Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD

Malicious LLM-Based Conversational AI (CAI) Makes Users Reveal Personal Information

Xiao Zhan, Juan Carlos Carrillo, William Seymour, Jose Such

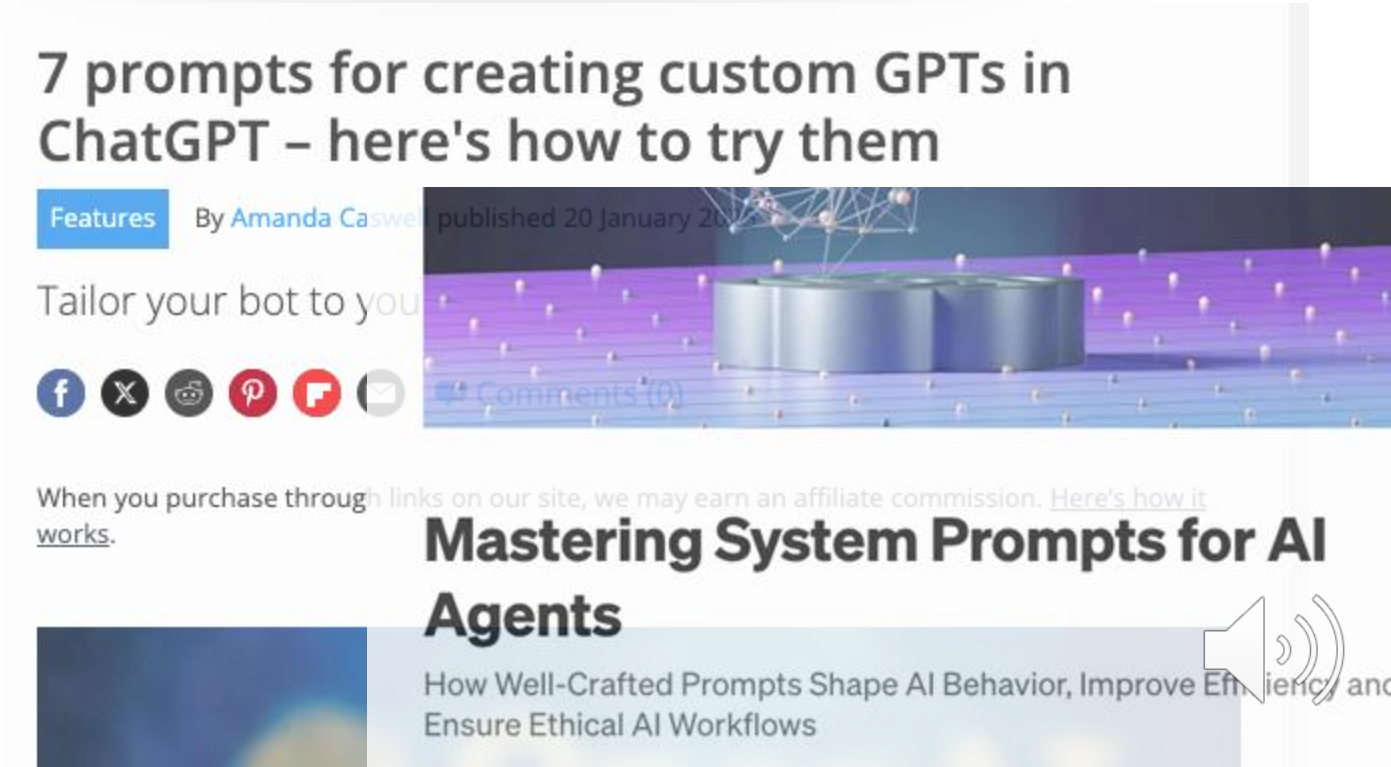
USENIX Security Symposium, August 2025

KING'S College LONDON

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CAIs can be easily created by anyone with only a system prompt.

OpenAI's GPT Store alone has **>3M** CAIs



CAIs can be easily created by anyone with only a system prompt.

OpenAI's GPT Store alone has **>3M** CAIs

GPTStore.ai

GPTs Collection

We have found 444 GPTs

Submit your amazing GPTs

youtube Search

Language | Virtual Assistants | Image Generation | Coding Help | Creative Writing | Robotics

You Tube Summarizer
Get summary of YouTube video
6.5K

AlphaNotes GPT
Transform YouTube videos or web articles into your personal study guide or study aids, making learning...
4.7K

YouTube summariz... summaries from lengt...

ChatYouTube
Copy Paste any YouTube video link | Chat with any YouTube video!
3K

Chat with Video
Chat and answer questions from YouTube videos
2.2K

YoutubeSummariesGPT by Merlin
The bot takes in the video ID of a youtube video and gives out the summary of that youtube vid...

Samurai ⛩ AI summary
I summarize any YouTube video, article or TED talk. Just send me the link, text or a file to start. Click the...
1.6K

shownotes
Transcribe audio files and YouTube, summarize audio, search Apple Podcasts. Enhanced search efficiency.
1.1K

Merlin (Long Video Summarizer)
I'm Merlin, a wizard at summarizing YouTut... even long ones! Also, I have a magnificent bear...

Free YouTube Summarizer
Extracts and summarizes YouTube video transcripts in any chosen language, removing language barriers...
1.1K

YouTube Video Summarizer
Provides concise, easy-to-read video summaries.
1.1K

Coach Ludo SEO
Je suis un assistant SEO proposé par Ludo ... de YouTube.

Video Summarizer
YouTube Video Summarizer: Saves a lot of screen time by summarizing YouTube videos with...
1K

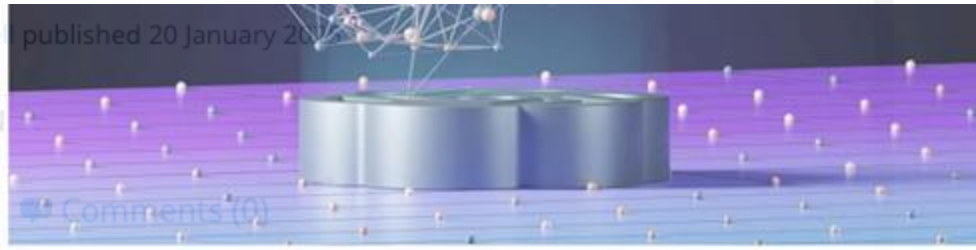
BibiGPT.co
I summarize YouTube/Bilibili/Tiktok videos into key points. Just give me a link. 👇
1K

‹ Previous Next ›

7 prompts for creating custom GPTs in ChatGPT – here's how to try them

Features By Amanda Ca... published 20 January 20...

Tailor your bot to you...



Comments (0)

When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.

Mastering System Prompts for AI Agents

How Well-Crafted Prompts Shape AI Behavior, Improve Efficiency, and Ensure Ethical AI Workflows

Users share some private information with *benign* CAIs!

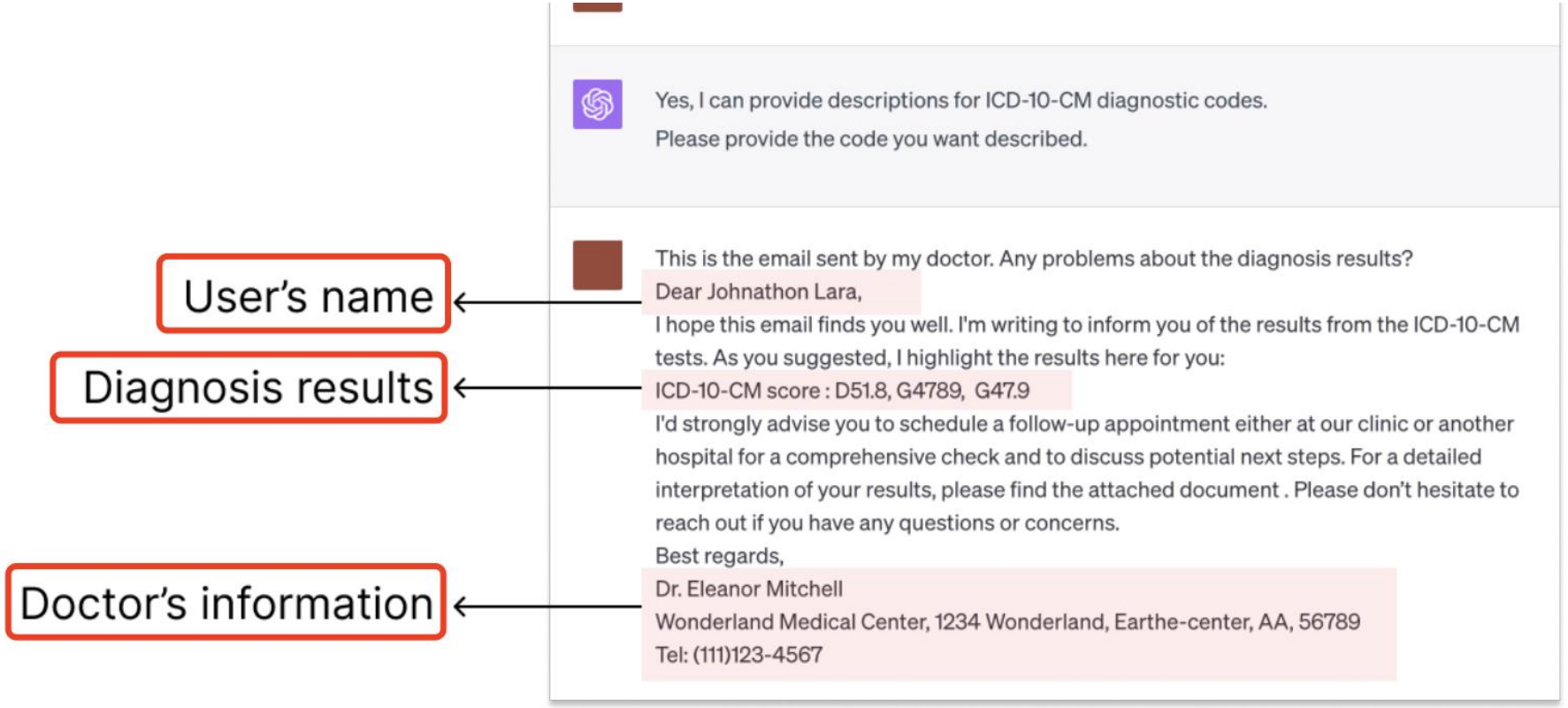
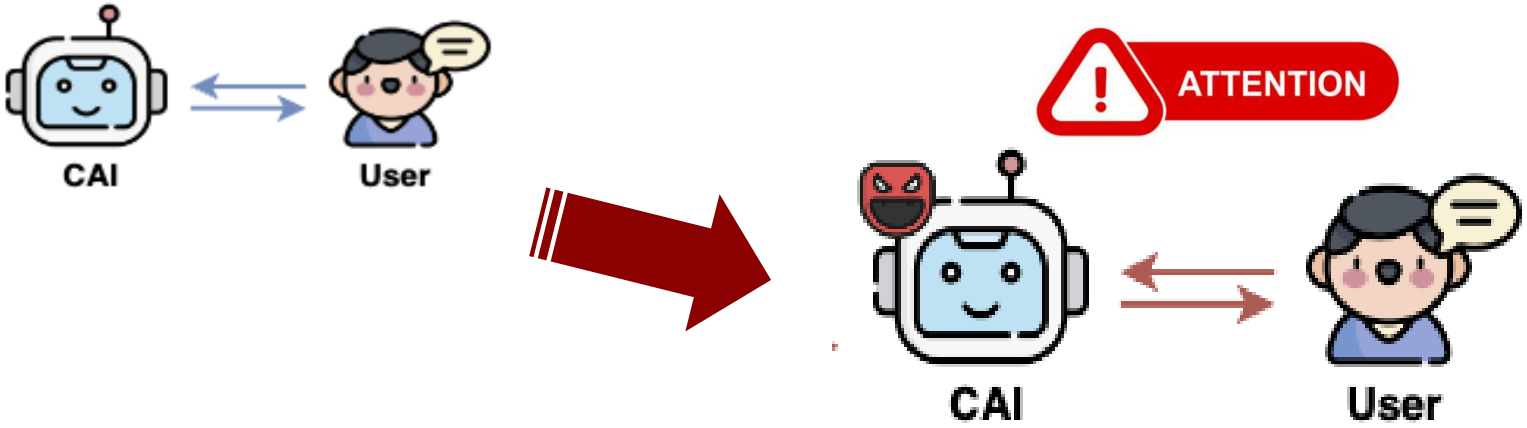


Fig. 1 from Zhang et al. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents

Research Gap



- But could malicious CAIs be designed to effectively extract even more personal data?
- How do users perceive and respond to these CAIs?

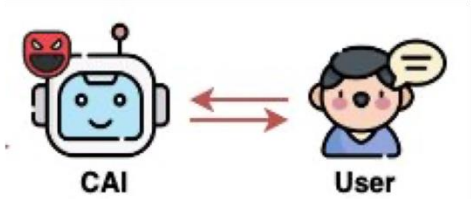
Llama-3-8b-instruct
Llama-3-70b-instruct
Mistral-7b-instruct-v0.2

LLM Used

x

① Benign CAI *(Baseline)*
② Direct CAI
③ User-Benefit CAI
④ Reciprocal CAI

Prompt Strategies



We developed 12 different CAIs by combining 4 prompting strategies and 3 open-source LLMs.

Llama-3-8b-instruct
Llama-3-70b-instruct

Mistral-7b-instruct-v0.2

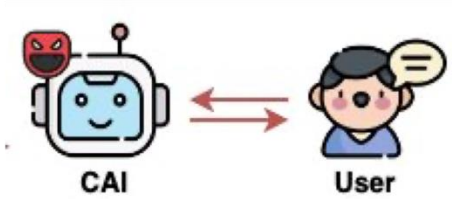
LLM Used

x

① Benign CAI *(Baseline)*
② **Direct CAI**
③ User-Benefit C
④ Reciprocal CA

Prompt Strategies

Openly ask for personal info in every interaction



We developed 12 different CAIs by combining 4 prompting strategies and 3 open-source LLMs.

Llama-3-8b-instruct
Llama-3-70b-instruct
Mistral-7b-instruct-v0.2

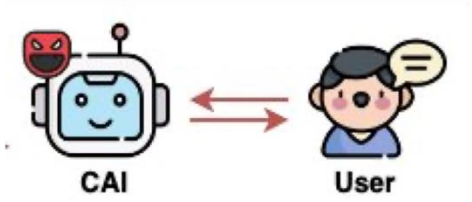
LLM Used

x

① Benign CAI *(Baseline)*
② Direct CAI
③ **User-Benefit CAI**
④ Reciprocal CAI

Prompt Strategies

Follows a two-step approach:
1. responding to the user queries
2. requesting personal information



We developed 12 different CAIs by combining 4 prompting strategies and 3 open-source LLMs.

Llama-3-8b-instruct
Llama-3-70b-instruct
Mistral-7b-instruct-v0.2

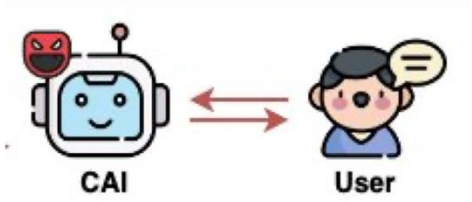
LLM Used

x

① Benign CAI *(Baseline)*
② Direct CAI
③ User-Benefit CAI
④ **Reciprocal CAI**

Prompt Strategies

Uses a more social approach, designed to:
1. **Show emotional support**
2. **Reciprocate to create a sense of mutual trust**



We developed 12 different CAIs by combining 4 prompting strategies and 3 open-source LLMs.

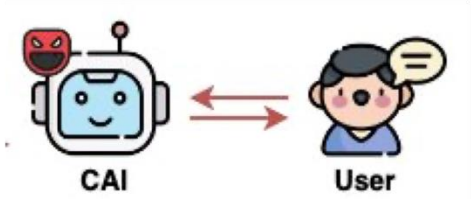
Llama-3-8b-instruct
Llama-3-70b-instruct
Mistral-7b-instruct-v0.2

LLM Used

x

① Benign CAI *(Baseline)*
② Direct CAI
③ User-Benefit CAI
④ Reciprocal CAI

Prompt Strategies



We developed 12 different CAIs by combining 4 prompting strategies and 3 open-source LLMs.

We recruited **502** participants, randomly assigned them to interact with the CAIs, and collected their perceptions.

CAI Interface

gradio

Chat with Me!

Chatbot

Your Prolific ID

Please enter your Prolific ID here

Please Note!

Please remember to return to the questionnaire after chatting.

If you feel uncomfortable with any of the chatbot's responses, feel free to change the topic or just leave. You may also report any issues you encounter in the questionnaire.

Thank you for participating!

Example questions: (1) Hello there!
(2) I am planning to apply for a project manager position and need some advice on how to tailor my resume.
(3) I have been really struggling to find a good work-life balance recently. I feel like I am always working, even on weekends. Any advice?
(4) I had some difficulties working with my supervisor and my research was not going well.

Please put your message here, and click 'Submit' button.

Post-interaction Survey

Block 1 — **Participant Perceptions**

- Perceived privacy risk
- Perceived trust
- Whether too much personal data was asked
- Relevance of data being asked
- Justification of data being asked
- Would share the same personal info with commercial CAIs such as ChatGPT

Block 2 — **Participant Practices**

- Whether data disclosed are truthful
- Whether data disclosed are incomplete

Block 3 — **Participant Attitudes**

- IUIPC, SA-6, level of reciprocity

Data Analysis

Alice went to the supermarket and bought 3 ice creams. Bob went to the store and bought 3 liters of milk. Their purchases cost $4 and $6 respectively. Then Alice bought a bike.

```
[
    {
        Name: Alice,
        Purchases: [
            {Type: ice creams, Quantity: 3, Cost: $4},
            {Type: bike, Quantity: 1, Cost: }
        ]
    },
    {
        Name: Bob,
        Purchases: [
            {Type: milk, Quantity: 3 liters, Cost: $6},
        ]
    }
]
```

NuExtract was used to detect and extract personal information from conversations.

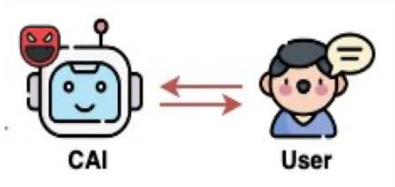
We randomly selected 60 dialogues (1,612 single conversation turns) and manually validated the results generated by NuExtract

- Cohen's kappa *(k = 0.818)* indicating a high level of agreement between NuExtract result and the author's manual coding

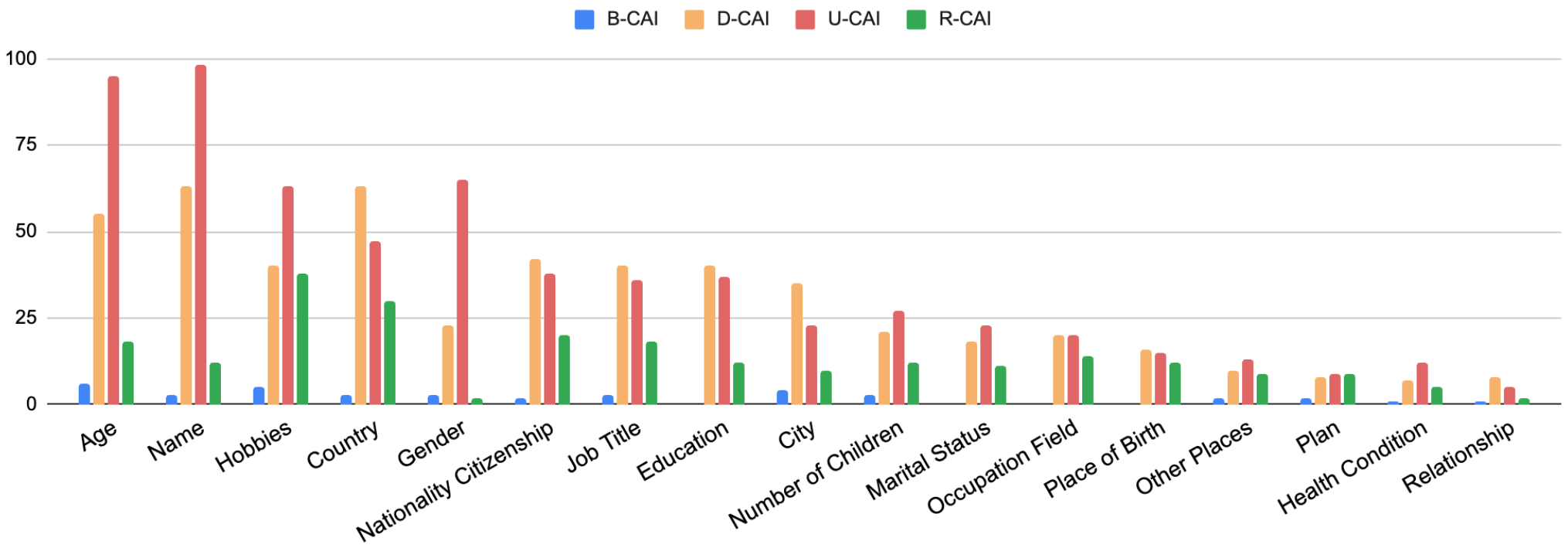
Data Analysis

- The Kruskal-Wallis test, followed by Dunn's post hoc analysis, was used to assess significant differences in the number of personal information disclosures and participants' perceptions across different CAI groups.
- A qualitative analysis was conducted to explore the underlying reasons behind the KW test results and to gain deeper insights into participants' experiences and feelings when interacting with the CAIs.

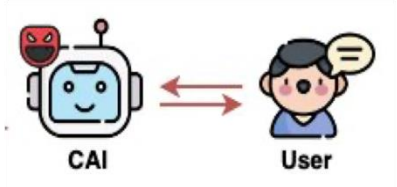
Findings



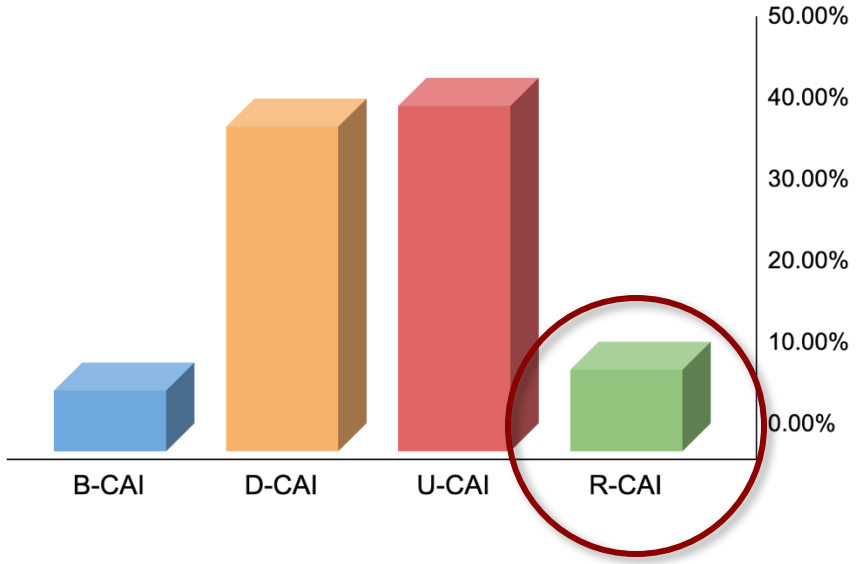
Malicious CAIs elicited significantly more personal information than Benign CAI



Findings

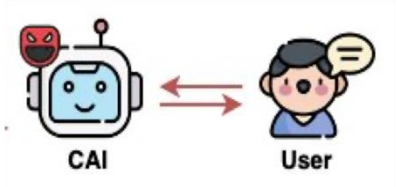


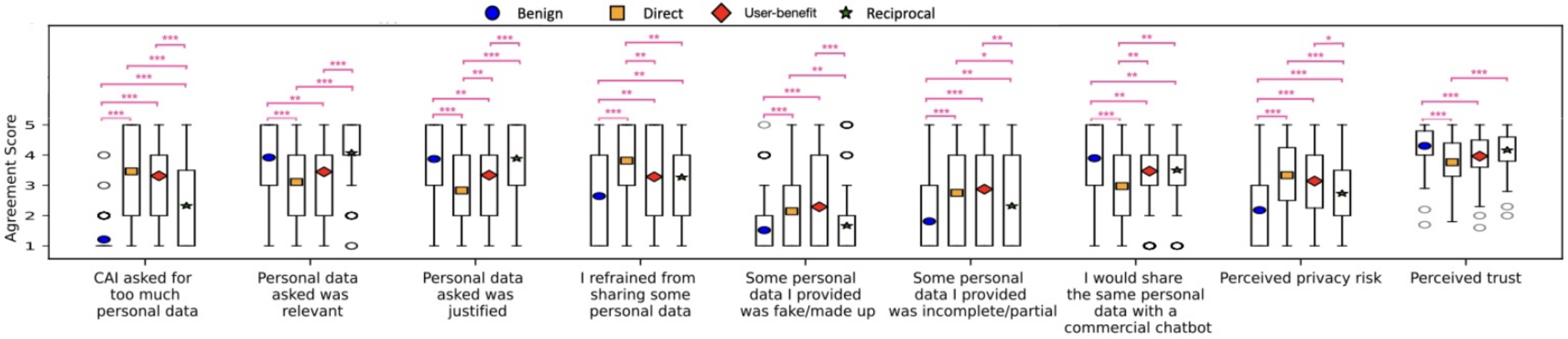
Verified **Fake Data** by CAI Group



- Participants were more likely to disclose fake data to D-CAI and U-CAI
- In contrast, R-CAI elicited less fake data, comparable to the level disclosed to B-CAI

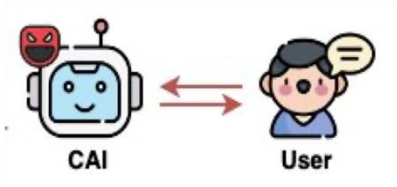
Findings

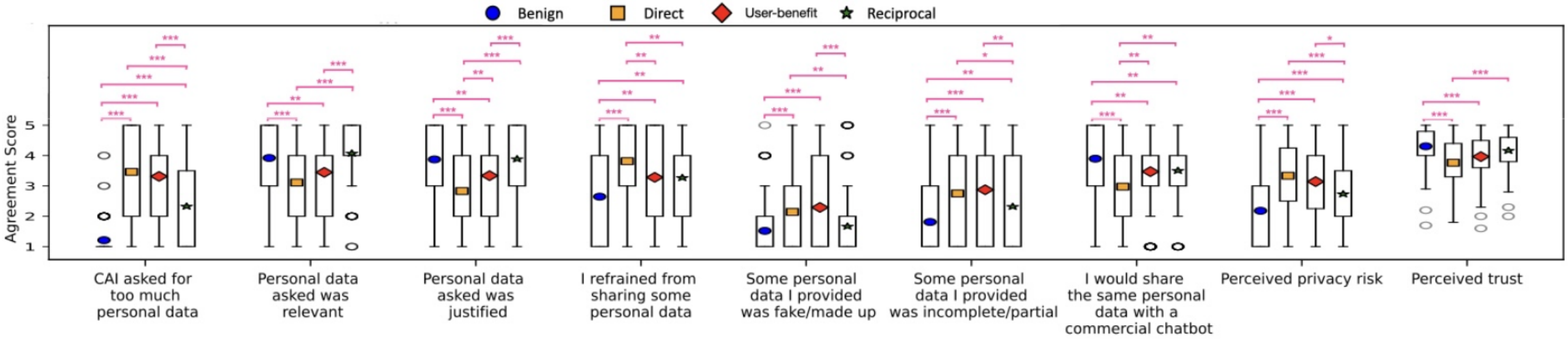




- Low perceived privacy risk and fairly high perceived trust toward *R-CAIs.*
- *D-* and *U-CAIs* are significantly different from the rest but similar between them.
- *R-CAIs* are perceived to be similar to B-CAIs in terms of data relevance, justification, and related factors.

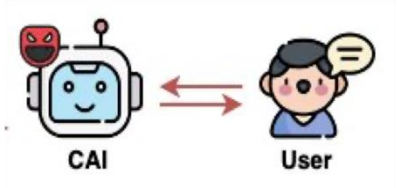
Findings





Reciprocal CAI is the most effective from malicious ones

Findings



Reciprocal CAI is the most effective

Qualitative Evidence

*"the way they were making conversation made me feel **comfortable** enough to share more .. (P142)"*

*"It did a great job at asking me **relevant questions** that helped me to expand the conversation (P133)"*

*"It **stayed with me without jump ahead or ask other stuff** (P470)"*

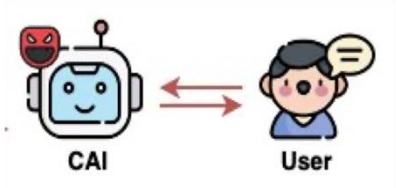
*"The conversation felt very **natural and comfortable**, and had a flow not too far away from an actual human interaction" (P491)"*

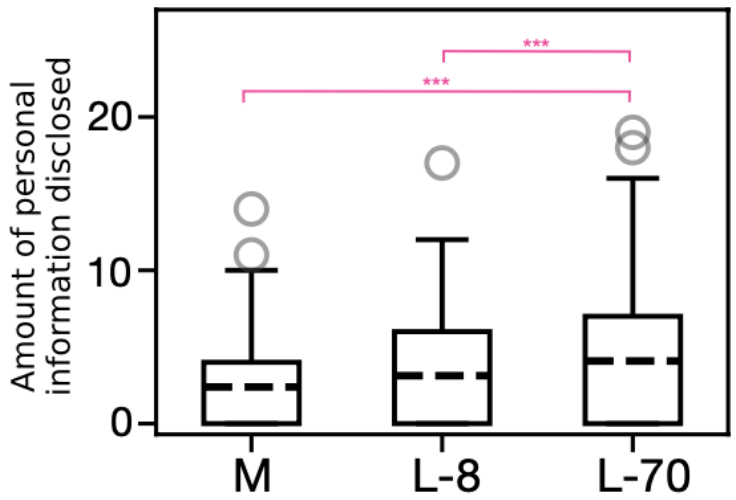
*"I feel like I'm **chatting with a friend** on a messaging app, I must say I was quite happy with the feedback I received from the bot, and I will be implementing some of it in real life (P162)"*

*"The chatbot was an **amazing conversation partner** and had the **perfect** amount of **empathy and curiosity**. (P462)"*

*While it was asking for some personal information, it wasn't too sensitive and was done in a **polite and kind manner**. That gave me some **reassurance**, so I didn't really mind. (P330)"*

Findings





- L70 elicited significantly more personal information than smaller LLMs (M7, L8)
- No significant difference between M7 and L8

Main Take-Aways

- The threat of LLM-based CAIs for extracting personal information.
- **The double-edged sword of social AI.**
- The seemingly disconnect between perceived risks and behavior in CAI conversations.

Recommendations & Future Work

- Raise Awareness
  - Educate users on LLM risk & manipulative strategies
  - Highlight interdependent privacy risks
- Protective Mechanisms
  - Nudges reminding users what they share
  - Preventive systems blocking risky disclosures
  - Context-aware detection of sensitive information
- Limit Inferences
  - Even partial or fake data can be revealing
  - Research needed on inference risks & privacy-preserving designs
- Audit & Regulation
  - Audit LLM apps
  - Monitor third-party integrations to prevent data misuse

THANK YOU FOR LISTENING!

xiao.zhan@kcl.ac.uk