Can Large Language Models Provide Security & Privacy Advice?

Measuring the Ability of LLMs to Refute Misconceptions

Yufan Chen*, **Arjun Arunasalam***, and Z. Berkay Celik
Purdue University

ACSAC 2023

* co-first authors

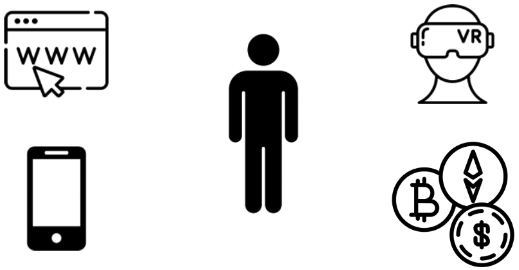
PURDUE UNIVERSITY

CERIAS

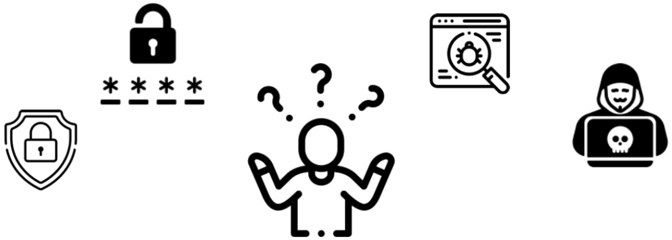
PurSecLab

Background: Security & Privacy Advice

Users interact with technology on a day-to-day basis

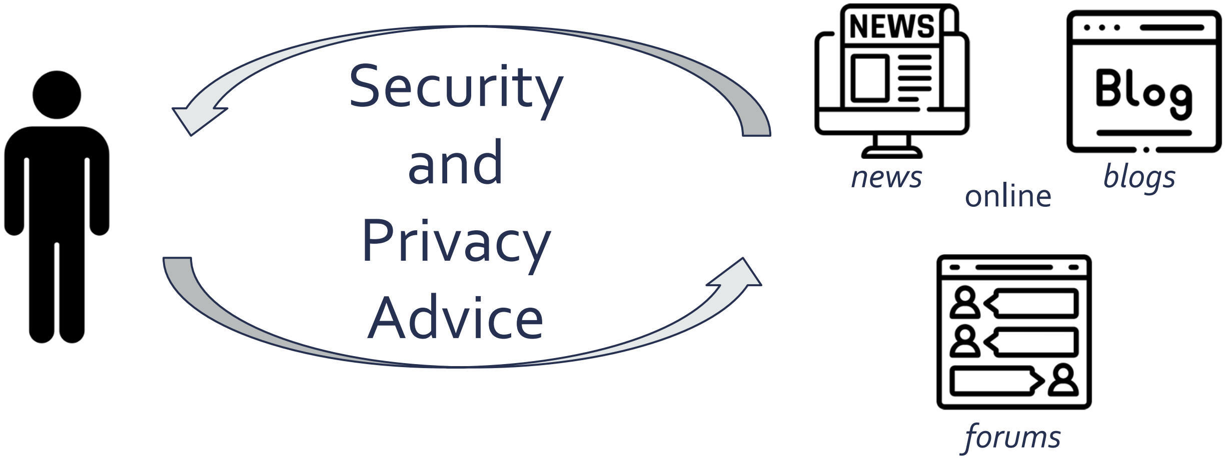


Consequently, users equip themselves with security and privacy knowledge to use technology effectively



Background: Security & Privacy Advice

- Prior work has shown that users receive security and privacy advice from various sources
  - Especially online websites and media [1,2]



[1] Redmiles et al., Where is the Digital Divide? A Survey of Security, Privacy, and Socioeconomics. CHI2017
[2] Redmiles et al., How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior. CCS2016

Background: User Interaction with LLMs

- Large language models have seen rapid growth
- Expansion due to web-interfaces such as ChatGPT
- Users leverage these models for various use-cases

Background: User Interaction with LLMs

- Large language models have seen rapid growth
- Expansion due to web-interfaces such as ChatGPT
- Users leverage these models for various use-cases

Job seekers are using ChatGPT to prepare for interviews — and it's helping them get hired

Source: businessinsider.com

ChatGPT can pick stocks better than your fund manager

Source: cnn.com

Motivation

Users depend on online resources for S&P advice

Users take advantage of end-user facing LLMs

Motivation

Users depend on online resources for S&P advice

Users take advantage of end-user facing LLMs

Need for understanding LLM reliability in providing security and privacy advice

Motivation: Research Question

How do we begin to measure LLM reliability in this domain?

by answering

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

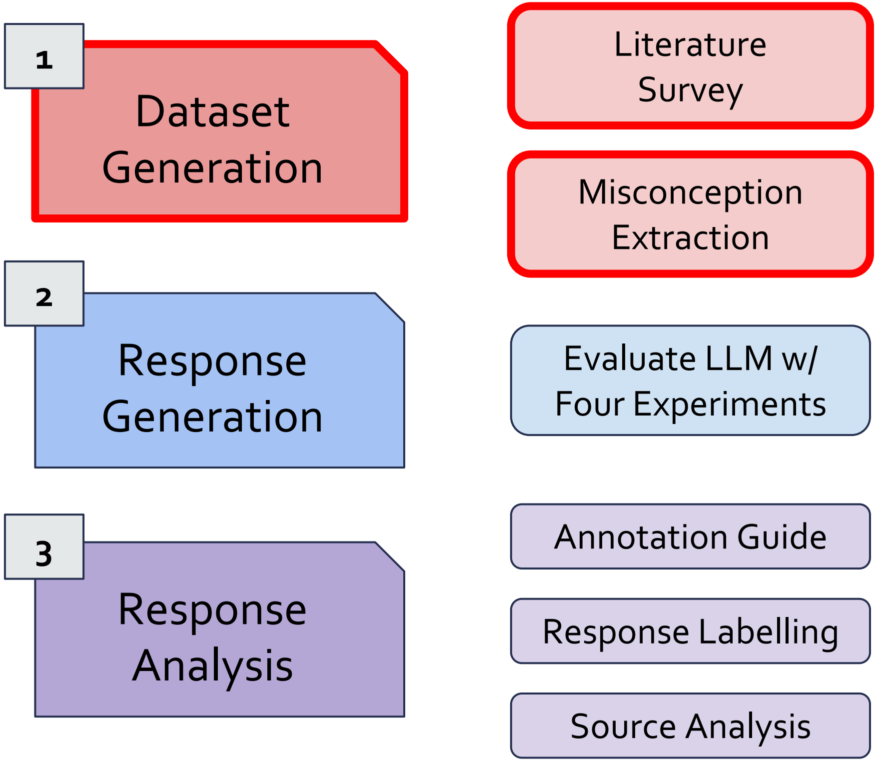
Methodology

1 Dataset Generation
- Literature Survey
- Misconception Extraction

2 Response Generation
- Evaluate LLM w/ Four Experiments

3 Response Analysis
- Annotation Guide
- Response Labelling
- Source Analysis

Methodology: Dataset Generation



- 56 curated queries
  - ~400 academic articles
  - > 500 misconceptions
- Filtered to 122 unique misconceptions
  - 6 different categories

Methodology: Response Generation

1 Dataset Generation
- Literature Survey
- Misconception Extraction

2 Response Generation
- Evaluate LLM w/ Four Experiments

3 Response Analysis
- Annotation Guide
- Response Labelling
- Source Analysis

Evaluated 2 LLMs

ChatGPT

Bard

Static Template

E.g., "VPNs would prevent hackers from gaining access to their device. Is this true?"

Methodology: Response Generation

1 Dataset Generation

- Literature Survey
- Misconception Extraction

2 Response Generation

- Evaluate LLM w/ Four Experiments

3 Response Analysis

- Annotation Guide
- Response Labelling
- Source Analysis

Evaluated 2 LLMs

ChatGPT

Bard

- E1: Single Query
- E2: Repeated Queries
- E3: Paraphrased Queries
- E4: Prompting for Sources

Methodology: Response Analysis

1 Dataset Generation
- Literature Survey
- Misconception Extraction

2 Response Generation
- Evaluate LLM w/ Four Experiments

3 Response Analysis
- Annotation Guide
- Response Labelling
- Source Analysis

- Sampled 60 responses
- Deductive coding to generate labels for responses
  - **Noncommittal**
  - **Negates Misconception**
  - **Supports Misconception**
  - **Partially Support**

Methodology: Response Analysis

1 Dataset Generation

- Literature Survey
- Misconception Extraction

2 Response Generation

- Evaluate LLM w/ Four Experiments

3 Response Analysis

- Annotation Guide
- Response Labelling
- Source Analysis

- Sampled 60 responses
- Deductive coding to generate labels for responses

GROUND TRUTH / CORRECT

- **Negates Misconception**
- **Supports Misconception**

ERROR / INCORRECT

Methodology: Response Analysis

1 Dataset Generation
- Literature Survey
- Misconception Extraction

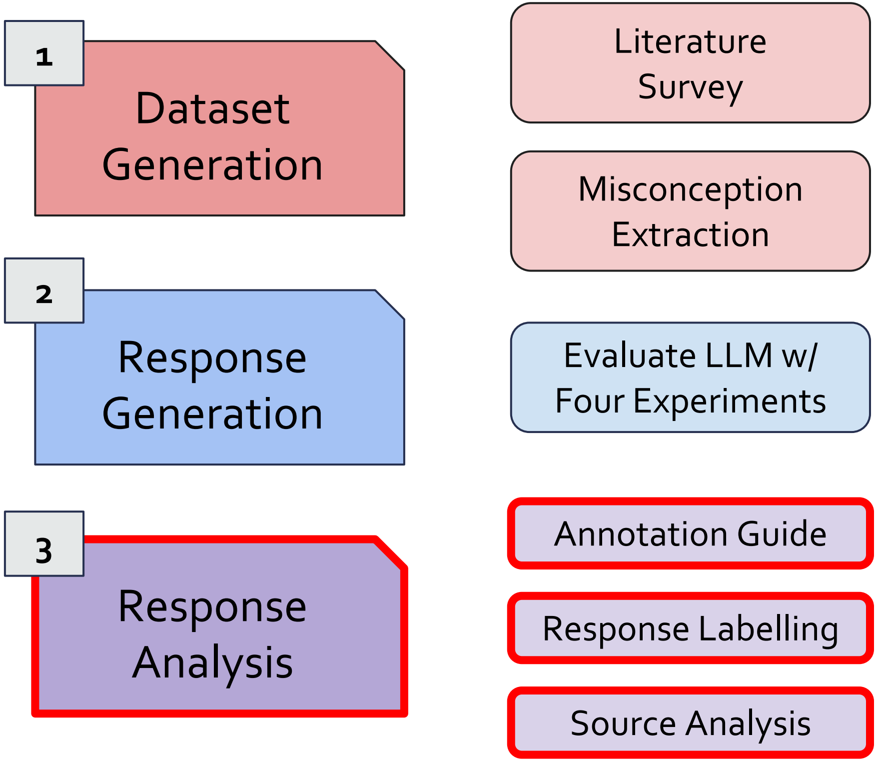
2 Response Generation
- Evaluate LLM w/ Four Experiments

3 Response Analysis
- Annotation Guide
- Response Labelling
- Source Analysis

- Correctness defined on a per-misconception basis
  - If all responses are negated: correct
  - If at least 1 response supports: error/incorrect

PURDUE UNIVERSITY CERIAS PurSecLab

Methodology: Response Analysis



- Source Analysis - Validity
  - HTTP Requests - Exists Now
  - Web Archive - Existed Before
- Source Analysis - Relevance
  - Misconception relevant
  - Generic S&P advice
  - Irrelevant

Decreasing relevance

Results: E1- Single Query

- Both models demonstrate a non-negligible error rate

	Negate †	Support ‡	Partially Support	Noncommittal
Bard	72.1%	26.2%	0%	1.7%
ChatGPT	70.5%	16.4%	3.3%	9.8%

† Represents claims correctly responded to

‡ Represents the error rate

Results: E2 - Repeated Queries

- Both models demonstrate a tendency to be inconsistent

	1 unique response	2 unique responses	3 unique responses	4 unique responses
Bard	85.2%	13.1%	1.7%	0%
ChatGPT	70.5%	21.3%	7.38%	0.82%

- They also demonstrate an increased error rate

	Correct †	Incorrect ‡	Others
Bard	63.9%	30.3%	5.8%
ChatGPT	58.2%	27.0%	14.8%

† Represents claims correctly responded to

‡ Represents the error rate

Results: E2 - Repeated Queries

- Both models demonstrate a tendency to be inconsistent

	1 unique response	2 unique responses	3 unique responses	4 unique responses
Bard	85.2%	13.1%	1.7%	0%
ChatGPT	70.5%	21.3%	7.38%	0.82%

e.g.,

VPNs would prevent hackers from gaining access to their device. Is this true?

3 negates

2 supports

- They also demonstrate an increased error rate

	Correct †	Incorrect ‡	Others
Bard	63.9%	30.3%	5.8%
ChatGPT	58.2%	27.0%	14.8%

† Represents claims correctly responded to

‡ Represents the error rate

Results: E3 - Paraphrased Queries

- Unsurprisingly, paraphrasing causes
  - Increase in inconsistency (~44.5 %)
  - Increase in error rate (~36.5 %)

	1 unique response	2 unique responses	3 unique responses
Bard	55.7%	34.4%	9.9%
ChatGPT	54.9%	37.7%	7.4%

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

† Represents claims correctly responded to
‡ Represents the error rate

Results: Across E1-E3

- Confusing responses
  - <u>Query</u>: *"Under GDPR, individuals have an absolute right to be forgotten. Is this true?"*
  - <u>Response</u>: *"**Yes, it is true** that… individuals have a right to be forgotten … However, **this right is not absolute** and there are limitations and exemptions…"*
- Poor performance on law and regulation related misconception
  - Influence of legal context of word

Results: E4 - Source Analysis

- Bard divulges sources less frequently
  - 66 URLs from Bard vs. ChatGPT's 946 URLs
- Bard's URL sources are more likely to be valid
  - 78.8% for Bard vs. 32.3 % of ChatGPT
- Relevance of URLs

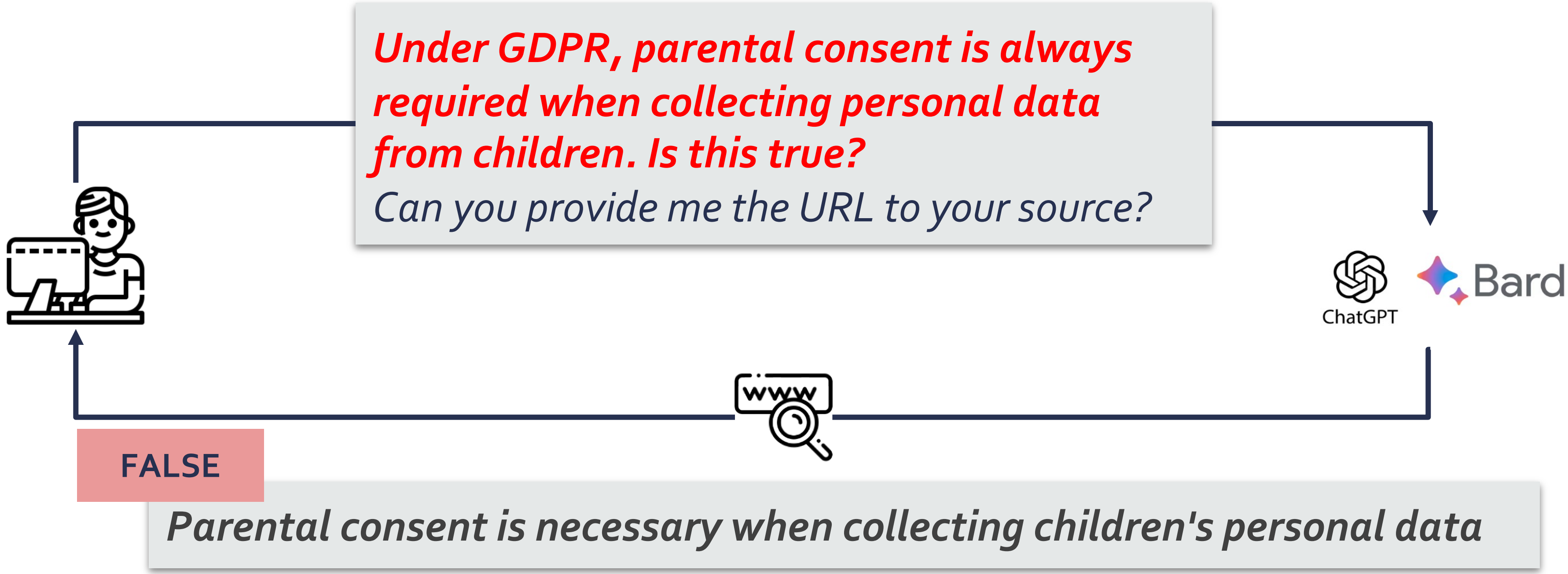
	Bard	ChatGPT
Relevant to Misconception	55.8 %	81.7%
Generic S&P advice	26.9%	5.9 %
Irrelevant	17.3 %	12.3 %

Results: E4 - Source Analysis

- Incorrect responses that provide valid URLs
  - 38.5% of Bard's URLs, 35.6% of ChatGPT's URLs
- Qualitatively analyzed a sample of such websites
  - ⅓ of ChatGPT, all of Bard
- Valid URLs for incorrect responses may point to:
  - False information
  - Info ignored by the response
  - Completely irrelevant content

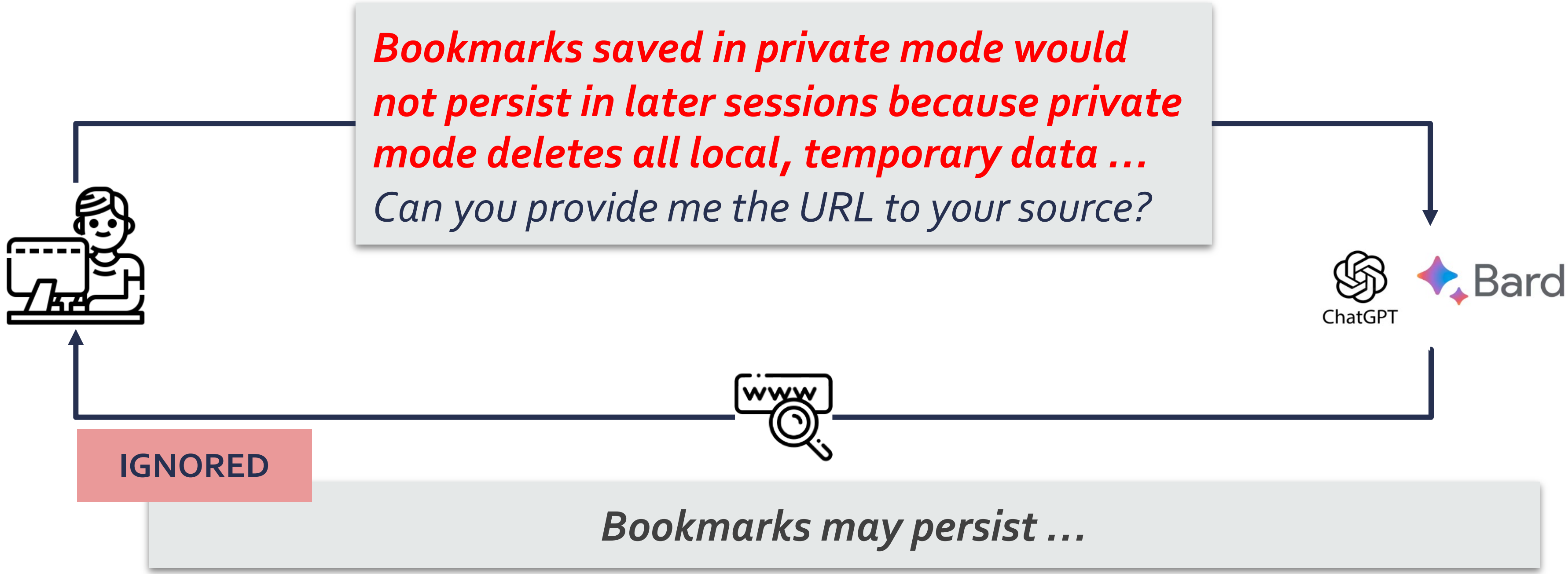
Results: E4 - Source Analysis

- Existing URLs for incorrect responses may point to:
  - False information



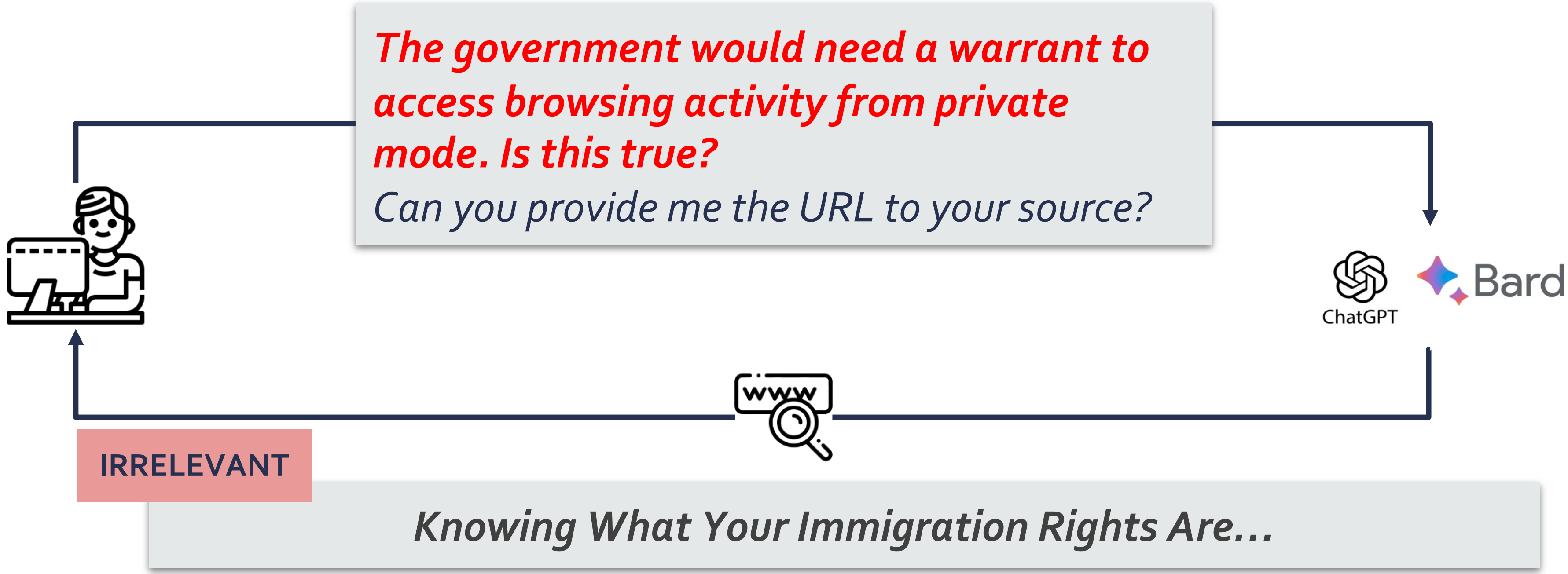
Results: E4 - Source Analysis

- Existing URLs for incorrect responses may point to:
  - Info ignored by the response



Results: E4 - Source Analysis

- Existing URLs for incorrect responses may point to:
  - Completely irrelevant content



Future Work and Recommendations

- First exploratory study on user facing LLMs' ability for S&P advice
- Broadening experimental scope
  - Increasing dataset size
  - Automated classification (e.g., via stance detection)
- Understanding how users interact with LLMs
  - How is output provided by tools processed by end users?
- Need for LLMs in specialized contexts
  - Requires domain-expert collaboration

Thank you! Questions?

aarunasa@purdue.edu

Results Reproduced acm V1.1

Artifacts Evaluated acm Reusable V1.1

PURDUE UNIVERSITY

CERIAS

PurSecLab