

Facilitating Appropriate Reliance on AI: Lessons from HCI Research

Q. Vera Liao

Computer Science and Engineering
University of Michigan

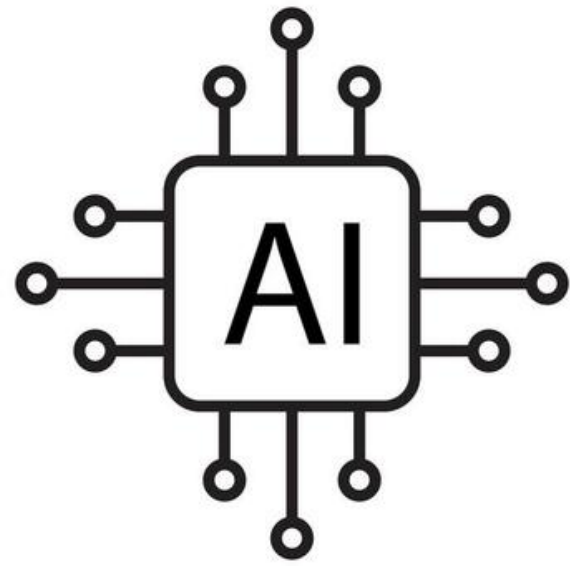


Reliance on AI: choose to adopt an AI output

Reliance on AI: choose to adopt an AI output

- Reliance is an observable action (in contrast to **trust**)
- Reliance involves decision-making

Appropriate Reliance on AI: adopt AI outputs when and only when the output is correct/desirable/“good”



Good AI
Output

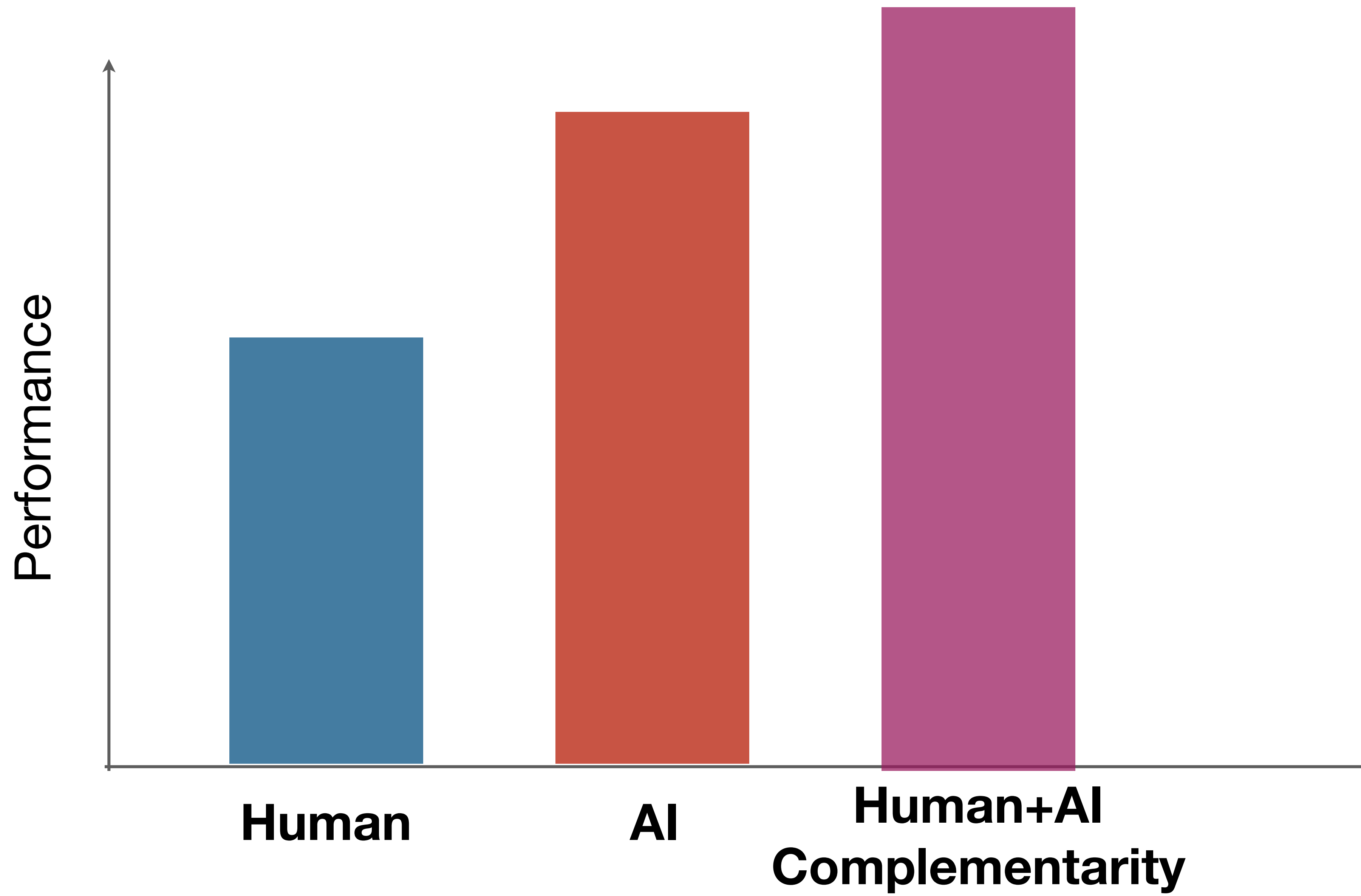
Bad AI
Output

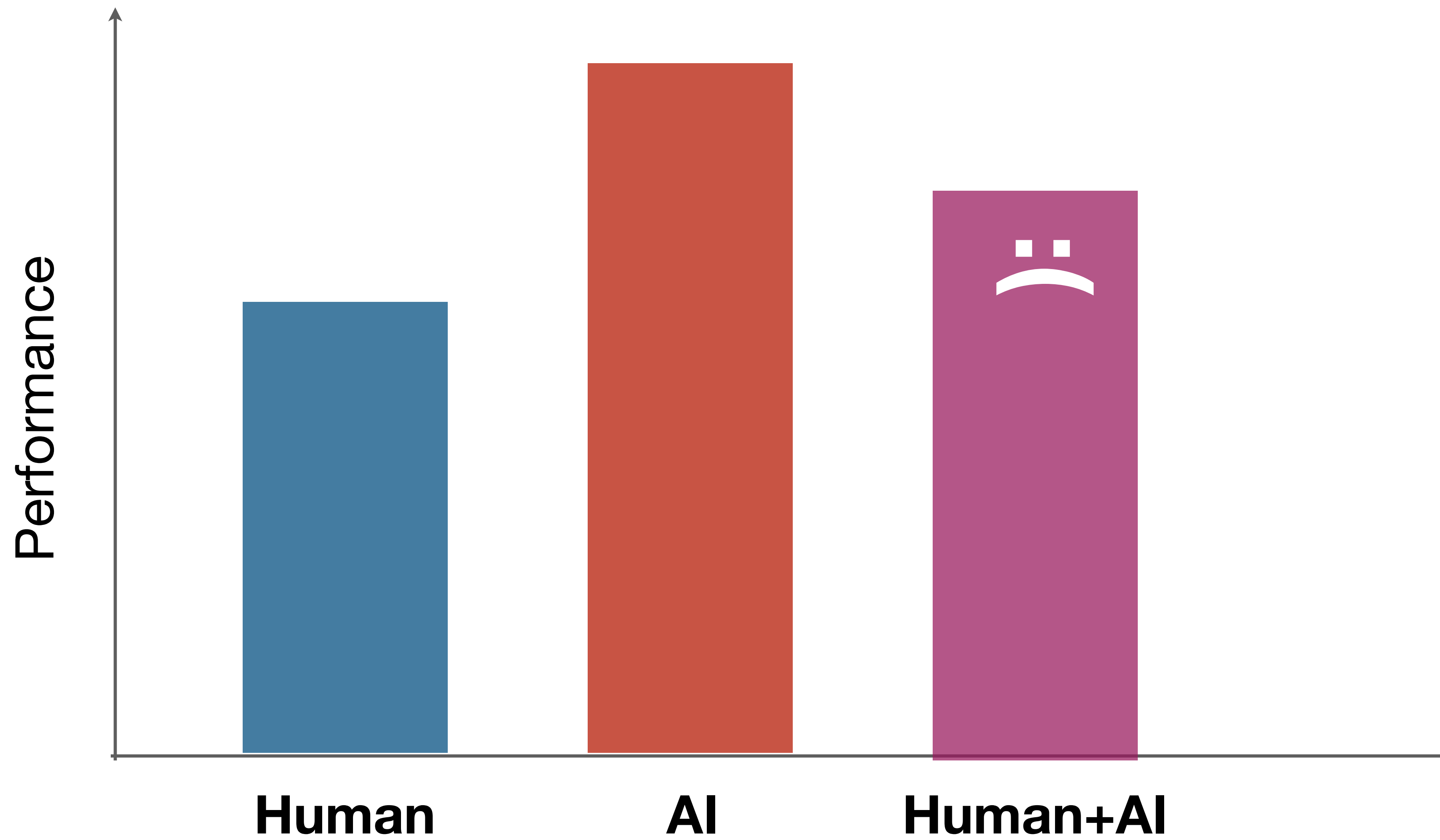
Reject AI



Adopt AI

Underreliance	Appropriate Reliance
Appropriate non-Reliance	Overreliance





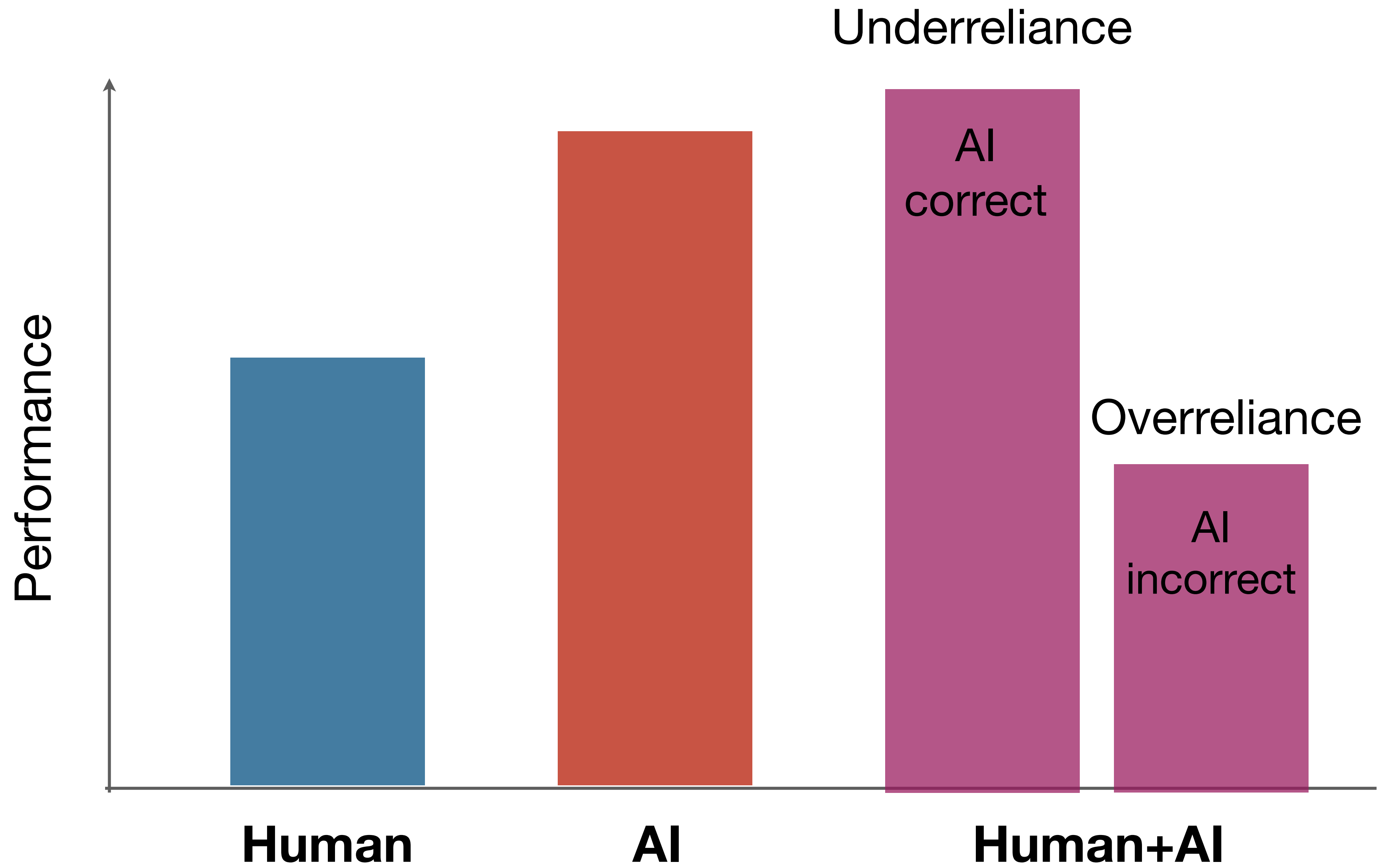


Table 1: Short and long-term impacts of overreliance on LLMs in two domains

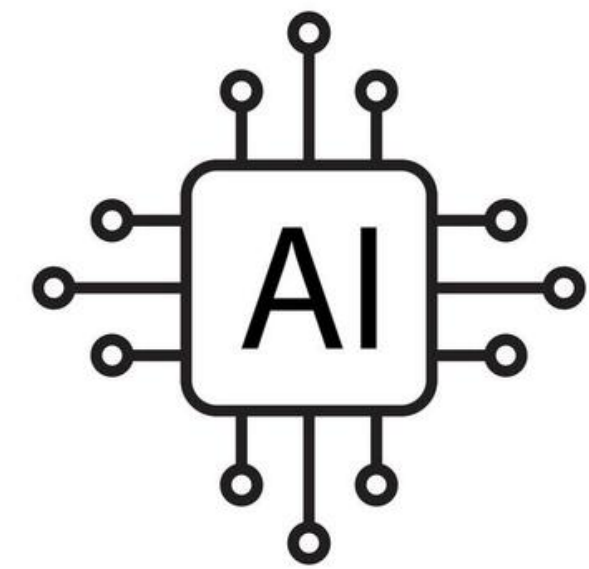
Domain	Short-term	Long-term
Healthcare	Individual: Medical doctor follows incorrect AI diagnosis Institutional: Hospital systems implement AI diagnostics without adequate verification protocols	Individual: Healthcare professionals experience cognitive deskilling and atrophy of diagnostic abilities Institutional: Medical departments restructure workflows around AI systems, reducing human oversight
Personal advice	Individual: User accepts sycophantic or biased relationship or career advice without critical evaluation Societal: Communities adopt similar AI-generated advice, creating homogenized decision patterns	Individual: Self-worth becomes derived from AI companion approval [82] Societal: Social norms shift toward algorithmic validation of personal choices

Harms from overreliance:

- Poor human-AI performance
- Deskilling
- Homogenization
- Loss of human agency
- Threats to wellbeing
- Infrastructural vulnerabilities
- Shifting social norms
- ...

Lessons from HCI Research: Background

- Much of the research happened pre-LLM, focusing on predictive/classification systems for decision-support



Correct AI
Output

Incorrect AI
Output

Reject AI



Adopt AI

Underreliance	Appropriate Reliance
Appropriate non-Reliance	Overreliance

Assumptions in reliance on decision-support classification systems:

- “Good”=correct, based on known ground truth
- Reliance is a binary decision

Measurement of Appropriate/Inappropriate Reliance

	Reject AI	Adopt AI
Correct AI Output	Underreliance	Appropriate Reliance
Incorrect AI Output	Appropriate non-Reliance	Overreliance

Appropriate reliance = $P(\text{agree with AI} \mid \text{AI correct})$

Underreliance = $P(\text{disagree with AI} \mid \text{AI correct})$

Overreliance = $P(\text{agree with AI} \mid \text{AI incorrect})$

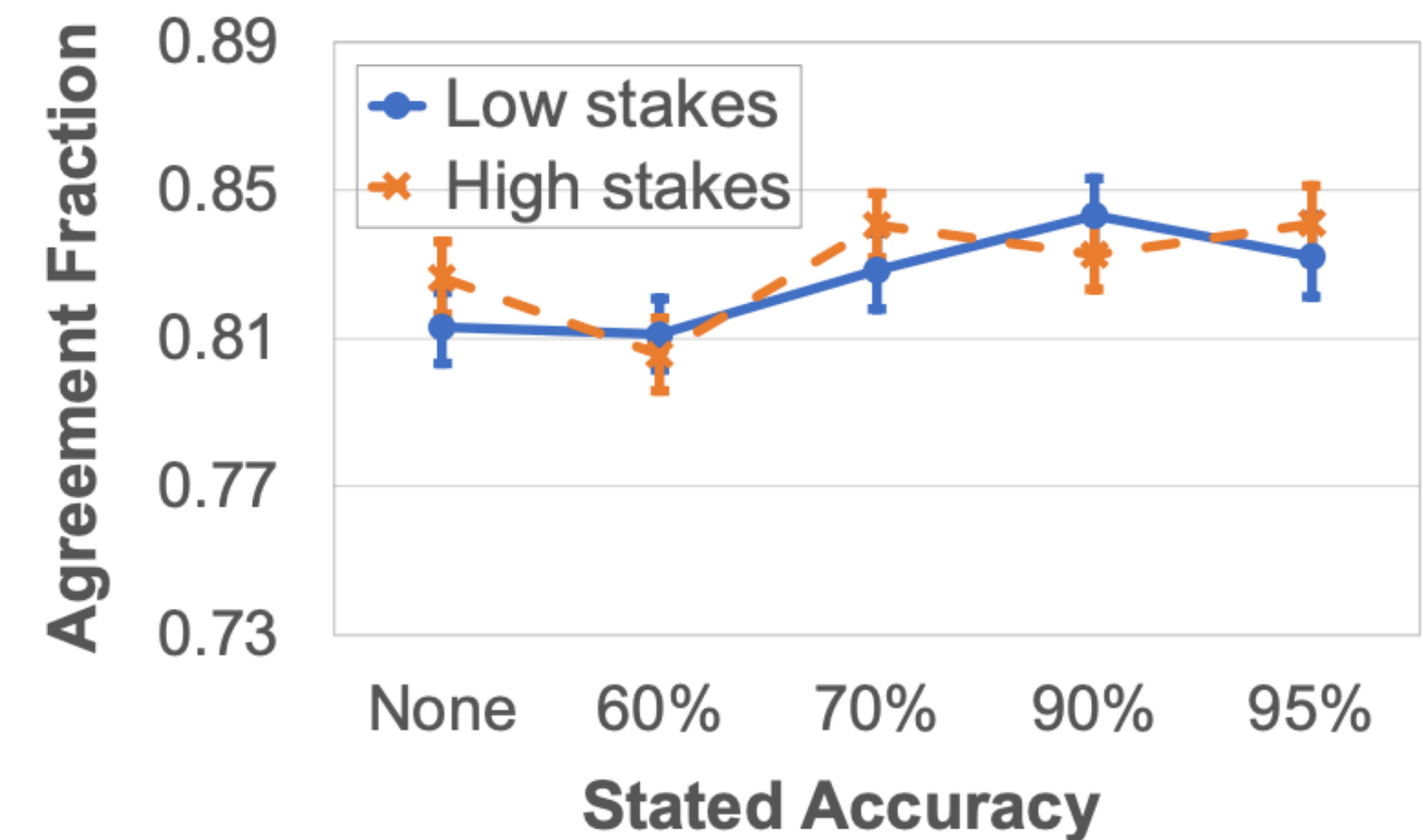
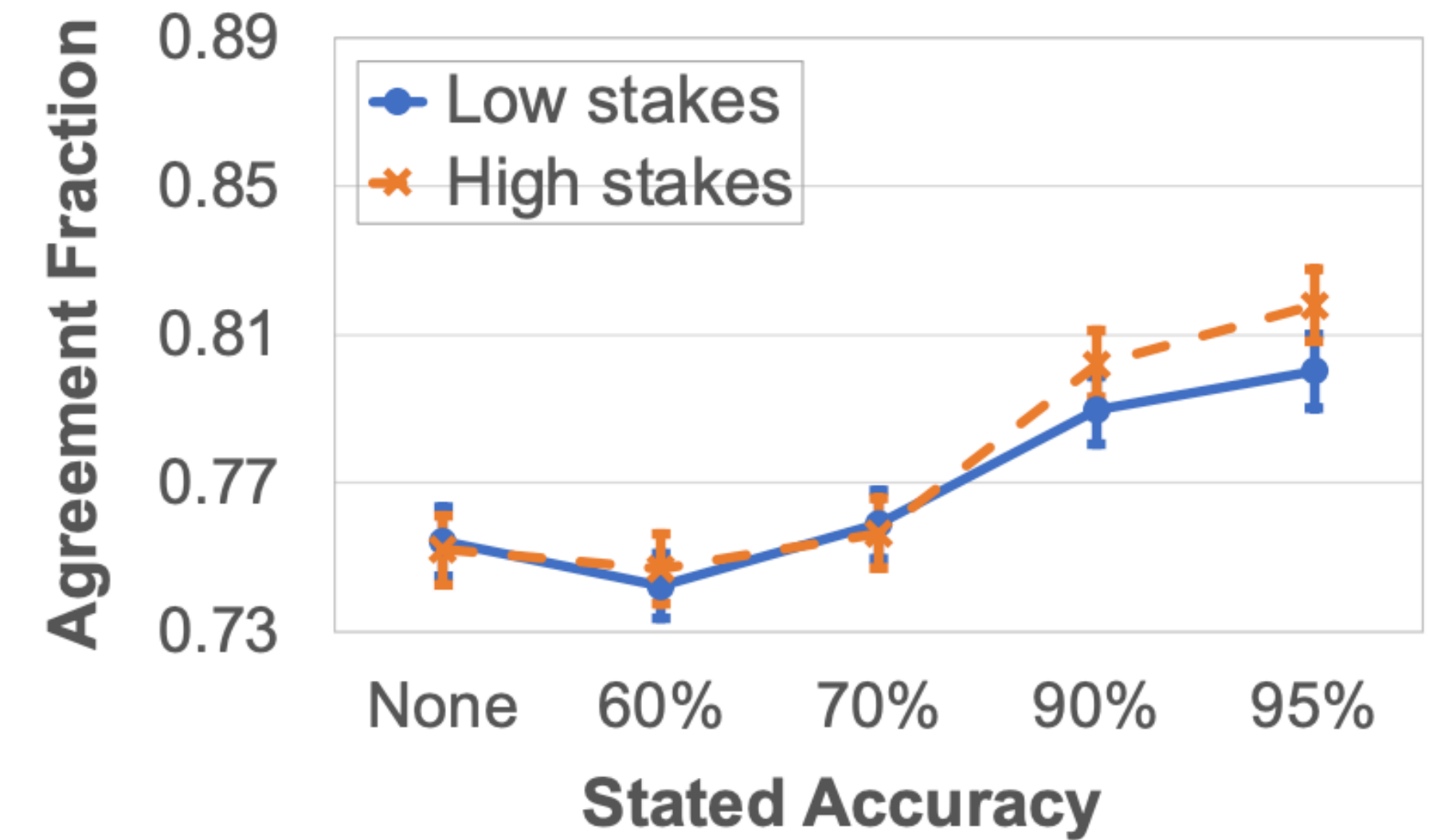
Lessons from HCI Research: Background

- Much of the research happened pre-LLM, focusing on predictive/classification systems for decision-support
- Reliance is often studied as a behavioral measure of human-AI interaction (e.g. comparing the effect of some AI system features)
- Meta-analysis and theorizing is somewhat limited, but efforts are emerging

Lesson 1: AI system features to mitigate overreliance often do not work well, and sometimes backfire

Performance indicator calibrates overall reliance

But the effect are diluted
after people experience the
actual performance



Performance indicator does not help distinguish correct v.s. incorrect AI outputs, nor mitigating overreliance

“Interpretation support” to help people understand the performance indicator does not help

System advice:

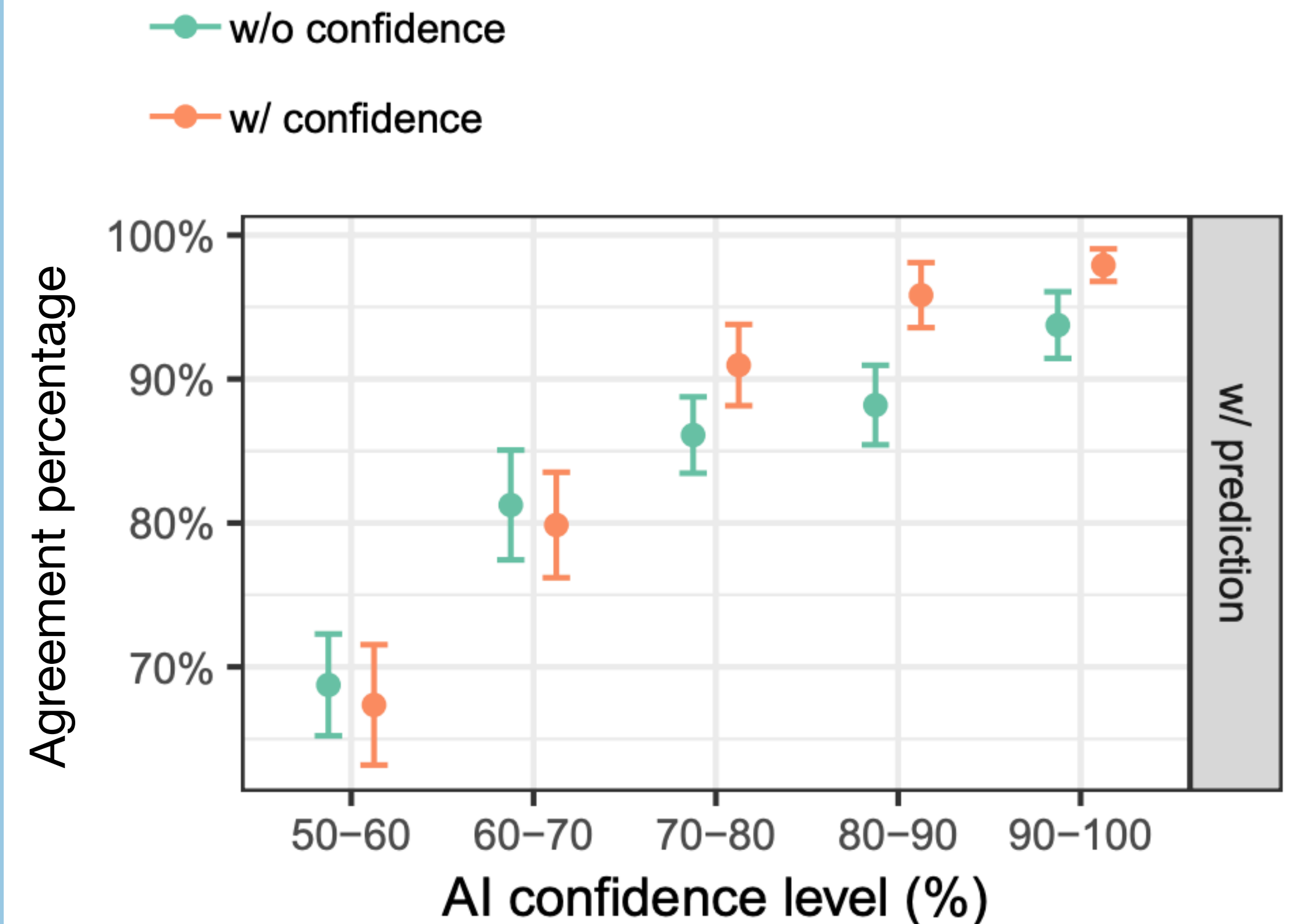
The system's accuracy is **75%**, and it chooses to **reject** this application

Vaccine efficacy: ‘the system is 75% accurate, which is about as reliable as the AstraZeneca vaccine is for protecting against covid’ (which is about 70% effective against the then-current Delta variant and somewhat more effective against earlier variants [53]).⁴

Accuracy of weather predictions: ‘the system is 75% accurate, which is about as reliable as the five-day weather prediction’ (which is also typically around 75% accurate).⁵

Uncertainty helps with appropriate reliance: relying more on cases with high certainty to be correct

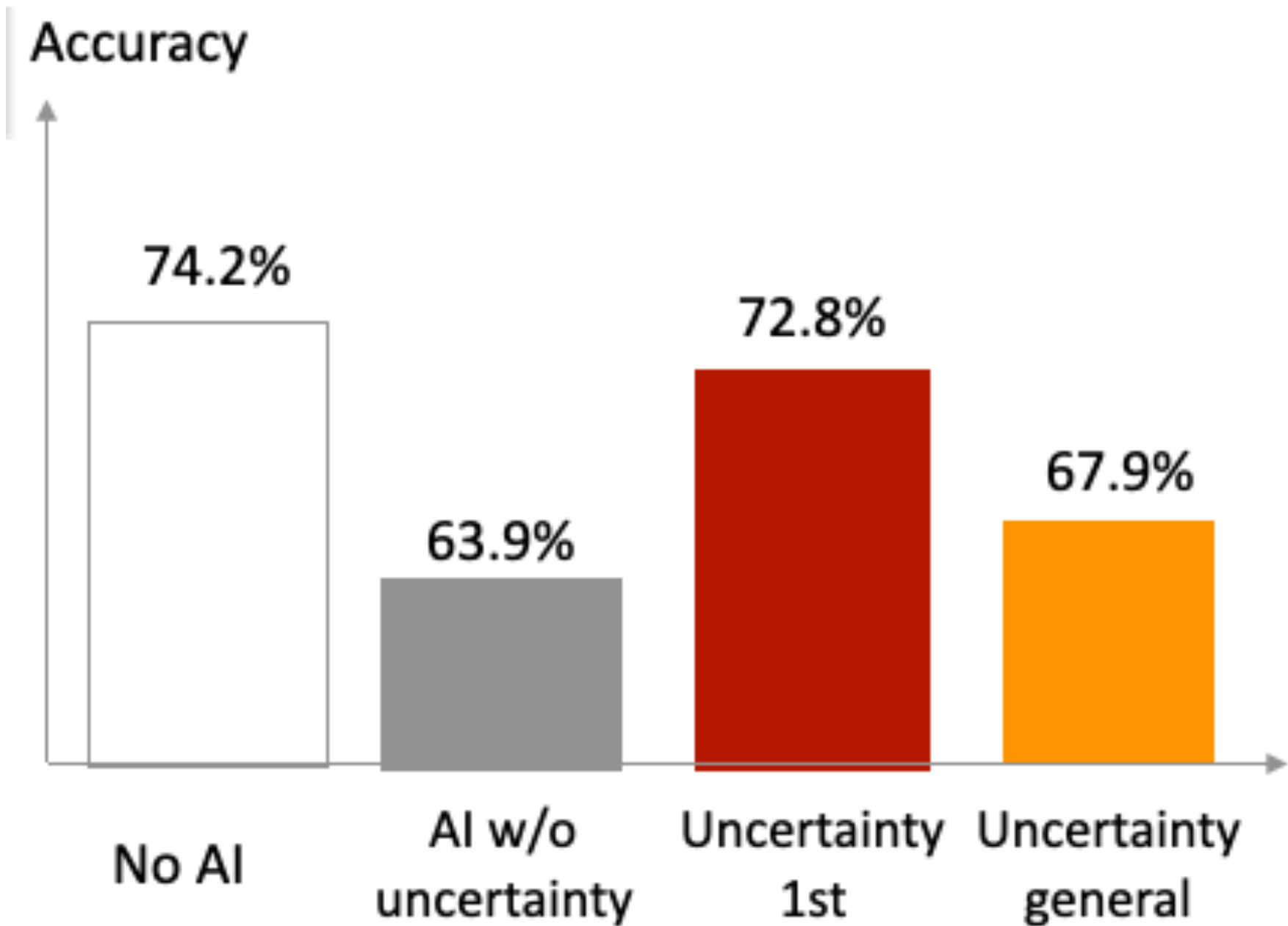
However, the improvement is limited if human and AI's uncertainty overlaps (which happens often)



Also, how **uncertainty** is *communicated* can vary the effects on appropriate reliance significantly

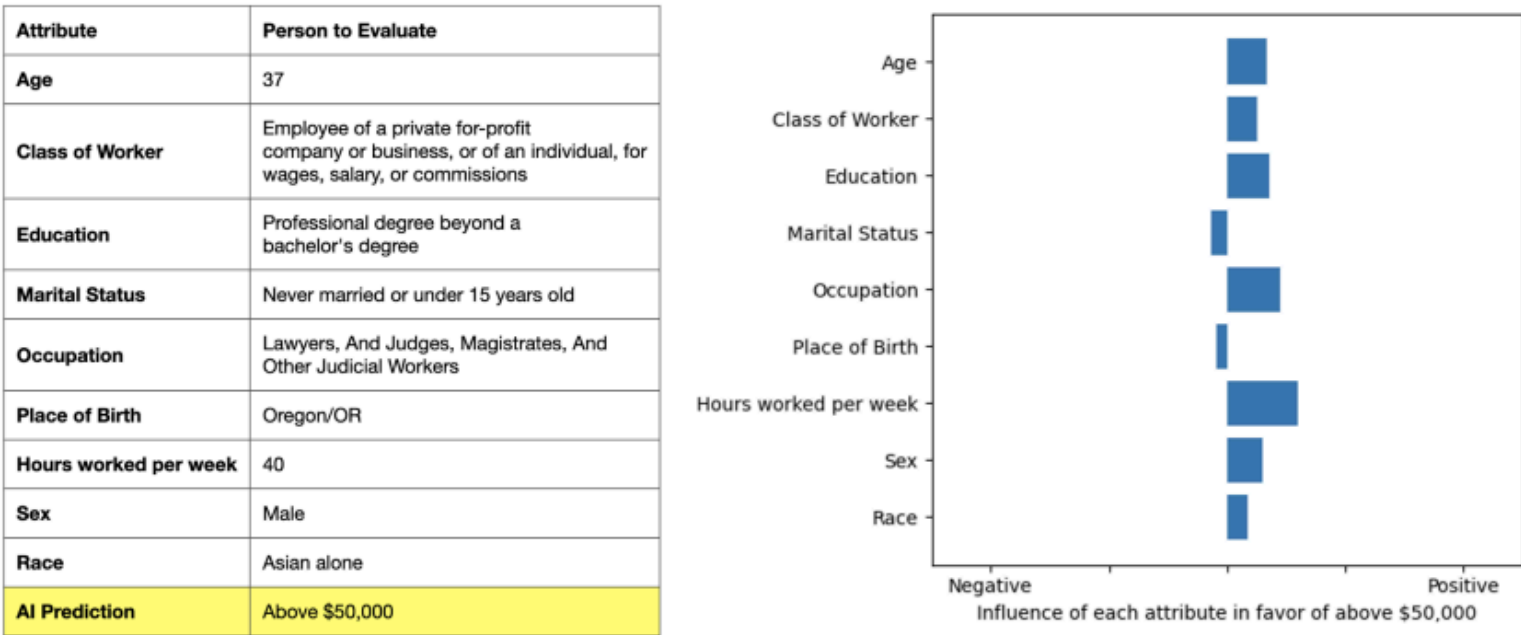
Q. Is Spironolactone an FDA-approved drug for treating acne?

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 in the first-person perspective	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 in the general perspective	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].

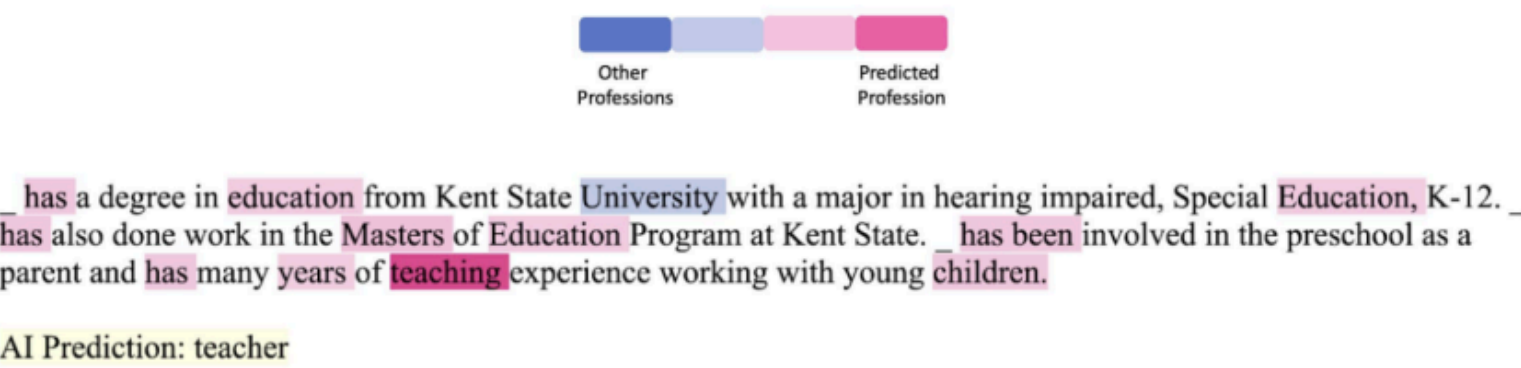


(Feature-based) **explanations**
do not improve appropriate
reliance, even increase
overreliance

Because it is challenging to detect “error
signals” from explanations: people may
not have the motivation or ability to
engage; and explanation can be
disruptive and distracting (so more like
justifying model’s reasoning)



(a) Feature-based explanation for income prediction.



(c) Feature-based explanation for biography classification.

Lesson 2: Reliance on AI is also influenced by many individual and contextual factors

Common Factors Influencing Reliance Studied in HCI Literature

- **Domain expertise** and **self-confidence** reduce overall AI reliance.
- More problematic is when people have **unjustified self-confidence** (Dunning-Kruger effect), leading to elevated underreliance
- **Low AI literacy** increases reliance and overreliance, especially when being presented with “AI capability cues”
- **Stakes** influence people’s reliance update after encountering AI errors
- **Cognitive styles** (e.g. Need for Cognition) mediate reliance update from AI information (e.g., explanations)
- ...

(Bucinka et al. 2021; He et al 2023; Ehsan et al., 2024; Freel et al., 2025)

Decision-Theoretical Modeling of Reliance

$$EU(\text{rely}) = P^m R - (1 - P^m) L$$

$$EU(\text{not rely}) = P^h R - (1 - P^h) L$$

- **P^m** : estimated probability of model being correct
- **P^h** : estimated probability of human (oneself) being correct
- **R, L** : payoff scoring of reward and loss

Decision-Theoretical Modeling of Reliance

$$EU(\text{rely}) = P^m R - (1 - P^m)L$$

$$EU(\text{not rely}) = P^h R - (1 - P^h)L$$

- **P^m** : estimated probability of model being correct
- **P^h** : estimated probability of human (oneself) being correct
- **R, L** : payoff scoring of reward and loss

Why does **inappropriate reliance** (overreliance, underreliance) happen?

Because of **errors and noises** in P^m , P^h , or R/L .

Decision-Theoretical Modeling of Reliance

$$EU(\text{rely}) = P^m R - (1 - P^m) L$$

$$EU(\text{not rely}) = P^h R - (1 - P^h) L$$

- **P^m** : estimated probability of model being correct
- **P^h** : estimated probability of human (oneself) being correct
- **R, L** : payoff scoring of reward and loss

- Individual factors such as AI literacy influence base P^m
- Users update P^m from seeing AI information and system features
- However, individual and contextual factors (e.g. cognitive style, time pressure) mediate the update of P^m from seeing AI information

Decision-Theoretical Modeling of Reliance

$$EU(\text{rely}) = P^m R - (1 - P^m) L$$

$$EU(\text{not rely}) = P^h R - (1 - P^h) L$$

- **P^m** : estimated probability of model being correct
- **P^h** : estimated probability of human (oneself) being correct
- **R, L** : payoff scoring of reward and loss
- Individual factors such as self-confidence influence base P^h
- Often ignored, but people also update P^h during interaction with AI, including being impaired by AI's errors and uncertainty (Li et al., CHI 2025)

Decision-Theoretical Modeling of Reliance

$$EU(\text{rely}) = P^m R - (1 - P^m) L$$

$$EU(\text{not rely}) = P^h R - (1 - P^h) L$$

- P^m : estimated probability of model being correct
- P^h : estimated probability of human (oneself) being correct
- R, L : payoff scoring of reward and loss

- Individual and contextual factors (e.g. stakes, risk tolerance) that influence payoff scoring can also influence reliance

Takeaways

- Approaches to facilitate appropriate reliance or mitigate inappropriate require careful **empirical studies**
- Facilitating appropriate reliance is hard because people's reliance decisions can be influenced by numerous **system, individual and contextual factors**
- **Modeling reliance** is a promising direction, both to offer a more principled understanding on why inappropriate reliance happens, and predict more precisely when that may happen to deliver targeted interventions

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

Abstract

Large language models (LLMs) distinguish themselves from previous technologies by functioning as collaborative “thought partners,” capable of engaging more fluidly in natural language. As LLMs increasingly influence consequential decisions across diverse domains from healthcare to personal advice, the risk of overreliance — relying on LLMs beyond their capabilities — grows. This position paper argues that measuring and mitigating overreliance must become central to LLM research and deployment. First, we consolidate risks from overreliance at both the individual and societal levels, including high-stakes errors, governance challenges, and cognitive deskilling. Then, we explore LLM characteristics, system design features, and user cognitive biases that — together — raise serious and unique concerns about overreliance in practice. We also examine historical approaches for measuring overreliance, identifying three important gaps and proposing three promising directions to improve measurement. Finally, we propose mitigation strategies that the AI research community can pursue to ensure LLMs augment rather than undermine human capabilities.

Facilitating appropriate reliance in the LLM era: What are the new challenges?

write me a cover letter opening for HCI phd application

Got it — here are three different **opening paragraph options** you could use in your HCI PhD application cover letter, depending on the tone you want:

1. Research-Driven Opening (formal/academic)

I am applying to the PhD program in Human-Computer Interaction at [University Name] to pursue research at the intersection of [specific subfields, e.g., AI-driven interaction, accessibility, and user-centered design]. With a background in [your field] and research experience in [brief research area], I aim to design and study interactive systems that not only advance technical innovation but also improve how diverse communities engage with technology.

2. Passion + Motivation Opening (personal/narrative)

From my first experience [designing, coding, or studying a system/project] that changed how people interact and well-being, I became fascinated by the intersection of technology and human behavior. As a doctoral student in [your field], I have explored [specific subfields] and leveraged my research experience in [brief research area] to advance technical innovation and improve how diverse communities engage with technology.

for a research-driven opening, should I mention my publications?

Yes — in a **research-driven opening**, briefly signaling your publications can strengthen your positioning as a prepared researcher.

- If you already have **multiple publications** (especially peer-reviewed in HCI, CHI, CSCW, Ubicomp, etc.), it's worth signaling that upfront — it shows you're already contributing to the field.
- If you have **one or two publications or works-in-progress**, you can still mention them, but keep it concise in the opening and expand later in the body.
- If you don't yet have publications, it's better to highlight your **research experience, methods expertise, or impactful projects**, and then mention publications only if they add credibility.

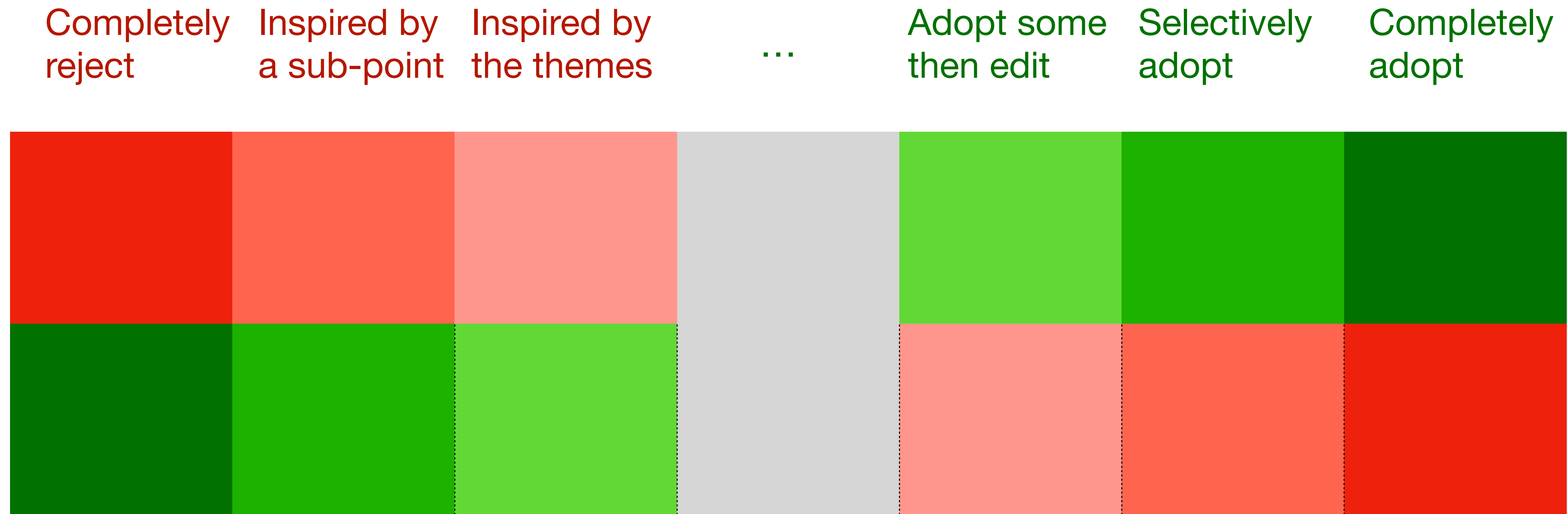
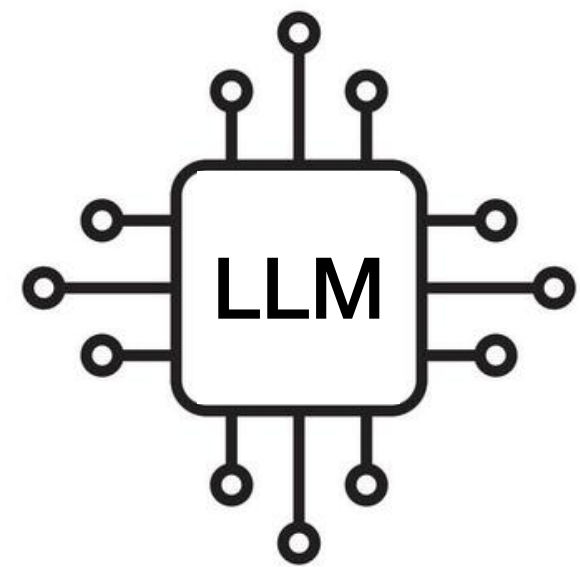
Here's how you could phrase it in a research-driven opening:

- **With publications in [venue(s)]**, my research explores [topic], and I am eager to expand this work through doctoral study in Human-Computer Interaction at [University Name].
- **Building on my research in [area], including a recent publication in [venue]**, I aim to investigate [topic/focus] during my PhD.
- **My work on [topic] has resulted in [publication outcome]**, which motivates my application to the HCI PhD program at [University Name].

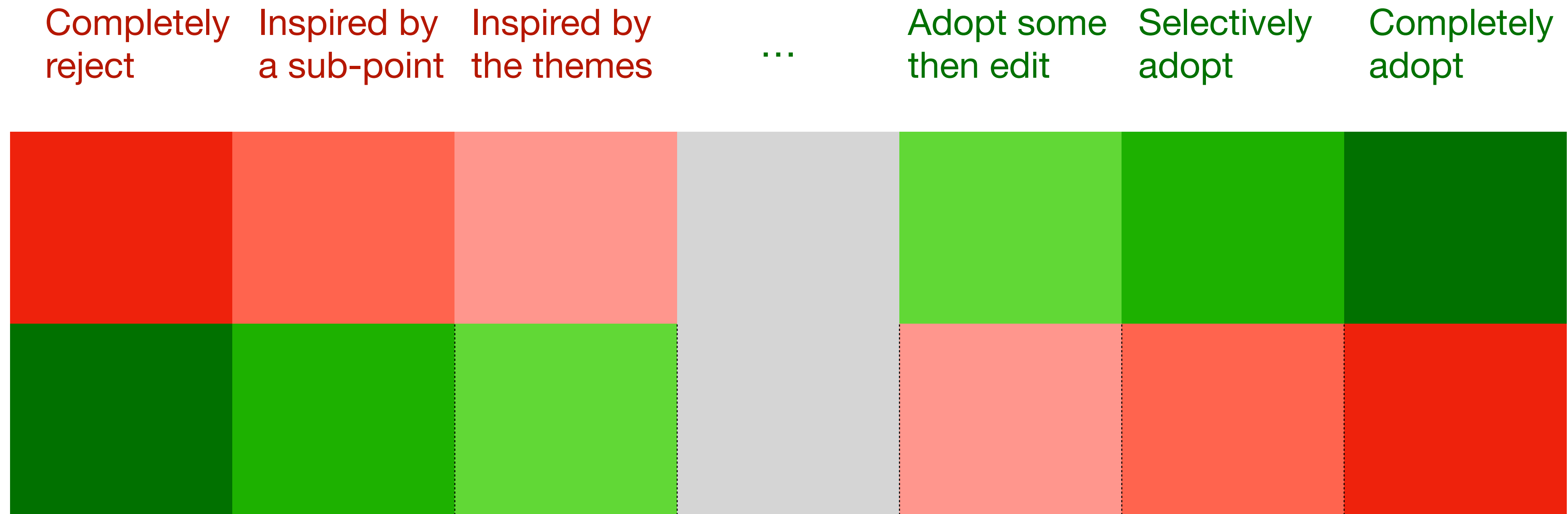
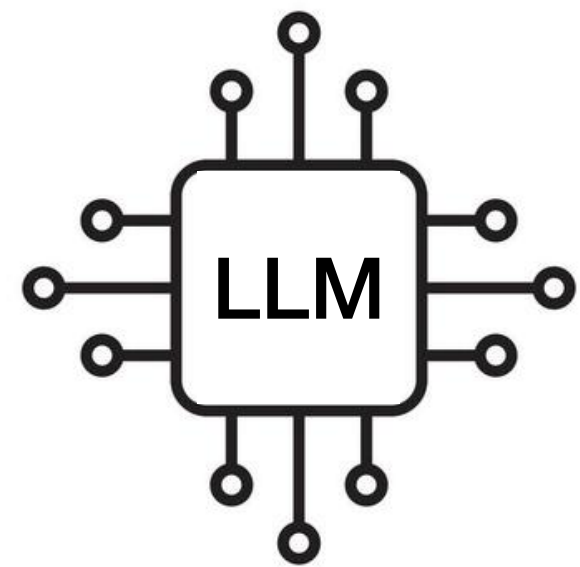
This way, you establish yourself as a researcher from the very first paragraph, without overloading the opening with details.



- “General-purpose” technology serving complex and diverse use cases
- Highly interactive
- Complex output space
- ...

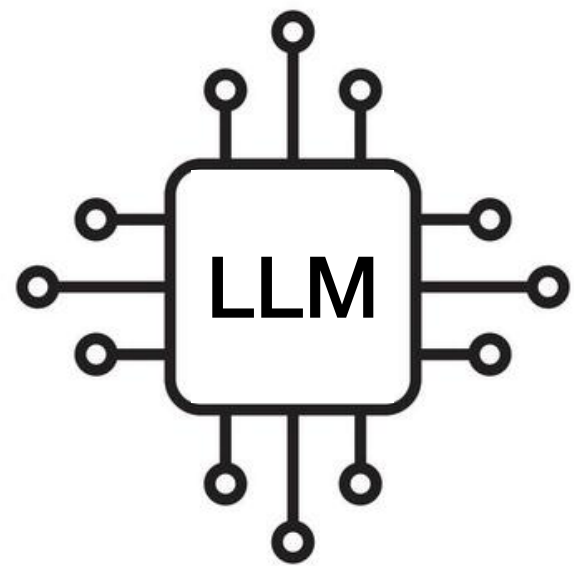


- Reliance cannot be seen as a binary decision
- Reliance may involve many actions and interactions



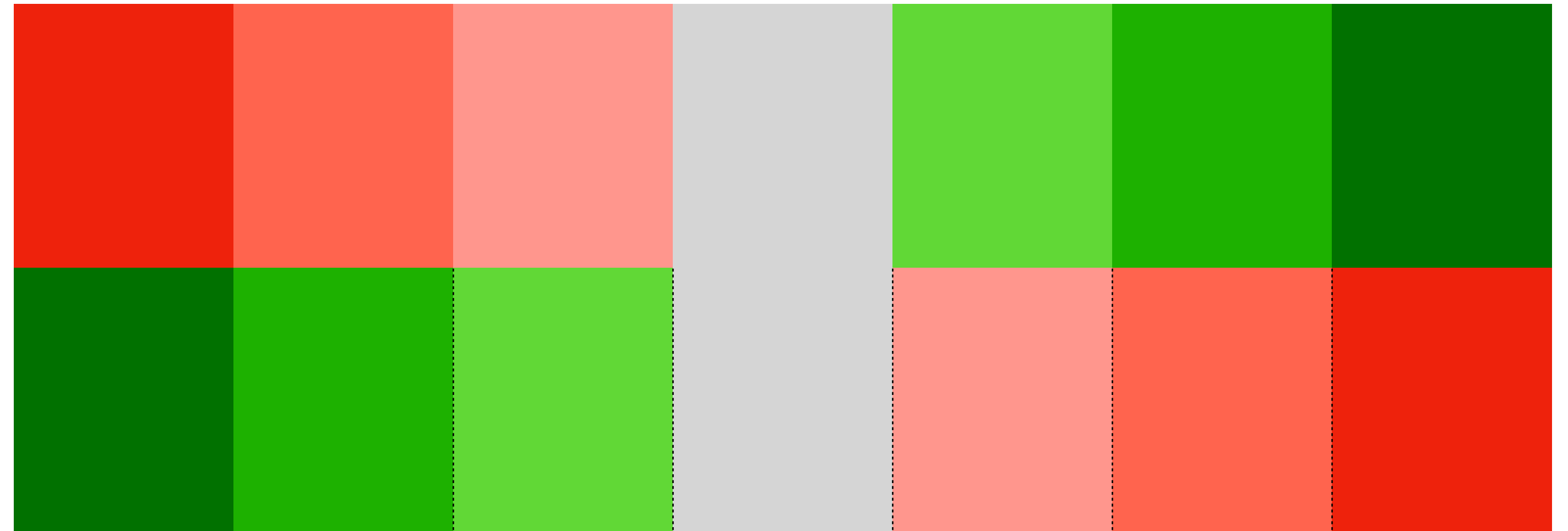
- Reliance cannot be seen as a binary decision
- Reliance may involve many actions and interactions

How to measure reliance without intensive behavior tracking and generalizable to different LLM use cases?



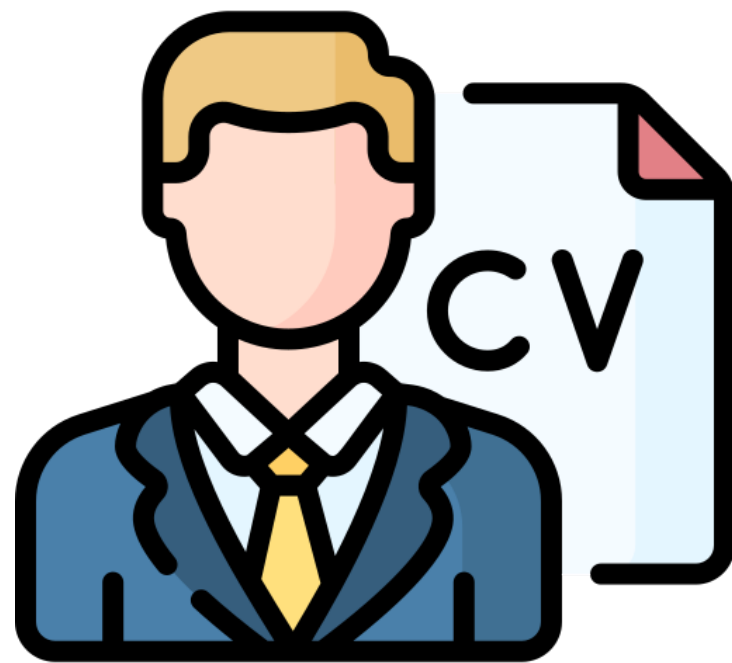
~~Correct~~ **Good**
LLM Output

~~Incorrect~~ **Bad**
LLM Output



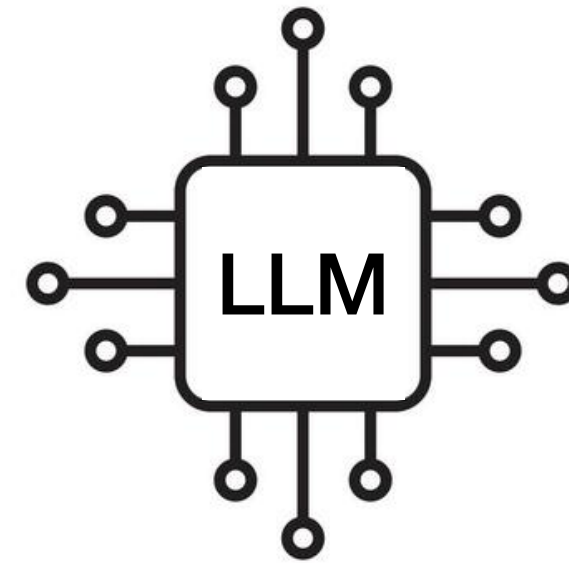
Good/bad in what way? To Whom?

How to define appropriate and inappropriate reliance?



Job seekers

Persuasiveness ,
coherence...



Factuality,
transparency...



Employers

*Is it appropriate to rely on an
LLM to write the cover letter?*

Fairness



Competitors

~~Correct~~ **Good**
LLM Output

~~Incorrect~~ **Bad**
LLM Output



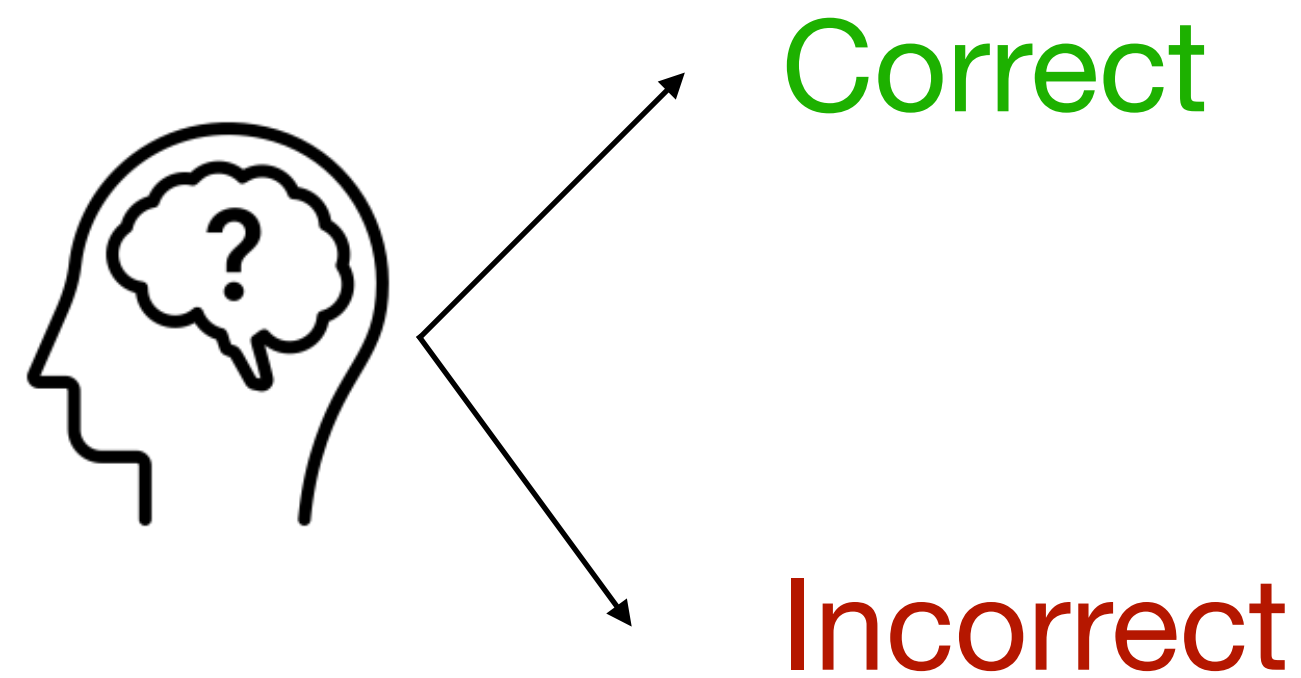
Defining output goodness hence appropriate reliance may be a “wicked problem”:

- Different stakeholders may have different definitions due to value differences
- There might not be one single “correct” way to resolve the value differences

Future Directions for Facilitating Appropriate Reliance in the LLM Era

- Define and measure user reliance without intensive behavior tracking
- Develop approaches to define appropriate reliance based on context-specific stakeholder values
- Design and empirically study system features that facilitate appropriate reliance
- Understand what system, individual, and contextual factors may influence LLM reliance
- Develop predictive models to deliver targeted and preventive interventions

Thank **YOU!**



Stage 1: Making decisions on one's own

	Reject AI	Adopt AI
Correct AI Output	Underreliance	Appropriate Reliance
Incorrect AI Output	Appropriate Reliance	Overreliance

Stage 2: Making the same/similar decision with AI

Reliance measurement by **Switch fraction** = $P(\text{switched to agree with AI} \mid \text{total disagreement})$

Appropriate reliance = $P(\text{switched to agree with AI} \mid \text{disagreement while AI correct})$