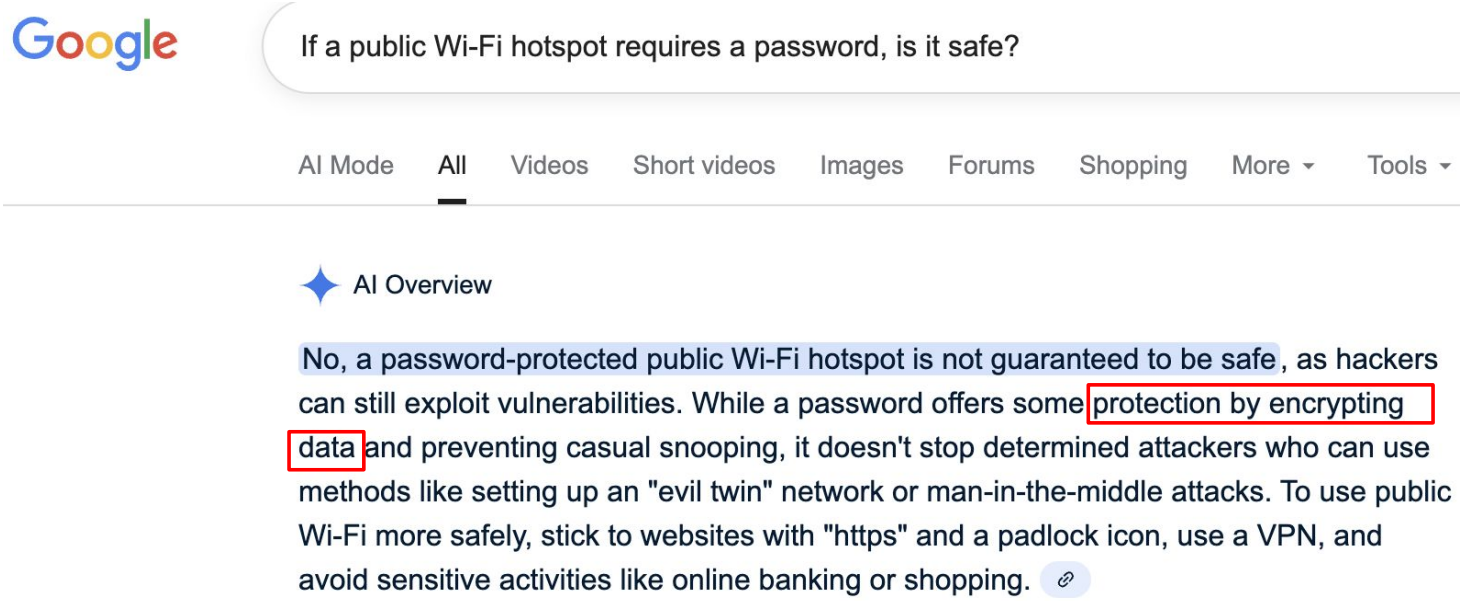
Experimental Aspects of Evaluation of LLM Responses to End-user Security Questions

Vijay Prakash, New York University; Kevin Lee, JPMorganChase; Arkaprabha Bhattacharya, Cornell University; Danny Yuxing Huang*, New York University; Jessica Staddon*, Northeastern University

NYU

LLM Integration in Popular User Information Sources

From search engines to Q&A platforms like Reddit and Quora!



NYU

Motivation

Users are known to go online for digital security advice [Redmiles S&P 2016]

Using LLMs is natural evolution

Even experts don’t have consensus on prioritizing advice [Redmiles USENIX 2020]

Hallucinations, reasoning errors, and concept drift in LLMs

NYU

Research Questions

Information quality in LLM response?

Patterns of deficiencies?

Actionable suggestions for users and model developers?



NYU

How did LLMs Perform?

Provide quality information in user-friendly language

But often with errors and deficiencies

Generalizable issues across three models we evaluated

NYU

What are Issues in LLMs Response?

Content issues

Over-prescription of security technologies and products

VPNs (19 responses)

A good **VPN can protect** you from **malicious websites and phishing** attempts.

Failure to leverage research findings

Outdated password guidance; HTTPS overreliance—even when widely used in phishing attacks

Communication issues

LLM-splaining: Unnecessarily verbose or polite (“Great question!”) delaying direct advice

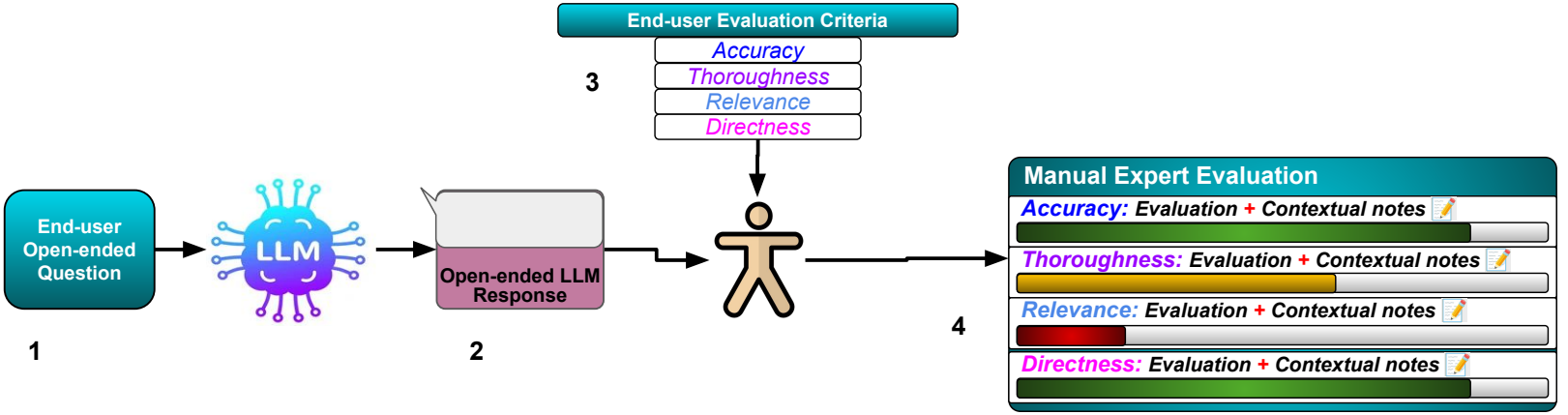
Question: Can I trust Google Password Manager?

Response starts with: Google Password Manager is a feature offered by Google that stores and manages passwords for various websites and applications.

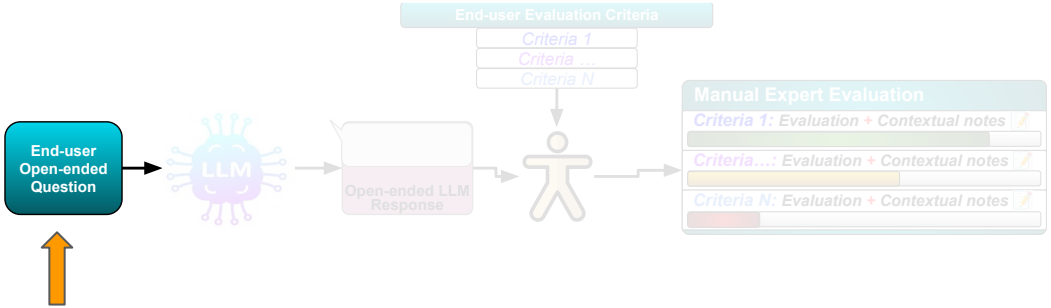
Learned, Lagged, LLM-splained: LLM Responses to End User Security Questions [Generative AI (for) Security]
Friday: 10:30 AM - 12 PM

NYU

Evaluation Approach



End-user Questions: Trusted Organisations



Trusted organizations: mission tied to user security (e.g., government agencies, universities) or a business that depends on customer security (e.g., financial services, security vendors)

Government agencies: NIST, FTC, CISA, NCSC (UK), CCCS (Canada), etc.

University guides; and more

FEDERAL TRADE COMMISSION
CONSUMER ADVICE

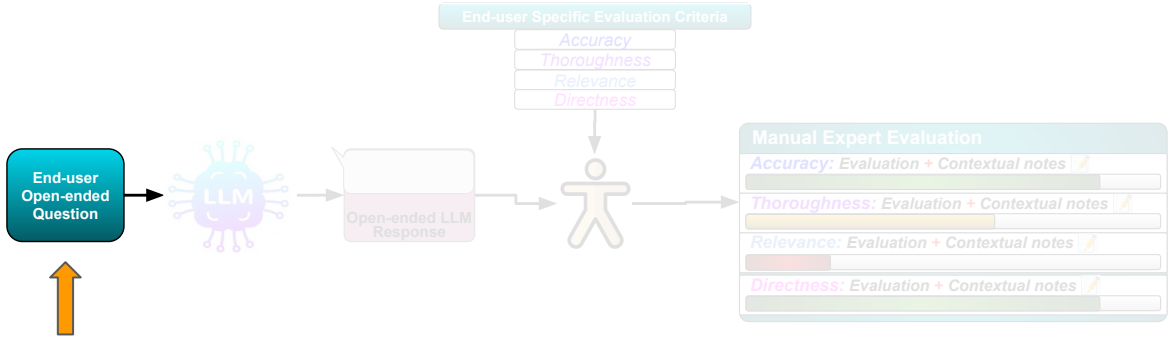
How To Recognize and Avoid Phishing Scams

National Cyber Security Centre

Why keep your devices up to date?

NYU

End-user Questions: Security Topics Discovery

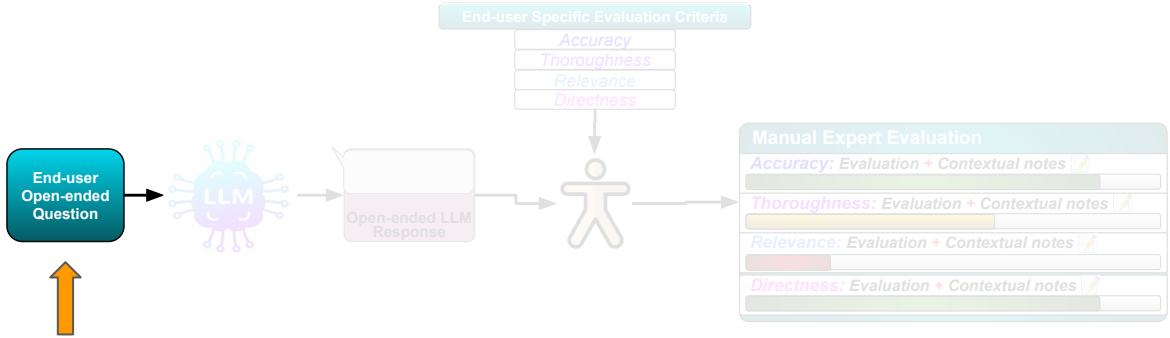


Category	Example Security Topic
Authentication	*Password manager*
Social Engineering	*Phishing*
Software Updates	*General Software Updates*
Antivirus/Antimalware	*AV Capabilities and Rationale*
Safe Browsing	*Compromised / Unsafe Web Pages*
Secure Wi-Fi/Network	*Public Wi-Fi*
Smart Devices Security	*Smart home*

Total 26 security topics across 7 categories

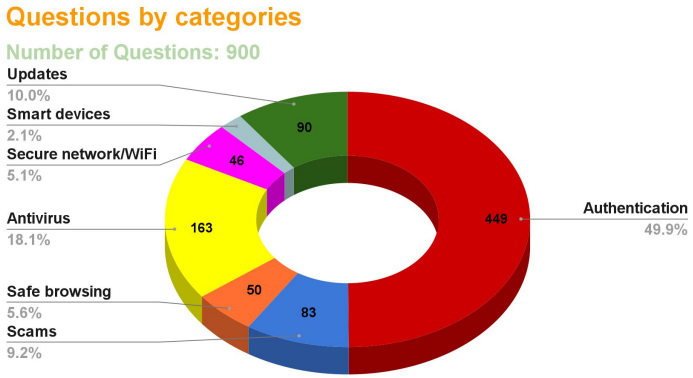
NYU

End-user Questions: Manually Mining FAQs

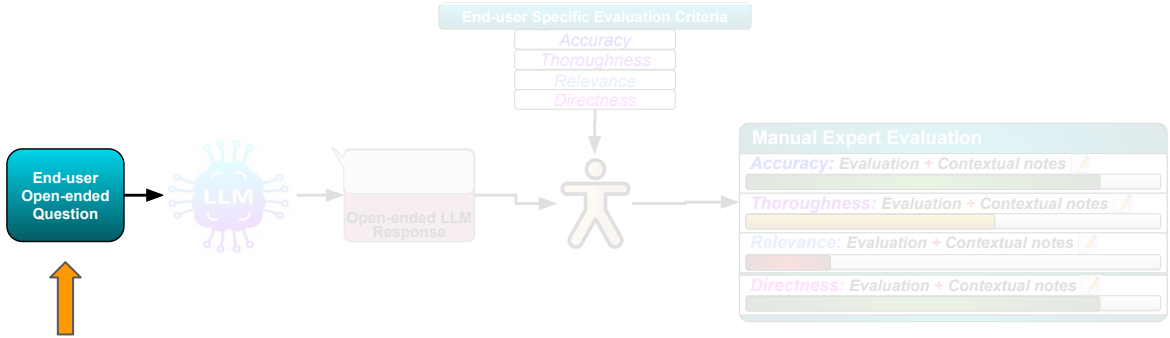


7 security categories

Manually collected 900 questions from FAQ section of 50 trusted organisations that represent common user concerns



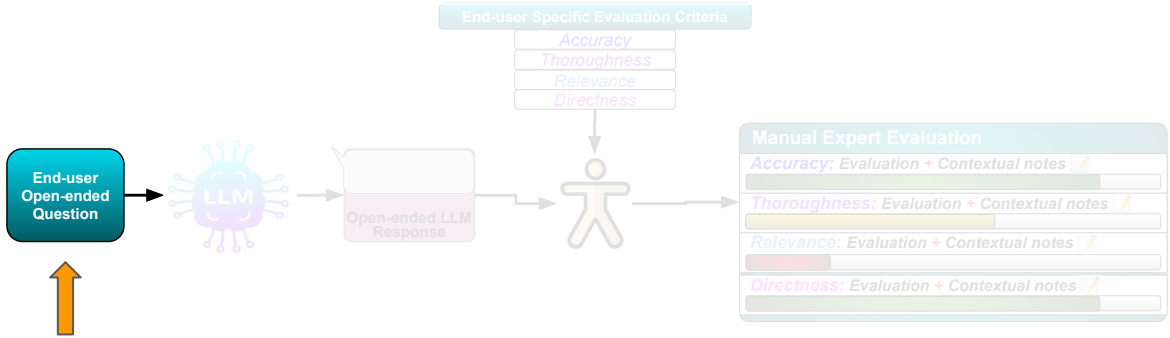
End-user Questions: Authoritative Sources



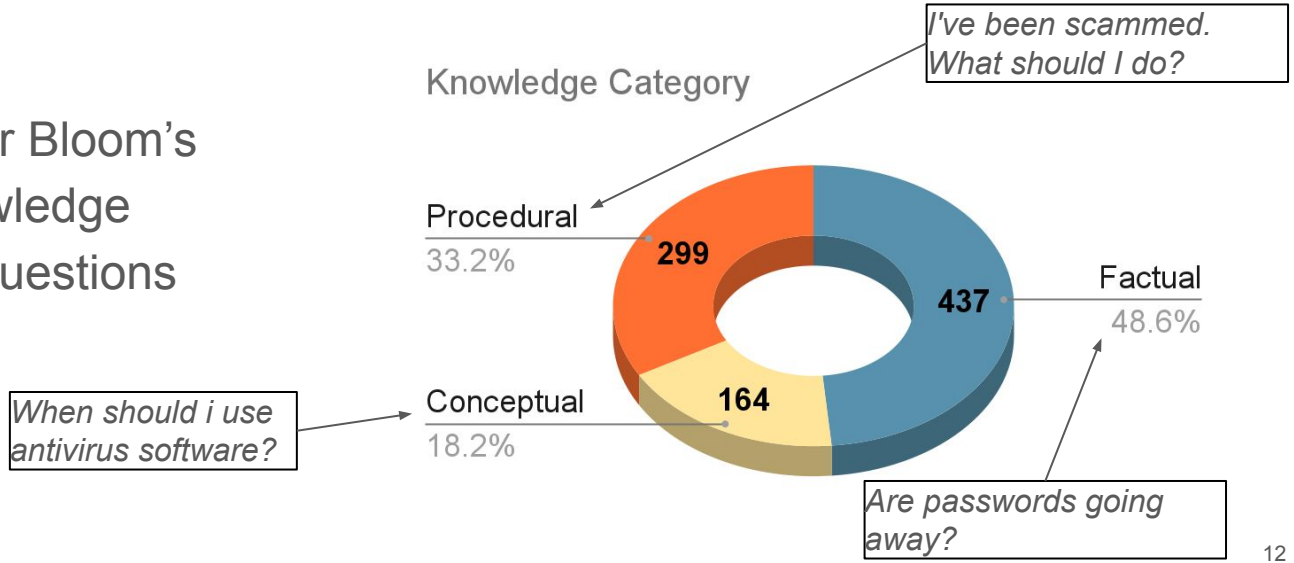
Authoritative sources to assess the accuracy and completeness of LLM responses

On average, 1.7 authoritative sources per question

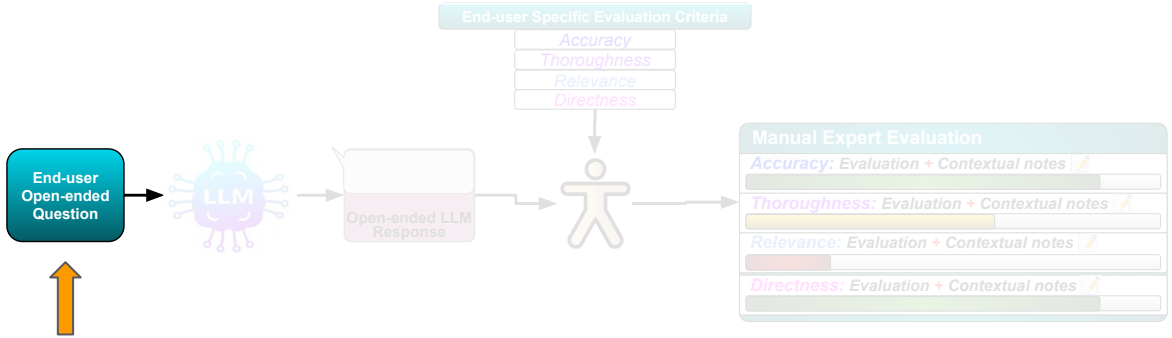
End-user Questions: Knowledge Categories



Knowledge category (per Bloom's taxonomy): Type of knowledge required to answer the questions



End-user Questions: Themes

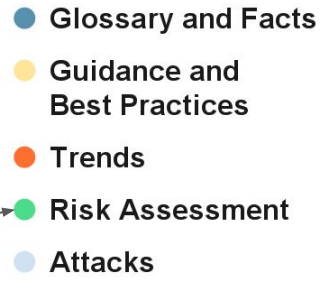


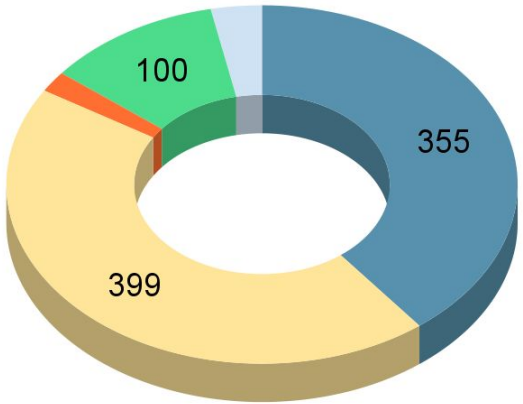
Question count

Question themes: Based on security information sought

What are recent scam trends?

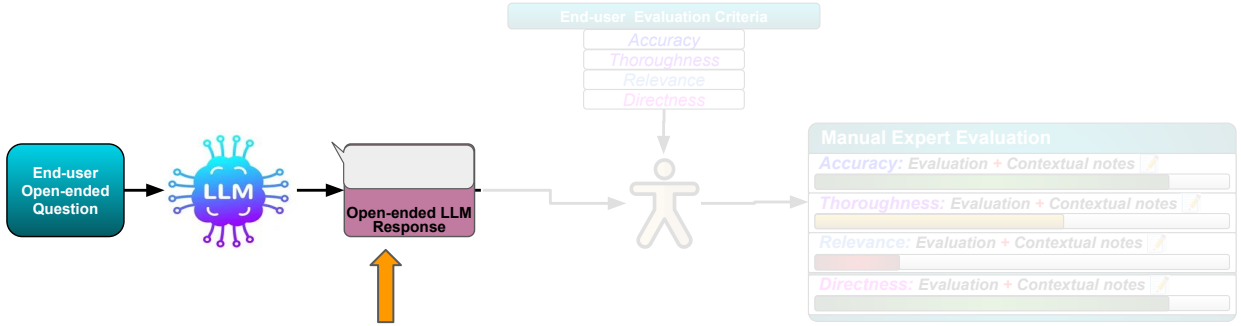
Can I trust Google password manager?





NYU

Response Generation Settings



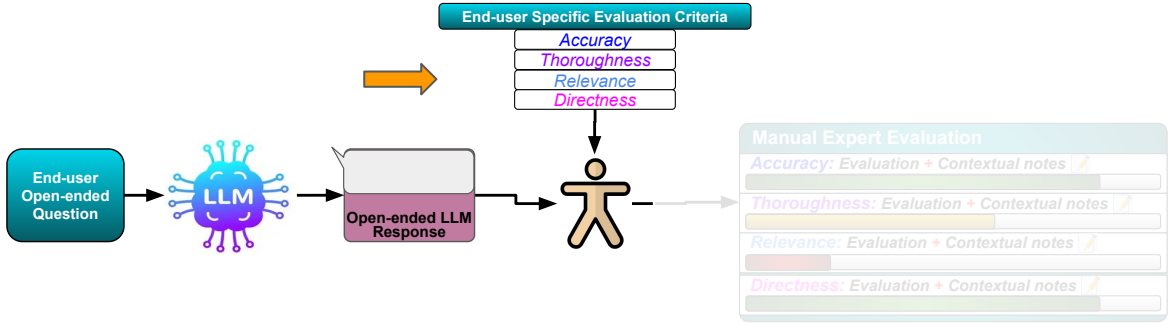
Models: GPT-4, Gemini-1.5, Llama-3

No prompt engineering, zero-shot, single turn

```
[
  {
    "role": "system",
    "content": "You are a helpful assistant."
  },
  {
    "role": "user",
    "content": "Hello!"
  },
  {
    "role": "assistant",
    "content": "Hello! How can I assist you today?"
  },
  {
    "role": "user",
    "content": "-> Questions go here. <-"
  }
]
```

NYU

Evaluation Criteria



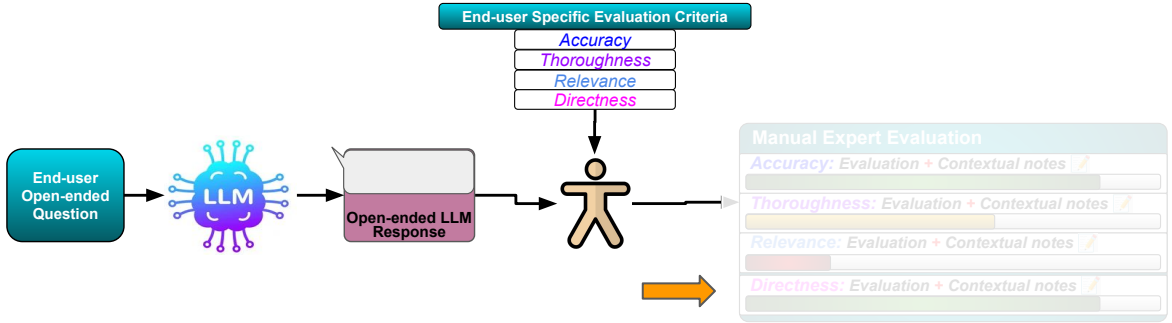
Evaluation criteria: Information integrity framework

- Accuracy
- Thoroughness
- Relevance
- Directness

3 point likert scale: completely, partially, minimally

2 point: Direct, Not Direct

Manual Expert Evaluation: Reliability Experiment

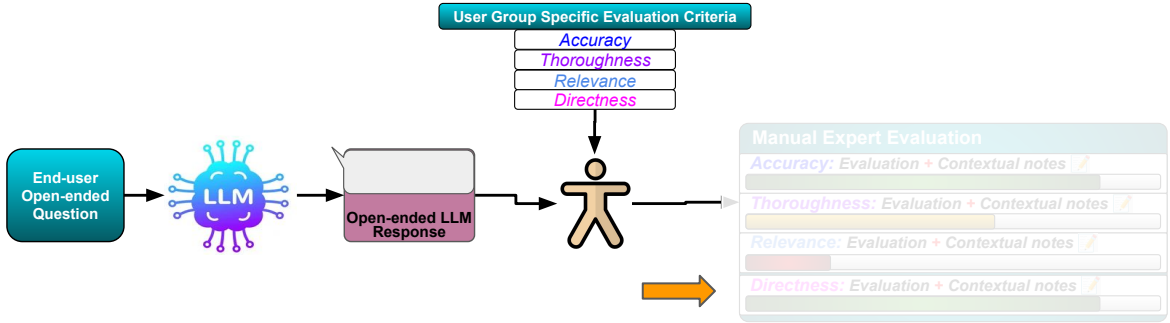


Reliability for paraphrased questions

- Question rephrasing
  - 16 Original Questions
  - Paraphrased 274 times combined using Google's "People also ask","More to ask", "Questions related to your search"
  - Manual verification
  - 10 questions and their rephrased 42 questions
- Generated responses
- Manually response similarity assessment by 2 researchers: completely, mostly, insufficiently similar
- Found not sensitive to paraphrasing as responses were mostly semantically similar

NYU

Manual Expert Evaluation: Reliability Experiment

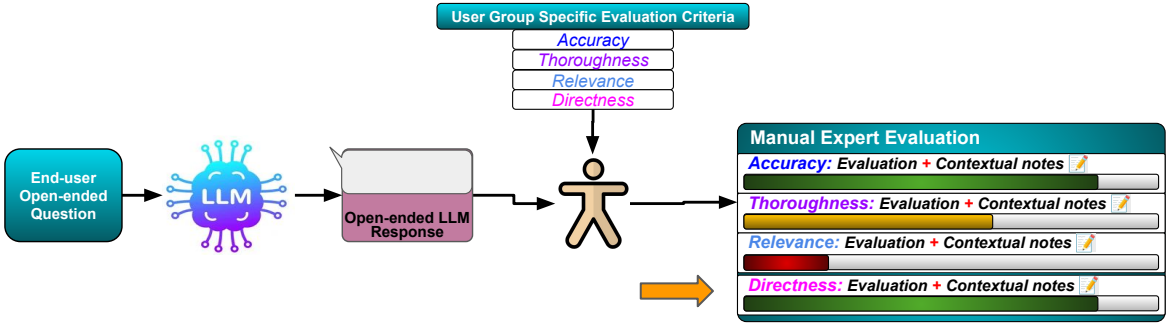


Reliability for repeated questions

- 10 questions repeatedly presented to LLMs 5 times to generate responses
- Manually response similarity assessment by 2 researchers: completely, mostly, insufficiently similar
- Found not sensitive to repetitions as responses were mostly semantically similar

NYU

Manual Expert Evaluation of GPT responses



Expert-led manual assessment

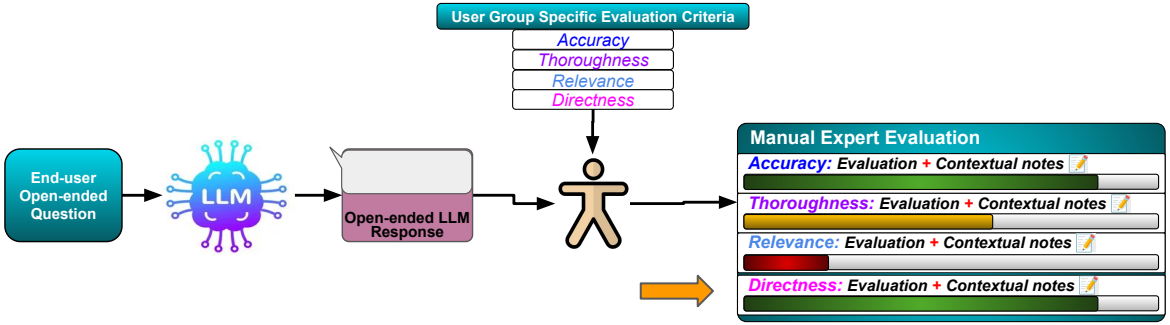
72 responses used to established IRR among 5 researchers with substantial level of agreement

Fleiss's κ: Accuracy 0.73, Thoroughness 0.63, Directness 0.76, 100% agreement on Directness

Thematic analysis of imperfect answers

NYU

Confirmation of Generalization of Issues in Gemini and Llama



GPT was evaluated on all 900 question

- 172 responses were imperfect

Found themes of issues were confirmed in Gemini and Llama

- Selective evaluation till saturation
- Evaluation of 304 responses of Llama and Gemini confirmed all issues

NYU

Meta-discussion

- Did you use experimentation artifacts borrowed from the community? No
- Did you attempt to replicate or reproduce results of earlier research as part of your work? Partially
- What can be learned from your methodology and your experience using your methodology?
- What did you try that did not succeed before getting to the results you presented?
- Did you produce any intermediate results including possible unsuccessful tests or experiments?

Future Directions

- Multi-turn evaluation
- Employ prompt engineering with or without few-shot
- Evaluation in privacy domain
- Automated evaluation of LLMs in security & privacy domain