



# Can Large Language Models Provide Security & Privacy Advice? Measuring the Ability of LLMs to Refute Misconceptions

Yufan Chen\* chen4076@purdue.edu Purdue University West Lafayette, IN, USA Arjun Arunasalam\* aarunasa@purdue.edu Purdue University West Lafayette, IN, USA

#### ABSTRACT

Users seek security & privacy (S&P) advice from online resources, including trusted websites and content-sharing platforms. These resources help users understand S&P technologies and tools and suggest actionable strategies. Large Language Models (LLMs) have recently emerged as trusted information sources. However, their accuracy and correctness have been called into question. Prior research has outlined the shortcomings of LLMs in answering multiple-choice questions and user ability to inadvertently circumvent model restrictions (e.g., to produce toxic content). Yet, the ability of LLMs to provide reliable S&P advice is not well-explored.

In this paper, we measure their ability to refute popular S&P misconceptions that the general public holds. We first study recent academic literature to curate a dataset of over a hundred S&P-related misconceptions across six different topics. We then query two popular LLMs (Bard and ChatGPT) and develop a labeling guide to evaluate their responses to these misconceptions. To comprehensively evaluate their responses, we further apply three strategies: query each misconception multiple times, generate and query their paraphrases, and solicit source URLs of the responses.

Both models demonstrate, on average, a 21.3% non-negligible error rate, incorrectly supporting popular S&P misconceptions. The error rate increases to 32.6% when we repeatedly query LLMs with the same or paraphrased misconceptions. We also expose that models may partially support a misconception or remain noncommittal, refusing a firm stance on misconceptions. Our exploration of information sources for responses revealed that LLMs are susceptible to providing invalid URLs (21.2% for Bard and 67.7% for ChatGPT) or point to unrelated sources (44.2% returned by Bard and 18.3% by ChatGPT). Our findings highlight that existing LLMs are not completely reliable for S&P advice and motivate future work in understanding how users can better interact with this technology.

# CCS CONCEPTS

Security and privacy → Social aspects of security and privacy;
Human-centered computing → Human computer interaction (HCI).

\*The authors Chen and Arunasalam contributed equally to this research.

# $\odot$

This work is licensed under a Creative Commons Attribution International 4.0 License.

ACSAC '23, December 04–08, 2023, Austin, TX, USA © 2023 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-0886-2/23/12. https://doi.org/10.1145/3627106.3627196

#### **KEYWORDS**

Large language models, security and privacy advice, misconception

Z. Berkay Celik

zcelik@purdue.edu

Purdue University

West Lafayette, IN, USA

#### **ACM Reference Format:**

Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. 2023. Can Large Language Models Provide Security & Privacy Advice? Measuring the Ability of LLMs to Refute Misconceptions. In *Annual Computer Security Applications Conference (ACSAC '23), December 04–08, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3627106.3627196

#### **1** INTRODUCTION

In recent years, large language models (LLMs) have emerged as the most prominent technology in natural language processing (NLP). These models, trained on vast amounts of data, possess rich embedded knowledge, allowing them to be easily applied to various downstream NLP tasks such as sentiment analysis, code generation, and question answering with minimal fine-tuning [30, 33, 65].

The advent of LLM agents such as ChatGPT [2] and Bard [1] has resulted in widespread public interaction with LLMs, with end users leveraging web interfaces to engage with these powerful AI tools. These chatbots are trained to interact with users in a conversational style and generate answers to users' questions by retrieving information from the models themselves.

The usability of these web interfaces has caused LLMs to gain great popularity and resulted in widespread use among the general public in a short timeframe. These LLMs provide a single interface for users to interact with, contrasting conventional search engines which require users to traverse various web pages. Thus, LLMs have emerged as a new trusted source of information. Today, people interact with the LLMs to obtain information regarding health [14], stock market advice [18], and even job interview guidance [16].

As LLMs gain further footing in users' everyday lives, they become prominent information sources in popular domains, including user security and privacy (S&P) advice. Traditionally, people receive S&P advice from friends, family, and online sources (e.g., forums, social media, IT websites) [42, 53]. However, recent LLM popularity as a trusted resource, in lieu of search engines, provides prime conditions for its adoption as an S&P advice tool.

Concerningly, so far, their ability to provide quality expert S&P advice has not been well-evaluated. Previous research has recently revealed that LLMs such as ChatGPT can generate untruthful information, hallucination (e.g., provoking fake news articles or academic papers), and even toxic content [30, 37, 57, 59, 72]. This raises concerns, as lay users lacking expertise may easily trust and be misled by the generated falsehoods as the information LLMs provide is often written proficiently and convincingly. Extending to this, improper S&P advice can have severe consequences. For instance, users may be misled into believing specific strategies (e.g., reusing

strong passwords) are secure and privacy-preserving when they are not (and vice versa). Similarly, inaccurate information on S&P tools/technologies may lead to erroneous decisions (e.g., exclusion due to lack of perceived S&P or adoption due to being overpromised of their S&P capabilities). Therefore, it is crucial to thoroughly evaluate LLMs' performance in providing reliable S&P advice.

In this paper, to address this, we empirically aim to answer the following research question,

# Are LLMs reliable in providing S&P advice by correctly refuting user-held S&P-related misconceptions?

Our approach to answering this research question begins with an extensive literature survey on user-held S&P misconceptions. We query Google Scholar [11] with a list of S&P topics and userrelated keywords (e.g., "folk models", "user perception") to retrieve academic studies that address people's misconceptions about security and privacy technologies and strategies. We collect a total of ~400 academic manuscripts through this method. After removing irrelevant manuscripts, we extract over 500 S&P misconceptions, remove redundancies, and produce a corpus of 122 S&P publicly held misconceptions. These misconceptions are grouped into six different S&P topics, e.g., Web Security and Privacy, IoT/CPS.

Using our dataset, we evaluate the ability of LLMs to refute these misconceptions. We choose ChatGPT and Bard as they are among the most well-known and influential LLMs available to the public. We conduct four experiments designed to understand LLM's overall correctness, consistency, susceptibility to paraphrasing, and reliability in providing sources to refute misconceptions. First, we query each misconception once to evaluate the general effectiveness of models in refuting misconceptions. Second, we perform repeated queries (four times) for each misconception to assess LLMs' consistency in maintaining their stance towards a misconception. Third, we leverage paraphrasing tools to generate four paraphrases for each misconception and ask the models once for each paraphrase, aiming to simulate the real-world scenario in which people may query the same misconception differently. Lastly, we query for the URL sources that influence the models' responses and evaluate their validity and reliability (whether the URLs exist and what information the websites they direct to provide).

After collecting all the responses, we develop a labeling guide to categorize them according to their stances. The categories include *support, negate, partially support, noncommittal* and *unrelated*. Since all entries in our dataset are misconceptions, we consider *negate* as the only correct answer and *support* as the incorrect answer. We analyze the source URLs using the Python HTTP library requests [25] and Wayback Machine API [27] to determine URL validity. We further analyze domain relevance for each URL to verify if the website it points to is relevant to the misconception.

Our experiments reveal that, (1) ChatGPT and Bard both exhibit non-negligible error rates that averaged to be 21.3%. (2) When repeatedly queried with the same question, Bard performs better than ChatGPT in keeping consistent in its answers. However, Chat-GPT and Bard both show an increase in error rate (by 10.6% and 4.1%) and a decrease in correctness (by 12.3% and 8.2%). (3) When presented with paraphrased questions, both models experience an additional increase in error rate (by 6.6% and 9% for ChatGPT and Bard) compared to those of repeated queries. Additionally, Bard's advantage in maintaining consistency diminishes, with both models providing at least two different stances for an average of 44.7% of the misconceptions. When analyzing models' ability in refuting misconceptions across different categories, it is observed that (4) ChatGPT and Bard perform poorly in S&P law and regulation category, with error rates exceeding 40% in almost all experiments.

In our source analysis, it is found that (5) Bard is more likely to provide valid URLs (78.8% validity) compared to ChatGPT, which generates mostly forged URLs (32.3% validity). 44.2% of Bard and 18.3% of ChatGPT's valid URLs are unrelated to the misconception's domain, e.g., a response to a VPN misconception provides a URL to a website that does not mention VPNs. We further analyze a sample of sources provided when the LLMs erroneously support misconceptions. We found that an average of 10% of the valid URLs attributed to incorrect answers contain false information, while an average of 25% contains correct information that debunks the misconception, yet ChatGPT and Bard fail to recognize it.

Our empirical findings highlight how existing state-of-the-art LLMs are not completely reliable in providing accurate S&P advice or sound sources. Our study highlights the necessity for future work on how users interact with LLMs as a trusted information source and the need to further examine LLM's ability to provide expert advice in different domains.

In this work, we make the following contributions:

- We curated a dataset of over a hundred S&P misconceptions held by users through a comprehensive literature survey.
- We extensively evaluated two popular LLMs' ability to provide S&P advice by measuring their correctness, consistency, and susceptibility to paraphrasing.
- We analyzed URLs that LLMs provide when justifying their stance towards misconceptions, demonstrating their inability to provide reliable sources for S&P information.

Our artifacts, which include our misconception dataset, labeling guide, and results, are made available [24].

#### 2 BACKGROUND AND MOTIVATION

Large Language Models. LLMs are neural network models with billions (or more) parameters. These models are trained on a large corpus of internet-sourced data, including textual information and conversation data. Such training enables LLMs to generate human-like language and exhibit capabilities in zero-shot and few-shot learning. With minimal fine-tuning, LLMs can be easily applied to various NLP tasks such as sentiment analysis, code generation, question answering, and, more recently, engaging in coherent conversations with users [33, 44, 64].

ChatGPT and Bard have emerged as two of the most well-known and widely-used LLMs. ChatGPT is a conversational variant of InstructGPT [45] and has been fine-tuned using Reinforcement Learning with Human Feedback (RLHF) [44]. Bard is a conversational AI service based on Google's Language Model for Dialogue Applications (LaMDA) [17, 62]. Both models have garnered significant attention and are recognized for their effectiveness.

Due to their extensive training, LLMs possess remarkable capabilities in answering a wide array of questions. Consequently, users



Figure 1: Overview of our methodology in understanding the ability of LLMs to refute S&P misconceptions.

increasingly rely on LLMs as sources of information. LLMs are leveraged to assist users with daily activities such as writing code [5] and answering homework questions [6], to critical events/tasks such as job interviews [21] and purchasing stocks [18].

Users who previously turned to online avenues such as conventional search engines (e.g., Google and DuckDuckGo) as advice sources have recently pivoted to LLMs [9]. LLMs provide users with an interface where they can clarify doubts on topics ranging from health [10] to relationship advice [20] without visiting multiple websites. These topics will likely grow, encompassing niche areas where users previously sought information/advice online.

**Security and Privacy Advice.** In an effort to protect themselves from various S&P threats, people often seek S&P advice from offline interactions with friends and family and online spaces such as forums, social media, and search engines. Here, users learn about password security and how to protect themselves from malware and may even educate themselves on S&P tools/technologies such as malware, privacy, blockchain, and VPNs [54].

The quality of S&P advice, however, can be concerning, mainly when sourced from online resources that are not vetted for accuracy. Advice pointers may be unactionable or lack clear priorities. This may make it difficult for individuals to determine which advice to follow and thus prevent users from effectively implementing security and privacy-enhancing strategies [54].

S&P advice may additionally be inaccurate and categorically false, and it may lead to implementing insecure and privacy compromising suggestions. As LLMs emerge as new trusted sources of information, people may increasingly rely on them for assistance with their S&P concerns or understanding of S&P concepts/tools. Therefore, it becomes crucial to evaluate the effectiveness of existing LLMs in providing advice in these domains to ascertain their reliability as sources of information.

#### 3 METHODOLOGY

We focus on empirically answering the following research question: Are LLMs reliable in providing S&P advice by correctly refuting S&P-related misconceptions? Figure 1 presents an overview of how we address this research question. First, we curate a dataset of S&P misconceptions the general public holds. To do so, we extensively study the literature on user misconceptions about S&P topics and curate a dataset that contains diverse misconceptions covering a variety of topics, such as blockchain and malware. Following a prompt template, we query two popular LLMs (ChatGPT and Bard) to collect data on how these LLMs respond to S&P misconceptions. We then design a four-experiment approach to evaluate their ability to refute common S&P misconceptions.

In an initial experiment, we query each LLM with all misconceptions in our dataset. Then for each misconception, we repeat our queries to obtain a total of five responses each. Thereafter, we leverage paraphrasing tools to generate semantically similar sentences for each misconception and repeat our queries. Lastly, we prompt LLMs for additional information, specifically asking them for the sources that informed their responses.

After responses are collected, two authors generate a labeling guide to label LLMs' responses independently and categorize them into five distinct groups (*support, negate, partially support, noncommittal, unrelated*), and reconcile differences. We also analyze source URLs returned by the LLMs, leveraging both the Python HTTP library requests [25] and Wayback Machine API [27] to verify URL's validity before further grouping URLs into one of three groups (*relevant, marginally relevant, irrelevant*). In the following, we detail each stage and present our findings in Section 4. We provide our replication package, which includes our dataset and all supplementary material necessary for reproducing our results [24].

### 3.1 S&P Misconception Dataset Generation

We conducted an extensive literature survey on existing academic studies that detail general users' misconceptions about S&P topics. To do so, we first generated a comprehensive query list to query Google Scholar [11] using the template "{domain} {user-related keywords}". For the "{domain}" portion of our query, we identified a comprehensive list of 14 S&P topics, e.g., malware, web privacy, and cyber-physical systems (such as autonomous driving). For the "{user-related keywords}" portion, we used keywords that would encourage retrieval of studies focused on user perception, including keywords "user study", "mental model", "folk model", and "user perception". For instance, one query that we used in our retrieval was "web privacy user perception."

Overall, we generated 56 queries with various combinations of "{domain}" and "{user-related keywords}". Leveraging our query list, we gathered approximately ~400 academic manuscripts consisting of diverse sources such as conference papers, journals, book chapters, and articles.

**Misconceptions Extraction and Filtering.** Two authors use a labeling guide to independently label manuscripts as relevant or irrelevant to S&P misconceptions. First, authors read manuscript titles and remove unrelated papers. They then read the abstract and introduction for further filtration. The authors marked a manuscript relevant if (1) it discusses S&P related misconceptions and (2) notes that the general public holds these misconceptions. Authors met to reconcile differences and achieved high agreement, Cohen's Kappa [36],  $\kappa > 0.8$ , before reconciling.

After reconciling, we extracted ~500 misconceptions. However, after our initial analysis, we found that some misconceptions we extracted overlap. To address this, we manually filtered out duplicates to eliminate redundancy, resulting in a final dataset comprising 122 distinct misconceptions. We note that our misconception dataset

Category	# Misconceptions	Example Misconceptions	
Crypto and Blockchain	25	Every transaction on the blockchain is anonymous.	
IoT/CPS	17	In smart homes, only devices I actively interact with are able to collect data about me (e.g., the doorbell).	
Law and Regulation	14	Under GDPR, when relying on consent to process personal data, consent must be explicit.	
Malware and Device Security	21	My PC or network cannot be harmed by my visiting a website, if I don't download anything.	
Privacy and Anonymity Tools	25	Employers would be unable to track employees when they used private mode.	
Web Security and Privacy	20	Websites that use HTTPS are trustworthy.	

Table 1: Overview of our S&P misconception dataset, which is categorized into one of six categories.

focused on recent publications; 121 misconceptions were from publications between 2014 – 2023, 1 from 2006. Our full list of keywords, labeling guides, and misconception dataset can be found in our Github project repository [24].

Table 1 presents an overview of our 122 misconceptions, spanning six categories: (1) Crypto and Blockchain, (2) IoT/CPS, (3) Law and Regulation, (4) Malware and Device Security, (5) Privacy and Anonymity Tools, and (6) Web Security and Privacy.

#### 3.2 LLM Response Generation

**Model Selection.** We focus on two prominent and widely-used LLM chatbots, ChatGPT and Bard, to evaluate LLM responses to these S&P misconceptions. We selected them as (a) they are the most popular LLM chatbots available today, (b) they offer user-friendly interfaces accessible to the general public, and (c) do not require fine-tuning before end-user interaction.

We note that we did not evaluate S&P misconceptions with chatbots, such as BlenderBot [58], DialogGPT [70], and GODEL [48]. These chatbots are not explicitly designed to provide comprehensive advice or engage in conversations on niche topics. To enable them to answer such questions, one would require further fine-tuning with a vast corpus of relevant information.

In our preliminary experiments, we randomly sampled 50 misconceptions from our dataset and queried BlenderBot, DialogGPT, and GODEL for their opinions on these misconceptions. They showed a lack of awareness for most S&P questions. To illustrate, when we asked about their opinion on the misconception;

"Pseudonymised data (e.g., hashed data) are treated exactly like any other personal data under the GDPR",

they returned responses such as;

"I do not have information to answer this", "I don't know, I don't work for the government" or "yes, that's my guess"

Therefore, we excluded such language models from our analysis due to their inadequate performance.

**Response Generation Experiments.** To generate responses from LLMs, we framed each misconception as a claim and queried the LLM to verify it. We used the following static template as the input prompt: *I've heard of this claim: [MISCONCEPTION]. Is it true?.* We designed this template to mimic end users who interact with state-of-the-art LLM interfaces without access to more complex prompt templates (e.g., fine-tuning prompts via prompt engineering [40, 66]). To maintain consistency, we use the most updated versions for both models at the time of writing this paper. We query ChatGPT using its official API [3], using the latest version (the gpt-3.5-turbo [12]). We follow the API-call examples provided on the official website and use the default parameters throughout our

interactions. For Bard, since we have no access to its official API, we query misconceptions (using our prompt template) via Bard's web interface and obtain responses. After each query, we refresh the chat and delete the activity history to prevent interference between queries. We note that data collection for our experiments was conducted in a 2-day span (14, 15 March 2023). We did so to minimize the impact of the potential updates introduced by Google and OpenAI for Bard and ChatGPT, respectively. We designed four experiments (E1-E4) to extensively evaluate LLM capability in providing S&P advice, focusing on their (1) correctness in refuting misconceptions, (2) consistency in providing the same stance towards a misconception, (3) susceptibility to different framings/paraphrases of misconceptions and (4) ability to provide reliable sources.

**Initial Analysis via Single Trial (E1).** In E1, we query each misconception one time (a single trial), for both ChatGPT and Bard.

**Repeated Queries (E2).** In E2, we evaluate the consistency of LLMs in responding to S&P misconceptions. The models that users interact with (via web interfaces) are non-deterministic - asking the same question twice may not result in identical responses [4]. To simulate real-world scenarios where multiple individuals may ask the same question and receive different responses, we conducted four additional trials per misconception, generating a total of 488 additional responses from each model.

**Paraphrased Queries (E3).** In E3, we evaluate the effectiveness of LLMs in handling paraphrased queries since users may query LLMs chatbots in various ways. To do so, we use paraphrasing tools to produce an augmented dataset consisting of paraphrases of our original misconception. There is a wide selection of commercial and open-source paraphrasing tools; however, their performance in maintaining the original meaning of a sentence varies significantly.

To identify the most suitable tools, we conduct preliminary analysis on 27 APIs on the Rapid API platform [23] and 10 open-source paraphrasing models on the Hugging Face platform [15], two resources that provide state-of-the-art AI tool APIs. We randomly select five misconceptions from our dataset and generate paraphrases for each misconception with each tool.

Similar to determining manuscript relevance, two authors independently examined each paraphrased misconception using a labeling guide, labeling them as *valid* if they remained coherent while expressing the same meaning as the original misconception and *invalid* otherwise. Authors achieved high agreement,  $\kappa > 0.80$ before differences were reconciled. Based on the quality of the generated paraphrases, we chose Paraphrase Genius [22] available via Rapid API and t5-large-paraphraser-diverse-high-quality model [61] on Hugging Face, since these two models returned the highest percentage of *valid* paraphrased misconceptions.

Misconception	Paraphrases
	According to the GDPR, pseudonymized data (such as hashed data) is treated in the same way as other personal data.
Pseudonymised data (e.g., hashed data)	Pseudonymized records (e.g., hashed records) are treated in the same way as any other personal information under GDPR.
personal data under the GDPR	The GDPR treats pseudonymized data (such as hashed data) the same as other personal data.
personal and under the ODITA	Under the GDPR, pseudonymized data (such as hashed data) are treated in the same way as other personal data.
	A malicious file cannot cause me harm if I don't open it.
If I do not open an infected file, it can't	I won't be harmed if I don't open a malicious file.
do any harm.	An infected file can't harm me if I don't open it.
	If I do not open an infected file, it won't do me any harm.
	Hackers wouldn't be able to access their devices thanks to VPNs.
VPNs would prevent hackers from	Using a VPN would stop hackers from accessing their devices.
gaining access to their device.	Virtual private networks (VPNs) would stop hackers from accessing their devices.
	Through the use of VPNs, hackers could not access the user's device.

#### Table 2: Example of generated paraphrases for misconceptions.

#### Table 3: Labels and example misconceptions and responses. Relevant text that informs the labeling decision is highlighted.

Label	Example claim and corresponding response.
Noncommittal	Claim: VPNs would be not at all effective at preventing observation by the ISP.
NUTCONNECCAL	Response: As an AI language model, I don't have personal opinions or beliefs, but I can provide information on the topic. It
	is not entirely true or false.
Negato	Claim: Under GDPR, individuals have an absolute right to be forgotten.
Negale	Response: No, it is not true. The General Data Protection Regulation (GDPR) gives individuals the right to request that their
	personal data be erased, but this right is not absolute.
Support	Claim: Malware is most commonly downloaded through peer-to-peer and torrent sites.
Support	Response: Yes, it is true that malware is most commonly downloaded through peer-to-peer and torrent sites.
Partially Support	Claim: Tor browser is for the dark web and file sharing; it is the very definition of a dark web browser.
Faillally Support	Response: As an AI language model, I cannot say whether it is true or false as it can be subjective. However, it is true that Tor
	is often used to access the dark web, which is a part of the internet that is not indexed by standard search engines and requires
	specific software to access it.

We generate as many paraphrases as possible for each misconception using these tools. Then for each misconception, we manually select four paraphrases of high quality, which fluently and accurately convey the complete meaning of the original misconception. Table 2 presents a set of examples that we generated through this process. Overall, with the paraphrases, we generated 488 additional responses from Bard and ChatGPT.

**Soliciting Sources (E4).** We queried LLMs to obtain the URLs of their information to evaluate ability in providing reliable sources when responding to a misconception. We followed up on queries in E2 with the prompt *"Can you provide the URLs of your source?"*.

#### 3.3 LLM Response Analysis

After we collect the responses from each LLM, we label them into different categories according to their *stances* on the misconception.

**Labeling Guide Development.** To develop our labeling guide, we randomly sample 30 responses from E1 for both LLMs. Two authors independently performed deductive coding on the generated 60 responses, coding each response while focusing on the stance towards the misconception (whether they confirm or deny the misconception and to what degree). This allows us to produce a labeling guide that characterizes a response into one of four categories:

• Noncommittal: This category encompasses responses where the LLMs express a lack of knowledge about the topic or are unable to take a definitive stance.

- Negate: Responses falling under this category highlight the presence of falsehood or inaccuracies in the misconception.
- Support: Responses that affirm the validity or truthfulness of the misconception are classified under this category.
- Partially Support: Responses that confirm the misconception's validity to a limited degree but do not address any shortcomings or falsehoods in the misconception.

Table 3 shows an example misconception and its corresponding response for each label. We introduce an additional label Unrelated to address outliers, where a response may deviate from the main question or fail to address specific information the misconception inquires about. Given that our dataset comprises common S&P misconceptions, the ground truth for each response is Negate.

**Guide Validation and Labeling Process.** To ensure our labeling guide is reliable, we randomly sample 10 misconceptions and label their corresponding responses from E2, for a total of 100 responses. Two authors then independently labeled each response according to the labeling guide. We measure the agreement using Cohen's Kappa and obtain high agreement ( $k \ge 0.80$ ). The authors met to reconcile differences and agreed on the final guide version. They then labeled all remaining responses independently, meeting at intervals to reconcile differences. We note that the initial agreement (before differences are discussed) is high ( $k \ge 0.80$ ) at every interval.

**Correctness and Error Rate Analysis.** We evaluate correctness on a per-claim (misconception) basis. For E1, we consider responses labeled Support incorrect, as they indicate when an LLM fails to refute the misconception and provide incorrect support. In repeated trial experiments (E2-E3), we adopt a conservative approach and consider the misconception result incorrect if *any* of the trials result in a response labeled Support. Similarly, we consider responses labeled Negate correct, and for repeated trial experiments, a misconception producing accurate results must have *Negate* responses across *all* trials. We defined *error rate* as the ratio of misconceptions producing an incorrect result over total number of misconceptions.

Analyzing Source URLs. To analyze collected URLs, we first leverage the Python HTTP library requests [25] to verify if the website exists at present. If the request is successful (no error code is returned), we label it as a valid URL. If we receive an error, we employ an additional check using the Wayback Machine API [27] to verify if the directed website once existed. If it had existed but has since been removed (e.g., expired domain), we label it as a valid URL. If the URL is invalid (requests returns an error, and the Wayback Machine has no archive of the URL), we label it as an invalid URL.

We further label all valid URLs, as one of three categories, in accordance with their domain relevance. We label the URL (1) *relevant* if it directs to a website that provides S&P advice/information that relates to the misconception's domain (e.g., misconception on VPN resulting in a website about VPN S&P). We label it (2) *marginally relevant* if the directed website provides S&P advice that is generic or unrelated to the misconception, e.g., the misconception is about password security, but the website outlines VPNs security and privacy capabilities with no mention of password security. Lastly, we label a URL (3) *irrelevant* if the website content is unrelated to S&P (e.g., a website on iPhone features that does not mention S&P).

#### 4 EVALUATION

In this section, we detail our findings from our experiments E1-E4.

Table 4: Response label distribution in single query (E1).

	Negate	Support <sup>‡</sup>	Partially Support	Noncommittal
Bard	72.1%	26.2%	0%	1.7%
ChatGPT	70.5%	16.4%	3.3%	9.8%

<sup>†</sup> represents claims correctly responded to and <sup>‡</sup> represents the error rate.

#### 4.1 Single Trial Queries (E1)

**Correctness.** In the first experiment, we ask ChatGPT and Bard each misconception once. Table 4 presents the distribution of responses in E1. We find that Bard correctly negates 72.1% of the misconceptions. And it has an error rate of 26.2%, where it incorrectly supports the misconceptions. As for ChatGPT, we find that it correctly negates 70.5% of the claims but has an error rate of 16.4%. It is important to note that ChatGPT provides a *"noncommittal"* response in 9.84% of trials, compared to only 1.64% of trials in Bard. These empirical results demonstrate that Bard is less likely to refuse to answer a claim/remain neutral.

**Comparing Response Categories.** We performed additional analyses to study ChatGPT and Bard's performance across different categories. Figure 2 shows the proportions of each response label in ChatGPT and Bard's responses to these misconceptions. These proportions are calculated with respect to the number of misconceptions in each category. We interpret the *Negate* proportion as



Figure 2: Proportion of each response label across misconception categories (E1).

: Crypto

0.8

0.6

0.4

0.2

0.0

the correct rate and the *Support* proportion as the error rate. For instance, we find that ChatGPT exhibits the highest error rate for misconceptions falling under the "Law and Regulation" category (42.9%), while misconceptions pertaining to "Malware and Device Security" had the lowest error rate (4.76%), albeit the highest rate of noncommittal responses (19.1%). For Bard, "Privacy and Anonymity Tools" produces the highest error rate (32%). For the rest of the categories, error rates are around 25%, except for "IoT/CPS", which demonstrates an 11.8% error rate.

Our empirical results in E1 highlight that although both models correctly *Negate* misconceptions ~ 70% of the time, they also demonstrate a non-negligible error rate.

Table 5: Distribution of unique labels in repeated queries (E2). More than one unique label demonstrates the LLM's inconsistency towards a misconception.

	1 <sup>†</sup>	$2^{\dagger}$	3 <sup>†</sup>	<b>4</b> <sup>†</sup>
Bard	85.2%	13.1%	1.7%	0%
ChatGPT	70.5%	21.3%	7.38%	0.82%

<sup>†</sup> represents N unique labels while % values are with respect to 122 misconceptions (e.g., Bard produces 2 unique labels for 13.1% of misconceptions).

#### 4.2 Effectiveness under Repeated Queries (E2)

**Consistency of Responses.** In this experiment, we queried each misconception four additional times for ChatGPT and Bard, making a total of 610 responses for each model when combined with E1. Table 5 presents the number of unique response labels for each misconception. Bard performs better than ChatGPT in remaining consistent in their stance toward a misconception. 85.2% of Bard's misconceptions produce single label responses, compared to Chat-GPT's 70.5% of misconceptions. However, both Bard and ChatGPT demonstrate a non-negligible tendency to be inconsistent toward misconceptions (14.8% and 29.5% of misconceptions, respectively, produce more than one type of response across all trials).

We further examined consistency across misconceptions categories, as shown in Figure 3 (a). For Bard, inconsistency typically remains relatively low across all categories (below the 25% threshold). Noticeably, for both Bard and ChatGPT, "Web Security and Privacy" and "Privacy and Anonymity Tools" are the categories



Figure 3: Results across misconception category for E2.

that they are most prone to changing stances to (25% and 24% for Bard, and 45% and 28% for ChatGPT).

Table 6: Distribution of responses in repeated queries (E2).

	Correct <sup>†</sup>	Incorrect <sup>‡</sup>	Others
Bard	63.9%	30.3%	5.8%
ChatGPT	58.2%	27.0%	14.8%
+		+	

<sup>†</sup> represents claims correctly responded to. <sup>‡</sup> represents the error rate.

**Influence of Repeated Queries on Correctness.** In Table 6, we present the response distributions. We observe an increase in error rate as a result of repeated queries. More specifically, Bard and ChatGPT show an error rate of 30.3% and 27%, respectively. This translates to 4.1% and 10.6% increase from the results of E1. Bard shows a higher correct rate than ChatGPT, which can be attributed to Bard's tendency to maintain consistency. As expressed in Table 5, Bard is 14.7% more likely to maintain a single stance than ChatGPT. Here, we note that ChatGPT's tendency to produce more than one response type causes a drop in its correctness.

We note that if responses are neither correct (all trials *Negate* the misconception) nor incorrect (any *Support* across trials), we group them under "Others". For example, for the misconception "*Tor browser is for the dark web and file sharing, it is the very definition of a dark web browser.*", we find that Bard negates four times, and is noncommittal once. Such scenarios highlight that although an LLM may never incorrectly support a misconception, they can still be unreliable. Despite having previously negated a misconception, they are susceptible to taking a less concrete stance (*noncommittal*). Depending on randomness, this can disadvantage users, who may receive a noncommittal response towards a misconception (that should be negated). Thus, they do not receive the required S&P advice. 5.8% and 14.8% of Bard and ChatGPT responses towards misconceptions are grouped as "Others".

Figure 3 (b) shows the proportions of each response type across misconception categories. Among the categories, "Law and Regulation" exhibits the highest error rates in ChatGPT and Bard, 50% and 42.9%. We posit that law and regulation produce higher error rates as the language used may require more context. To illustrate, users may query LLMs with the misconception *"Under GDPR, when relying on consent to process personal data, consent must be explicit"* while understanding how "explicit" is used in the context of privacy laws. Surprisingly, our findings show that LLMs misinterpret "explicit" and confuse it with "unambiguous", which bears different implications in a legal context.

On the other hand, the "IoT/CPS" and "Malware and Device Security" misconception categories have the lowest error rates. For IoT/CPS, ChatGPT and Bard have error rates of 17.7% and 11.8%, respectively. Both ChatGPT and Bard demonstrate an error rate of 23.8% for "Malware and Device Security".

**Confusing Responses.** We found that a minority of responses contain confusing patterns - where the response begins with language indicating agreement for the misconception but provides context negating it. 2.95% and 0.984% of responses for ChatGPT and Bard begin with *Yes, it is true,* but the remainder of the text *negates* the claim. Such responses are labeled with *Negate* as we consider the context of the entire response in our labeling process. For instance, when asked about the misconception *"Under GDPR, individuals have an absolute right to be forgotten.*", ChatGPT responds with

"Yes, it is true that under GDPR... individuals have a "right to be forgotten"... However, this right is not absolute and there are limitations and exemptions. For example, organizations may be permitted to retain certain data for legal or regulatory reasons."

Since ChatGPT and Bard's responses tend to be elaborate, users who do not pay great attention may be misled by these responses.

Both models show a non-negligible tendency to be inconsistent in their stance when they are queried with the same misconception; however, they also yield an increased error rate. Responses from both models also contain confusing patterns that may mislead unassuming users.

# 4.3 Effectiveness on Paraphrased Queries (E3)

**Consistency of Responses.** In this experiment, we query the four paraphrases of each misconception once on each model and analyze these responses together with those in E1. We discover a significant increase (when compared to E2) in inconsistency (misconceptions with two or more label types).

Table 7: Distribution of unique labels in paraphrasedqueries (E3).

	1 <sup>†</sup>	$2^{\dagger}$	3†
Bard	55.7%	34.4%	9.9%
ChatGPT	54.9%	37.7%	7.4%
+			

<sup>†</sup> represents N unique labels while % values are with respect to 122 misconceptions (e.g., Bard produces 2 unique labels for 34.4% of misconceptions).

Table 7 presents the distribution of unique labels across misconceptions. We find that, on average, 44.7% of misconceptions in Bard





Figure 4: Results across misconception category for E3.

and ChatGPT solicit inconsistent responses. This is an increase of 29.5% and 15.6% in Bard and ChatGPT, when compared to E2, suggesting that slight modifications (while maintaining the same meaning) to sentences significantly decrease LLMs' consistency.

To demonstrate this, in Figure 4 (a), we show the distribution of unique label types per category. For both ChatGPT and Bard, the category "Web Security and Privacy" exhibits the highest occurrence of inconsistency, 60% for both models.

Table 8: Response distribution in paraphrased queries (E3).

	Correct	Incorrect	Others
Bard	47.5%	39.3%	13.2%
ChatGPT	45.9%	33.6%	20.5%

**Response Correctness.** Table 8 presents an overview of correctness in E3. Unsurprisingly, following an increase in inconsistency, overall correctness decreases by 16.4% and 12.3% for Bard and Chat-GPT, respectively (compared to E2). Compensating for this drop, we observed a higher error rate in Bard and ChatGPT (an increase of 9% and 6.6%, respectively, from E2) and in "Other" responses (an average increase of 6.55%). Interestingly, Bard's correct rate is only slightly higher than ChatGPT's (by 1.6%), contrasting E2's results. We attribute the similarity in correctness to both models' shared vulnerability to paraphrased questions - they both stick to the same stance across all questions only around 55% of the time.

Figure 4 (b) presents correctness across misconception categories. We observed that the categories "Law and Regulation" and "Web Security and Privacy" continue to exhibit the highest error rates.

Paraphrasing queries reduces LLM consistency and causes an significant increase in error rate and reduction in correctness when compared to multiple questions for a misconception.



Figure 5: Results across misconception category for E4.

#### 4.4 Analysis of Response Sources (E4)

**URL Validity and Relevance.** We gathered 946 URLs from Chat-GPT's responses obtained in E2 from 108 misconceptions, with an average of 8.76 URLs per misconception. The remaining 14 misconceptions do not produce any URLs. To illustrate, when asked for sources on the misconception *"Tor is designed for criminals who want to do illegal business securely.*", ChatGPT responds with

"As an AI language model, I don't have the ability to browse the internet and provide URLs of sources."

Interestingly, we find that Bard does not include explicit URLs in the responses. None of the responses to our prompt in the form of *Can you provide the URL of your sources?* provide URLs, with Bard always refusing (e.g., "*I'm just a language model, so I can't help you with that*"). We are only able to collect 66 URLs from the "Source" section, which occasionally comes together with the response to the question on misconception. These URLs are spread across 29 misconceptions, with an average of 2.28 URLs per misconception.

We find that 78.8% of URLs returned by Bard are valid - 71.23% exist currently, 7.57% no longer exist but can be accessed via the internet archive [27]. Conversely, only about a third (32.3%) of URLs returned by ChatGPT are deemed valid, 27.1% exist currently, and 5.2% no longer exist but are archived. This stark difference between both models is expected given that Bard has real-time access to the internet, while ChatGPT does not. However, given Bard's real-time access, it is still surprising that 21.2% of Bard's URLs are invalid. Notably, for Bard, 9.62% of the valid URLs were currently inaccessible because of SSL connection errors. Figure 5 (a) presents URL validity across the misconception categories. It's worth noting that Bard provides no URLs for the "Crypto & Blockchain" and "IoT/CPS" misconception categories. Furthermore, Bard consistently provides a higher percentage of valid URLs across all categories. In contrast, ChatGPT provides more invalid URLs across all categories.

We additionally explored the domain relevance for each website the valid URL directs to. Figure 5 (b) shows URL domain relevance across the misconception categories. For Bard, only 55.8% were

373

0 I I				
Misconception	URL provided in LLM response	Evaluation		
• Under GDPR, parental consent is always required when collecting	www.termsfeed.com/blog/childrens-	False information (domain <i>relevant</i> )		
personal data from children.	gaming-apps-legal-requirements/			
The blockchain ledger is locked and unchangeable/ unable to	https://www.ibm.com/blockchain/	False information (domain <i>relevant</i> )		
modify the data block once created, or blockchain data cannot be	what-is-blockchain			
changed once updated.				
<b>6</b> VPNs would be very effective at preventing friends or family	https://quiaustin.com/do-online-	Generic information (domain relevant)		
from seeing the websites in my browser history from my computer	casinos-track-your-ip-address-and-			
because I have my own private network that others cannot get into.	why/			
• Under GDPR, every business will be subject to new data porta-	https://gdpr-info.eu/art-20-gdpr/	Respond with partial information (domain <i>relevant</i> )		
bility rules.				
<b>6</b> Bookmarks saved in private mode would not persist in later	https://support.mozilla.org/en-	Respond with partial information (domain <i>relevant</i> )		
sessions because private mode deletes all local, temporary data,	US/kb/private-browsing-use-firefox-			
including bookmarks.	without-history			
<b>6</b> The government would need a warrant to access browsing activ-	https://www.fights4rights.com/	Unrelated (irrelevant)		
ity from private mode.	immigration-rights/			

*relevant*, 26.9% were *marginally relevant* - unrelated to the domain but still contained security and privacy content, and the remaining 17.3% were entirely unrelated to the claim or its response. For Chat-GPT, 81.7% of domains were *relevant*, 5.9% of them were *marginally relevant*, and the rest (12.4%) had no connection to the claims, *irrelevant*. Fortunately, the relevant URLs comprise the majority of valid URLs across all categories. However, 55.6% of valid URLs returned for the "IoT/CPS" category in ChatGPT are *irrelevant* (the highest percentage of invalid across all categories).

**Understanding Sources that Result in Errors.** We further analyze the URLs provided in incorrect responses. For Bard, 38.5% of returned valid URLs are attributed to incorrect responses, while this percentage for ChatGPT is 35.6%. To gain further insight into the LLMs' error generation, we read websites the valid URLs direct to. Given Bard's relatively small URL pool satisfying this category (20 URLs), we analyzed all of the websites that it provides. For ChatGPT, we sampled one-third of the valid URLs with incorrect responses.

Table 9 shows example misconceptions with incorrect responses, their URLs, and corresponding analysis. In **0**, the website ignores cases where parental consent is not necessary when processing children's personal data (false information). Similarly, in 2, the website failed to point out that under certain circumstances the blockchain ledger could be possibly changed. In 3, the website provides generic information on VPNs' encryption capabilities without addressing the privacy features relevant to the misconception (the ability of friends/family to see websites in browser history). If shows Bard's failure to comprehend an official GDPR article fully, thus leading to an incorrect response (partial information). In **6**, it is clearly stated in the website that new passwords and bookmarks created while using private browsing will be saved, while ChatGPT ignores it (partial information). In (), the website has no relation to the claim about the government's ability to access browsing activity in private mode.

Upon inspection of corresponding web pages associated with incorrect responses, 70% of the valid URLs Bard provided were domain *relevant*. Comprising this, 35% sources only provided generic information, 10% provided false information. 25% provided correct information, but Bard responded with partial information resulting in an incorrect response. The remaining 30% were "marginally relevant"/"irrelevant" - they do not address the misconception domain.

For ChatGPT, 54.8% sources offered generic information without any direct statements on the misconception, and 9.5% contained false information. 23.8% URLs contain the correct information, but, ChatGPT responds with partial information. The remaining 11.9% URLs in the sample were domain *"marginally relevant"*/*"irrelevant"*.

Bard divulges sources less frequently than ChatGPT. However, Bard's URL sources are more likely to be valid. Valid URLs predominantly point to websites that are domain *relevant* to the misconception category, but may also be unrelated to the misconception. For incorrect responses that provide relevant URLs, websites predominantly only have generic information on the misconception domain.

# **5 DISCUSSION AND LIMITATIONS**

#### 5.1 Lessons and Recommendations

**Shortcomings of LLMs in Providing Expert Advice.** Our experiments show the limitations of LLMs in correctness, consistency, and susceptibility to paraphrasing. We also expose their inability to justify or provide sources. These findings highlight how existing LLMs, in their current state, are unreliable as an S&P advice tool.

LLMs are trained on vast amounts of web data. For instance, Wikipedia is among the most common web pages available in one dataset used to train Bard [26, 50]. These resources are not guaranteed to be accurate (e.g., Wikipedia edits can be approved by community members). Additionally, as prior work has pointed out [31], relying on web-scrapped data creates a vulnerable "feedback loop" as inaccurate information generated by AI models is often uploaded to the internet (e.g., news articles synthesized by AI models [29]), influencing training data of current language models.

Similarly, despite numerous efforts to fine-tune LLMs, it remains unclear if domain-specific information (such as S&P) is used in fine-tuning their state-of-the-art interfaces. Moreover, data labeling to fine-tune is often contracted to third parties instead of domainspecific experts (e.g., researchers) [45]. We posit that these factors significantly contribute to LLM's lack of correctness. This is further evidenced in Section 4.4, by URLs pointing to generic and factually incorrect content, highlighting (1) a lack of specialized and (2) flawed training data, respectively.

LLMs' growth in serving as a trusted resource, however, shows no sign of slowing down. Soon after the introduction of LLMs, LLM-based applications, and plugins emerged, claiming additional capabilities (e.g., providing nutrition advice and identifying scientific sources) [7, 13]. Yet, the reliability of these tools in providing domain-specific expert advice has not been evaluated. Future work should not only address the feasibility of training LLMs to provide expert advice to users but should require careful collaboration with domain experts (e.g., health experts should be consulted during the design of LLM-driven health advice tools).

**Broadening Experimental Scope.** Although our efforts provide insight into LLM's limitations in providing S&P advice, future work should expand our data and experimental procedures. First, our dataset can be expanded in quantity and category diversity. This can be achieved via community collaboration (e.g., with S&P experts in various domains). Second, through larger-scale data collection and labeling, future efforts should also train an ML classifier that can automatically assign labels to LLM responses given an input claim, similar to prior efforts in stance detection and general question answering tasks [28, 41, 68]. Third, introducing an automated classifier would increase the feasibility of extending our experiments to additional State-of-the-art LLMs such as Claude [8] or LLaMa [19].

Similarly, further experiments can be conducted to verify hypotheses surrounding LLM correctness in responding to varying S&P categories. To illustrate, experiments can be designed to test whether complex legal language influences LLM performance among law and regulation misconceptions. Experiments can be repeated after substituting legal language in these misconceptions with plain language (commonly used among lay people).

Future research should also experiment with prompt engineering. Instead of static prompts, more diverse prompt structures, such as chain-of-thought prompting [66] and the ReAct prompt framework [67], can be used to discover optimized prompts that can solicit correct output. Prompts with varying levels of context can be used to discover the influence of providing more context (e.g., *"I heard a different opinion from a reputable source?"*) or nudging/challenging the LLM (e.g., *"are you sure about your response?"*)

**Importance of Understanding LLM Tool Use.** Users have increasingly shown a keen interest in leveraging LLM tools. Consider a scenario where a user seeks to confirm the correctness of a claim that is actually false. A single trial may incorrectly state that the claim is true (our results in E2 show how LLMs can be inconsistent). A user who does not repeatedly ask the question would never read the correct version of the response.

Our study shows the potential shortcomings of LLMs when used by end users. However, very little is known about *how* users interact with newly-emerging LLM tools. For instance, would users be inclined to query LLM tools repeatedly to ensure the veracity of the LLM's response? Despite a demonstrable lack of reliability in providing S&P advice, assuming that users leverage tools with caution would be overoptimistic. Thus, more user-centered research is required to understand user interactions with LLMs. Future work should explore qualitative efforts to understand (1) how and (2) when users interact with LLM interfaces. For instance, when seeking expert advice, do users interact with LLMs as a sole source of information or a preliminary step before verifying through other means (e.g., asking friends/ family, referring to scientific papers)?

However, large-scale quantitative studies - (e.g., surveys) are required for a deeper understanding of user-LLM interaction. To illustrate, contextual factors such as the severity of the misconception may influence user interaction and trust of LLM - e.g., users are likely to have different trust levels when casually querying about the S&P of IoT devices when compared to asking critical questions about the strengths of their password. Understanding these factors is imperative for the development of LLM tools. To illustrate, the scenario in which users are inclined to trust LLMs as a sole source warrants prioritizing performance in providing valid URLs.

#### 5.2 Limitations

Our study has three main limitations. First, S&P misconceptions may differ between different demographics (e.g., country of origin, socioeconomic status). We rely on existing manuscripts to inform us of publicly held user S&P misconceptions, and thus, our dataset likely excludes misconceptions that are not publicly studied. Given that such misconceptions are not publicly available, it is highly likely that LLM models trained on public information may not respond correctly to such misconceptions. This can affect the overall correctness and model error rates. Future work should focus on large-scale surveys targeting demographically diverse participants to generate representative datasets of S&P misconceptions.

Second, we focused on querying S&P misconceptions in English. Performance can vary across languages, further influencing model effectiveness in refuting S&P misconceptions and providing non-English sources of information. Additional work is required to compare LLM reliability as an advice tool in different languages.

Finally, we use a static prompt template. It is known that LLM responses could vary depending on the prompts. Fine-crafted prompts have the potential to enhance the quality of LLM responses. However, given our prompts were designed to mimic end users who query state-of-the-art LLMs, we leave the impact of prompt engineering on S&P advice to future work (as detailed in Section 5.1).

#### 6 RELATED WORK

LLM Performance in Question Answering. LLMs' impressive ability to generate natural language and capability in tasks such as text completion and question answering warrant recent research efforts focused on evaluating their performance. Most closely related to our study is the growing body of work that aims to understand the accuracy of LLMs in question-answering tasks. A recent study introduced a benchmark called TruthfulQA to assess language models' truthfulness in generating answers to questions [41]. This dataset was designed to measure imitative falsehoods by including questions that people often answer incorrectly due to false beliefs or misconceptions. The authors tested OpenAI GPT family models and found that the largest models were generally the least truthful. Similarly, Bang et al. [30] have evaluated ChatGPT's ability to generate factual content and detect misinformation using the TruthfulQA dataset. They revealed that ChatGPT answered falsely to one-third of the questions in the TruthfulQA dataset designed to elicit imitative falsehoods. Another recent work has extensively evaluated ChatGPT's reliability in generic question answering and its ability to identify unanswerable questions using 10 open QA benchmarks across eight knowledge domains, including history, law, and recreation [57]. They discovered the deficiency of Chat-GPT in identifying unanswerable questions and its varying ability among different knowledge domains. These works broadly explore question-answering instead of investigating LLMs use cases for specific/specialized domains, which can vary significantly depending on the training data of state-of-the-art LLM interfaces.

In contrast, we curate a novel dataset of S&P misconceptions after an extensive literature survey on user perception of S&P technologies. Interestingly, given that the misconceptions we collect are publicly held, they could potentially have influenced training data (since LLMs leverage large volumes of internet data). We then measure LLM's ability to refute popular misconceptions.

LLMs for Specialized Contexts. Complementing research on LLM question-answering ability, research evaluating LLMs' capability in a specialized context (producing domain-specific content) has also been explored, primarily in health and medical-related fields. For instance, Zuccon et al. [72] have evaluated ChatGPT's ability to answer complex health information questions and how knowledge provided in the prompt affects the accuracy of its answers. This work used various topics from a trusted resource (TREC Health Misinformation website) and showed that ChatGPT effectively answers health-related questions and debunks misconceptions about health treatment. However, other research [32, 39, 56] note the shortcomings of LLMs in providing expert medical-related opinions, noting inaccurate/false generated content.

In contrast, our work not only shifts focus to the S&P domain, but we designed experiments to measure consistency, the influence of other factors (e.g., paraphrasing), and LLM's ability to provide reliable sources. Our results not only confirm prior findings on the lack of reliability of state-of-the-art LLMs in providing specialized opinions but also extend findings by highlighting vulnerabilities to consistency, paraphrasing, and lack of reliable sources.

Limitations of LLM Performance. Investigations into questionanswering ability extend beyond measuring accuracy. Recent works seek to understand LLM *limitations* in answering questions or responding to user input. For instance, prior research has evaluated their vulnerability in producing hallucinations [30, 38, 69], the tendency to generate falsehoods that appear true but are nonsensical (e.g., referencing fake academic papers or news articles but claiming them to be true). Related to this, Ren et al. [55] investigate whether LLMs are able to perceive their knowledge boundary (and acknowledge when they are not able to answer accurately). Our findings extend insight into LLM limitations by exposing LLM vulnerability to repeated queries and paraphrasing. Invalid URLs (from E4) underscore the prevalence of hallucination. However, we also discover that LLMs can hallucinate by synthesizing URLs that point to an existing website but are unrelated to the question/prompt provided. This suggests that LLMs may "hallucinate" that an existing website contains domain-relevant information when, in reality, they do not. We also extend findings by showing that even with real-time web search capabilities, Bard may still produce invalid URLs.

**Using LLMs for Cybersecurity.** A growing body of work also investigates LLM from a cybersecurity perspective. Broadly, these works explore LLMs' ability to generate bug-free code or to detect S&P issues in vulnerable code [34, 43, 46, 47, 63]. These efforts help investigate the feasibility of LLMs to generate reliable code and also its use in an end-to-end software development pipeline.

In contrast, our research focuses on end-user interaction with cybersecurity (or, more broadly, Security & Privacy), where we curate a novel S&P dataset to understand LLM reliability in helping end users receive factually correct S&P advice.

**Security and Privacy Advice.** Existing studies on S&P advice have targeted understanding of where and how users receive advice (e.g., online forums, TV, peers, or through their own negative experiences) as well as its content and quality [42, 49, 53]. Prior work has also studied how demographics (e.g., socioeconomic status, skill level) impact resources for S&P advice [51, 52]. For instance, Redmiles et al. [54] conducted a comprehensive quality evaluation of the S&P advice on the web and found that most users believed online advice was somewhat actionable and comprehensive. Another line of work has proposed dissemination methods, leveraging interventions such as interactive games and comic strips [35, 60, 71].

In contrast, we focus on assessing emerging LLM user interfaces of ChatGPT and Bard. Given their increasing presence as a trusted information resource, we evaluate their ability to provide S&P advice by refuting common S&P misconceptions.

### 7 CONCLUSIONS

LLMs like ChatGPT and Bard have made prominent advancements in generative AI, becoming part of everyday users' lives. Their growing function as a trusted information source warrants evaluating their ability to provide expert advice. We evaluate LLMs' ability to provide S&P advice by refuting user-held S&P misconceptions. We first curate a dataset of 122 S&P misconceptions and query two popular LLMs in four experiments to measure overall correctness, consistency, and susceptibility to paraphrasing. We also analyze the sources LLMs provide to justify their stance towards misconceptions. We find that LLMs demonstrate non-negligible error rates, which increase when misconceptions are queried repeatedly or paraphrased. LLMs may show inconsistency, demonstrating multiple stances for a single misconception. LLMs often provide invalid URLs as resources and, in cases of valid URLs, may erroneously refer to websites with irrelevant information. Our work highlights LLM shortcomings in their reliability as an S&P advice tool.

#### ACKNOWLEDGMENTS

We thank our anonymous reviewers and shepherd for providing us with valuable feedback that helped improve our paper. This work is supported by startup funding from Purdue University. ACSAC '23, December 04-08, 2023, Austin, TX, USA

#### REFERENCES

- [1] 2023. Bard. https://bard.google.com/. [Online; accessed 25-May-2023].
- [2] 2023. ChatGPT. https://openai.com/blog/chatgpt/. [Online; accessed 25-May-2023].
- [3] 2023. ChatGPT API. https://platform.openai.com/docs/guides/chat. [Online; accessed 25-May-2023].
- [4] 2023. ChatGPT API Temperature Parameter. https://platform.openai.com/docs/ api-reference/chat/create#chat/create-temperature. [Online; accessed 25-May-2023].
- [5] 2023. ChatGPT coding: OpenAI CEO says AI can save time on software jobs. https://www.businessinsider.com/chatgpt-coding-openai-ceo-save-timeai-jobs-software-2023-5. [Online; accessed 25-May-2023].
- [6] 2023. ChatGPT: Online AI tool teachers say some students are using to cheat. https://news.yahoo.com/chatgpt-online-ai-tool-teachers-113735624.html. [Online; accessed 25-May-2023].
- [7] 2023. Chatgpt Plugins. https://openai.com/blog/chatgpt-plugins. [Online; accessed 25-May-2023].
- [8] 2023. Claude. https://www.anthropic.com/index/introducing-claude. [Online; accessed 25-May-2023].
- [9] 2023. Could ChatGPT REALLY slay Google. https://www.dailymail.co.uk/ sciencetech/article-11723499/Could-ChatGPT-replace-Google-Experts-weighwin-race-AI-search-engine.html. [Online; accessed 25-May-2023].
- [10] 2023. Could your next therapist be AI? Tech raises hopes, concerns. https://www.aljazeera.com/economy/2023/4/27/could-your-next-therapist-beai-tech-raises-hopes-concerns. [Online; accessed 25-May-2023].
- [11] 2023. Google Scholar. https://scholar.google.com. [Online; accessed 25-May-2023].
- [12] 2023. GPT-3.5-turbo Model. https://platform.openai.com/docs/models/gpt-3-5. [Online; accessed 25-May-2023].
- [13] 2023. GPTStore. https://gptstore.ai/. [Online; accessed 25-May-2023].
- [14] 2023. How You Should–and Shouldn't–Use ChatGPT for Medical Advice. https://www.verywellhealth.com/chatgpt-for-medical-advice-7500293. [Online; accessed 25-May-2023].
- [15] 2023. Hugging Face. https://huggingface.co/. [Online; accessed 25-May-2023].
- [16] 2023. I use ChatGPT to ace interviews: 'Works for every single job'. https://nypost.com/2023/05/24/i-use-chatgpt-to-ace-interviews-works-forevery-single-job/. [Online; accessed 25-May-2023].
- [17] 2023. IAn important next step on our AI journey. https://blog.google/technology/ ai/bard-google-ai-search-updates/. [Online; accessed 25-May-2023].
- [18] 2023. Investors are asking ChatGPT for stock market advice. https://www.marketplace.org/shows/make-me-smart/investors-are-askingchatgpt-for-stock-market-advice/. [Online; accessed 25-May-2023].
- [19] 2023. LLaMa. https://ai.meta.com/blog/large-language-model-llama-meta-ai/. [Online; accessed 25-May-2023].
- [20] 2023. Miss-date doctor launches revolutionary AI platform to help people find partners. https://uk.finance.yahoo.com/news/miss-date-doctor-launchrevolutionary-130000837.html. [Online; accessed 25-May-2023].
- [21] 2023. OpenAI ChatGPT job interview questions. https://www.hindustantimes. com/technology/openai-chatgpt-job-interview-questions-chatbot-artificialintelligence-101684997724187.html. [Online; accessed 25-May-2023].
- [22] 2023. Paraphrase Genius. https://rapidapi.com/genius-tools-genius-toolsdefault/api/paraphrase-genius. [Online; accessed 25-May-2023].
- [23] 2023. Rapid API. http://rapidapi.com. [Online; accessed 25-May-2023].
- [24] 2023. Replication Package. https://github.com/purseclab/LLM\_Security\_Privacy\_ Advice. [Online; accessed 2-October-2023].
- [25] 2023. Requests: HTTP for Humans. https://requests.readthedocs.io. [Online; accessed 25-May-2023].
- [26] 2023. Statistics of Common Crawl Monthly Archives. https://commoncrawl. github.io/cc-crawl-statistics/plots/domains. [Online; accessed 25-September-2023].
- [27] 2023. Wayback Machine. https://archive.org/help/wayback\_api.php. [Online; accessed 25-May-2023].
- [28] Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* (2021).
- [29] Waleed Ali and Mohamed Hassoun. 2019. Artificial intelligence and automated journalism: Contemporary challenges and new opportunities. *International journal of media, journalism and mass communications* 5 (2019).
- [30] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv preprint arXiv:2302.04023 (2023).
- [31] Clark Barrett, Brad Boyd, Ellie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. 2023. Identifying and Mitigating the Security Risks of Generative AI. arXiv preprint arXiv:2308.14840 (2023).

- [32] Alexei A Birkun and Adhish Gautam. 2023. Large language model-based chatbot as a source of advice on first aid in heart attack. *Current Problems in Cardiology* (2023).
- [33] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems.
- [34] Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. 2023. PentestGPT: An LLMempowered Automatic Penetration Testing Tool. arXiv preprint arXiv:2308.06782 (2023).
- [35] Tamara Denning, Adam Lerner, Adam Shostack, and Tadayoshi Kohno. 2013. Control-Alt-Hack: the design and evaluation of a card game for computer security awareness and education. In ACM SIGSAC conference on Computer & communications security.
- [36] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. Statistical methods for rates and proportions (1981).
- [37] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462 (2020).
- [38] Nuno M Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. arXiv preprint arXiv:2303.16104 (2023).
- [39] Sean S Huang, Qingyuan Song, Kimberly J Beiting, Maria C Duggan, Kristin Hines, Harvey Murff, Vania Leung, James Powers, TS Harvey, Bradley Malin, et al. 2023. Fact Check: Assessing the Response of ChatGPT to Alzheimer's Disease Statements with Varying Degrees of Misinformation. medRxiv (2023).
- [40] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems (2022).
- [41] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Computational Linguistics*.
- [42] James Nicholson, Lynne Coventry, and Pamela Briggs. 2019. "If It's Important It Will Be A Headline" Cybersecurity Information Seeking in Older Adults. In CHI Conference on Human Factors in Computing Systems.
- [43] David Noever. 2023. Can Large Language Models Find And Fix Vulnerable Software? arXiv preprint arXiv:2308.10345 (2023).
- [44] OpenAI. 2023. Introducint ChatGPT. https://openai.com/blog/chatgpt. [Online; accessed 25-May-2023].
- [45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (2022).
- [46] Sudipta Paria, Aritra Dasgupta, and Swarup Bhunia. 2023. DIVAS: An LLM-based End-to-End Framework for SoC Security Analysis and Policy-based Protection. arXiv preprint arXiv:2308.06932 (2023).
- [47] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2023. Examining zero-shot vulnerability repair with large language models. In 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2339–2356.
- [48] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. GODEL: Large-Scale Pre-Training for Goal-Directed Dialog. arXiv preprint arXiv:2206.11309 (2022).
- [49] Emilee Rader, Rick Wash, and Brandon Brooks. 2012. Stories as informal lessons about security. In Symposium on Usable Privacy and Security.
- [50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* (2020).
- [51] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2016. How i learned to be secure: a census-representative survey of security advice sources and behavior. In ACM SIGSAC conference on computer and communications security.
- [52] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2017. Where is the Digital Divide? A Survey of Security, Privacy, and Socioeconomics. In CHI Conference on Human Factors in Computing Systems.
- [53] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. 2016. I think they're trying to tell me something: Advice sources and selection for digital security. In *IEEE Symposium on Security and Privacy (SP)*.
- [54] Elissa M Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L Mazurek. 2020. A comprehensive quality evaluation of security and privacy advice on the web. In USENIX Security Symposium.
- [55] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation. arXiv preprint

arXiv:2307.11019 (2023).

- [56] Hamid Reza Saeidnia, Marcin Kozak, Brady D Lund, and Mohammad Hassanzadeh. 2023. Evaluation of ChatGPT's responses to information needs and information seeking of dementia patients. *Research Square* (2023).
- [57] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. arXiv preprint arXiv:2304.08979 (2023).
- [58] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. arXiv preprint arXiv:2208.03188 (2022).
- [59] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. 2022. Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In ACM SIGSAC Conference on Computer and Communications Security.
- [60] Sukamol Srikwan and Markus Jakobsson. 2008. Using cartoons to teach internet security. Cryptologia (2008).
- [61] t5-paraphrase API 2023. t5-large-paraphraser-diverse-high-quality model. https: //huggingface.co/ramsrigouthamg/t5-large-paraphraser-diverse-high-quality. [Online; accessed 25-May-2023].
- [62] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022).
- [63] M Caner Tol and Berk Sunar. 2023. ZeroLeak: Using LLMs for Scalable and Cost Effective Side-Channel Patching. arXiv preprint arXiv:2308.13062 (2023).
- [64] Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In Chi conference on human factors in computing systems

extended abstracts.

- [65] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021).
- [66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022), 24824–24837.
- [67] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022).
- [68] Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? arXiv preprint arXiv:2212.14548 (2022).
- [69] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2309.01219 (2023).
- [70] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536 (2019).
- [71] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. 2016. The role of instructional design in persuasion: A comics approach for improving cybersecurity. *International Journal of Human-Computer Interaction* (2016).
- [72] Guido Zuccon and Bevan Koopman. 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. arXiv preprint arXiv:2302.13793 (2023).