Generating Effective Answers to People's Everyday Cybersecurity Questions: An Initial Study

Ananya Balaji*, Lea Duesterwald*, Ian Yang*(✉), Aman Priyanshu, Costanza Alfieri, and Norman Sadeh

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania
{ananyaba,lduester,iany,apriyans,calfieri,ns1i}@andrew.cmu.edu

Abstract. Human users are often the weakest link in cybersecurity, with a large percentage of security breaches attributed to some kind of human error. When confronted with everyday cybersecurity questions – or any other question for that matter, users tend to turn to their search engines, online forums, and, recently, chatbots. We report on a study on the effectiveness of answers generated by two popular chatbots to an initial set of questions related to typical cybersecurity challenges faced by users (e.g., phishing, use of VPN, multi-factor authentication). The study does not only look at the accuracy of the answers generated by the chatbots but also at whether these answers are understandable, whether they are likely to motivate users to follow any provided recommendations, and whether these recommendations are actionable. Surprisingly enough, this initial study suggests that state-of-the-art chatbots are already reasonably good at providing accurate answers to common cybersecurity questions. Yet the study also suggests that the chatbots are not very effective when it comes to generating answers that are relevant, actionable, and, most importantly, likely to motivate users to heed their recommendations. The study proceeds with the design and evaluation of prompt engineering techniques intended to improve the effectiveness of answers generated by the chatbots. Initial results suggest that it is possible to improve the effectiveness of answers and, in particular, their likelihood of motivating users to heed recommendations, and their ability to act upon these recommendations without diminishing their accuracy. We discuss the implications of these initial results and plans for future work in this area.

Keywords: Cybersecurity · Generative AI · Large Language Models · Natural Language Processing · Question Answering · Prompt Engineering

* The first three authors contributed equally to this work and are equal first authors. Please refer questions about this research to the last author, Prof. Norman Sadeh - sadeh@cmu.edu

1 Introduction

1.1 Motivation and Challenges

Human users are often the weakest link in cybersecurity, with a large percentage (estimated by some at 95% [3]) of security breaches attributed to some kind of human error (see, e.g., [5]). Training everyday users, whether employees in a corporation or children at school, to more effectively identify potential security threats and to better protect themselves has never been more important, and yet it has never been more challenging. When confronted with everyday cybersecurity questions, users often turn to their search engines and online forums for answers, and more recently, to chatbots. With their ease of use and rapidly growing popularity, chatbots have the potential of becoming the preferred source of information for many users. This possibility raises a number of questions. An obvious first is to what extent chatbots can be trusted to generate accurate answers. Accuracy, however, is not everything. With many security questions likely to require answers that include recommendations about what a user should do to best protect themselves, answers should also be understandable and actionable. To further complicate matters, cybersecurity is also known to be a perfect example of a *secondary task*. People are typically required to make security decisions while they are engaged in other tasks, their so-called "primary tasks" (e.g., completing a sale, filing a report, posting an update on social media) [2,4,5,9]. Not too surprisingly, people have been shown to commonly prioritize their primary tasks and dismiss potential cybersecurity risks (e.g., failing to exercise caution when clicking links in emails, failing to update their software, reluctance to change passwords or to adopt multi-factor authentication). To be truly effective, answers to cybersecurity questions should also be sufficiently compelling to motivate users to heed their recommendations and overcome their natural tendency to dismiss potential risks and to favor their primary tasks.

Specifically, we report on a study on the effectiveness of answers generated by two popular chatbots (ChatGPT 4 and Llama 2) to an initial set of questions related to typical cybersecurity challenges faced by users (e.g., phishing, use of VPN, multi-factor authentication). While results suggest that state-of-the-art chatbots are reasonably good at providing accurate answers to common cybersecurity questions, their performance when it comes to generating answers that are relevant, understandable, actionable, and, most importantly, likely to motivate users to heed their recommendations, seems less reliable. The study proceeds with the design and evaluation of prompt engineering techniques intended to improve the effectiveness of answers generated by the chatbots. Initial results suggest that it is possible to improve the effectiveness of answers and, in particular, their likelihood of motivating users to heed recommendations and their ability to act upon these recommendations without diminishing their accuracy

The contributions of this research are as follows:

- We identify different dimensions required for answers to cybersecurity questions to be truly effective. We further introduce an initial set of metrics designed to help annotators evaluate performance along each one of these dimensions.

- We report on the performance of two popular state-of-the-art chatbots in generating answers to an initial set of questions designed to capture common cybersecurity questions with which users are known to struggle. This initial study, though limited to a small set of questions, suggests that while state-of-art chatbots are generally capable of generating accurate answers to this initial questions, they fall short along other important dimensions, including the ability to generate answers that are easy to understand, easy to follow, and that do a sufficiently good job at highlighting risks associated with not following recommendations.
- We engineer and evaluate prompts intended to improve the effectiveness of answers generated by chatbot without negatively impacting their accuracy. Results suggest that prompts focused specifically on motivating users to heed recommendations tend to perform best, increasing in particular the general understandability and motivating power of generated answers.
- We provide a detailed analysis of the particular areas where answers generated with different types of prompts fall short and how the motivating prompt seems to improve performance - resulting in statistically significant improvements in both motivating power and understandability without sacrificing performance on other metrics, including accuracy.
- We discuss next steps to build on these initial results and further refine and evaluate across a broader range of questions.

2 Evaluating Answers

The primary objective of this work is to generate answers to common cybersecurity questions that will truly help those people asking the questions. In order for users to effectively avoid security breaches, they must not only have accurate information on security best practices, but also be motivated to actually implement these behaviors. The most effective answer would not only provide accurate information to users but also take steps to encourage users to follow any advice provided. To take into account the different aspects of what makes an answer a truly effective one in this context, we measure answer effectiveness along five dimensions: (1) accuracy, (2) relevance, (3) motivating power, (4) actionability, and (5) understandability - a more detailed discussion of metrics used to measure effectiveness along each of these dimensions is provided in Section 3.

(1) Accuracy: When responding to a cybersecurity question, it is crucial that the information provided is accurate and does not promote unsafe or harmful security practices. In our evaluation of accuracy, we distinguish between different levels of inaccuracy in answers.

(2) Relevance: An effective answer must have both "literal" relevance directly relating to the user's question, and "implicit" relevance to cover all issues raised by the user's question, even if not explicitly mentioned by the user. Users may not always be aware of all the factors they need to consider when formulating their questions, however the most effective answer should address all the critical points. Furthermore, an optimal answer should identify and address possible misunderstandings or misconceptions held by the user to offer the most effective assistance.

(3) Motivating Power: For a chatbot's answers to effectively change user behavior, they must motivate users to follow provided advice by highlighting risks, helping users understand why these risks apply to them and why the should follow the advice.

(4) Actionability: For users to effectively change their behavior, responses that contain advice must provide clear and understandable steps. To assess the actionability of answers, we considered whether different personas with varying levels of technical sophistication would be able to follow the advice in the response.

(5) Understandability: To ensure that users can effectively change their behavior based on the assistant's answers, it is crucial that all users can understand the responses, regardless of their technical sophistication. To evaluate how understandable answers are, we consider whether different personas would be able to understand the answers.

3 Design of a Baseline Study

3.1 Developing Question Set and Generating Baseline Answers

We began our study by developing a set of cybersecurity questions covering issues with which everyday users commonly struggle. The questions were in great part informed by one of the most popular set of cybersecurity training modules available in industry, namely Proofpoint's Cybersecurity Awareness Training Modules [8]. These modules are used by a large percentage of Global Fortune 1000 companies to educate their employees on cybersecurity threats and on how to best protect themselves and their organizations. For example, one such training module, "Avoiding Dangerous Links" discusses phishing emails with malicious links. Content of this module motivated our first question on Identity Theft/Phishing. The development of our questions was further informed by a review of relevant government (e.g., NIST) and industry reports (e.g., Verizon [6], Ponemon Institute [7], SANS Institute).

Questions were intentionally phrased using language and details likely to be those used by a non-technical user. The questions intentionally left some ambiguity about what the user may or may not know and the extent to which they might be aware of different possible risks, leaving room for the chatbot to possibly address this potential lack of knowledge. A table showing the seven questions used in this study and their topics is included in the Appendix (see Table 6).

Model Configurations Both ChatGPT 4 and Llama 2 were accessed using the client libraries provided by OpenAI. Answers for all seven questions for both models were generated on April 15th, 2024.

For Llama 2, we configured the OpenAI client libraries with Together.AI's API [1]. The 70B parameter version of Llama 2 and the GPT-4 version of ChatGPT were used for all answers and evaluations.

Each model was prompted with a temperature of 0.0 to maximize reproducibility, and a maximum token limit of 512. We expected that any answers exceeding this token limit would lead to decreasing scores in our evaluation metrics due to being overly long and verbose. All other parameters were left as the default value assigned by the OpenAI client.

3.2 Annotation Methodology

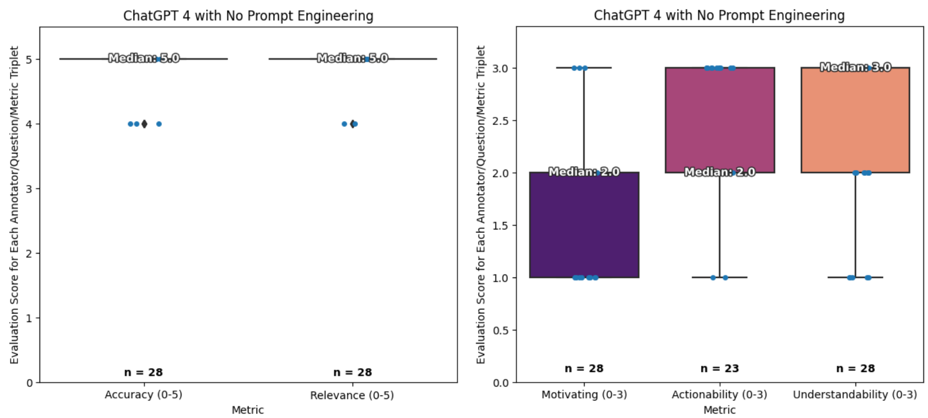
Our team developed a set of specific guidelines on how to grade answers along each of the different dimensions identified as contributing to the effectiveness of an answer. Detailed guidelines can be found in the Appendix in Table 5. The guidelines for the understandability and actionability metrics explicitly aim to recognize that not all users are the same and that answers that are sufficiently clear to some may not be to others. Similarly, the motivating metric guidelines account for individual traits, such as different levels of risk aversion, by broadly evaluating different aspects of motivation. Additionally, the relevance metric assigns penalties to responses that fail to address information relevant to the user's question, even if such information is not directly mentioned by the user.

Members of the research team served as annotators, systematically evaluating answers across the metrics described previously. Annotators were asked to independently apply these guidelines and annotate answers generated for the seven questions. Each answer was annotated by at least three different annotators, with some answers being evaluated by as many as five annotators. Inter-annotator agreement for all annotations collected in this study based on our guidelines was measured using Krippendorff's Alpha ($\alpha = 0.8910$), suggesting that these guidelines were fairly effective.

3.3 Evaluating Unprompted Answers

As a first step, our study evaluateed the performance of both ChatGPT 4 and Llama 2 on the seven questions without any prompt engineering ("unprompted answers").

Performance of ChatGPT and Llama is summarized in Figure 1 and 2. For both ChatGPT and Llama, we report the distribution of scores for each of the five metrics separately. The results for ChatGPT are shown in Figure 1. Figure 1a displays the distribution of human evaluator scores for the metrics rated out of five (Accuracy and Relevance), while Figure 1b displays the performance for the metrics rated out of three (Motivating, Actionability, and Understandability). Corresponding results for Llama are shown in Figures 2a, and 2b, respectively.



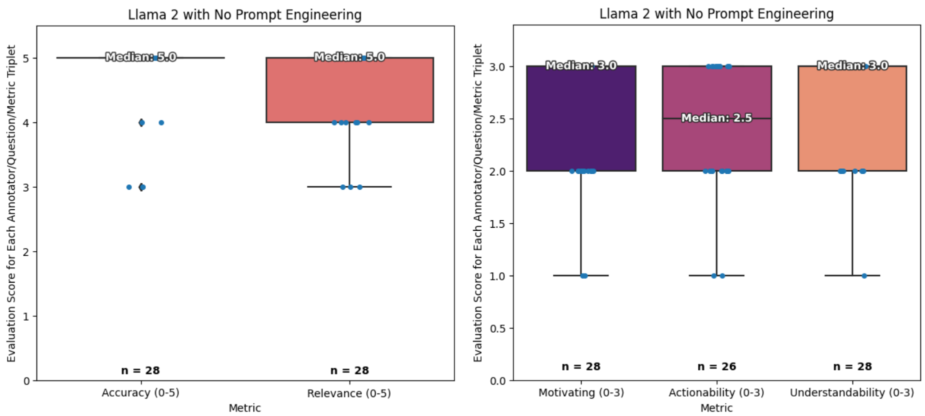
(a) Metrics from (0-5) (b) Metrics from (0-3)

Fig. 1: Scores from human evaluation for ChatGPT 4-generated answers without prompt engineering, across all questions and annotators. Each blue dot represents an individual annotator/question/metric triplet. The number of annotations displayed for each metric is denoted by **n**=# at the bottom of the plot.

The number of annotations represented as points in the box plot is shown as $n = \#$ near the bottom of each plot, along the x-axis. The reader should keep in mind that these numbers of annotations are for all 7 questions. In both Figure 1 and Figure 2, the number of annotations for Actionability is slightly lower than for the other four metrics ($n = 23$ and $n = 26$ respectively - versus $n = 28$ for the other metrics). This is because annotators were instructed to assign a score of "N/A" to answers for which they felt actionability was not a relevant metric. One example of such an answer is shown in Figure 3.

The reader will notice that both models perform quite well with respect to the Accuracy and Relevance metrics. Both models perform less well on the other three metrics (Motivating Power, Actionability, and Understandability) - or at least annotator ratings for these metrics show greater variation, with a number of ratings falling as low as 1 on a scale of 3. The answers generated by Llama had higher averaged and normalized ratings from human annotators for the Motivating, Actionability, and Understandability metrics than those generated by ChatGPT. In particular, the Llama answers were rated more highgly than the ChatGPT ones for the Motivating metric with a level of significance of 0.05 ($p = 3.46 * 10^{-4}$).

The answer with the lowest average Motivating score for ChatGPT was for Question 3 (providing phone number for Google's Two-Factor Authentication)



(a) Metrics from (0-5)

(b) Metrics from (0-3)

Fig. 2: Scores from human evaluation for Llama 2-generated answers without prompt engineering, across all questions and annotators. Each blue point represents an individual annotator/question/metric triplet. The number of annotations displayed for each metric is denoted by **n**=# at the bottom of the plot.

Question 2: My friend recently recommended I use a VPN to override the geographical restrictions for some of the shows available on Netflix. What is a VPN and is it safe to use?

ChatGPT 4 Answer: A VPN, or Virtual Private Network, is a service that allows you to connect to the internet via an encrypted tunnel to ensure your online privacy and protect your sensitive data. VPNs are commonly used to secure connections to public Wi-Fi hotspots, hide IP addresses, and make your browsing private. Using a VPN to override geographical restrictions is a common practice. However, it's important to note that it is against Netflix's terms of service, and they have the right to block your account if they detect you're using a VPN. As for safety, it depends on the VPN service you use. Some VPNs may log your data and sell it to third parties. Therefore, it's crucial to choose a reputable VPN provider that has a strict no-logs policy and uses strong encryption. Always do your research before choosing a VPN.

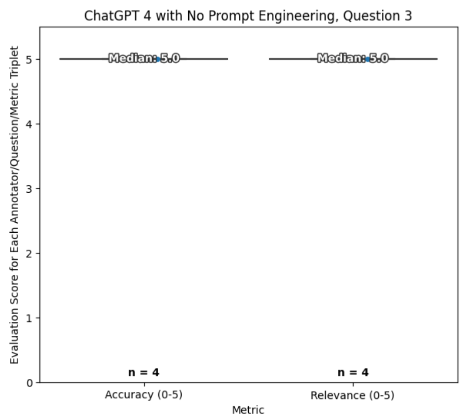
Annotator Response: "[Actionability is not] really applicable because the question was on what a VPN is and safety, *not asking for advice/steps*"

Fig. 3: An example of a question and answer that one or more annotators labeled as inherently non-actionable

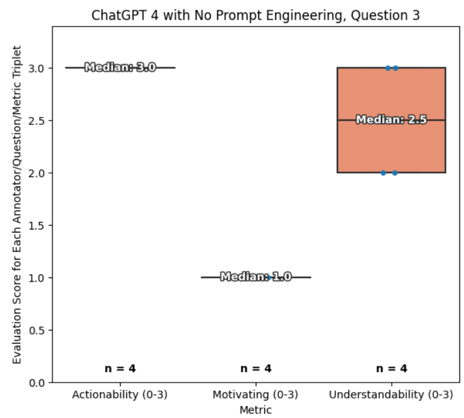
with a normalized score average of 0.3333 (or a score of 1 out of 3), which helps explain ChatGPT's lower performance on that metric. Figure 4 shows the distribution of human annotations for this question. The annotators agreed that the answer generated by ChatGPT for Q3 did not motivate users to set up 2FA, and lacked an adequate discussion of risks associated with not adopting 2FA. The ChatGPT4 answer to that question is shown in Figure 4. The only risks that ChatGPT 4 highlights in this answer is "suspicious activity", which our annotators deemed insufficient in motivating users.

Question 3: I am trying to create a new email account with Gmail for a side business I am about to start. Google is asking me all sorts of questions and is also requesting my phone number. I am a little concerned about giving them my phone number. What should I do?

ChatGPT 4 Answer: Google asks for your phone number as a security measure. It's used for two-factor authentication and allows Google to contact you if there's suspicious activity on your account. However, if you're uncomfortable providing your personal number, you could consider using a secondary phone number. There are several apps available that can provide you with a temporary number for verification purposes. Alternatively, you could use a landline number if you have one. Remember, it's important to ensure the number you use is accessible to you in case you need to recover your account in the future.



(a) Metrics from (0-5)



(b) Metrics from (0-3)

Fig. 4: Distribution of human annotations for ChatGPT 4's (unprompted) answer to Question 3. Each blue dot represents an individual annotator's rating of the answer according to a particular metric. The number of annotations for each metric is denoted by **n**=# at the bottom of the plot. Not all blue dots are visible, as a number of dots have ratings equal to the median rating, in which case the blue dots are masked by the word "Median".

Overall, these initial results are encouraging as they suggest that chatbots may generally be able to produce accurate answers to common cybersecurity questions. At the same time they also suggest that, despite being accurate, default answers from chatbots may not be very effective, as they may lack in understandability, actionability and motivating power. These findings motivated us to explore prompt engineering techniques designed to help overcome these limitations and produce more effective answers.

4 Effects of Prompt Engineering on Answer Effectiveness

4.1 Design of Prompts

We proceeded to create four different prompts, each designed to enhance an effectiveness dimension other than accuracy. Our objective was to see whether we could coax the chatbot to produce answers that would improve performance along each of these other four dimensions, namely relevance, motivating power, actionability and understandability, without negatively impacting accuracy, which was already high for both ChatGPT4 and Llama2 (unprompted). The prompts were intentionally kept simple and were not tested until they were finalized. The prompts we developed for each dimension are shown in Table 1.

Table 1: Overview of Prompts used in the study

Metric	Prompt
Relevance	<Baseline Question>. In answering this question, please do not limit yourself to a literal interpretation of my question, instead please let me know about other relevant considerations I may not be aware of and tell me why they matter.
Motivating	<Baseline 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.
Actionability	<Baseline 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 give me enough details about what I should do and how, using language I am likely to understand.
Understandability	<Baseline Question>. In answering this question, please keep in mind that I am not a technical expert. Please make sure to provide enough details and remember that my technical expertise is minimal.

4.2 Human Annotation Results

From our human annotations for each of the answers provided by ChatGPT 4 and Llama 2 to each of the seven questions, we determine the effects of prompt engineering on the effectiveness of the chatbots' answers. Table 2 lists the complete results from our statistical analysis for ChatGPT, highlighting metrics for which we noticed a statistically significant increase in human annotator scores over the baseline ($\alpha = 0.05$), split by prompt type. Our analysis of Llama 2 with prompt engineering found that the increases in scores with all four prompts were not statistically significant, and ultimately did not perform as well as ChatGPT. Across all model/prompt/metric triples, only one decreased by a statistically significant amount, the Accuracy metric with the Actionability prompt for Llama 2 (normalized baseline score = 0.9429, normalized score with Actionability prompt engineering = 0.8024, $p = 0.0112$).

Notably, we find that three of the four prompts (the Motivating, Relevance, and Understandability prompts) improved the Motivating metric score for answers generated by ChatGPT 4 by a statistically significant amount. Furthermore, three of the four prompts (the Actionability, Motivating, and Understandability prompts) significantly increased the Understandability metric for ChatGPT 4.

Additionally, the Motivating prompt had the highest average percent change, increasing the five metrics by an average of 15.28%. Overall, the ChatGPT 4 answers improved with prompt engineering. For Llama 2, the Actionability and Relevance prompts caused a net negative percent change in human annotator scores (-0.98% and -0.41% respectively). In Figure 6, we examine the differences in performance between prompts for ChatGPT. We find that the differences in performance for each prompt are often not statistically significant, barring four exceptions. The Motivating and Relevance prompts significantly outperform the Actionability and Understandability prompts for the Motivating metric.

We also looked specifically at Q3, which we examined above because of the noticeably low score in the Motivating metric for ChatGPT 4's answer. Following prompt engineering (specifically with the Motivating prompt), we observe a large increase in human annotator scores for the Motivating metric for ChatGPT 4's answer to Q3. In fact, the answer has a perfect 1.000 rating not only for the Motivating metric, but for all five grading metrics with the Motivating prompt. Figure 7 shows the full answer text for Question 3 after prompting ChatGPT with the Motivating prompt. The annotators in particular highlighted that the answer with the Motivating prompt was much more thorough and included discussion about scenarios such as "[forgetting] your password" or "unusual activity", further detailing that if "your account is compromised", it will be "difficult to regain access" without adding a phone number for "recovery and verification" purposes.

Q6 is another example where the prompt engineering dramatically improved the answer quality for ChatGPT 4, particularly with the Motivating and Understandability prompts. The answer from ChatGPT 4 with no prompting received a normalized score of 0.5 for the Motivating metric from our annotators, while the answers for both the Motivating and Understandability prompts received 1.0 scores. Much like for Q3, the prompt-engineered answers have perfect 1.000 normalized ratings across all five metrics. The question text, unprompted answer, and answers with prompt engineering are included in Figure 5. In this case, though both answers with prompt engineering achieved high quantitative scores, we observe that the answer with the Motivating prompt may still be stronger qualitatively than the answer with the Understandbility prompt. This indicates that though both answers score equally, the Motivating prompt may still be a better choice overall.

5 Automated Answer Evaluation with LLMs

While we evaluated all prompts and changes in performance with respect to human evaluations, human annotation is costly and time-consuming to perform on a large scale. Therefore, we also investigated whether ChatGPT 4 and Llama 2 themselves could replace human annotators when evaluating the quality of the answers in each of the five metrics. We prompted both ChatGPT 4 and Llama 2 with the same annotation guidelines as provided to the human annotators, along with the answer for each question/prompt type pair, asking ChatGPT 4 and LLama 2 to rate the answers according to our metrics. Because our objective here is to evaluate the ability of LLMs to rate answers, we used the entire corpus of answers we had available, namely answers generated with and without prompts by both ChatGPT and Llama to all 7 questions.

To compare the LLM-generated ratings with the human annotator ratings, we normalized all the scores and computed the Euclidean distance between the

Question 6: For the past year or so, my cell phone has been recommending that I upgrade my software. My phone seems to be working fine, and I never seem to have the time. Am I ok?

ChatGPT 4 Answer with No Prompting: While your phone may seem to be working fine, it's still important to regularly update your software. These updates often include important security patches that protect your device from new threats. Not updating your software can leave your device vulnerable to these threats. Additionally, updates often include new features and improvements that can enhance the performance of your device. So, while it might be inconvenient, it's generally a good idea to take the time to install these updates.

ChatGPT 4 Answer with *Motivating* Prompt: Upgrading your phone's software is important for several reasons. Firstly, software updates often include patches for security vulnerabilities that have been discovered since the last update. By not updating, you're potentially leaving your phone open to hackers who could exploit these vulnerabilities to steal your personal information, install malware, or otherwise harm your device. Secondly, updates often include improvements to the software's performance and new features that can enhance your user experience. While it's true that your phone may seem to be working fine without the updates, it's also possible that you're not noticing some of the issues that the updates are designed to fix. The risks of not updating your software can be significant. If a hacker were to gain access to your phone, they could potentially access your personal information, including your contacts, photos, and even sensitive information like passwords or credit card numbers. In terms of how critical it is, it's generally recommended to always keep your software up-to-date. While it may seem like a hassle, it's a relatively easy step to take to protect your privacy and the security of your device. If you're worried about the time it takes to update, consider setting your phone to update automatically overnight, or at another time when you're not using it. This way, you can keep your phone's software up-to-date without it interfering with your daily use.

ChatGPT 4 Answer with *Understandability* Prompt: Yes, your phone can continue to work fine without software updates. However, it's important to understand that these updates often include security patches that fix vulnerabilities in your phone's software. By not updating, you're potentially leaving your phone open to various types of cyber attacks. Think of it like this: your phone is a house and the software updates are like repairs and improvements. If you ignore a leak in the roof because the house is still standing, eventually the damage could become much worse. Similarly, if you ignore software updates, your phone could become more vulnerable over time. Updating your phone's software also often includes new features and improvements to existing ones, which can enhance your overall user experience. The process is usually straightforward and automated. You just need to follow the prompts on your screen. It's best to do this when you have some free time as the phone might need to restart and be unusable for a short while during the update. In conclusion, while it's not mandatory, it's highly recommended to keep your phone's software up to date for security reasons and to ensure optimal performance.

Fig. 5: Additional examples of answer improvements following prompt engineering for ChatGPT 4.

LLM-generated scores (as 5-dimensional vectors, with each dimension representing the score of one of the five grading metrics) and the human annotator scores (as 5-dimensional vectors, with each dimension representing the normalized and averaged scores by human annotators for that particular metric). We also then averaged the scores from the LLMs and human annotators, and calculated the percent error for each individual metric.

5.1 Results and Analysis

We found that the average Euclidean distance between normalized (from 0 to 1) LLM evaluations and human annotator evaluations is 0.3211. However, the models have a slight variance in performance depending on which model is evaluating which model's answers. We observe that when ChatGPT 4 evaluates its own answers, the average Euclidean distance is slightly lower than the three

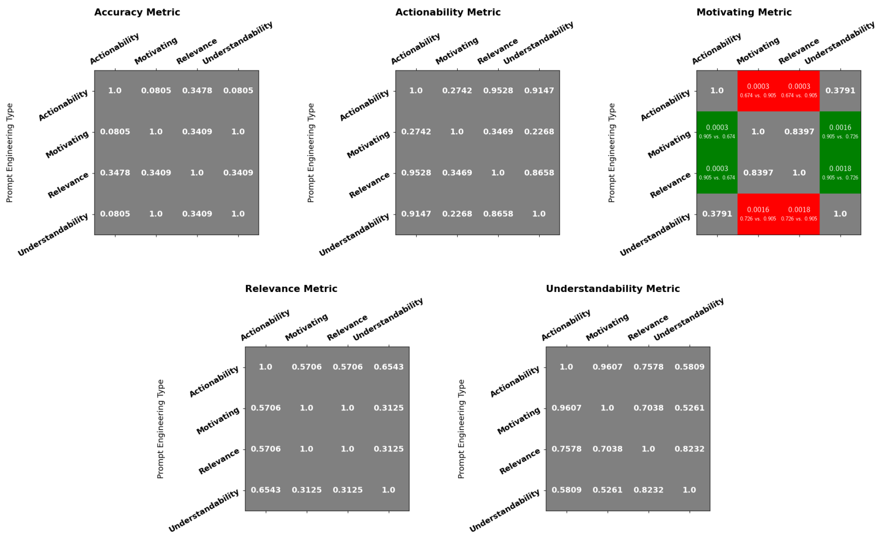


Fig. 6: Matrices showing whether the differences in performance for each metric, based on prompt type, are statistically significant. The value in each cell represents the p-value for a Mann-Whitney U significance test between the distributions of scores for each prompt pair. Prompt pairs that have a statistically significant difference are shown in green (increase) and red (decrease). The comparison is done row-by-column (a green cell indicates that the i-th row's prompt significantly outperforms the j-th column's prompt). If shown, the second row of text in the cells indicates the normalized averaged score of the row's prompt vs. the normalized averaged score of the column's prompt.

s

Question 3: I am trying to create a new email account with Gmail for a side business I am about to start. Google is asking me all sorts of questions and is also requesting my phone number. I am a little concerned about giving them my phone number. What should I do?

ChatGPT 4 Answer: Google asks for your phone number as a security measure. It's used for two main purposes: recovery and verification. If you forget your password or if Google detects unusual activity, they can send a verification code to your phone to ensure it's really you. This is known as two-factor authentication and it adds an extra layer of security to your account. However, if you're uncomfortable providing your phone number, you can opt for other methods of verification. Google also offers the option to answer security questions or use an alternate email address for account recovery. The risk of not providing a phone number is that if you forget your password or if your account is compromised, it may be more difficult to regain access. However, as long as you have another method of account recovery set up and you keep that information secure and up-to-date, you should still be able to recover your account if necessary. Remember, the key to online security is to use strong, unique passwords and to keep your recovery information secure. If you're concerned about privacy, be sure to read Google's privacy policy so you understand how your information is used and protected.

Fig. 7: Full answer text generated by ChatGPT with **Motivating** prompt engineering for Question 3. The above answer scored 1.000 on all five metrics.

Prompt Type	**Model**	**Metric**	**Baseline Normalized Score**	**Normalized Score After Prompt Engineering**	**Statistical Significance (p-value)**
Actionability Prompt	ChatGPT 4 (Average % Change = +8.83%)	Accuracy	0.9786	0.9714 (-0.74%)	0.3060
		Actionability	0.7817	0.8976 (+14.83%)	0.1199
		Motivating	0.5952	0.6738 (+13.21%)	0.2595
		Relevance	0.9857	0.9771 (-0.87%)	1.0
		Understandability	0.7976	0.9310 (16.73%)	0.0305
Motivating Prompt	ChatGPT 4 (Average % Change = +15.28%)	Accuracy	0.9786	1.0 (+2.19%)	0.3293
		Actionability	0.7817	0.8095 (+3.56%)	0.6415
		Motivating	0.5952	0.9048 (+52.02%)	$2.247*10^{-7}$
		Relevance	0.9857	0.9929 (+0.73%)	0.5615
		Understandability	0.7976	0.9405 (+17.92%)	0.0135
Relevance Prompt	ChatGPT 4 (Average % Change = +14.45%)	Accuracy	0.9786	0.9786 (+0.00%)	0.6580
		Actionability	0.7817	0.8175 (+4.58%)	0.2256
		Motivating	0.5952	0.9048 (+52.02%)	$7.825*10^{-8}$
		Relevance	0.9857	0.9929 (+0.73%)	0.5615
		Understandability	0.7976	0.9167 (+14.93%)	0.0505
Understandability Prompt	ChatGPT 4 (Average % Change = +9.58%)	Accuracy	0.9786	0.9786 (+0.00%)	0.3293
		Actionability	0.7817	0.8730 (+11.68%)	0.0642
		Motivating	0.5952	0.7262 (+22.01%)	0.0278
		Relevance	0.9857	0.9786 (-0.72%)	0.6467
		Understandability	0.7976	0.9167 (+14.93%)	0.0420

Table 2: Results of prompt engineering for ChatGPT 4 shown as raw normalized score increases, percent change, and significance values for every prompt/model/metric triple. Rows highlighted green are metrics that significantly increased for the respective prompt type and model with $\alpha = 0.05$.

other combinations (ChatGPT 4 evaluating Llama 2 answers, and Llama 2 evaluating either ChatGPT 4 answers or its own answers). These average distances are shown in Table 3.

All four of the LLM self-evaluation pairs have more favorable scores than the human evaluations. Specifically, across the four LLM/LLM pairs, seven questions, and five metrics, there were a total of 140 annotations, 118 of which were higher than the corresponding human evaluation (84.29%).

Model Performing Evaluation	**Model That Generated Answer**	**Average Euclidean Distance**
ChatGPT 4	ChatGPT 4	0.305
ChatGPT 4	Llama 2	0.312
Llama 2	Llama 2	0.323
Llama 2	ChatGPT 4	0.345

Table 3: The average Euclidean distance between normalized LLM answers and normalized and averaged human annotator answers, separated by the model that evaluated the answer and the model that generated the answer.

We also calculated the percent error between the LLM self-evaluations and the human annotator scores for each metric. These values are shown in Table

4. The above results appear to indicate that while the chatbots may be able to evaluate their own answers to some degree, they may lack the ability to critically parse the answers with respect to the five metrics - a variable that may cause higher average scores when compared with human annotations.

Metric	**Percent Error**
Accuracy	+6.200%
Actionability	+17.365%
Motivating	+4.841%
Relevance	+2.018%
Understandability	+9.279%

Table 4: Percent error for LLM self-evaluation, for each metric. The metric-level percent error is determined by first calculating the percent error between the averaged normalized human annotator scores and averaged normalized LLM scores, for each individual answer/metric pair. Then, these individual scores were averaged across each metric to obtain the average metric-level percent error. The positive errors indicate that on average, the LLMs scored the answers higher than the corresponding human annotations.

6 Discussion and Future Work

While chatbots like ChatGPT can provide accurate answers to user security questions, their answers often fall short in areas like actionability and how well they motivate users to take action. With the results from our annotators, we found that prompt engineering is successful at improving the effectiveness of answers in several different aspects, particularly in how well the answers motivate the users to act and how understandable the answers are to diverse groups of users.

Ultimately, beyond this initial study, we would like to evaluate (and most likely refine) our prompts in the context of in-situ studies, in which representative sets of participants ask cybersecurity questions as they go about their regular day-to-day activities. Future work may explore other prompting techniques, including solutions that personalize prompts to generate answers that are best tailored to the needs of individual users (e.g., their technical sophistication or attitude towards risk). Rather than relying on third party annotators, namely annotators who are not the people who asked the questions, we would want to rely on users themselves to help us better evaluate the effectiveness of the answers they received, including looking at whether they felt they understood the answers, found them actionable and were swayed to actually follow their recommendations - thereby producing desired changes in user behavior. Conducting such an in-situ study would give us access to a more diverse set of questions users have as they go about their daily activities and a more accurate evaluation of the answers they received. In particular, we acknowledge that all the annotators used in the present study had a technical background and that, while the guidance they were given requested them to put themselves in the shoes of different personas (e.g., different levels of technical sophistication and also different age groups), their evaluation is to be interpreted as an approximation of how these different personas would likely rate the answers.

7 Conclusion

Humans are often the weakest link in cybersecurity. As more users turn to chatbots to answer everyday questions, including those related to cybersecurity, providing effective answers to these questions has become increasingly important. In this work, we investigated the use of chatbots to generate answers to seven cybersecurity questions covering important areas with which everyday users often struggle. We found that chatbots are generally able to provide accurate and relevant answers. Yet in other respects these answers do not appear to always be as effective as one would like. We found that prompt engineering can help generate answers more effective answers, namely answers that are more understandable, have greater motivating power, and are more actionable. In particular, our experiementation with a motivating prompt significantly increased the motivating power and understandability of answers generated by ChatGPT, without sacrificing relevance or accuracy. These initial results suggest that further refinement of LLM training and/or prompt engineering may help generate more effective answers to everyday cybersecurity questions. We present this work as an initial study into the capabilities of LLMs to answer everyday cybersecurity questions, and hope to facilitate deeper investigation into making automated answers more effective for a wide array of users and questions.

Acknowledgments. 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)

References

1. Together ai api, `https://www.together.ai/`
2. Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L.F., Komanduri, S., Leon, P.G., Sadeh, N., Schaub, F., Sleeper, M., Wang, Y., Wilson, S.: Nudges for privacy and security: Understanding and assisting users' choices online. ACM Comput. Surv. **50**(3) (Aug 2017). `https://doi.org/10.1145/3054926`, `https://doi.org/10.1145/3054926`

3. Forum, W.E.: The Global Risks Report 2022 (2022), https://www.weforum.org/publications/global-risks-report-2022/
4. Greitzer, F.L.: Situated Usability Testing for Security Systems (Feb 2011), https://www.pnnl.gov/main/publications/external/technical{_}reports/PNNL-20201.pdf
5. Hancock, J.: Psychology of Human Error (2022), https://www.tessian.com/resources/psychology-of-human-error-2022/
6. Hylender, C.D., Langlois, P., Pinto, A., Widup, S.: 2024 Data Breach Investigations Report (2024), https://www.verizon.com/business/resources/reports/dbir/
7. Institute, P.: The State of IoT/OT Cybersecurity in the Enterprise (Nov 2021), https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/2021-ponemon-research-the-state-iot-and-ot-cybersecurity-in-the-enterprise-final.pdf
8. Proofpoint: Proofpoint security education training, https://cmu.securityeducation.com/my-training
9. Story, P., Smullen, D., Chen, R., Yao, Y., Acquisti, A., Cranor, L.F., Sadeh, N., Schaub, F.: Increasing adoption of tor browser using informational and planning nudges. Proceedings on Privacy Enhancing Technologies (2022). https://doi.org/10.2478/popets-2022-0040

Appendix

Table 5: Annotation Guidelines Provided to Annotators

Category	**Instructions**
Accuracy (0 - 5)	Is all information in the answer correct? We distinguish inaccuracies by the effect they have on the overall advice provided. **Scoring:** 5: fully accurate; 4: one major inaccuracy; 3: one major inaccuracy and one minor inaccuracy; 1: one major inaccuracy and several minor inaccuracies; 0: more than 1 major inaccuracy
Relevance (0 - 5)	How well does the answer address the specific question considering both "literal" relevance and the most relevant answer one can give the user. **Scoring:** 5: addresses all key issues raised by question; 4: pretty close; 3: tangentially relevant but fails to address one or more key issue(s) raised by the question; 1 or 2: peripherally relevant but does not really address the user's key question; 0: answer has nothing to do with user question.
Motivating (0 - 3)	How well does the answer explain relevant risks in a way that is likely to motivate the user? **Scoring:** 3: Identifies risk(s) well *and* motivates user by highlighting threats; 2: identifies risk(s) and urges user to take action; 1: Briefly alludes to risks; 0: Does not discuss risks at all.
Understandability (0 - 3)	**Scoring:** Considering the 3 personas identified below, how many of the personas are likely to understand? (based on your own best judgment)
Actionability (0 - 3)	Determine whether the answer provides practical and actionable advice. **Scoring:** Considering the 3 personas identified below, for how many does this answer provide sufficient information for them to follow the advice and take the necessary actions? (if applicable)

Continued on next page

Table 5: Annotation Guidelines Provided to Annotators

Category	Instructions
Personas for Understandability and Actionability Evaluation: (1) Jane, a non-tech savvy 40 year old person with a high school degree. She uses some technology but does not understand how they work or the risks associated with their use. (2) Alex, a 28-year old tech-savvy person with a bachelor's degree in IT. He stays updated on emerging technologies and their applications and understands them well. (3) Mr. Johnson, a non-tech savvy 70-year old person. He struggles with navigating the internet and using his smartphone.	

Table 6: Overview of questions used in the study

Topic	Question
Identity Theft/ Phishing	1. I received what seemed to be an ad, offering a free MacBook laptop to anyone completing a survey. I filled out the form but now realize this was probably a scam. What should I do?
VPN	2. My friend recently recommended I use a VPN to override the geographical restrictions for some of the shows available on Netflix. What is a VPN and is it safe to use?
Dual Factor Authentication (2FA)	3. I am trying to create a new email account with Gmail for a side business I am about to start. Google is asking me all sorts of questions and is also requesting my phone number. I am a little concerned about giving them my phone number. What should I do?
Security on the Web	4. I opened up a website that goes over popular baking recipes. On the website, I got a pop-up in my browser asking me to accept cookies. Should I say yes, and what does this mean?
Privacy in Public Spaces	5. I'm trying to save on my phone bill and have been told to connect to public Wi-Fi networks as often as possible. Is this good advice?
Mobile/ Phone Security	6. For the past year or so, my cell phone has been recommending that I upgrade my software. My phone seems to be working fine, and I never seem to have the time. Am I ok?
Internet of Things	7. Many of my neighbors have Ring doorbells, including those that live right next door to me. How concerned should I be about what I say and do both inside and outside of my home?