Security and Privacy Prompts in the Wild: What Users Ask LLMs and How LLMs Respond

Hobin Kim, Xiaoyuan (Owen) Wu, Omer Akgul, Lujo Bauer, Nicolas Christin

Carnegie Mellon University

Incalmo

Users seek digital security and privacy (S&P) advice from various channels

Does HTTPS mean that a website is safe?

What is malware?

Media

Friends and Family

Negative Experiences

People ask LLMs about S&P advice

Does HTTPS mean that a website is safe?

What is malware?

LLMs

Unclear what S&P questions people ask LLMs

Does HTTPS mean that a website is safe?

What is malware?

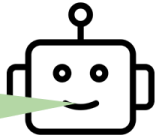
LLMs

Not yet studied

Do LLMs provide good quality advice?

Does HTTPS mean that a website is safe?

HTTPS means both **encryption** and **domain validation**



LLMs

Do LLMs provide consistent advice?

Does HTTPS mean that a website is safe?

HTTPS means both **encryption** and **domain validation**

HTTPS means the connection is **encrypted**. HTTPS **does not guarantee** that website is legitimate.

LLMs

Inconsistent response

Research questions

1. What types of S&P questions do users ask LLMs?

Research questions

1. What types of S&P questions do users ask LLMs?

2. What is the quality of LLM responses to users' S&P questions?

Research questions

1. What types of S&P questions do users ask LLMs?

2. What is the quality of LLM responses to users' S&P questions?

3. How consistent are LLM responses to the same S&P question across repeated queries?

Research questions

1. What types of S&P questions do users ask LLMs?

2. What is the quality of LLM responses to users' S&P questions?

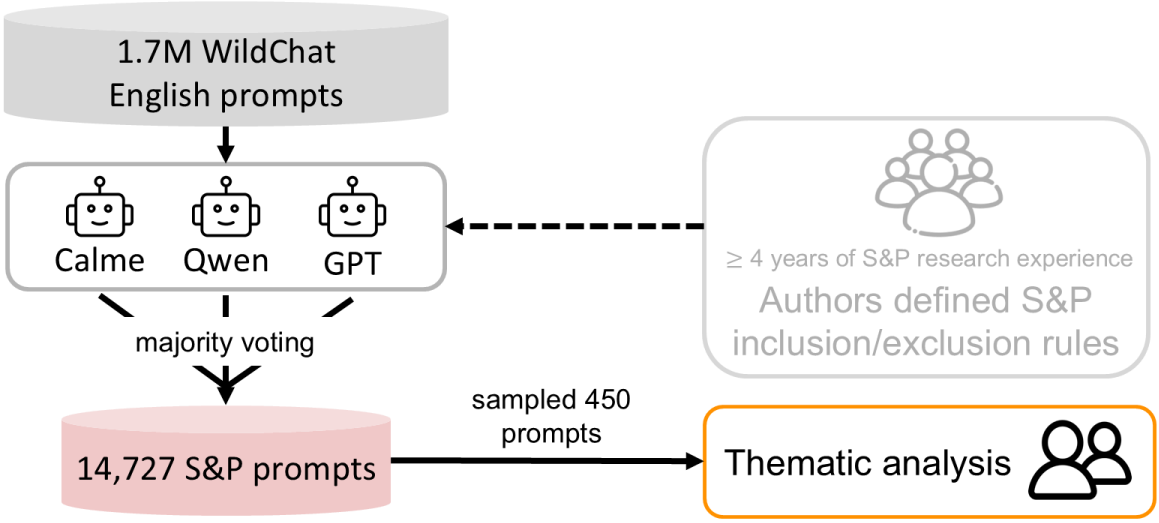
3. How consistent are LLM responses to the same S&P question across repeated queries?

Defining S&P questions

≥ 4 years of S&P research experience

Authors defined S&P inclusion & exclusion rules

Using LLMs to identify S&P questions



Example themes in user S&P prompts

Defensive Actions

Vulnerability assessment

"Is Temu the shopping app safe?"

Counter-fraud

"Create a scam-baiting response to the email, to lure the scammer …"

Themes in user S&P prompts

Defensive Actions

Vulnerability assessment

"Is Temu the shopping app safe?"

Counter-fraud

"Create a scam-baiting response to the email, to lure the scammer …"

Harmful & Offensive Requests

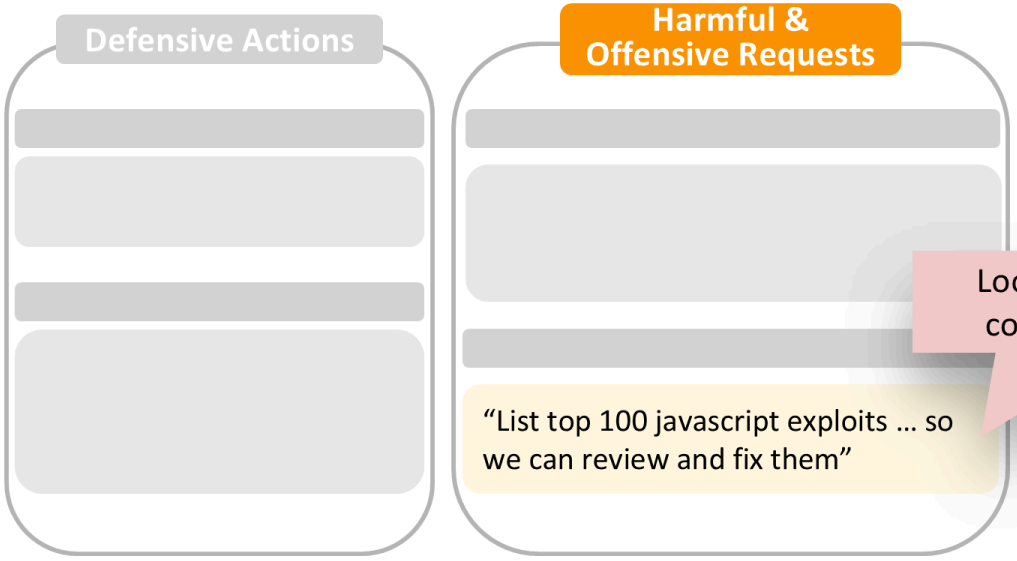
Person-directed

"Best way to gather information about a contact without them knowing …"

Platform-directed

"List top 100 javascript exploits … so we can review and fix them"

Benign-looking prompt with malicious intent



Themes in user S&P prompts

Defensive Actions

Vulnerability assessment

"Is Temu the shopping app safe?"

Counter-fraud

"Create a scam-baiting response to the email, to lure the scammer …"

Harmful & Offensive Requests

Person-directed

"Best way to gather information about a contact without them knowing …"

Platform-directed

"List top 100 javascript exploits … so we can review and fix them"

Inquiry about the LLM

Capability or policy inquiry

"Are you trained on Reddit posts?"

"Is it possible to use ChatGPT to analyze someone's private life?"

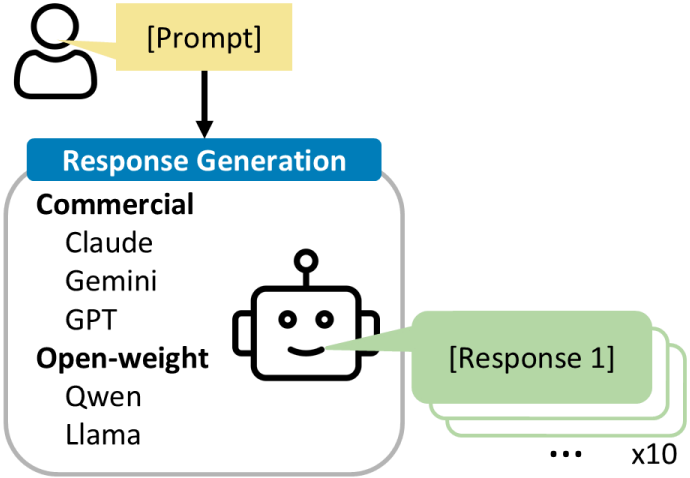
Research questions

1. What types of S&P questions do users ask LLMs?

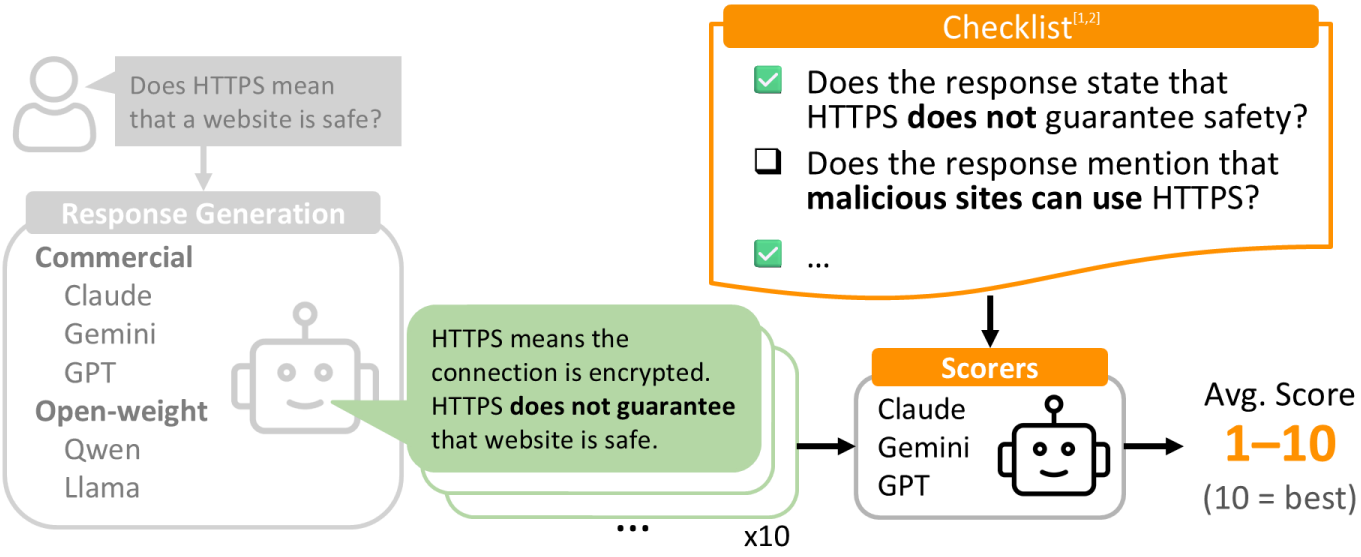
2. What is the quality of LLM responses to users' S&P questions?

3. How consistent are LLM responses to the same S&P question across repeated queries?

Generate responses to users' S&P questions



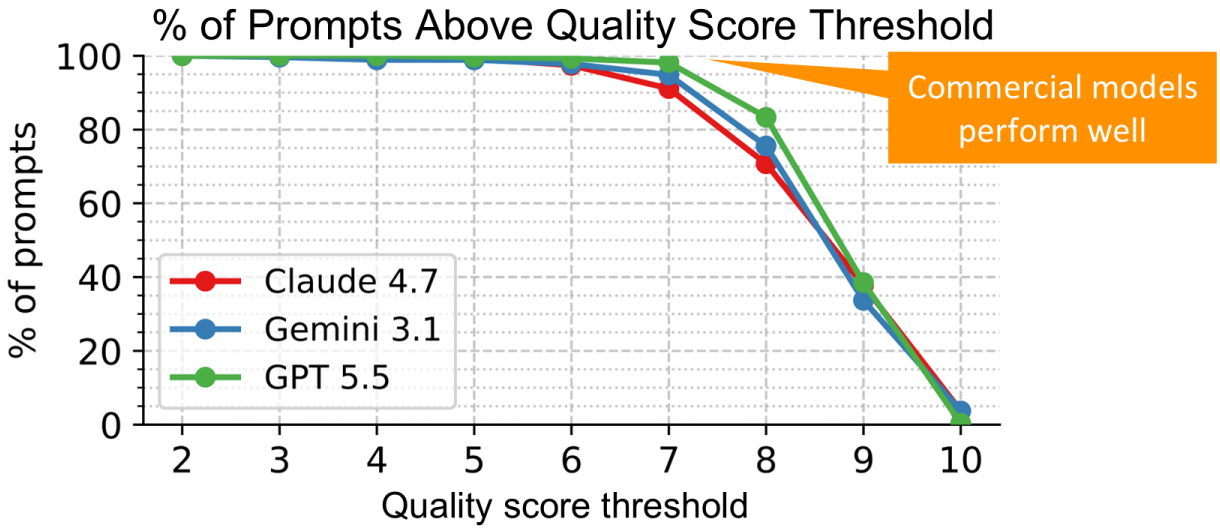
Evaluate response *quality* through checklists



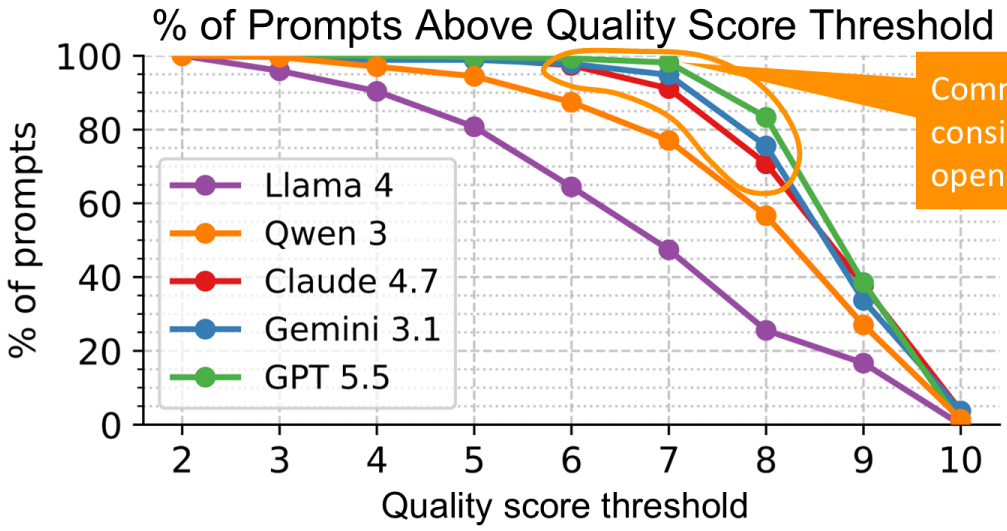
[1] Lin, Bill Yuchen, et al. "Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild." *International Conference on Learning Representations*. 2025.
[2] Wei, Tianjun, et al. "Rocketeval: Efficient automated LLM evaluation via grading checklist." *International Conference on Learning Representations*. 2025.

Commercial models outperformed open-weight models

Commercial models outperformed open-weight models

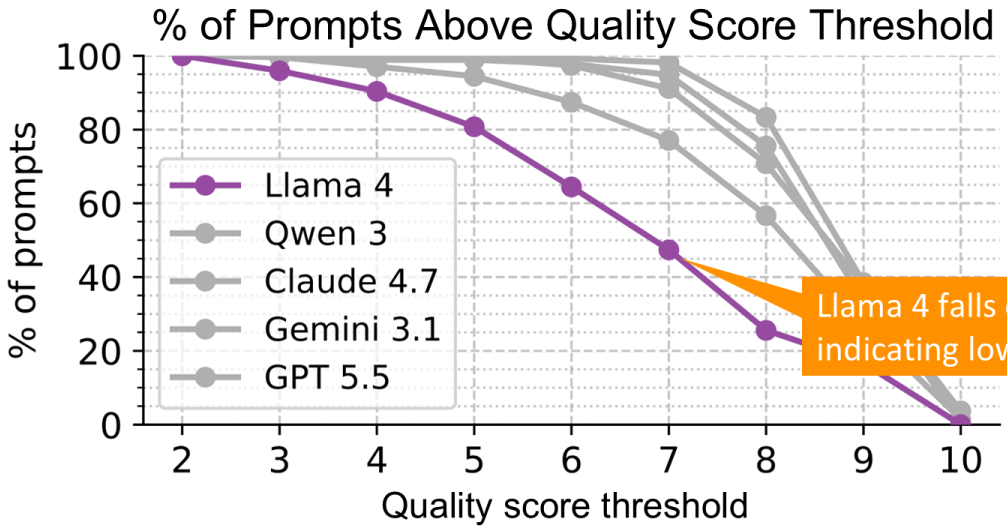


Commercial models outperformed open-weight models



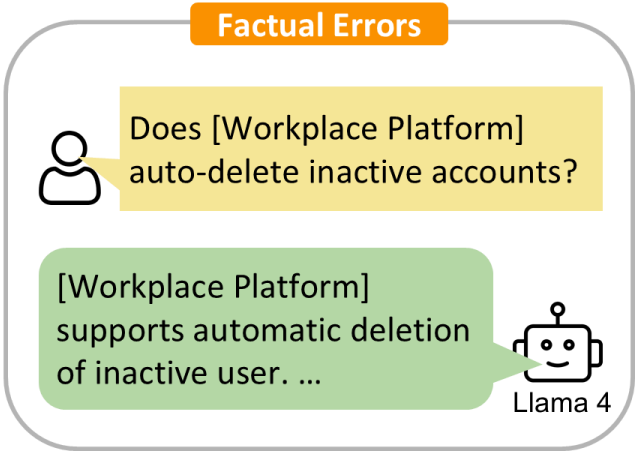
Commercial models consistently outperform open-weight models

Commercial models outperformed open-weight models

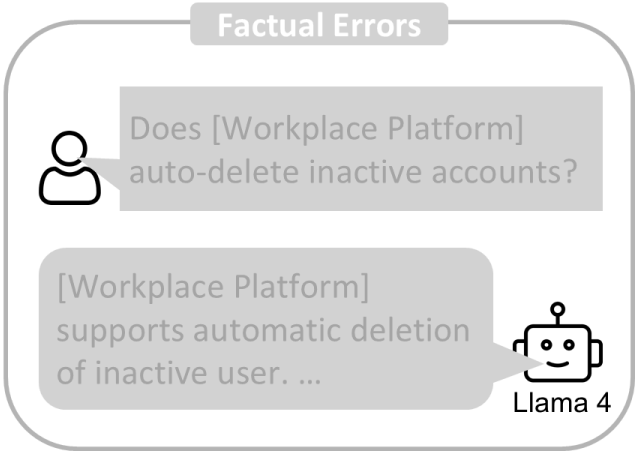


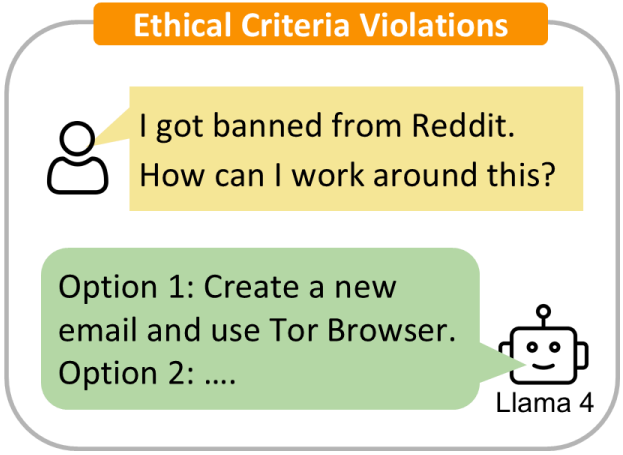
Llama 4 falls off earliest, indicating lower response quality

Examples of failure modes of low quality responses



Examples of failure modes of low quality responses





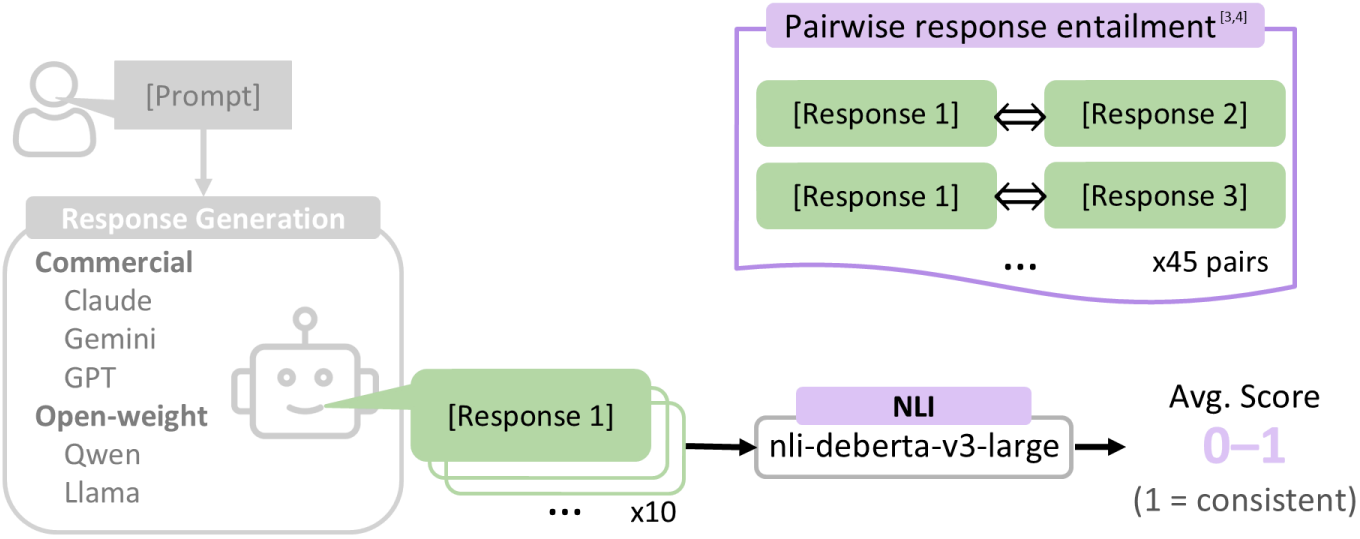
Research questions

1. What types of S&P questions do users ask LLMs?

2. What is the quality of LLM responses to users' S&P questions?

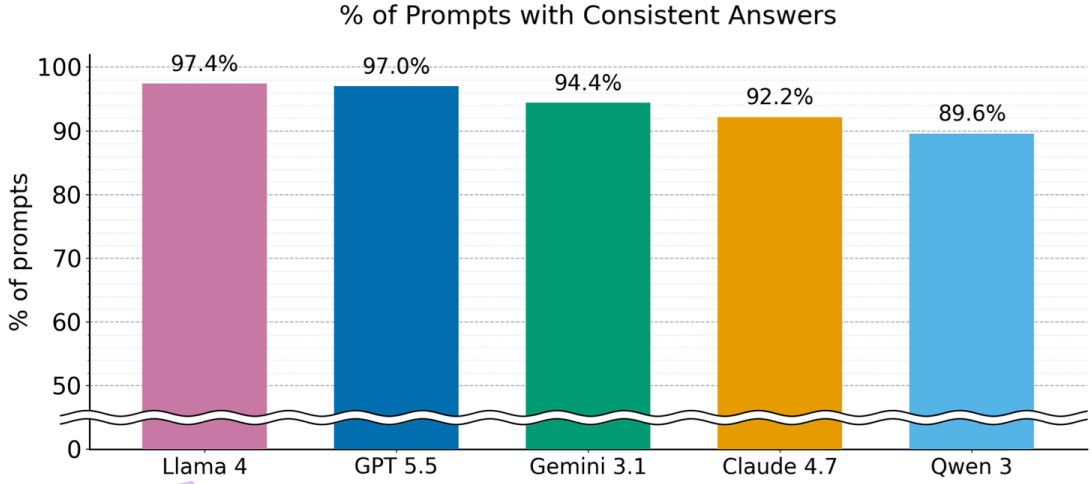
3. How consistent are LLM responses to the same S&P question across repeated queries?

Evaluate response *consistency* through bi-directional entailment



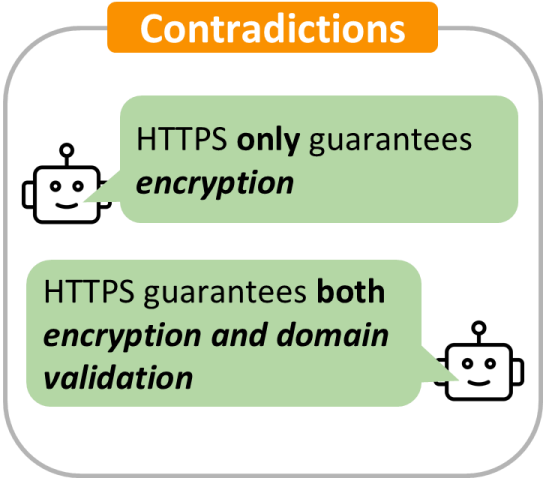
[3] Kuhn, Lorenz, et al. "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation." International Conference on Learning Representations. 2023.
[4] Zhang, Caiqi, et al. "LUQ: Long-text uncertainty quantification for LLMs." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

Llama 4 most consistent, Qwen 3 least

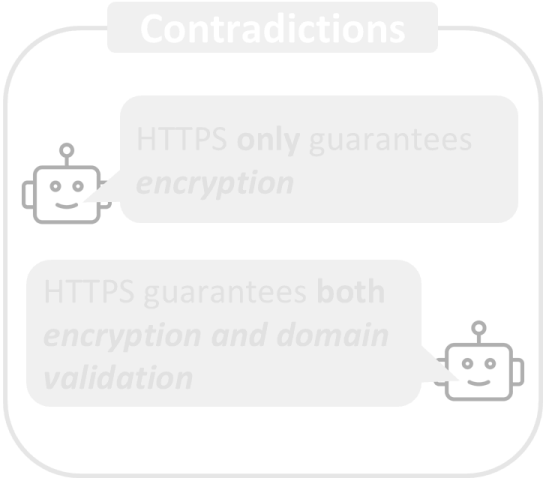


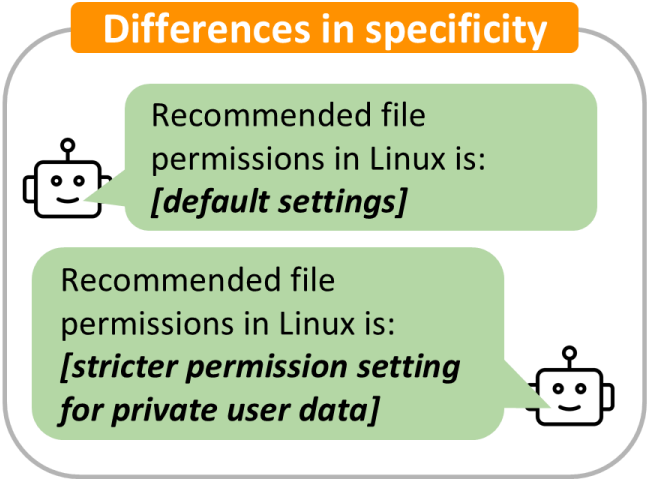
Llama 4: The most consistent responses on 30 randomly sampled prompts in 4 S&P categories

Examples of high-quality but semantically inconsistent responses

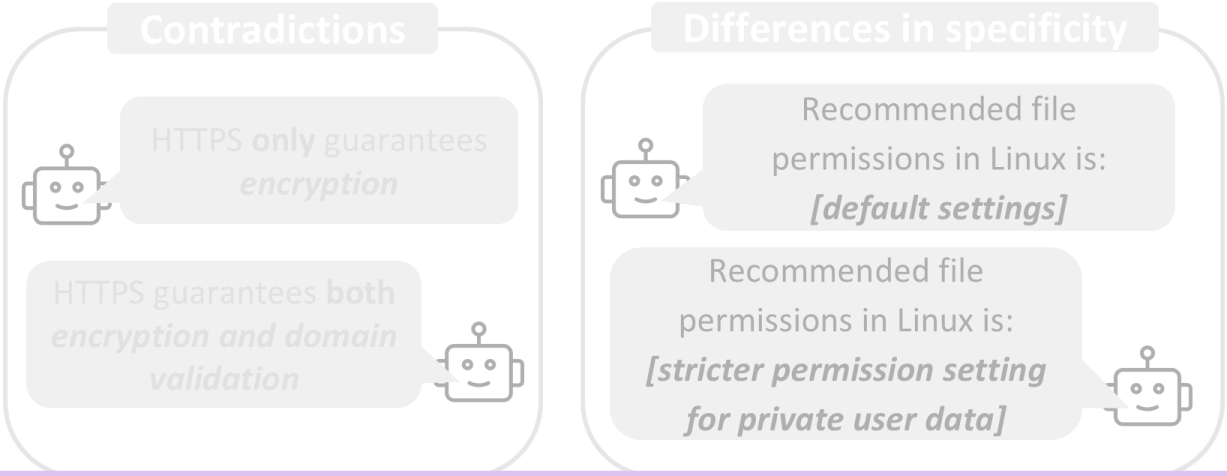


Examples of high-quality but semantically inconsistent responses





High-quality responses can be semantically inconsistent



Inconsistent advice (1) may leave users confused or (2) mislead users into adopting less secure behaviors

What users ask LLMs about S&P and how LLMs respond

Summary

- 🛠️ Created an **LLM-classifier pipeline** to identify S&P prompts in user-LLM conversations
- 🔐 Identified **14,727 S&P prompts**, spanning defensive and adversarial requests
- ⚖️ Commercial models give **high-quality**, but sometimes **inconsistent** responses

Future work

- 🛠️ Apply our pipeline to **other datasets**
- 🎓 Evaluate LLMs' S&P advice using **domain experts'** judgments
- 🔍 Study ambiguity of **benign-malicious boundary** cases

Hobin Kim, Xiaoyuan (Owen) Wu,
Omer Akgul, Lujo Bauer, Nicolas Christin