

Fostering Appropriate Reliance on Large Language Models

Sunnie S. Y. Kim

sunniesuhyoung.github.io



PRINCETON
UNIVERSITY



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



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



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

Risks from inappropriate reliance on AI



“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)



“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)

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.

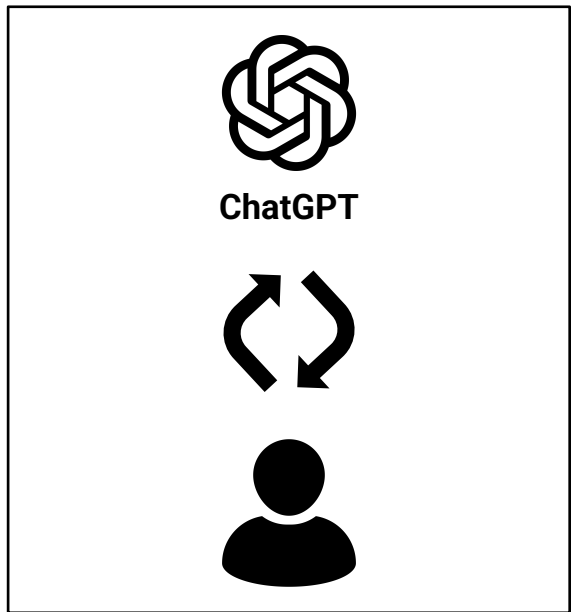
Kim, Liao, Vorvoreanu, Ballard, Vaughan. *FACCT* 24

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:

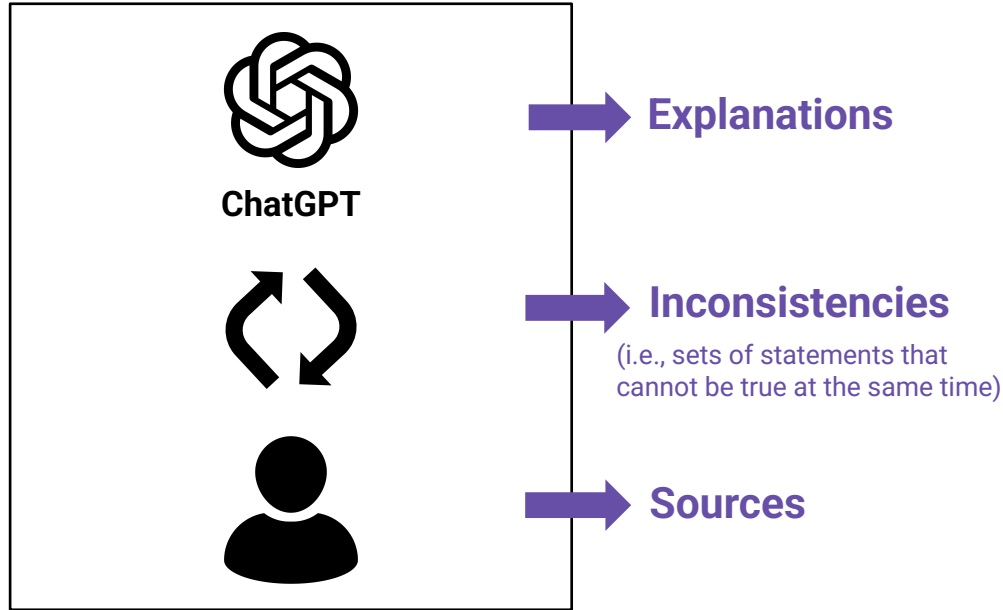
Determine the correct answer to the question

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?)

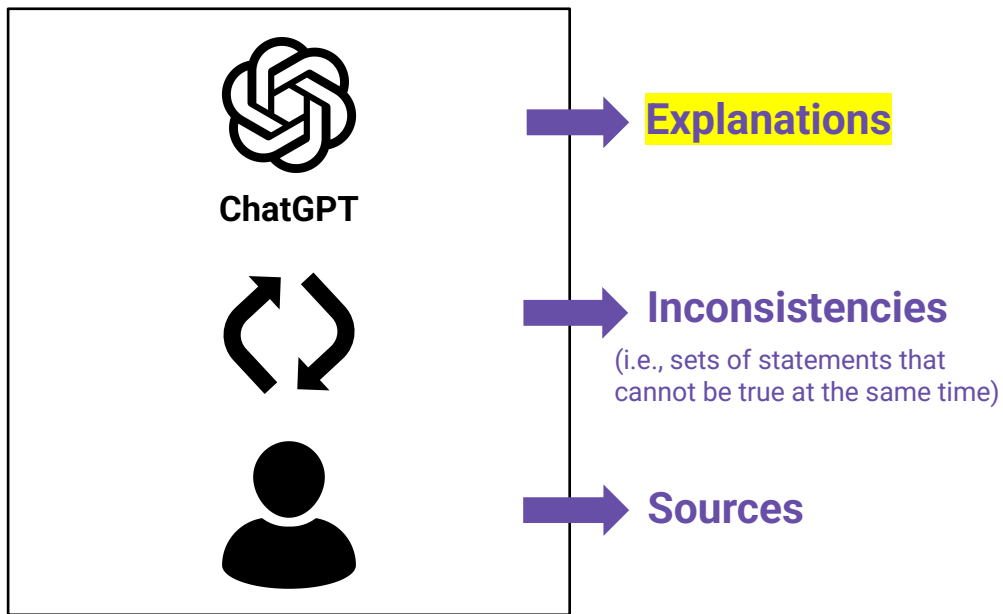
Math question (e.g., Sue puts one grain of rice on the first square of a Go board and puts double the amount on the next square. How many grains of rice does Sue put on the last square?)

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?

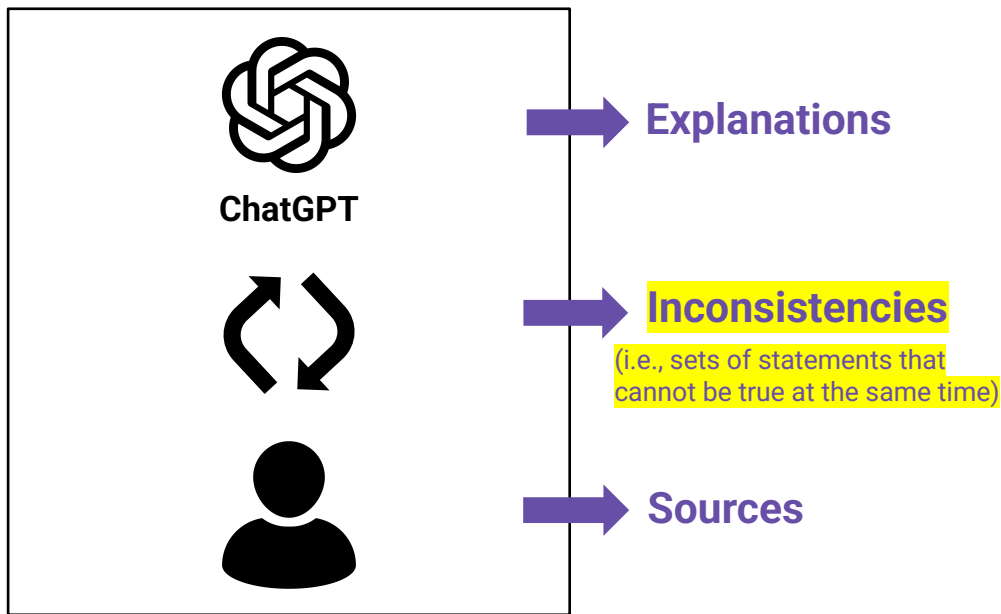
LLM response (example)

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?

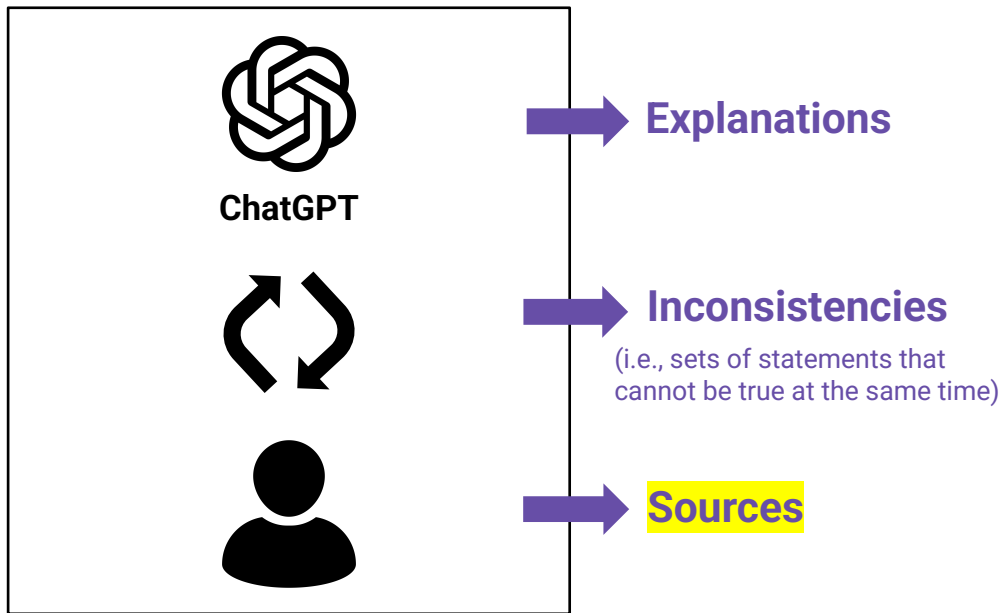
LLM response (example)

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?

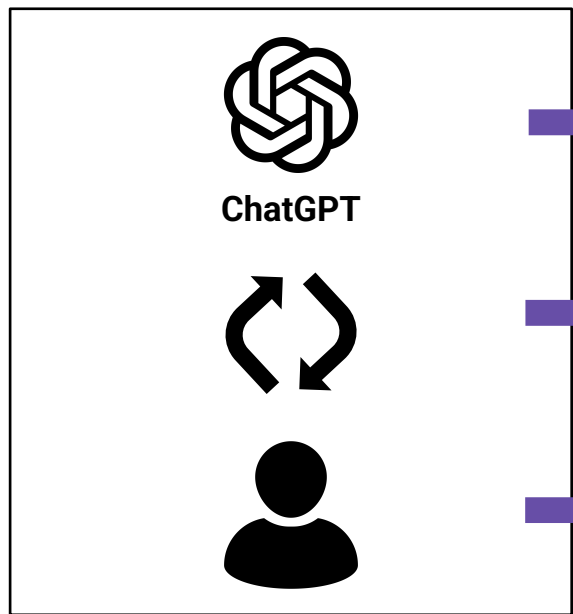
LLM response (example)

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:

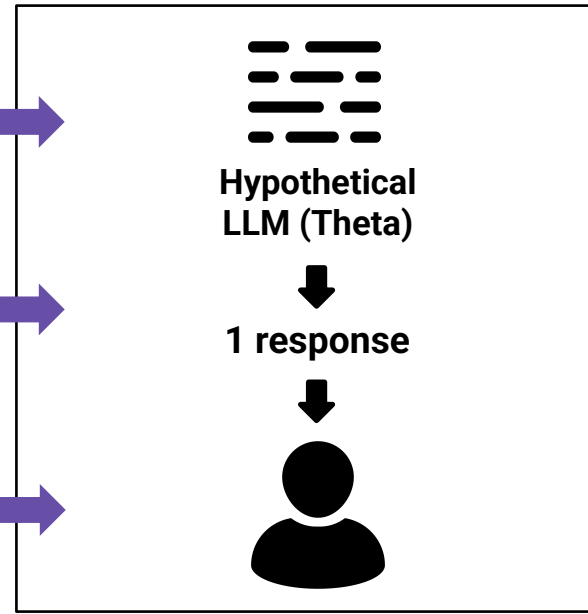
1. <https://www.worldometers.info/world-population/brazil-population/>
2. <https://worldpopulationreview.com/continents/south-america>
3. https://en.wikipedia.org/wiki/List_of_South_American_countries_by_population

Study 1: Think-Aloud Study



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

Study 2: Controlled Experiment



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

Task: Factual question answering

Task 5 / 8

Question: Do more than two thirds of South America's population live in Brazil?

Theta's response:

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

What do you think the correct answer to the question is?

No

Yes

How confident are you in your answer? (1: Not confident at all, 7: Extremely confident)

1

2

3

4

5

6

7

Please rate Theta's response. (1: Strongly disagree, 7: Strongly agree)

1 2 3 4 5 6 7

Theta's response offers good justification for its answer.

☐ ☐ ☐ ☐ ☐ ☐ ☐

Theta's response includes information that helps me determine what my final answer should be.

☐ ☐ ☐ ☐ ☐ ☐ ☐

If you could ask a follow-up question to Theta, what would it be?

I would ask the following question:

I'm satisfied with the current response and would not ask a follow-up question



Variables & Analyses

Dependent variables

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

**Drawn from prior work
in HCI and psychology**

Independent variables

- AI_Correct, AI_Explanation, AI_Sources (+ AI_Inconsistencies)

Analyses

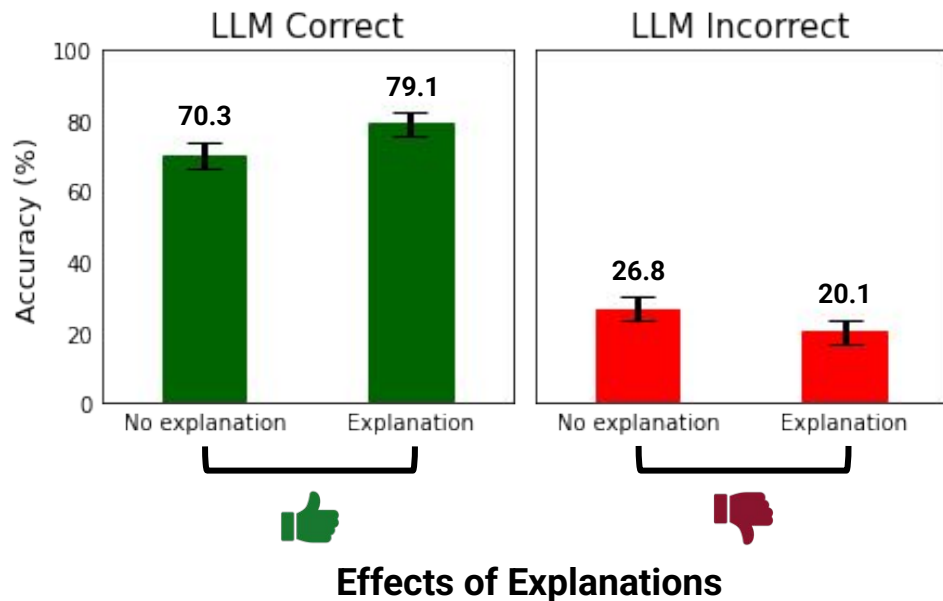
- Main analysis: $DV \sim AI_Correct * AI_Explanation * AI_Sources + (1|PID) + (1|QID)$
- Additional analysis of the effects of inconsistencies in explanations
- Additional analysis of the relationship between SourceClick and other DVs
- Qualitative analysis of free-form responses and follow-up questions

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



Tension between subjective ratings and appropriate reliance

Explanations tend to increase

Confidence

JustificationQuality

Actionability

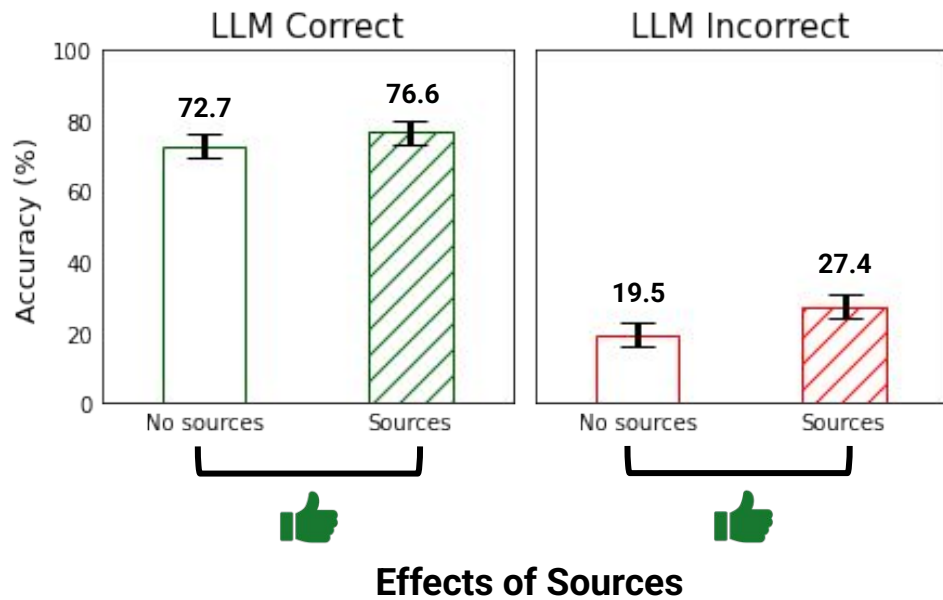
even when the LLM answer is incorrect

Optimizing LLMs for “helpfulness” and “preference” may exacerbate overreliance

Should explanations always be provided?

Key results

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



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

Liu et al. *Findings of EMNLP 23*

Wu et al. *arXiv 24*

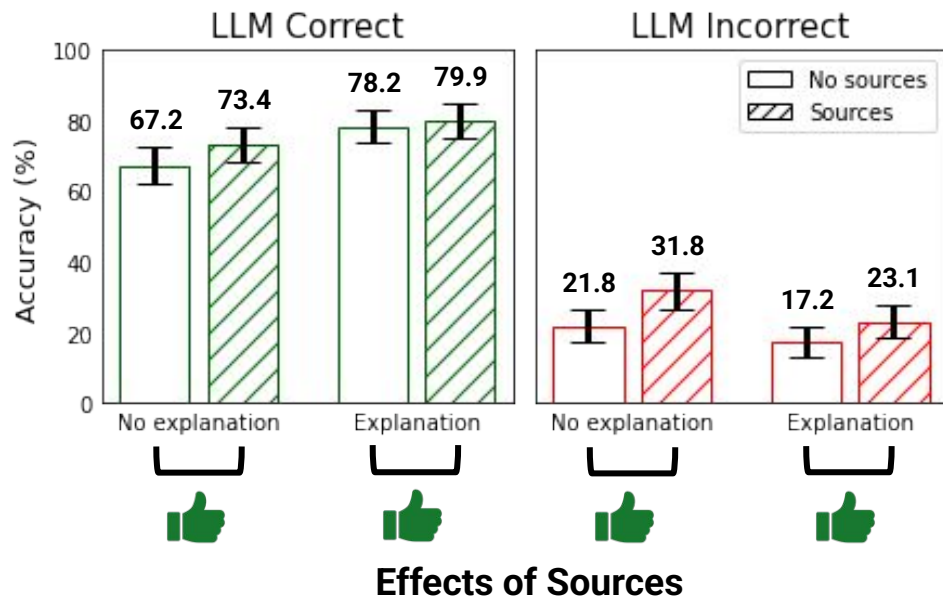
Investigate the effects of source quality

Provide (accurate and relevant) sources

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

Key results

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



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

Liu et al. *Findings of EMNLP 23*

Wu et al. *arXiv 24*



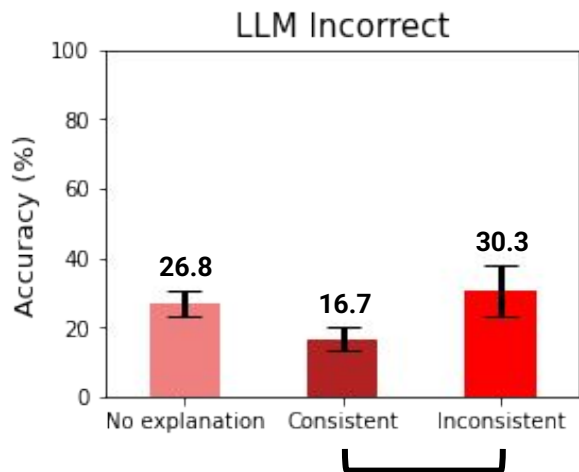
Investigate the effects of source quality

Provide (accurate and relevant) sources

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

Key results

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



Effects of 
Inconsistencies

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)

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 Hu¹ Xinyang Lu¹ Yifei Li⁴
Jie Fu³ 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 Li²
Raquel Fernández¹ Barbara Plank^{9,3} Rico Sennrich^{7,8} Chrysoula Zerva^{5,6} Wilker Aziz¹

¹University of Amsterdam ²TU Darmstadt & hessian.AI ³IT University of Copenhagen
⁴Pioneer Centre for Artificial Intelligence ⁵Instituto de Telecomunicações
⁶Instituto Superior Técnico & LUMIS (Lisbon ELLIS Unit) ⁷University of Zurich
⁸University of Edinburgh ⁹LMU Munich & Munich Center for Machine Learning

Reducing Conversational Agents' Overconfidence Through Linguistic Calibration

Sabrina J. Mielke^{1,2} Arthur Szlam² Emily Dinan² Y-Lan Boureau²

¹Department of Computer Science, Johns Hopkins University, USA ²Facebook 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
University of Oxford

sylin07@gmail.com

Jacob Hilton
OpenAI

jhilton@openai.com

Owain Evans
University of Oxford

owaine@gmail.com

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

2. LLMs expressing uncertainty w/wo using personal pronouns

Not Uncertain

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].

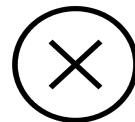
Uncertain 1st

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].

Uncertain General

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].

No AI



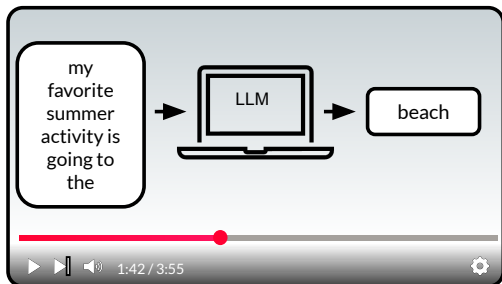
N=404 between-subjects experiment

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

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

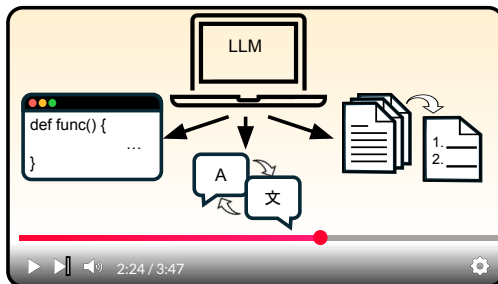
3. Portraying LLMs as machines vs. tools vs. companions

LLMs as *Machines*



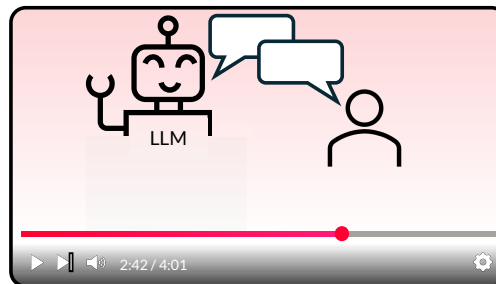
"[LLMs] predict what word comes next..."

LLMs as *Tools*



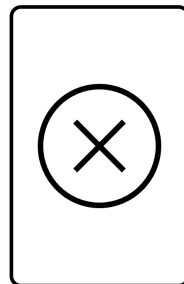
"LLMs have many use cases..."

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

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

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



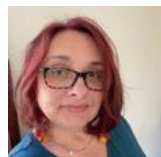
Allison
Chen



Amaya
Dharmasiri



Judy
Fan



Mickey
Vorvoreanu



Stephanie
Ballard



<https://sunniesuhyoung.github.io>



sunniesuhyoung@gmail.com