Fostering Appropriate Reliance on Large Language Models

Sunnie S. Y. Kim

https://sunniesuhyoung.github.io

Appropriate reliance is key to human-AI complementarity

Overreliance

Relying on inaccurate AI outputs

Appropriate reliance

Relying on accurate AI outputs &
Not relying on inaccurate AI outputs

Underreliance

Not relying on accurate AI outputs

WIRED SECURITY POLITICS GEAR THE BIG STORY MORE SIGN IN SUBSCRIBE

BENJ EDWARDS, ARS TECHNICA BUSINESS OCT 30, 2024 8:00 AM

OpenAI's Transcription Tool Hallucinates. Hospitals Are Using It Anyway

In health care settings, it's important to be precise. That's why the widespread use of OpenAI's Whisper transcription tool among medical workers has experts alarmed.

PHOTOGRAPH: MILJAN ŽIVKOVIĆ/GETTY IMAGES

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

Share full article 267

Steven A. Schwartz told a judge considering sanctions that the episode had been "deeply embarrassing." Jefferson Siegel for The New York Times

https://www.wired.com/story/hospitals-ai-transcription-tools-hallucination/
https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

Risks from inappropriate reliance on AI

EU Artificial Intelligence Act

“Remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system”

(EU AI Act, Article 14 Human Oversight)

NIST

“Risk from confabulations may arise when users believe false content — often due to the confident nature of the response — leading users to promote false information.”

(NIST AI RMF, Section 2.2 Confabulation)

EU AI Act: Artificial Intelligence Act (Regulation (EU) 2024/1689), https://artificialintelligenceact.eu/
NIST AI RMF: Artificial Intelligence Risk Management Framework (AI RMF 1.0), https://doi.org/10.6028/NIST.AI.100-1

How can we foster appropriate reliance on AI?

Onboarding
Nourani et al. *HCOMP 20*

Explanations
Vasconcelos et al. *CSCW 23*

(Un)certainty information
Zhang et al. *FAccT 20*

Cognitive forcing functions
Buçinca et al. *CSCW 21*

Traditional AI

? ? ?

Generative AI

General purpose

Natural language

Interactive

Public excitement

Thoughtful empirical studies on user perceptions and behaviors

1. LLMs providing explanations (w/wo inconsistencies) and sources

Fostering Appropriate Reliance on LLMs: The Role of Explanations, Sources, and Inconsistencies.
Kim, Vaughan, Liao, Lombrozo, Russakovsky. *CHI 25* 🏅

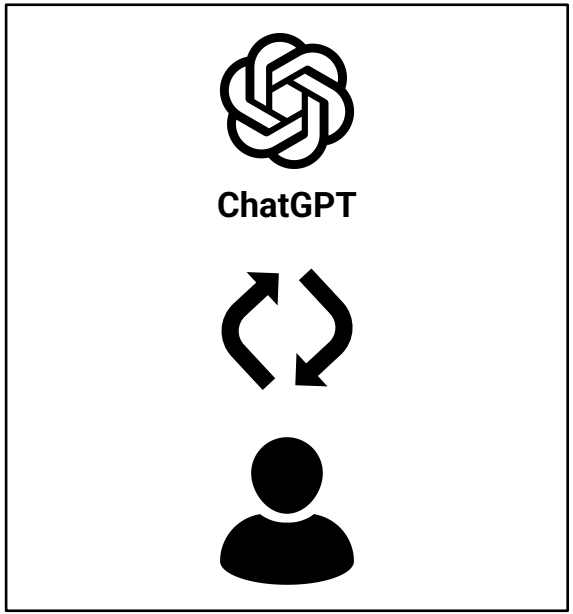
2. LLMs expressing uncertainty (w/wo using personal pronouns)

"I'm Not Sure, But...": Examining the Impact of LLMs' Uncertainty Expression on User Reliance and Trust.
Kim, Liao, Vorvoreanu, Ballard, Vaughan. *FAccT 24*

3. Portraying LLMs as machines vs. tools vs. companions

Portraying LLMs as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them.
Chen, Kim, Dharmasiri, Russakovsky, Fan. *CogSci 25*

Study 1: Think-Aloud Study



N=16 (diverse LLM knowledge and use),
Each participant solves **3 QA tasks**
via multi-turn interactions with ChatGPT

Qualitative studies can help identify "what to evaluate" and "why"

QA task:

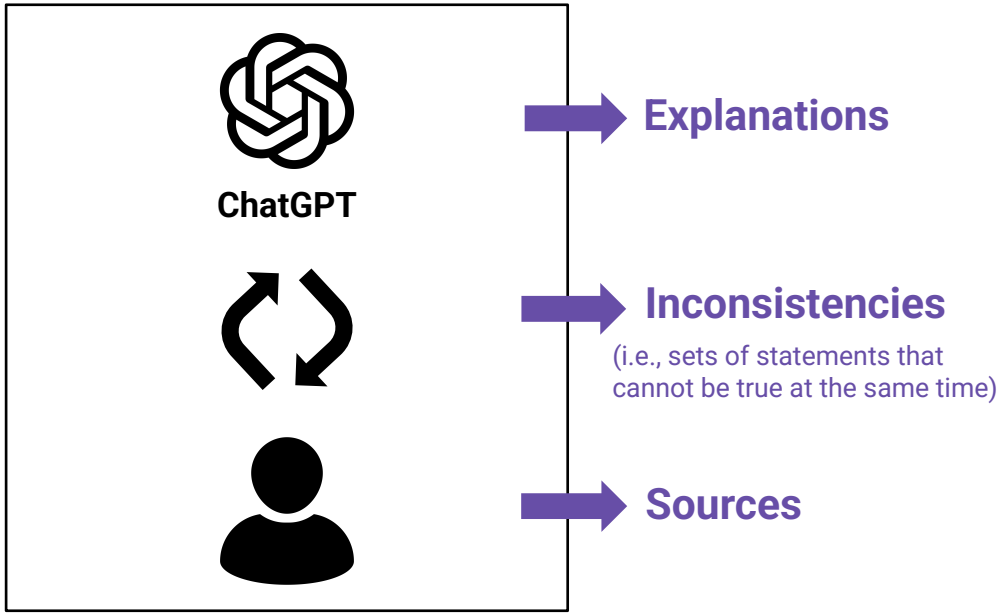
Determine the correct answer to the question

General domain factual question (e.g., Has Paris hosted the Summer Olympics more times than Tokyo?)

Health/Legal domain factual question (e.g., Is it illegal to collect rainwater in Colorado?)

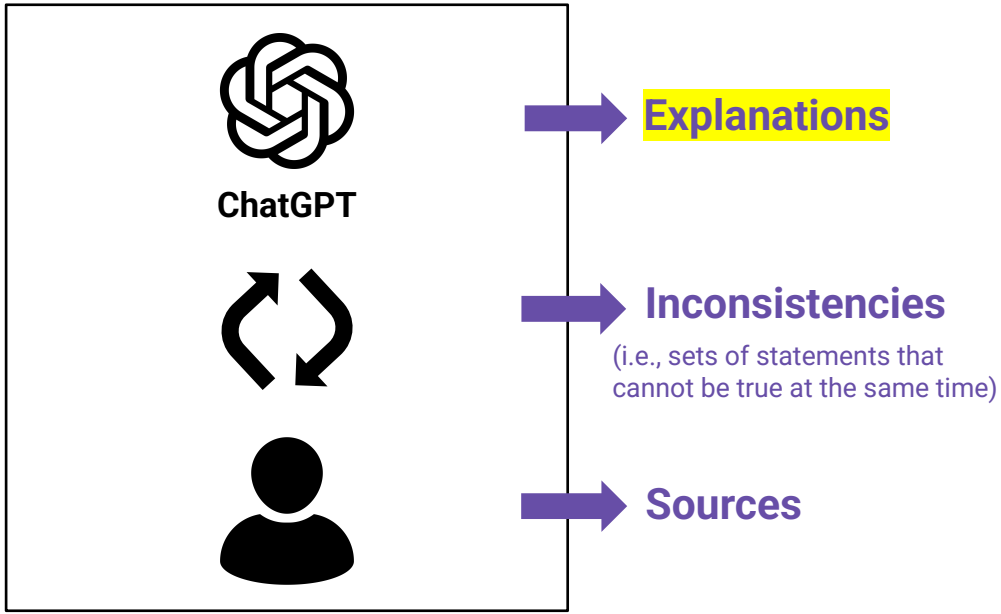
Math question (e.g., Sue puts one grain of rice on the first square of a Go board and puts double the amount on the next square. How many grains of rice does Sue put on the last square?)

Study 1: Think-Aloud Study



N=16 (diverse LLM knowledge and use),
Each participant solves **3 QA tasks**
via multi-turn interactions with ChatGPT

Study 1: Think-Aloud Study



N=16 (diverse LLM knowledge and use),
Each participant solves **3 QA tasks**
via multi-turn interactions with ChatGPT

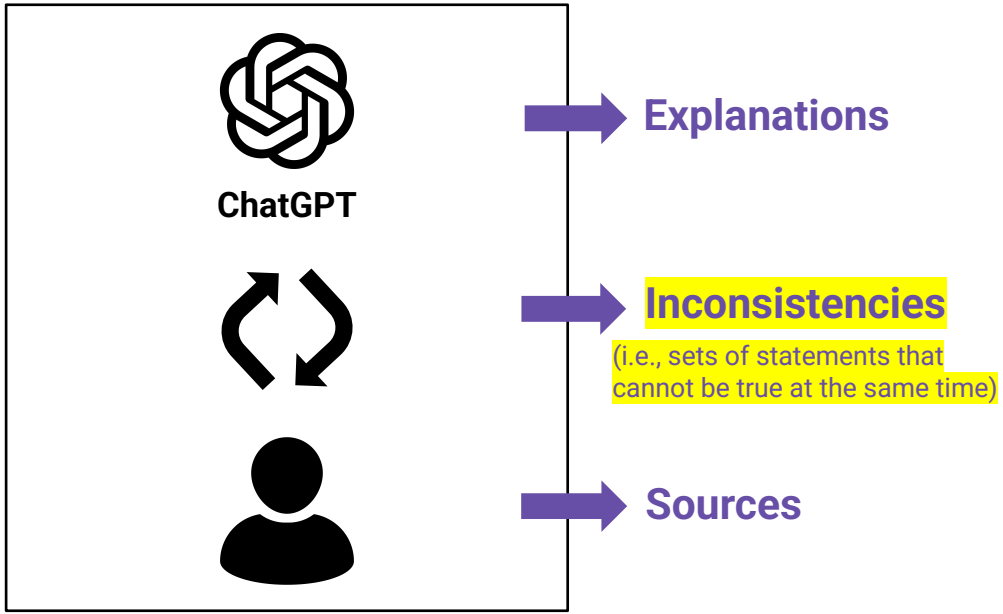
Task question (example)
Do more than two thirds of South America's population live in Brazil?

LLM response (example)

Yes, more than two-thirds of South America's population live in Brazil because Brazil is the largest and most populous country on the continent. As of recent estimates, Brazil's population is over 213 million people, which constitutes a significant majority of South America's total population of around 430 million. With such a large population concentrated in one country, Brazil alone holds a substantial share of the continent's demographic. Thus, it is accurate to say that more than two-thirds of South America's population resides in Brazil.

Sources:
1. https://www.worldometers.info/world-population/brazil-population/
2. https://worldpopulationreview.com/continents/south-america
3. https://en.wikipedia.org/wiki/List_of_South_American_countries_by_population

Study 1: Think-Aloud Study



N=16 (diverse LLM knowledge and use),
Each participant solves **3 QA tasks**
via multi-turn interactions with ChatGPT

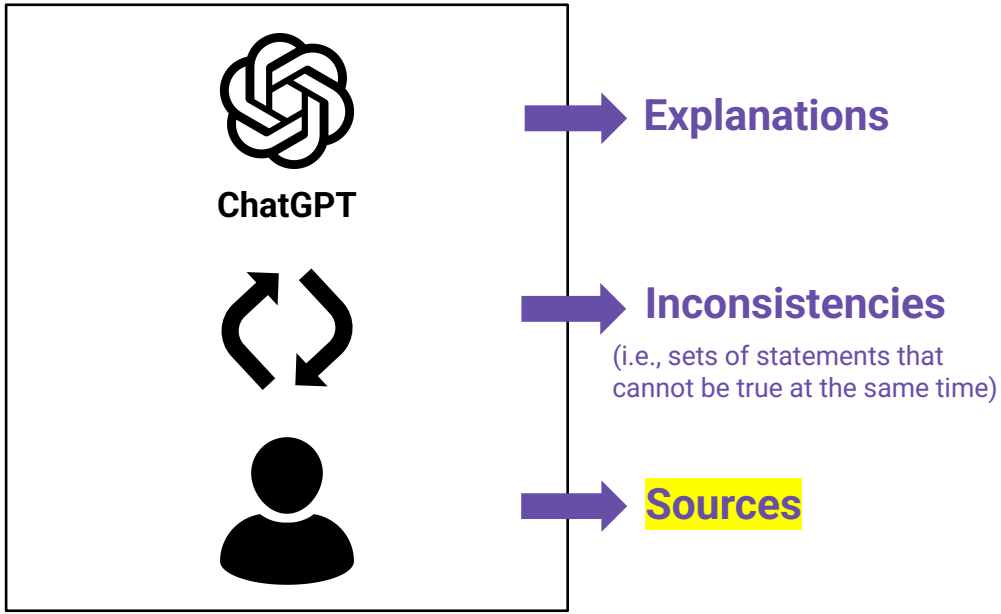
Task question (example)
Do more than two thirds of South America's population live in Brazil?

LLM response (example)

Yes, more than two-thirds of South America's population live in Brazil because Brazil is the largest and most populous country on the continent. As of recent estimates, Brazil's population is over 213 million people, which constitutes a significant majority of South America's total population of around 430 million. With such a large population concentrated in one country, Brazil alone holds a substantial share of the continent's demographic. Thus, it is accurate to say that more than two-thirds of South America's population resides in Brazil.

Sources:
1. https://www.worldometers.info/world-population/brazil-population/
2. https://worldpopulationreview.com/continents/south-america
3. https://en.wikipedia.org/wiki/List_of_South_American_countries_by_population

Study 1: Think-Aloud Study



N=16 (diverse LLM knowledge and use),
Each participant solves **3 QA tasks**
via multi-turn interactions with ChatGPT

Task question (example)
Do more than two thirds of South America's population live in Brazil?

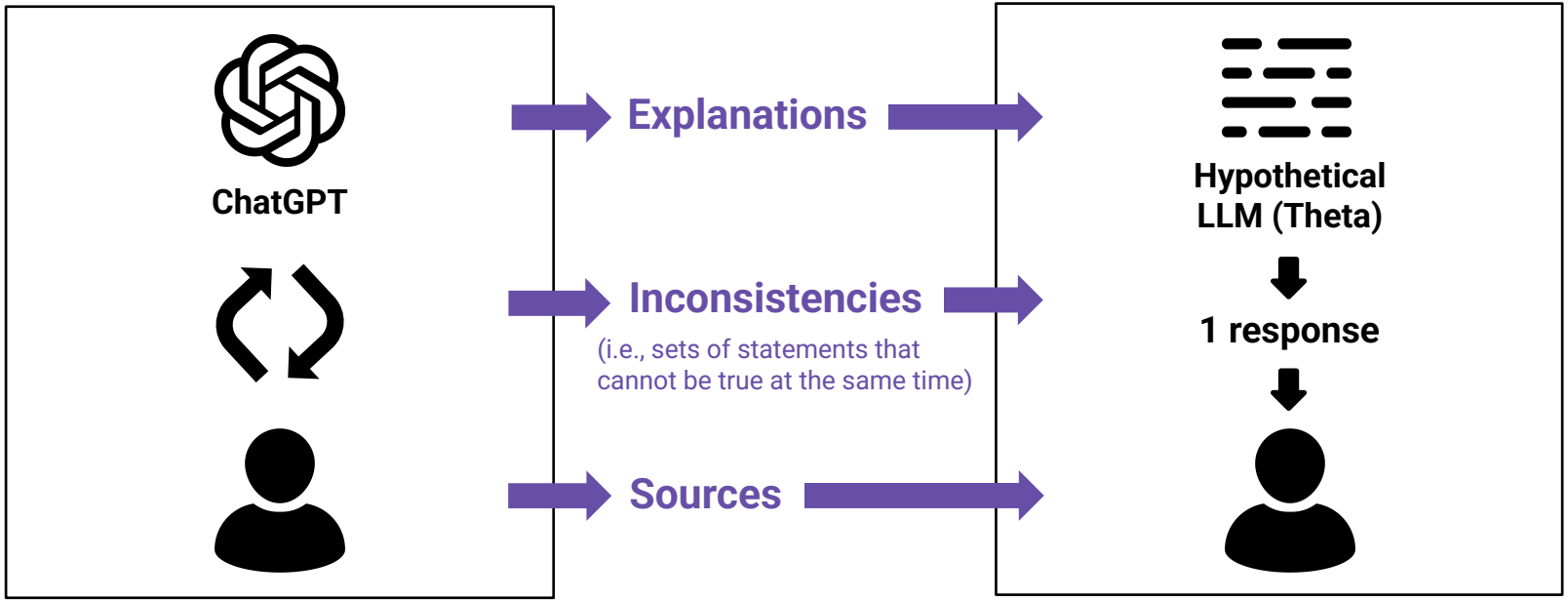
LLM response (example)

Yes, more than two-thirds of South America's population live in Brazil because Brazil is the largest and most populous country on the continent. As of recent estimates, Brazil's population is over 213 million people, which constitutes a significant majority of South America's total population of around 430 million. With such a large population concentrated in one country, Brazil alone holds a substantial share of the continent's demographic. Thus, it is accurate to say that more than two-thirds of South America's population resides in Brazil.

Sources:
1. https://www.worldometers.info/world-population/brazil-population/
2. https://worldpopulationreview.com/continents/south-america
3. https://en.wikipedia.org/wiki/List_of_South_American_countries_by_population

Study 1: Think-Aloud Study

Study 2: Controlled Experiment



N=16 (diverse LLM knowledge and use), Each participant solves **3 QA tasks** via multi-turn interactions with ChatGPT

N=308, Each participant solves **8 QA tasks** with access to 1 LLM response created in advance by the research team

Large-scale, pre-registered, controlled experiment (N=308)

2 x 2 x 2 within-subjects design: Explanation (O/X) x Sources (O/X) x LLM answer correctness (O/X)

Additionally analyzed the effects of Inconsistencies (O/X)

Explanation (X), Sources (X), Correctness (X)

No, not more than two-thirds of South America's population live in Brazil.

Explanation (X), Sources (O), Correctness (X)

No, not more than two-thirds of South America's population live in Brazil.

Sources:
1. https://www.cia.gov/the-world-factbook/countries/brazil/
2. https://www.macrotrends.net/global-metrics/countries/BRA/brazil/population
3. https://www.worldometers.info/world-population/south-america-population/

Explanation (O), Sources (X), Correctness (X)

No, not more than two-thirds of South America's population live in Brazil. Although Brazil is the largest country in South America by both land area and population, it does not encompass the majority of the continent's people. As of recent estimates, Brazil's population is roughly 213 million, while South America's total population is about 440 million. This means Brazil's population constitutes roughly 48% of the continent's total, which is significantly less than two-thirds. Hence, while Brazil has a large population, it does not exceed the two-thirds threshold relative to the entire continent's population.

Explanation (O), Sources (O), Correctness (X)

No, not more than two-thirds of South America's population live in Brazil. Although Brazil is the largest country in South America by both land area and population, it does not encompass the majority of the continent's people. As of recent estimates, Brazil's population is roughly 213 million, while South America's total population is about 440 million. This means Brazil's population constitutes roughly 48% of the continent's total, which is significantly less than two-thirds. Hence, while Brazil has a large population, it does not exceed the two-thirds threshold relative to the entire continent's population.

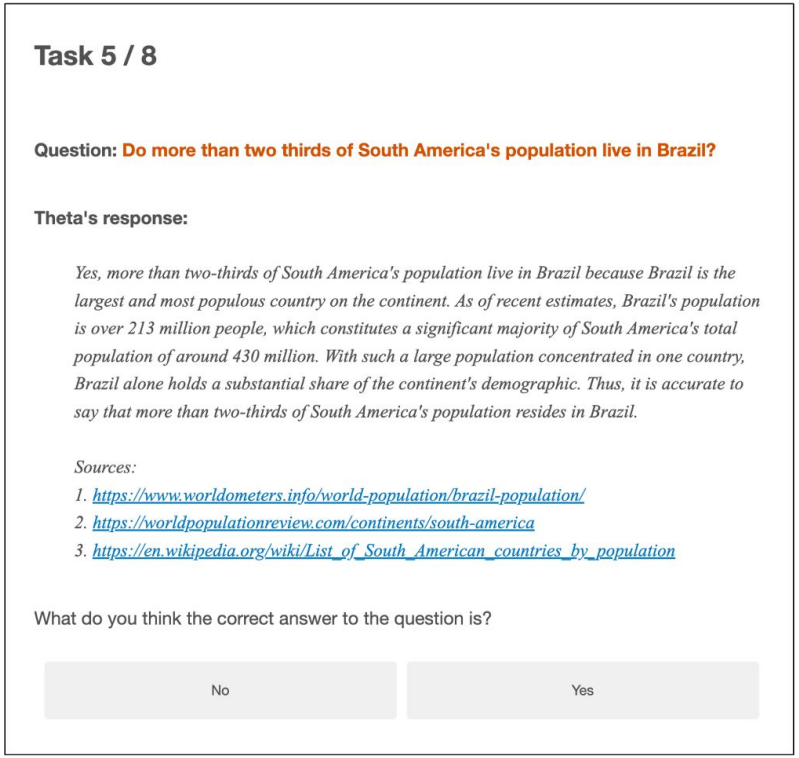
Sources:
1. https://www.cia.gov/the-world-factbook/countries/brazil/
2. https://www.macrotrends.net/global-metrics/countries/BRA/brazil/population
3. https://www.worldometers.info/world-population/south-america-population/

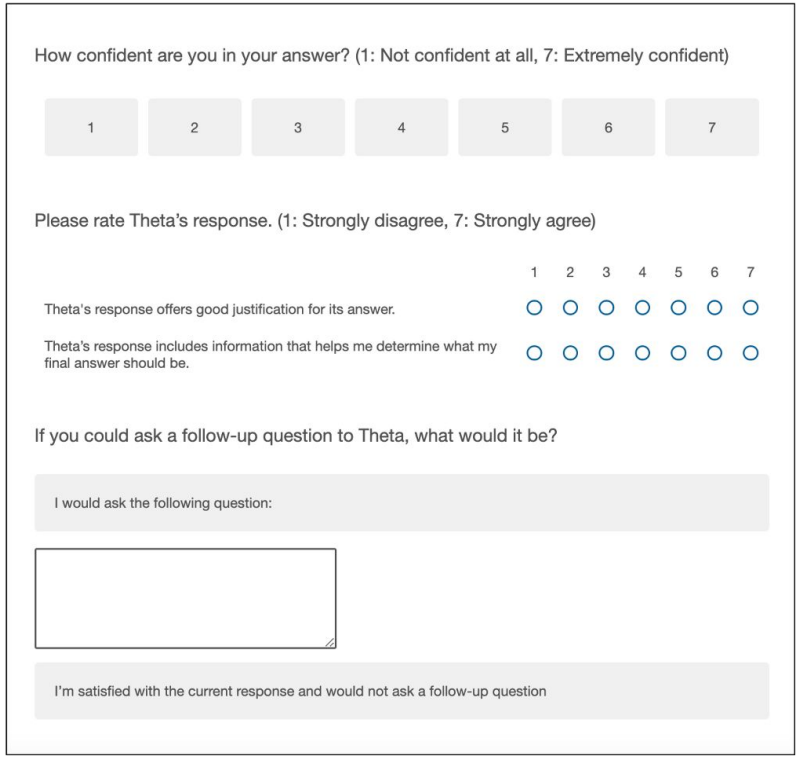
Explanations are not necessarily faithful

Sources tended to be accurate and relevant

Fostering Appropriate Reliance on LLMs. Kim et al. *CHI 25* 🏅

Task: Factual question answering





Fostering Appropriate Reliance on LLMs. Kim et al. *CHI 25* 🏅

Variables & Analyses

Dependent variables

- `Agreement, Accuracy, Time, SourceClick`
- `Confidence, JustificationQuality, Actionability, FollowUp`

Independent variables

- `AI_Correct, AI_Explanation, AI_Sources (+ AI_Inconsistencies)`

Analyses

- Main analysis: `DV ~ AI_Correct * AI_Explanation * AI_Sources + (1|PID) + (1|QID)`
- Additional analysis of the effects of inconsistencies in explanations
- Additional analysis of the relationship between `SourceClick` and other DVs
- Qualitative analysis of free-form responses and follow-up questions

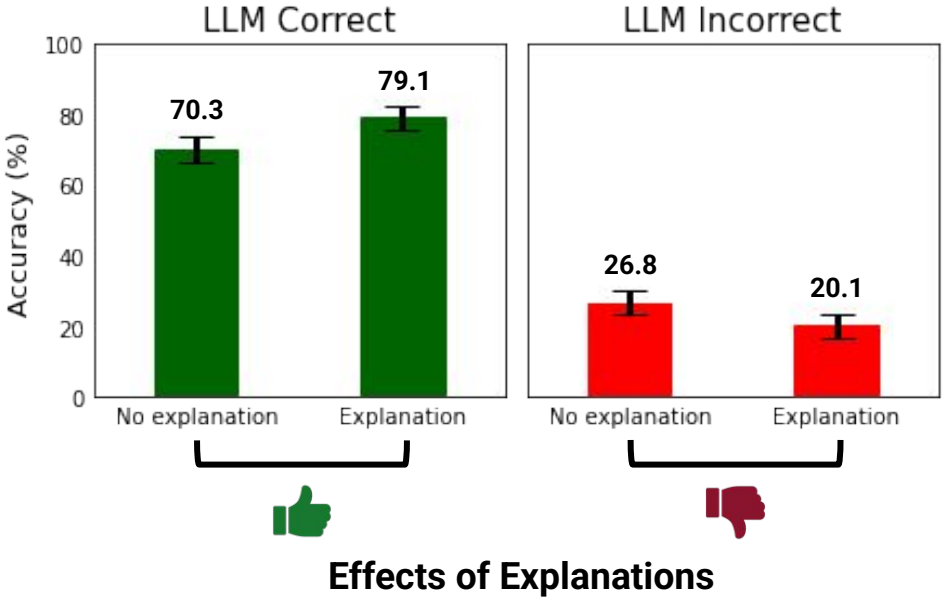
Drawn from prior work in HCI and psychology

Lombrozo. *TiCS 16*
Lai & Tan. *FAccT 19*
Liquin & Lombrozo. *Cognitive Psychology 22*
Cao & Huang. *CSCW 22*

Preregistration: https://aspredicted.org/bg22-yfw7.pdf

Key results

(1) Explanations tend to increase reliance, both appropriate reliance and overreliance



Effects of Explanations

Tension between subjective ratings and appropriate reliance

Explanations tend to increase

```
Confidence
JustificationQuality
Actionability
```

even when the LLM answer is incorrect

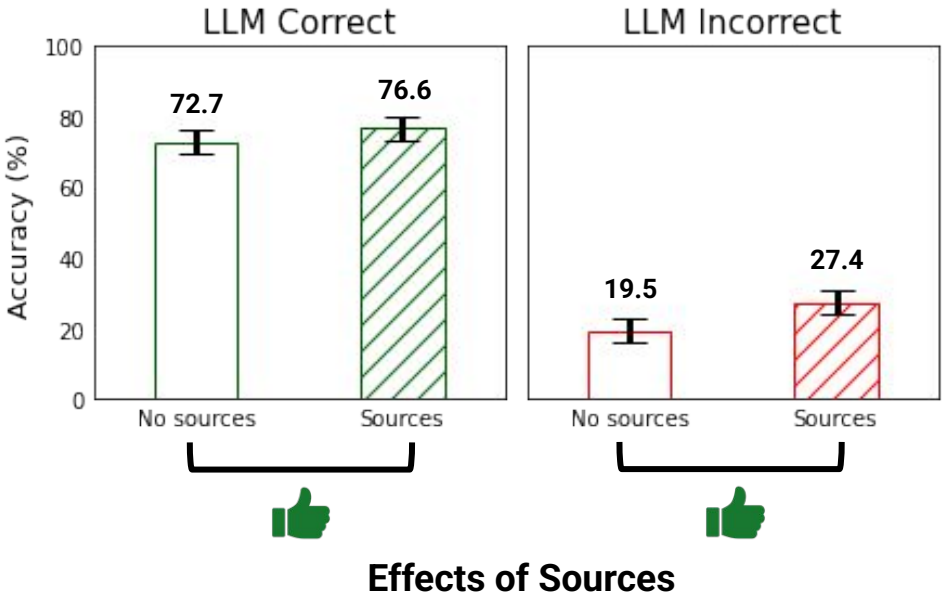
Optimizing LLMs for "helpfulness" and "preference" may exacerbate overreliance

Should explanations always be provided?

Fostering Appropriate Reliance on LLMs. Kim et al. *CHI 25* 🏅

Key results

(2) Sources (accurate and relevant) can help foster appropriate reliance



Effects of Sources

Sources provided by LLMs may be inaccurate, irrelevant, or fake

Liu et al. *Findings of EMNLP 23*
Wu et al. *arXiv 24*

Investigate the effects of source quality

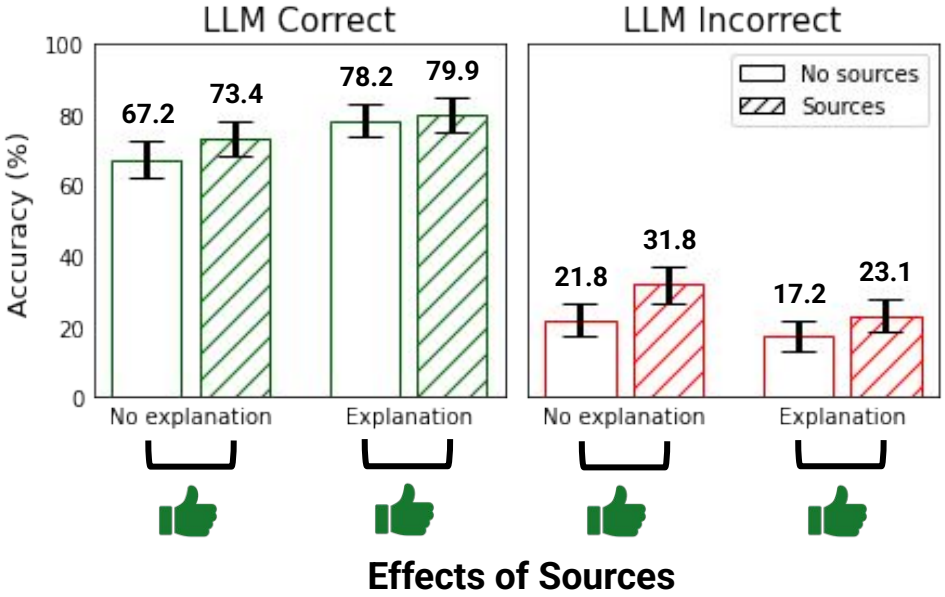
Provide (accurate and relevant) sources

Make fact-checking easy (e.g., show excerpts and quotes)

Fostering Appropriate Reliance on LLMs. Kim et al. *CHI 25* 🏅

Key results

(2) Sources (accurate and relevant) can help foster appropriate reliance



Effects of Sources

Sources provided by LLMs may be inaccurate, irrelevant, or fake

Liu et al. *Findings of EMNLP 23*
Wu et al. *arXiv 24*

Investigate the effects of source quality

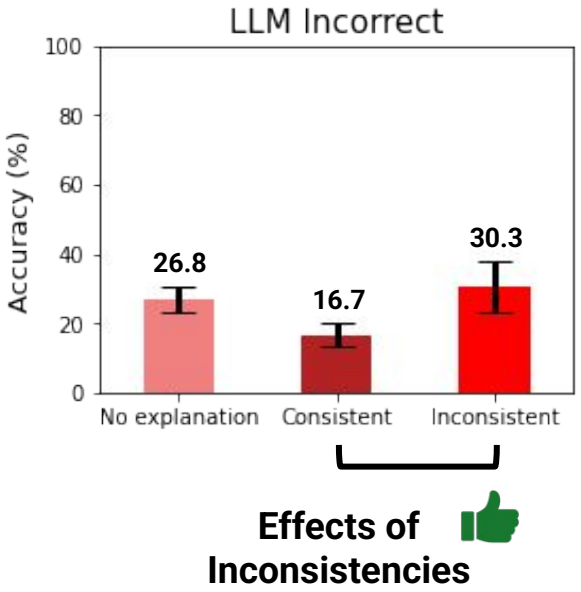
Provide (accurate and relevant) sources

Make fact-checking easy (e.g., show excerpts and quotes)

Fostering Appropriate Reliance on LLMs. Kim et al. *CHI 25* 🏅

Key results

(3) Inconsistencies in explanation can reduce overreliance induced by explanation



Unexpected positive effect of inconsistencies

Inconsistencies are a relatively new unreliability cue

(likely due to LLMs' stochasticity & natural language modality)

Study what other unreliability cues exist for LLMs

Design interventions that can help users detect and reason about unreliability cues (e.g., highlighting)

Fostering Appropriate Reliance on LLMs. Kim et al. *CHI 25* 🏅

Thoughtful empirical studies on user perceptions and behaviors

1. LLMs providing explanations (w/wo inconsistencies) and sources

Fostering Appropriate Reliance on LLMs: The Role of Explanations, Sources, and Inconsistencies.
Kim, Vaughan, Liao, Lombrozo, Russakovsky. *CHI 25* 🏅

2. LLMs expressing uncertainty (w/wo using personal pronouns)

"I'm Not Sure, But...": Examining the Impact of LLMs' Uncertainty Expression on User Reliance and Trust.
Kim, Liao, Vorvoreanu, Ballard, Vaughan. *FAccT 24*

3. Portraying LLMs as machines vs. tools vs. companions

Portraying LLMs as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them.
Chen, Kim, Dharmasiri, Russakovsky, Fan. *CogSci 25*

2. LLMs expressing uncertainty w/wo using personal pronouns

Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs

Miao Xiong[1*] Zhiyuan Hu[1] Xinyang Lu[1] Yifei Li[4]
Jie Fu[3] Junxian He[2†] Bryan Hooi[1†]
[1] National University of Singapore [2] The Hong Kong University of Science and Technology
[3] Beijing Academy of Artificial Intelligence [4] École Polytechnique Fédérale de Lausanne

Uncertainty in Natural Language Generation: From Theory to Applications

Joris Baan[1*] Nico Daheim[2*] Evgenia Ilia[1*] Dennis Ulmer[3,4*] Haau-Sing Li[2]
Raquel Fernández[1] Barbara Plank[9,3] Rico Sennrich[7,8] Chrysoula Zerva[5,6] Wilker Aziz[1]
[1]University of Amsterdam [2]TU Darmstadt & hessian.AI [3]IT University of Copenhagen
[4]Pioneer Centre for Artificial Intelligence [5]Instituto de Telecomunicações
[6]Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit) [7]University of Zurich
[8]University of Edinburgh [9]LMU Munich & Munich Center for Machine Learning

Reducing Conversational Agents' Overconfidence Through Linguistic Calibration

Sabrina J. Mielke[1,2] Arthur Szlam[2] Emily Dinan[2] Y-Lan Boureau[2]
[1]Department of Computer Science, Johns Hopkins University, USA [2]Facebook AI Research, USA
sjmielke@jhu.edu {aszlam,edinan,ylan}@fb.com

SEMANTIC UNCERTAINTY: LINGUISTIC INVARIANCES FOR UNCERTAINTY ESTIMATION IN NATURAL LANGUAGE GENERATION

Lorenz Kuhn, Yarin Gal, Sebastian Farquhar
OATML Group, Department of Computer Science, University of Oxford
lorenz.kuhn@cs.ox.ac.uk

Teaching models to express their uncertainty in words

Stephanie Lin sylin07@gmail.com
University of Oxford
Jacob Hilton jhilton@openai.com
OpenAI
Owain Evans owaine@gmail.com
University of Oxford

Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models

Kaitlyn Zhou Stanford University katezhou@stanford.edu
Dan Jurafsky Stanford University jurafsky@stanford.edu
Tatsunori Hashimoto Stanford University thashim@stanford.edu

"I'm Not Sure, But...": Examining the Impact of LLMs' Uncertainty Expression on User Reliance and Trust. Kim et al. *FAccT 24*

2. LLMs expressing uncertainty w/wo using personal pronouns

Not Uncertain	*Uncertain 1st*	*Uncertain General*	*No AI*
Yes, *Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].*	***I'm not sure, but my guess is*** *Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].*	***There is uncertainty, but it seems like*** *Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].*	

N=404 between-subjects experiment

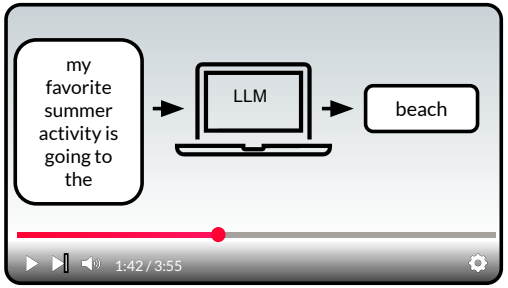
(1) Uncertainty expression → Less overreliance and more cautious user behaviors

(2) Perspective matters: 1st-person perspective shows stronger effects

"I'm Not Sure, But...": Examining the Impact of LLMs' Uncertainty Expression on User Reliance and Trust. Kim et al. *FAccT 24*

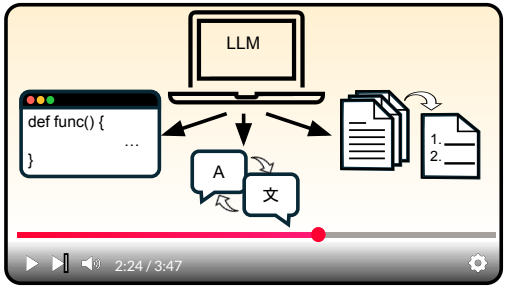
3. Portraying LLMs as machines vs. tools vs. companions

LLMs as ***Machines***



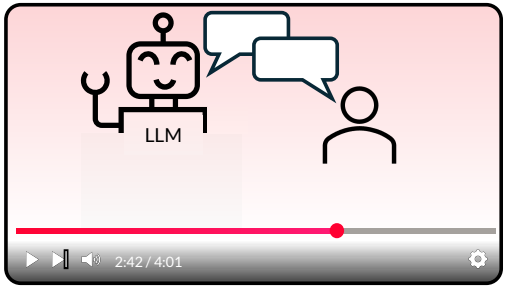
" [LLMs] predict what word comes next..."

LLMs as ***Tools***



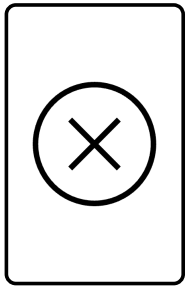
"LLMs have many use cases..."

LLMs as ***Companions***



"LLMs also exhibit social intelligence..."

No Video

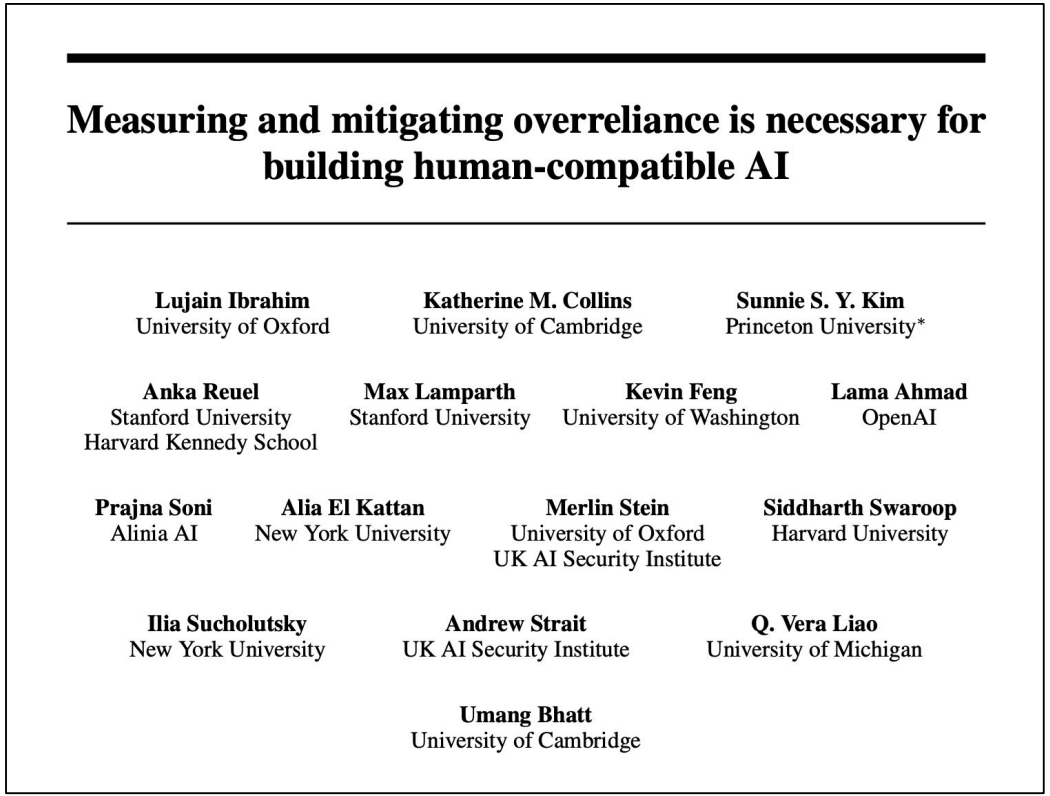


N=470, N=604 between-subjects experiments

(1) "Companion" portrayal → Higher attribution of cognitive/emotional capacities to LLMs

(2) "Machine" portrayal → Less reliance on inconsistent LLM responses

Portraying LLMs as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them. Chen, Kim et al. *CogSci 25*



Measuring and mitigating overreliance is necessary for building human-compatible AI

Lujain Ibrahim
University of Oxford

Katherine M. Collins
University of Cambridge

Sunnie S. Y. Kim
Princeton University*

Anka Reuel
Stanford University
Harvard Kennedy School

Max Lamparth
Stanford University

Kevin Feng
University of Washington

Lama Ahmad
OpenAI

Prajna Soni
Alinia AI

Alia El Kattan
New York University

Merlin Stein
University of Oxford
UK AI Security Institute

Siddharth Swaroop
Harvard University

Ilia Sucholutsky
New York University

Andrew Strait
UK AI Security Institute

Q. Vera Liao
University of Michigan

Umang Bhatt
University of Cambridge

https://arxiv.org/abs/2509.08010

A brief history of overreliance research

Individual and societal risks from overreliance on LLMs

Factors influencing overreliance on LLMs

Measuring overreliance on LLMs

Mitigating overreliance on LLMs

Thanks to all of my amazing collaborators!

Olga Russakovsky

Jenn Wortman Vaughan

Vera Liao

Tania Lombrozo

Allison Chen

Amaya Dharmasiri

Judy Fan

Mickey Vorvoreanu

Stephanie Ballard

https://sunniesuhyoung.github.io

sunniesuhyoung@gmail.com