positive rankings, the treatment group ratings were better for all questions. This difference was statistically significant in 3 of the 6 questions. While even the base responses performed well, these results provide support that the prompting techniques applied to the questions in the treatment group were successful at improving answers across the understandability, motivating, and actionability metrics.

C. Alternate Answer Comparisons

The remaining questions on the evening surveys provided participants with an alternate answer and asked them to compare the alternate answer to the one they received during the day. For participants in the treatment group, the alternate answer was the one they would have received without prompt

engineering and for the control group, the alternate answer had prompt engineering applied. The results of the responses to these questions in the evening survey are shown in Table IV.

When asked how helpful the alternate answers were, both groups found the answers "somewhat" to "very helpful", with the treatment group finding their alternate answers somewhat less helpful than the control group (although this difference was not statistically significant - p=0.17). When we calculated the fraction of answers that were found to be "very helpful" for both groups, we found a statistically significant difference (p = 0.00577) with 63% of alternate answers generated with prompt engineering (the control group in this case) being perceived as very helpful and 54.3% of alternate answers generated without prompt engineering (treatment group in this case) being perceived as very helpful.

When comparing the helpfulness of the alternate answers to the ones they received during the day, participants in both groups found the alternate answers slightly more helpful. The control group however found the alternate answers (namely the answers with prompting) to be significantly more helpful (0.498 on a scale of -2 to +2, with +2 being "much more helpful") than the treatment group who indicated the alternate answers (the one without prompting) were very similar to the answers they actually received (0.062 on the same scale of -2 to +2, where 0 corresponded to "About the same"). The results were similar when we asked participants to compare the answers they received to the alternate answers in terms of how likely they would have been to follow the advice provided. Both groups indicated they would have been slightly more likely to follow the advice but participants in the treatment group indicated that the answers were about the same while participants in the control group generally gave the alternate answers higher rankings (0.571 vs. 0.089 where 0 corresponded to "About the same" and 2 corresponded to "Much more likely").

When comparing the helpfulness rankings that participants gave the alternate answers vs. the main answers, the main answer ratings were slightly higher than the alternate answers for both groups. However, preference for the main answers was stronger for the treatment group as compared to the control group (a difference in ratings of 0.241 vs. 0.060)

The ratings participants provided when comparing the answers they received to the alternate answers show the same support for prompt engineering techniques as the results ranking only the main answer. Participants who didn't receive prompting for their main answer were more favorable towards the alternate answer (which was generated with prompt engineering) than participants in the treatment group who benefited from prompt engineering in their main answers and were comparing to answers that were generated with no prompt engineering. This support for prompt engineering is shown in all 3 questions, with differences between groups being statistically significant in 2 of the 3 questions (see Table IV).

D. Topic Analysis of User Questions

1) Classification of Questions: We further analyzed the topics of the questions that participants asked throughout the user study. We used two different classification schemes to categorize the questions: (1) Splunk Inc.'s Top 50 Cybersecurity Threats (which includes only 48 threats) [15], and (2) a list of 25 cybersecurity topics automatically generated by prompting gpt-4o-mini with all of the user questions from our human study.

In the classification step, the complete list of categories was passed in to gpt-4o-mini as the system prompt. gpt-4o-mini was then prompted to assign a topic to each question individually (in a separate context window), based on which category was most relevant. If none of the categories applied sufficiently to the question, gpt-4o-mini was instructed to assign a topic of "Other". This step was performed twice, once for each of the classification schemes. The full list of topics for each scheme, along with the number of questions labeled with each topic, can be found in Table VII in Appendix A.

In particular, we notice that the users' questions from the study spanned 36 topics from the Splunk classification scheme, and all 25 of the automatically-generated topics. Furthermore, 396 of the 1045 total questions (37.90%) were given a topic of "Other" under the Splunk scheme; the same was true for 108 of the 1045 questions (10.33%) with the gpt-4o-mini-generated scheme. This indicates that even though the user questions largely pertained to common or prototypical cybersecurity topics, a substantial proportion concerned more niche or esoteric topics.

These results also support the claim that in situ, users have questions that span a diverse set of cybersecurity topics. As mentioned previously, in our prior work [3] we proposed seven initial questions that spanned various domains of cybersecurity. While there is no direct one-to-one mapping between the initial prototypical seven questions, a generous mapping of each of these questions onto one of the 25 gpt-4o-mini topics would only cover *28.90% of the questions* (302 out of 1045) asked by our study participants. The vast array of topics that participants covered in our user study also confirms that hoping to train users on all relevant security topics is unrealistic.

2) Impact of Prompt Engineering at the Topic Level: For a deeper analysis, we examined the effect of our prompt engineering at the topic level. We only performed this analysis for the topics generated automatically from our questions using gpt-4o-mini, as over a third of the questions were categorized as "Other" using the Splunk Inc. classification scheme and thus limited the number of questions we could analyze.

We first split the questions into those asked by users who were in the treatment group (answers were generated with prompt engineering), and those that were in the control group (answers were generated without extra prompt engineering). We then isolated the questions for which the users reported that they had followed the advice (responded "yes" to "did you follow the advice provided by the assistant?") or would definitely follow the advice in the future (responded "definitely yes" to "do you plan to follow the advice in the future?"). We

Survey Question	Treatment Group (alternate answers without prompting)	Control Group (alternate answers with prompting)	p-value
On a scale of 1 - 4 how would you rate the above alternate answer? (1 = not helpful at all, 2 = not terribly helpful 3 = somewhat helpful, 4 = very helpful)	3.46194	3.52259	0.16991
Fraction of participant answers that are "very helpful"	0.54331	0.63102	0.00577
If you had received the above alternate answer would you have found it more or less helpful? *(Much less helpful = -2, Somewhat less helpful = -1, About the same = 0, More helpful = 1, Much more helpful = 2)*	0.05774	0.49849	5.24E-14
If you had received the above alternate answer would you have been more or less likely to follow the advice provided (if applicable)? *(Much less likely = -2, Somewhat less likely = -1, About the same level of likelihood = 0, Somewhat more likely = 1, Much more likely = 2)*	0.08883	0.57072	2.04E-15
Base answer helpful rating - alternate answer helpful rating	0.24147	0.06024	3.33-05

TABLE IV: Results of the evening surveys for questions pertaining to the alternate answer provided in the evening survey. Results for the treatment group (prompting condition) and control group (no prompting condition) are averaged across all questions and all days. The leftmost column displays questions from the evening surveys (bolded) and statistics derived from answers directly provided to participants (not bolded). For questions with categorical values, the choices were mapped to numbers before means were calculated. These mappings are shown in blue and in italics. Rows highlighted in green are questions for which there was a significant difference between the results from the control and treatment groups (at a significance level of $\alpha = 0.05$).

find that these two responses indicate that the answer provided by the security assistant was motivating enough to compel the user to act upon the recommendation.

In Table V, we report the differences in proportions of answers with strong motivating power to the total number of questions for each topic, between questions from the treatment and control groups. Notably, of the 24 topics (including "Other") which had questions from both the treatment and control groups, 22 had answers that seemed to perform better with the prompt engineering, further providing evidence of effectiveness for our prompt engineering methodology.

VI. Discussion and Future Work

A. Security Assistant Found Helpful and Effective

In general, participants found the extension and answers they received helpful and indicated that they followed the advice provided by the extension or planned to follow it in the future. During the final interviews, when asked whether they followed the advice provided by the extension, many participants responded positively, indicating that they changed their behavior and took specific actions as a result of the response they received from the extension. One participant answered *"Yeah, absolutely ... pretty much all of [the answers] had helpful information and several of them had concrete things that I did do as a result of it."* Even participants who had a background in IT and already had knowledge about computer security indicated that found the responses helpful and that they served as reminders which prompted them to take action, e.g., *"I wasn't aware that mobile antivirus was a thing because I remember years ago, we didn't worry about it on android. So the fact that it told me about those three [antivirus softwares], the ones that I was familiar with, at least on the desktop side, made me go 'oh, yeah, that's a good answer. I'll go look at those and grab one of those. So it definitely was the answer that it gave me that that caused me to do it."* Results from both the evening surveys as well as the final interviews indicated that the extension responses were able to actually motivate users to take certain actions. These results provide support for the potential of a tool like this one to provide advice that is successful at actually changing user behavior.

Further, for participants who already had a level of expertise or indicated that they already knew the information that the extensions supplied, many indicated that the extension served as a helpful reminder or encouraged them to actually take specific actions that they might have known they should do but were not motivated to do before. Specifically one participant indicated *"When it said 'oh you should do the two factor authentication' I kind of just went through and made sure that everything was set the way it was so it was a nice reminder because I don't do security things on a set schedule"*.

Overall, the extension demonstrated potential for encouraging and motivating users. Even for participants who had a background in cybersecurity, it was able to serve as a helpful reminder and successfully influencing behavior change. This result supports the notion that such a tool could be both utilized and effective in practice.

B. PMT-Guided Prompt Engineering Improves Answer Effectiveness

To evaluate the extent to which the prompting strategy based on PMT we developed in our initial study [3] was effective with real users and their everyday cybersecurity questions, we split participants into a treatment group who received responses generated with prompting and a control group who received base answers with no prompting applied.

All results from the evening surveys provided support that prompt engineering improved response effectiveness with a majority of the results showing statistically significant differences between the two groups. Participants in the treatment group consistently provided higher rankings for survey questions assessing the understandability, actionability, and motivating power metrics. In final interviews, participants in

Topic (Generated by gpt-4o-mini)	Proportion of Prompt Engineered Answers that had Strong Motivating Power	Proportion of Non-Prompt Engineered Answers that had Strong Motivating Power
Software Updates	100.00% (5/5)	71.43% (5/7)
Social Engineering	100.00% (4/4)	50.00% (7/14)
Children and Online Safety	100.00% (3/3)	100.00% (2/2)
User Behavior and Awareness	100.00% (2/2)	57.14% (4/7)
Backup Strategies	100.00% (2/2)	55.56% (5/9)
Public Wi-Fi Risks	87.50% (7/8)	62.50% (10/16)
Two-Factor Authentication (2FA)	86.67% (13/15)	71.43% (10/14)
Social Media Security	83.33% (5/6)	50.00% (4/8)
Email Security	83.33% (5/6)	71.43% (5/7)
Data Privacy	78.95% (30/38)	47.76% (32/67)
Malware and Viruses	76.74% (33/43)	53.97% (34/63)
Incident Response	75.00% (6/8)	50.00% (8/16)
Identity Theft Prevention	75.00% (6/8)	69.23% (9/13)
Secure Browsing Practices	68.18% (15/22)	65.91% (29/44)
Network Security	65.79% (25/38)	35.80% (29/81)
Device Security	64.71% (22/34)	63.83% (30/47)
Password Security	62.86% (22/35)	65.91% (29/44)
Phishing Awareness	61.54% (8/13)	86.67% (26/30)
Security Tools and Software	58.33% (7/12)	51.85% (14/27)
Cybersecurity Education	50.00% (5/10)	50.00% (7/14)
Types of Cyberattacks	50.00% (4/8)	48.00% (12/25)
Cloud Security	42.86% (3/7)	25.00% (4/16)
Other	40% (16/40)	39.71% (27/68)
Emerging Technologies	28.57% (4/14)	25% (3/12)
Remote Work Security	No questions with prompt engineering	100.00% (2/2)
Legal and Ethical Considerations	No questions with prompt engineering	36.36% (4/11)

TABLE V: The proportions of answers that were rated by user study participants as having strong motivating power (either responded "yes" to "did you follow the advice provided by the assistant?" or responded "definitely yes" to "do you plan to follow the advice in the future?"), separated by the topic. The rows are sorted in descending order based on the proportion of highly motivating answers with prompt engineering. The rows highlighted in green represent topics where the proportion of answers rated as highly motivating was greater for the answers generated using prompt engineering than the answers generated without prompt engineering.

the control group demonstrated a preference for the alternate answer.

One participant indicated that they *"picked the alternate answer because it not only gave me the answer, but it gave me reasons why."* Another participant provided similar support for the more detailed responses generated with prompting, indicating that they *"really liked the alternative answers because it gave a more in-depth, like I said, a full circle response"*, even stating that they would wait for the alternate answer in the evening surveys, *"let me wait for the alternative answer because that is probably going to answer what my follow-up question would be"*. Participants tended to find that the extra details that prompt-engineered responses provided made the answers more helpful or actionable. Participants would also highlight specific aspects of the prompted responses that made them more likely to follow the advice such as the list format of the response *"Yeah, I like the list because it's more organized"*.

For participants who did not prefer prompt-engineered responses and reported the alternate and main answers as about the same, a common reason we observed in the final interviews was that they tended to find the responses too verbose and indicated that they already understood the base answer so the extra information provided by the alternate answer was unnecessary. In the final interviews, a participant from the control group indicated that the alternate answers were *"not really providing any additional helpful information kind of just maybe a little like extra words but not really saying anything more"*. These results suggest that responses could be further improved by generating answers that are personalized and adapt to the level of expertise and sophistication of their users.

To better understand the qualitative effect that prompting had on responses, we examined answers for which participants in the control group indicated that they found the alternate answer much more helpful or would have been much more likely to follow the advice and answers for which participants in the treatment group indicated they found the alternate answer (with no prompting) much less helpful or would have been much less likely to follow its advice. Table VI. contains 2 example questions asked by users in the control group as well as the answer and alternate answer they received. For both questions, the participant indicated that they would have been much more likely to follow the advice provided by the alternate answer (which had prompting applied). Both questions are examples of cases in which without prompting a chatbot may be able to provide an accurate answer, but the prompting applied to generate the alternate answer increased the motivating power of the answer, increasing the chance of changing user behavior.

Considering the first question about antivirus programs, the baseline answer gives an accurate overview of antivirus programs but does not go beyond listing a few examples of these programs. In contrast, the alternate answer includes additional statements to motivate the user such as "Not using antivirus software can leave your Mac vulnerable to malware that could steal personal information or damage your system" and includes an additional section, 'Important Considerations', which highlights the risks of not using antivirus programs thereby encouraging users to go actually download the software. Similarly, both answers for the second question on clearing the cache in an iPhone are accurate, but the alternate answer (with prompting) goes a step further, stating not just that clearing the cache is recommended for performance but also including another section 'Risks of Not Clearing Cache' to highlight potential risks and motivate users to take action and clear their cache on their mobile phone. For both questions, each answer is accurate, but the alternate answer includes additional details to motivate the user to actually follow the advice provided, thus making the user much more

likely to follow its advice. These results demonstrate the significance of prompt engineering in improving the effectiveness of answers provided by the cybersecurity QA assistant. Furthermore, these results indicate that principles from PMT can effectively be integrated into prompts for LLMs, making generated responses more effective at motivating users to act upon the given recommendations.

C. Limitations

One potential limitation of this study is that participants may have asked some questions solely to meet their daily two question quota and that these questions may not have truly arisen in the context of their everyday computer use. The final interviews suggest however that participants were generally asking a mix of genuine in-context questions along with some more general questions. Of the 10 participants who were interviewed in the final study, 6 indicated their questions were a mix of genuine questions they had in the moment and somewhat more general questions they might have been making up to meet their quota. One participant indicated that their questions were primarily in-context questions. Two indicated that they felt their questions were not truly genuine because they already had knowledge about computer security and already knew the answers to their questions. The last participant indicated their questions were primarily general questions because they believed the extension was an unsophisticated bot and wouldn't be able to handle more complex questions. These results suggest that while several of the questions were just general questions and were asked to meet the requirements of the study, many of the questions were asked in-context about topics participants were genuinely wondering about or asking for genuine advice. This interpretation is further reinforced by the large percentage of questions that participants reported as being helpful and the substantial percentage of questions where participants reported having followed the advice or planning to do so, especially in the prompting group.

D. Future Possible Directions

Answers collected in the evening surveys and the final interviews outline future possible directions for refining cybersecurity QA assistants. This includes opportunities to further improve their effectiveness, change user behavior and reduce human-related cybersecurity incidents. Our results generally support the use of prompting to improve answer effectiveness. This is not to say that better prompts could not be designed, including prompts tailored to individual users. Final interviews also identified specific features of answers that would their advice more likely to be followed.

To improve actionability, several participants mentioned wanting the responses to provide very specific next steps or provide specific links or websites to go to. One quote from a final interview indicated: *'If it would have been able to expand upon it and go in a detail or link videos and and detailed reasons or case studies, things like that.' 'So if it provided links to external or maybe pointers to next things to look at, do you think that would have been helpful as well?' 'Yeah, that would have been a lot better.'* Adding such information could make the advice provided in responses easier to follow. A few participants also indicated that they would like the responses to more openly encourage them to ask follow-up questions, and help ensure they understand the responses - perhaps even suggest follow-on questions they might want to ask, including possibly questions they would not necessarily come up with on their own. One participant identified this desire to have the extension suggest follow-on questions as they might not know what to ask next or related questions or potential concerns. *"What might be nice to have is, if I ask a question and it answers it, that it may kind of think ahead and say, 'would you like to know more about such and such that's related to the question?' because sometimes you'll get an answer and you don't really realize or you may not be thinking at the time that you need to know something else to go along with it. I think that would have be a nice feature, if it could think ahead for me and suggest something else for me."*

One of the most promising directions suggested by participants to improve the effectiveness of the cybersecurity assistant - suggested by almost all participants who completed final interviews, was the opportunity for generating personalized answers. An assistant that takes into account the knowledge of the user and other relevant traits and adjusts its responses accordingly would likely be even more helpful and more likely to positively impact user behavior. In final interviews, several participants indicated the desire for such a cybersecurity QA assistant capable of tailoring their answers to individual users, e.g., *"that would be a good feature, I think is to have it gauge what level you might be comfortable with and give more detailed and more higher level answers."*. A couple of participants also suggested having an extension that allowed users to select different levels of knowledge e.g., *"If you had it separated in by like three or four different categories, like you have no knowledge, you have basic knowledge, you have general or advanced, ... I think that would definitely help"*. Further, an ideal tool would adapt to the user as the conversation continues and learn what kinds of answers the user preferred or what kinds of answers motivated the user the most. This could include adjusting aspects of the answer such as technicality, language, format, etc.

VII. Concluding Remarks

Humans are often the weakest link in cybersecurity, being responsible for a majority of security breaches. LLM-powered chatbots open up an opportunity to mitigate these breaches with their ability to offer specific guidance and encourage users to adopt safer behaviors. While these chatbots, such as ChatGPT, seem capable of providing accurate responses to a broad range of everyday cybersecurity questions, an ideal cybersecurity assistant would not only provide accurate answers but also motivate users to act on the advice provided and alter user behavior. In this work, we introduced a cybersecurity question answering (QA) assistant, implemented as a Chrome extension and study its effectiveness in the context of

Question What free anti-virus programs are safe and recommended to download on Mac laptops?
Answer For Mac laptops, several reputable free antivirus programs provide good protection against malware and other security threats. Here are a few recommended options: 1. **Avast Security for Mac**: Offers malware protection, web and email shields, and a Wi-Fi security scan. 2. **AVG Antivirus for Mac**: Provides real-time malware protection and automatic updates, maintaining security against the latest threats. 3. **Sophos Home Free**: Includes real-time antivirus, parental web filtering, and remote management of up to three devices. 4. **Malwarebytes for Mac**: Known for its fast scanning and ability to catch and remove malware that other programs might miss. 5. **Bitdefender Virus Scanner for Mac**: Uses the same virus signatures as its paid versions, good for quick scans and easy to use. Ensure to download antivirus software directly from the official website or a trusted app store to avoid counterfeit software.
Alternate Answer 1. **Avast Security for Mac**: Avast offers effective malware protection along with web and email shields. It's user-friendly and doesn't heavily impact system performance. Not using antivirus software can leave your Mac vulnerable to malware that could steal personal information or damage your system. 2. **AVG Antivirus for Mac**: Similar to Avast, AVG provides real-time malware protection and is straightforward to use. It includes a triple-layer of security, ensuring thorough scanning and removal of threats. 3. **Sophos Home Free**: This option includes real-time antivirus protection and parental controls if needed. Sophos is known for its simplicity and minimal performance impact on your system. 4. **Malwarebytes for Mac**: While the free version only offers manual scanning, it's effective at detecting and removing malware. It's particularly useful for adware and potentially unwanted programs that can slow down your Mac. **Important Considerations**: - Regularly update the antivirus software to protect against the latest threats. - Even though Macs are less targeted than PCs, they are not immune to malware. Failing to use antivirus software can expose you to risks like identity theft, loss of data, and compromised personal information. Using any of these tools will help protect your Mac from common threats, making them a critical component of your digital security.
Question How can i clear my cache from my iPhone? is it recommended to do so for security reasons?
Answer Clearing the cache on your iPhone can help with privacy and security by removing stored data that could potentially be accessed by unauthorized users. Here's how to clear your cache: For Safari: 1. Open **Settings**. 2. Scroll down and tap **Safari**. 3. Scroll down and tap **Clear History and Website Data**. 4. Confirm by tapping **Clear History and Data**. For other browsers/apps: Check the app's settings for options to clear cache or browsing data. Is it recommended? Yes, periodically clearing your cache can help protect your privacy and improve device performance by removing old and unnecessary files.
Alternate Answer Here's a non-technical guide on how to clear the cache: 1. **For Safari Browser:** - Go to the Settings app. - Scroll down and tap "Safari." - Scroll down again and tap "Clear History and Website Data." - Confirm by tapping "Clear History and Data." 2. **For Other Apps:** - Go to the Settings app. - Tap "General," then "iPhone Storage." - Select an app whose cache you want to clear. - Tap "Offload App" to delete the app but keep its data, or "Delete App" to remove everything. Reinstall the app if needed. **Is it recommended?** Yes, it's recommended to clear your cache periodically for security reasons to eliminate stored data that could include sensitive information. This is especially important if you notice your device behaving oddly or if you frequently access sensitive information like banking details. **Risks of Not Clearing Cache:** Not clearing your cache can lead to: - Slower device performance. - Potential security risks if sensitive data is stored and accessed by unauthorized apps or individuals. - Buildup of corrupted files which might cause app malfunctions. Clearing your cache is a simple maintenance task that can help keep your iPhone running smoothly and more securely.

TABLE VI: Two example questions with the answer (generated with no prompting) and alternate answer (generated with prompting applied) for which participants indicated they would have been much more likely to follow the advice provided by the alternate answer.

a 10-day in situ study involving 51 participants. Participants were divided into two conditions, one which relied on an unprompted version of a state of the art chatbot (GPT-4) and the other where the same chatbot was enhanced with a prompt based on Protection Motivation Theory (PMT). This prompt had been reported to be particularly effective in an initial study conducted by our research group involving seven synthetic cybersecurity questions, in which answers were evaluated by our group [3]. In contrast, our study participants collectively generated over 1,000 questions in the context of their regular computer activities, and evaluated themselves the effectiveness of the answers they had received to their questions, including whether the answers were understandable, actionable, whether they had followed the advice offered in these answers, and whether the answers were helpful. Our findings generally support those reported in our initial study, but also provide significantly stronger evidence of the effectiveness of PMT-guided prompt engineering across a much richer set of questions, with these questions being submitted in situ and being evaluated by the very people who submitted them. Our results provide support for prompt engineering, guided by principles from Protection Motivation Theory, as a way of enhancing answer effectiveness and impacting user behavior. Across all results in the evening surveys evaluating the assistant's responses, answers generated with prompt engineering received higher scores, with most differences between the two conditions (with prompting versus without prompting) being statistically significant. Participants found that prompt engineering made the responses more detailed and easier to understand and follow.

More generally our findings provide evidence that a cybersecurity QA assistant based on state-of-the-art LLM technology would likely be found useful by a broad swath of users, would likely see actual use, and would be likely to positively alter cybersecurity behaviors. Specifically, participants in our study reported finding the assistant helpful, and reported that they tended to follow its advice, especially when enhanced with the PMT prompt. Over 40 percent of participants responded that they would be interested in continuing to use the assistant, "as is". Nearly another 50 percent responding "maybe" to that same question. A third of participants indicated they would expect to use such a tool either "every day" or "a few times per week", another third "a few times per month", and the rest "a few times per year".

Our final interviews also point to specific opportunities to further enhance the effectiveness of a cybersecurity QA assistant. A particularly promising area of exploration seems to relate to the generation of personalized answers.

ACKNOWLEDGMENT

This research has been supported in part by grants from the National Science Foundation under the SaTC program (grants CNS-1914486) and under the REU program, the latter in part through CMU's RE-USE Program (NSF grant 2150217). Additional support was also provided by CMU's Block Center under its Responsible AI initiative.

REFERENCES

[1] J. Hancock, *Psychology of Human Error*, 2022. [Online]. Available: https://www.tessian.com/resources/psychology-of-human-error-2022/

[2] W. E. Forum, *The Global Risks Report 2022*, 2022. [Online]. Available: https://www.weforum.org/publications/global-risks-report-2022/

[3] A. Balaji, L. Duesterwald, I. Yang, A. Priyanshu, C. Alfieri, and N. Sadeh, "Generating effective answers to people's everyday cybersecurity questions: An initial study," in *International Conference on Web Information Systems Engineering*, 2024. [Online]. Available: https://usableprivacy.org/static/files/balaji_wise_2024.pdf

[4] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson, "Nudges for privacy and security: Understanding and assisting users' choices online," *ACM Comput. Surv.*, vol. 50, no. 3, Aug 2017. [Online]. Available: https://doi.org/10.1145/3054926

[5] F. L. Greitzer, *Situated Usability Testing for Security Systems*, Feb 2011. [Online]. Available: https://www.pnnl.gov/main/publications/external/technical{_}reports/PNNL-20201.pdf

[6] P. Story, D. Smullen, R. Chen, Y. Yao, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub, "Increasing adoption of tor browser using informational and planning nudges," *Proceedings on Privacy Enhancing Technologies*, 2022.

[7] J. E. Maddux and R. W. Rogers, "Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change," *Journal of Experimental Social Psychology*, vol. 19, no. 5, pp. 469–479, 1983. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022103183900239

[8] P. Story, D. Smullen, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub, "From intent to action: Nudging users towards secure mobile payments," in *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020, pp. 379–415.

[9] E. M. Redmiles, N. Warford, A. Jayanti, A. Koneru, S. Kross, M. Morales, R. Stevens, and M. L. Mazurek, "A comprehensive quality evaluation of security and privacy advice on the web," in *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 89–108. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/redmiles

[10] Z. Ji, E. Choi, and P. Gao, "A knowledge base question answering system for cyber threat knowledge acquisition," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022, pp. 3158–3161.

[11] G. Agrawal, K. Pal, Y. Deng, H. Liu, and Y.-C. Chen, "Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23 164–23 172, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/30362

[12] J. Gennari, S. hon Lau, S. Perl, J. Parish, and G. Sastry, "Considerations for evaluating large language models for cybersecurity tasks," Carnegie Mellon University, Software Engineering Institute, Tech. Rep., 02 2024.

[13] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff *et al.*, "The prompt report: A systematic survey of prompting techniques," *arXiv preprint arXiv:2406.06608*, 2024.

[14] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, vol. 15, pp. 625–632, 2010.

[15] Splunk, *Top 50 Cybersecurity Threats*. San Francisco, California: Splunk Inc., 2023.

APPENDIX

A.

Topic from Splunk Inc. Top 50 Cybersecurity Threats (Including "Other")	**Number of Questions (out of 1045) Related to Splunk Topic**	**Topic from gpt-4o-mini Generated Topics from User Questions (Including "Other")**	**Number of Questions (out of 1045) Related to gpt-4o-mini-Generated Topic**
Other	396	Network Security	119
Malware	120	Other	108
Phishing	73	Malware and Viruses	106
Man-in-the-Middle Attack	59	Data Privacy	105
Account Takeover	55	Device Security	81
Web Session Cookie Theft	41	Password Security	79
IoT Threats	37	Secure Browsing Practices	66
Compromised Credentials	32	Phishing Awareness	43
Brute Force Attack	33	Security Tools and Software	39
Social Engineering Attack	28	Types of Cyberattacks	33
Ransomware	25	Two-Factor Authentication (2FA)	29
Spyware	15	Emerging Technologies	26
Router and Infrastructure Attacks	13	Incident Response	24
Zero-Day Exploit	13	Cybersecurity Education	24
Drive-by Download Attack	12	Public Wi-Fi Risks	24
Insider Threat	11	Cloud Security	23
Network Sniffing	10	Identity Theft Prevention	21
DoS Attack	10	Social Engineering	18
Business Email Compromise	7	Social Media Security	14
Credential Reuse Attack	6	Email Security	13
DNS Hijacking	6	Software Updates	12
Command and Control Attack	6	Legal and Ethical Considerations	11
SQL Injection	5	Backup Strategies	11
Privileged User Compromise	4	User Behavior and Awareness	9
Advanced Persistent Threat	3	Children and Online Safety	5
Suspicious Cloud Storage Activities	4	Remote Work Security	2
Cross-Site Scripting	3		
Masquerade Attack	3		
Spear Phishing	2		
Typosquatting	2		
Application Access Token	2		
Identity Theft	2		
Amazon Web Services (AWS) Attacks	2		
Credential Dumping	2		
Simjacking	1		
Supply Chain Attack	1		
Malicious Powershell	1		

TABLE VII: Comprehensive list of topics from both classification schemes (Splunk Inc. and automatically generated topics from gpt-4o-mini), along with the number of questions belonging to each topic from the 1045 user questions. The number of questions for each scheme are independent (a question is classified into one topic from the Splunk Inc. scheme and one topic from the gpt-4o-mini scheme; the second and fourth columns both add to 1045). Though the Splunk Inc. scheme has 48 cybersecurity threats, the user questions only pertained to 36 of them. Both schemes have the "Other" category included as well. Lastly, the number of questions includes questions from both the treatment and control groups.